Методология

ITEM RESPONSE THEORY: ОСНОВНЫЕ ПОНЯТИЯ И ПОЛОЖЕНИЯ

Вадим Аванесов
testolog@mail.ru

В статье рассмотрены основные понятия и исходные положения теории, называемой на Западе Item Response Theory (IRT). Показана неприемлемость двух распространённых утверждений относительно IRT: первое — о её безусловных достоинствах и преимуществах по сравнению с классической (статистической) теорией тестов, и второе — о «современности» IRT, что позволяет с лёгкостью переводить остальные теории измерений в разряд «устаревших». У IRT есть свои достоинства и недостатки, но она не может считаться более современной, чем другие теории, активно используемые сейчас в педагогических измерениях.
Приводится русскоязычная система определений и исходных положений IRT. Педагогический тест определён как единство трёх систем: педагогической, статистической и математической.

Введение

Item Response Theory (IRT) — английское название теории, используемой преимущественно в педагогических и психологических измерениях. Эта теория смогла привлечь к себе внимание классиков мировой теории педагогических измерений и психометрики, математиков, статистиков, программистов, педагогов и управленцев сферы образования многих стран мира. К настоящему времени за рубежом появились десятки тысяч научных исследова-

ний по IRT, возникла эффективная практика создания тестов, на её основе создаются адаптивные обучающие и контролирующие системы многих университетов и стран.

В России название IRT переводили такими словами, как «теория латентных черт», «теория характеристических кривых заданий», «теория моделирования и параметризации педагогических тестов», «современная» теория тестов и т.д. Столь заметные различия в переводах одного только названия IRT уже сами по себе являются свидетельством неблагополучия в понимании её сути. Не лучшим образом обстоит дело с переводом на русский язык исходных понятий и положений IRT.

По мнению некоторых авторов, IRT — это современная теория педагогических измерений, преодолевающая недостатки других теорий, она способна решить многие, если не все, проблемы повышения качества педагогических измерений. Но это не вполне верный ход мысли. Ни одна из известных теорий не исчерпала свой потенциал развития. Вопрос лучше ставить о сравнительных достоинствах и об ограничениях, присущих каждой теории. Например, расчёт так называемых параметров математических функций[1], описывающих свойства заданий, предполагает достаточно большое число испытуемых. При малом числе испытуемых получаемые значения параметров таких функций очень ненадёжны, а потому в таких случаях на результаты применения IRT нельзя полагаться в полной мере.

Хотя IRT решает ряд задач образовательной практики лучше других теорий, анализом ответов испытуемых на задания реального или *проектируемого теста*[2] занимаются все теории педагогических измерений. И в этом смысле прямой перевод смысла слов, образующих название данной теории, является не существенным. Здесь сложилась примерно такая же ситуация, как и с переводом слова или научного понятия «тест». Философы уже давно знают существенные различия между переводом смысла слов и смысла научных понятий. Употребление и толкование смысла слов имеет в своей основе обыденное сознание, в то время как опора на научные понятия свойственна только научному мышлению. Вот почему перевод названия и основных положений Item Response Theory на русский язык оказался непростым делом.

Смысл английских слов Item (задание), Response (ответ) и Theory (теория) наталкивает на идею перевода «теория ответа на задание». Но такой перевод бессодержателен, потому

1

Item Characteristic Curves, часто переводимые как «характеристические кривые заданий», что трудно признать удачным переводом.

2

Проектируемый тест можно определить как совокупность заданий, формальные свойства которых статистически не определены, а потому подлежат эмпирическому и статистическому исследованию. Это определение вводится здесь исключительно из прагматических соображений противодействия широкому распространению в практике т.н. «тестов», не имеющих научного обоснования качества их заданий. Пусть такие «тесты», расплодившиеся сейчас в стране в огромных количествах, получат более скромное и подходящее им название проектируемого теста, состоящего из заданий в тестовой форме. После проверки педагогических, статистичес-

что классическая теория тоже занимается анализом ответов на задания. В ней есть даже специальный раздел «Item Analyses». Таким образом, для выявления сущности IRT и используемых в ней моделей измерения простой перевод упомянутых слов, составляющих название этой теории, ничего не даёт.

Перевод IRT как «теория ответа на задание» может вызвать ироническую усмешку у методистов, и тем более методологов. Поскольку перевод сразу попадает в разряд псевдо- или квазинаучных теорий типа «зонтиковедение», «чемодановедение», «ЕГЭведение», «КИМология» и т.п. Последние два направления претендуют на ведущее место в обосновании текущей российской практики[3] псевдотестирования.

Есть один очень простой метод демаркации тестов от тестоподобных материалов. Если время проверки знаний превышает примерно 40 минут, то этот признак свидетельствует о потере одного из самых существенных свойств теста — кратковременности процесса контроля знаний. Когда, например, говорят о четырёхчасовых «тестах» ЕГЭ по русскому языку или по математике, легко видеть, что столь длительное время «тестирования» является убедительным признаком бюрократического выхолащивания самой сути тестового метода.

Проблема с русским названием IRT заключается в том, что английское название этой теории не точное, не полное, отчасти устаревшее и метафоричное. А потому прямо не переводимое, в принципе. Откуда следует необходимость искать обходные, смысловые варианты. Такой вариант нового названия IRT приводится в самом конце этой статьи.

Вместе с тем, дословный перевод IRT на русский язык имеет всё-таки некоторое отношение к пониманию сущности этой теории, хотя и не раскрывает её возможности в деле разработки качественных тестов, шкалирования, обоснования качества и эффективности тестовых результатов. А потому перевод IRT как «теория ответа на задание» является слишком узким и не адекватным её широким возможностям.

Здесь самое время затронуть важный вопрос состояния понятийного аппарата педагогических измерений. Не так давно мы были свидетелями попыток разработок терминологического словаря тестирования[4], из содержания которого с очевидностью проявляет себя стремление представить практическое тестирование как форму научной деятельности, якобы имеющую свой собственный понятийный аппарат, отличающийся от понятийного аппарата теории педагогических измерений.

ких и метрических свойств. Часть таких заданий может получить шанс называться тестовыми заданиями и попасть в настоящий тест. При данной логике тестовой деятельности будет совершаться меньше ошибок, чем это делается сейчас. Подробнее об этом см.: Аванесов В.С. «Основы педагогической теории измерений» // Педагогические измерения, № 1, 2004. С. 15–16.

3

В фундаментальной науке мелким и частным системам своеобразных знаний такого рода места нет. В прикладных же науках место под солнцем получают лишь те теории, которые имеют научный потенциал эффективного решения важных задач практики.

4

Основные принципы построения системы понятий и терминов педагогического тестирования // Стандарты и мониторинг в образовании № 2, 2003, с. 53–61.

На Западе большинство понятий практики и понятий теории давно и благоразумно разделились. И не случайно там практика называется *тестированием*[5], а наука, обеспечивающая качество тестовых результатов, имеет название «Педагогические измерения». До конца 70-х годов XX века эта наука называлась преимущественно теорией тестов. К началу XXI века чаще стало использоваться название Educational Measurements.

К настоящему времени об IRT написано уже немало, однако восприятие основ данной теории затруднено в силу спорности и неадекватности используемой лексики, избыточной математизации и недостаточной педагогичности текстов об IRT.

Возникновение IRT

Общим источником для создания IRT послужила так называемая логистическая функция вида $Y = \frac{e^x}{1+e^x}$, известная в биологической науке с 1844 года. С тех пор она широко применялась в биологии для моделирования прироста растительной массы или роста ряда организмов. Как модель психологического и педагогического измерения она начала применяться, начиная с 50-х годов XX столетия. У истоков развития моделей IRT лежали стремление визуализировать формальные характеристики тестовых заданий, попытки преодолеть недостатки классической теории тестов, повысить точность измерения и, наконец, стремление оптимизировать процесс контроля за счёт адаптации теста к уровню подготовленности студента с помощью компьютера.

В числе первых предпосылок к созданию IRT стали те результаты исследовательской работы A. Binet и T. Simon[6], в которых было отражено стремление авторов выявить как, образно говоря, «работают» те задания, которые они давали детям разного возраста. Расположив затем на координатной плоскости точки, где по оси абсцисс откладывался возраст (в годах), а по оси ординат — доля правильных ответов в каждой возрастной группе испытуемых, авторы увидели, что полученные точки, после усреднения по каждой группе, напоминают кривую, позже названную характеристической.

В 1936 году M.W. Richardson провела обширное эмпирическое исследование, опросив 1200 студентов по 803 заданиям, в процессе которого студенты, в зависимости от полученного ими тестового балла, были разделены на 12 групп, по сто человек в каждой. Она первой обратила внимание на различающуюся крутизну кривых тестовых заданий и выдвинула идею рассматривать меру кру-

5 Тестирование можно определить как практическую деятельность по применению тестов для объективированной оценки уровня и структуры подготовленности испытуемых.

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

тизны как примерную оценку дифференцирующей способности задания[7]. M.W. Richardson была, по-видимому, первой, осознавшей плодотворность использования усреднённых точек для графической презентации формальных характеристик заданий проектируемых тестов[8].

В 1942–43 годах по данной проблеме появились ещё две работы. Вероятность успешно выполнить задания W.A. Fergusson[9] выразил в виде

$$P_j = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\theta-\beta_j}{s_j}} e^{-\frac{u^2}{2}} du \,, \quad (1)$$

где θ — уровень знаний;

β_j, s_j — константы задания j.

Несколько дальше пошел D.N. Lawly, задавшийся целью создать такой метод оценки качества заданий, который не зависел бы от состава оцениваемой группы. Многолетние попытки привели к двум вариантам решения. Один из них заключался в рекомендации увеличивать выборку до тех пор, пока выборочная статистическая мера трудности задания q_j не станет сколь угодно близкой к значению параметра трудности того же задания j, получаемого на генеральной совокупности. Этот метод традиционно тривиален и расточителен, и потому его трудно признать эффективным для решения стоящей задачи. Второй результат, полученный D.W. Lawly в 1943 году, представлял собой попытку оценить устойчивость показателя трудности задания безотносительно к уровню подготовленности конкретной группы испытуемых. Построив по эмпирическим данным усреднённую кривую[10], похожую на рис. 1, он обнаружил, что при наличии для каждой кривой своего значения параметра трудности (β_j), результаты студентов со слабой подготовкой, на каждое задание, группируются в нижней части кривой (рис. 2), а проекции испытуемых с отличной подготовкой — в верхней части усреднённой кривой (рис. 3).

В качестве меры работоспособности задания, т.е., способности дифференцировать студентов по уровню их подготовленности, D.W. Lawly стал рассматривать параметр крутизны логистической кривой. Таким образом, каждому заданию теста ему удалось поставить в соответствие два параметра — трудность задания и крутизну кривой[11].

В своём первом монографическом исследовании F.M. Lord[12] использовал модель W.A. Fergusson, однако позже, под влиянием работы A. Birnbaum[13], он стал применять логистические кривые, которые оказались удобнее для расчётов.

7

Richardson Marion W. The Relation Between the Difficulty and the Difference Validity of a Test / Psychometrika, 1936, 1: 2, 33–49.

8

Richardson M.W. Notes on the Rationale of Item Analysis / Psychometrika , 1936,1: 169–76.

9

Fergusson G.A. Item Selection by the Constant Process / Psychometrika, 1942, v. VII, pp. 19–29.

10

Рисунки 1, 2 и 3 представлены из книги Baker, Frank. The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD 2001.

11

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. 1943 v. LXI, part III, p. 273–287, 43.

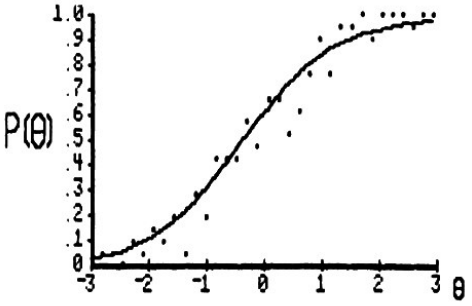

Рис. 1. Подбор графического образа задания по результатам его эмпирической апробации

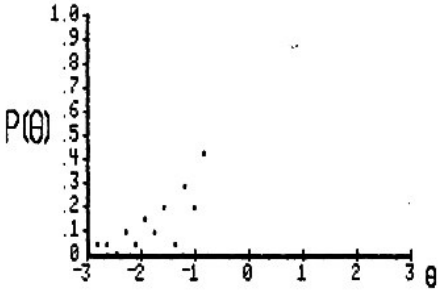

Рис. 2. Эмпирические доли правильных ответов в группах слабоподготовленных испытуемых

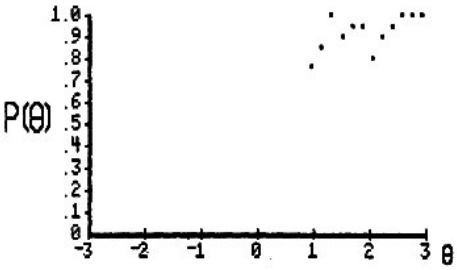

Рис. 3. Эмпирические доли правильных ответов у хорошо и отлично подготовленных испытуемых

Как и F.M. Lord, A. Birnbaum тоже начинал с совершенствования модели W.A. Fergusson, введя параметр a_j для оценки дифференцирующей способности задания. Вследствие этого выражение (1) приобрело вид

$$P_j = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{a_j(\theta-\beta_j)}{s_j}} e^{-\frac{u^2}{2}} du \ , \quad (2)$$

где θ — уровень знаний; β_j, s_j — константы задания j.

В одном техническом отчёте, написанном в 1952 г. по теме, далёкой от теории педагогических измерений, D.C. Haley[14] предложил такую модель для описания собранных им данных:

$$\varphi(x) = \frac{e^x}{1+e^x}. \quad (3)$$

Латентная составляющая IRT

IRT является психолого-педагогическим вариантом более общей методологии латентно-структурного анализа, развивавшегося, главным образом, в лабораториях военных ведомств США и университетов. Латентно-структурный анализ

12 *Lord F.M.* A Theory of Test Scores. Psychometric Monographs, 1952, 7. Richmond. Publ. by Psychometric Society. 84 pp.

13 *Birnbaum A.* Some Latent Trait Models and Their Use in Inferring an Examinee's Ability / In: F.M. Lord and M.R. Novick. Statistical Theories of Mental Test Scores. Reading, Mass: Addison Wesly, 1968. 568 p.

14 *Haley D.C.* / Cit in: Hambleton R.K. Swaminathan H. Item Response Theory: Principles and Applications, Boston 1985. 327 p.

(от англ. Latent Structure Analyses, LSA) нацелен на выявление латентных качеств (факторов) поведения посредством математико-статистических моделей измерения. Это направление работ в заметной мере обязано П. Лазарсфельду[15]. Первый этап исследований по LSA совпал с началом второй мировой войны. На этом этапе было проведено широкомасштабное социально психологическое исследование «American Soldier», призванное повысить боеспособность солдат за счёт выявления и устранения тех скрытых факторов, которые влияют на поведение солдат в боевой обстановке, а также факторов, которые помогают преодолеть страх[16].

LSA часто называют теорией, но с этим трудно согласиться, потому что это не теория какой-то предметной деятельности, а теория разработки и применения методов исследования эмпирических данных. Из того, что это теория методов, вытекает более точное, но не совсем привычное название методологии LSA[17]. Это действительно методология или, лучше сказать, методологический подход, в рамках которого появились и закрепились первые модели IRT. В литературе модель измерения определяется как структурное построение, позволяющее соединить латентные переменные с одним или большим числом наблюдаемых переменных[18].

Латентными называются интересующие положительные и отрицательные качества личности, не поддающиеся непосредственному измерению. Примерами являются «подготовленность студентов», «знание учебной дисциплины», «способность понимать», «интеллектуальное развитие» и многое другое. Попытки измерения подобных качеств личности[19] на уровне обыденного сознания оканчиваются словесными или численными оценками, содержащими в себе немалые погрешности.

Идея и методы измерения латентных качеств реализуются в тесной зависимости от эмпирических результатов. Именно на основе реально наблюдаемых данных ставится задача воссоздания непосредственно ненаблюдаемого качества, измеряемого с помощью модели. Эмпирическим определением латентной переменной величины является содержание заданий теста. По мнению И. Канта, ненаблюдаемый мир отличается от наблюдаемых явлений. Это утверждение принимается в качестве отправного положения в IRT.

Далее делаются такие предположения о том, что интересующее свойство личности:

- существует в *латентном состоянии*;
- оно устойчиво;

15 *Lazarsfeld P.F.* The interpretation and computation of some latent structures / Measurement and Prediction, N.V. John Wiley and Sons, 1950. p. 415–472.

16 *Stouffer S.A.* a.o. The American Soldier: Adjustment During Army Life. Prienceton, N.J., Prinston Univ. Press, 1949. 599 p.

17 *Lazarsfeld P.F., Henry N.W.* Latent Structure Analyses. Boston, Houghton Mifflin Co., 1969. 292 p.

18 *Bollen K.A.* Structural Equations with Latent Variables. P. 182, N.V. Wiley & Sons, 1989. 514 p.

19 Понятия «качество», «свойство», «признак» удобно рассматривать как обобщенный аналог английского понятия trait.

• имеется у данных испытуемых, в каких-то количествах;
• измеряемо с некоторой погрешностью.

Теория измерений — это научная форма организации знаний о неявно заданных свойствах объектов, о правилах и методах отображения этих свойств в числовую систему с отношениями. Если выясняется, что у кого-то из испытуемых нет проявлений данного свойства, то это даёт основания для исключения данного испытуемого из предполагаемой выборки лиц, обладающих данным свойством.

Имея общим объектом тестовый процесс, практика и теория имеют различные предметы исследования. Практики занимаются тестированием, куда входит создание тестов, получение результатов тестирования, обработка данных и интерпретация результатов. Научной основой практики тестирования является теория педагогических измерений. Научная работа концентрируется вокруг исследования проблем педагогических измерений: определение понятийного аппарата, развитие тестовых форм, исследование критериев оптимизации содержания тестов, разработка новых методов статистической и математической обработки данных, вопросы шкалирования результатов испытуемых и параметров заданий.

В самом общем виде величиной можно назвать всё то, что может быть больше или меньше, что может быть присуще объекту в большей или меньшей степени; числовая величина — такая, которая может быть выражена числом[20].

Классики педагогических измерений рассматривают уровень подготовленности испытуемых как непрерывную латентную величину. Они обозначают эту величину символом q. Значение испытуемых на этой величине обозначаются символом (θ_i). В настоящей статье преимущественно используется символика, введённая Ф. Лордом[21]. Уровень трудности заданий также рассматривается как непрерывная величина, обозначаемая символом θ. Значение меры трудности каждого задания на этой переменной величине обозначается символом β_j.

Ведущая идея, утверждения и ключевые понятия IRT

Для перевода смысла названия IRT на русский язык и понимания роли этой теории для науки и практики необходимо задаться вопросом о ведущей идее и ключевых понятиях IRT.

Ведущая идея IRT сводится к обоснованию возможности эффективного прогнозирования результатов тестирования на за-

20 *Петров Ю.А., Никифоров А.Л.* Логика и методология научного познания. М.:Изд-во Моск. ун-та, 1982. 249 с.

21 *Lord F.M.* Application of Item Response Theory to Practical Testing Problems. Hillsdale N.J. Lawrence Erlbaum Ass., Publ. 1980, 266 pp.

дания различного уровня трудности. Такой прогноз особенно необходим в системах профессионального отбора, адаптивного обучения и адаптивного тестового контроля. Прогноз основан на утверждениях.

Первое утверждение теории IRT — вероятность правильного ответа на задание *j* у хорошо подготовленного испытуемого должна быть больше вероятности правильного ответа у слабо подготовленного испытуемого: чем выше подготовка испытуемого, тем выше может быть вероятность правильного ответа на задание данного уровня трудности. Это утверждение иногда формулируется в обратном виде: чем ниже уровень подготовленности, тем меньшей может быть вероятность правильного ответа на задание того же фиксированного уровня трудности.

Второе утверждение теории IRT — о вероятности правильного ответа испытуемого фиксированного уровня подготовленности на задания теста, при строгом соблюдении правил тестирования, исключающих возможности списывания и других нарушений учебной этики, и при высоком качестве заданий.

Особенность применения IRT заключается в том, что ответы множества испытуемых на множество заданий теста прогнозируются на основе математических моделей при наличии эмпирически полученной матрицы *исходных тестовых баллов* X_{ij}, где индекс *i* указывает на номер испытуемого, а индекс *j* — на номер задания. X_{ij}. Обычная, традиционная практика давать значениям X_{ij} один балл, если ответ испытуемого *i* на задание *j* правильный, и ноль — если ответ неправильный. В последние годы стали шире применяться и другие оценки, что повышает качество измерений. Автор этой статьи рекомендует переходить на другие оценки, которые позволяют давать испытуемым новые формы тестовых заданий[22].

Иначе говоря, в IRT обычно имеет место математическое моделирование эмпирически получаемых результатов, что полезно для оценки соответствия качества получаемых педагогических измерений теоретическим положениям .

Ведущая идея IRT опирается на следующее понятие: *уровень трудности задания* (item difficulty).

В IRT принимаются во внимание несколько мер трудности заданий.

Первая мера — доля неправильных ответов испытуемых на каждое задание проектируемого теста (q_j). Это исходное значение меры трудности каждого задания, которое находят эмпирически, из матрицы тестовых результатов, по формуле

22

Аванесов В.С. Форма тестовых заданий. М.: Центр тестирования, 2005. 155 с. Применение тестовых форм в Rasch Measurement // Педагогические измерения. № 4, 2005 г.

$$q_j = \frac{W_j}{N}, \qquad (4)$$

где W_j — число неправильных ответов на задание под номером j и N — число испытуемых.

Вторая мера трудности — это отношение $\frac{q_j}{p_j}$, что является интересной и очень показательной мерой трудности задания, с высокой вариацией результатов. Она была предложена Г. Рашем, но не получила ни названия, ни развития в его трудах. Эту меру можно условно назвать *потенциалом трудности задания.*

Третья мера трудности заданий — это значение натурального логарифма отношение $\frac{q_j}{p_j}$, что даёт $\ln\frac{q_j}{p}$. Последнее значение далее корректируется для построения общей (единой) шкалы уровня трудности заданий и уровня подготовленности испытуемых. Это и есть процесс *шкалирования*, проводимый в наши дни с помощью западных статистических пакетов типа, например, Winsteps или RUMM 2020, после чего получается четвёртая мера трудности заданий.

Четвёртая мера трудности заданий. — это корректированные в процессе шкалирования значения $\ln\frac{q_j}{p}$. В качестве окончательной меры трудности заданий принимается именно эта мера. В IRT она называется *параметром трудности задания.* Скорректированные значения $\ln\frac{p_i}{q_i}$ называются *параметром подготовленности испытуемого* под номером *i.*

Шкалирование — это процесс присвоения значений на числовой оси испытуемым и заданиям, в зависимости от уровня проявления интересующего свойства. Для заданий принимается во внимание три формальных свойства: уровень трудности, дифференцирующая способность заданий (discriminant ability) и априорная вероятность угадать правильный ответ на задание со стороны неподготовленного испытуемого. В этом процессе используется общая средняя арифметическая, равная нулю, и общий показатель вариации, равный единице. Для испытуемых измеряемым свойством является уровень подготовленности.

Качества шкал можно оценивать по следующим критериям:

- уровень шкалы: номинальная, порядковая, интервальная и пропорциональная, предпочтительны две последние;
- наличие общей единицы измерения, что обеспечивает сравнимость результатов различных тестов;

• размах значений, пределы значений оценок и измерений, получаемых по разным шкалам, дисперсия — желательно иметь их совпадающими, что обеспечивает равноценность баллов, получаемых по той или иной шкале;
• совпадающие средние значения шкальных баллов, показатели асимметрии и эксцесса позволяют корректно сравнивать распределения результатов по разным тестам.

IRT позволяет организовать такой процесс шкалирования, который способствует получению двух сопоставимых шкал: одну — для испытуемых, другую — для заданий. С общим началом, общей единицей измерения и с общей средней арифметической, равной нулю. Это шкала логитов уровня подготовленности испытуемых и шкала логитов уровня трудности заданий. При применении двух- и трёхпараметрической модели педагогического измерения можно получить также шкалу уровня дифференцирующей способности каждого задания.

Уровень подготовленности испытуемого по интересующему свойству личности является обобщённым (интегральным) показателем уровня знаний, умений, навыков и представлений, а также компетентности личности в интересующей области. Уровень подготовленности представляется в виде концептуальной величины, выражаемой содержанием заданий, включённых в тест.

Для любого испытуемого под номером i исходный уровень подготовленности определяется из той же матрицы тестовых результатов, что и натуральный логарифм отношения доли правильных ответов испытуемого i к доле его неправильных ответов на тестовые задания. К числу таковых не относятся ответы испытуемого на так называемые *экстремальные* задания. К последним относят задания очень лёгкие, на которые отвечают все испытуемые, а также задания очень трудные, на которые не может ответить правильно ни один испытуемый. Такие задания в тест не включаются, удаляются из матрицы тестовых результатов как не адекватные уровню подготовленности испытуемых.

Симметрично, если в матрице *исходных тестовых баллов* выясняется, что какие-то испытуемые совсем не обнаруживают интересующее измеряемое свойство личности и они отвечают неправильно на все задания, а также если выявляются испытуемые, способные правильно решать все задания в течение отведённого времени, то такие испытуемые также называются *экстремальными*. Такие испытуемые исключаются из дальнейшего анализа тестовых результатов

как не соответствующие данному уровню трудности теста и уровню подготовленности испытуемых. В качественных педагогических измерениях уровень трудности заданий должен соответствовать уровню трудности заданий[23].

Проблема и предмет IRT

Проблема, которую призвана решать IRT — это повышение качества проводимых в практике педагогических измерений и улучшение интерпретации результатов. Это означает, что данная теория содержит возможности такого улучшения качества измерений, в сравнении с которым возможности других теорий воспринимаются как недостаточные. И это верно.

В число проблем IRT входит также поиск подходящих прогностических моделей, а также измерение уровней подготовленности испытуемых и уровня трудности заданий на одной и той же шкале. Измерение проводится на основе выборочных статистик с целью получить оценки параметров испытуемых и заданий — оценки, удовлетворяющие требованиям статистической науки: точности, эффективности, оптимальности и несмещённости. Чем лучше подобрана та или иная прогностическая функция, тем точнее оказываются оценки.

В фокусе исследований IRT — углубленная проверка формальных свойств заданий для повышения точности измерения, принятия решения о включении проверяемых заданий в тест. До момента возникновения IRT уже существовала теория измерений, которая по установившейся ещё в начале XX века привычке называлась Classical Test Theory (CTT). На самом деле CTT представляла собой первую научно-статистическую теорию педагогических и психологических измерений.

В фокусе CTT находились все статистические вопросы разработки тестов, начиная от концепции истинного и ошибочного компонентов измерения, количественных оценок качества всех используемых при разработке теста заданий, что охватывалось понятием Item Analysis — вплоть до оценок качества педагогических и психологических измерений по критериям надёжности и валидности результатов. Попытку расширения сферы применения этой теории на социологические измерения автор этой статьи предпринял в своей монографии[24].

Основной *предмет применения математических моделей IRT* — прогнозирование вероятности правильного ответа испытуемых на задания различной трудности. В IRT основной предмет — анализ не сумм бал-

23

Это и есть ахиллесова пята, обрекающая ЕГЭ на перманентную некачественность, если не будут приняты необходимые меры по научному переосмыслению всего комплекса вопросов, возникающих вокруг ЕГЭ. Главные из них, и первоочередные — тотальная принудительность ЕГЭ, структурная некачественность его оценок, отрицательные последствия на результаты образовательной деятельности, системная социальная противоречивость интересов различных групп населения, и повышенная коррупционность самого ЕГЭ.

24

Аванесов В.С. Тесты в социологическом исследовании. М.: Наука, 1982. 199 с.

лов испытуемого, т.е. не баллов проектируемого теста, а баллов испытуемых, полученных по каждому заданию.

Исходные аксиомы педагогических измерений сводятся к тому, что интересующее свойство личности:

- существует в латентном состоянии;
- оно устойчиво;
- имеется у данных испытуемых, в каких-то количествах;
- измеряемо с некоторой погрешностью[25].

Предмет педагогического измерения — измерение уже упоминавшихся формальных свойств заданий. Вторым предметом является уровень и структура подготовленности испытуемых по изучавшейся учебной дисциплине. Обычно измеряется подготовленность именно по отдельным учебным дисциплинам. Что связано с понятием гомогенного теста. Процессуально свойства заданий являются первоочередным предметом измерений, однако главным предметом измерения традиционно считается интересующее свойство личности.

В профессиональной западной литературе нередко можно встретиться со случаем, когда фактически разные измеряемые величины могут иметь одно и тоже общее название *ability*. Оно пришло из психометрики. В наши дни оно не обязательно должно переводиться как некая интересующая исследователей способность испытуемых. Применительно к педагогическим измерениям лучше опираться на понятие «подготовленность испытуемых», куда входят знания, умения, навыки, представления и компетенции. Всё это — русские аналоги общего и традиционного англоязычного термина ability.

Главные принципы определения предмета измерения:

- соответствия цели;
- формулирование в явном виде;
- актуальность и перспективность;
- технологичность метода, измеряющего интересующий предмет (свойство личности) у испытуемых;
- эффективность выделения предмета измерения;
- соответствие требуемым критериям (минимальной компетентности и уровню подготовленности к обучению в вузе).

В случае измерения способностей и знаний приходится давать более точные и дифференцированные названия каждому предмету измерения.

Цель, задачи и преимущества IRT

Цель педагогического измерения — определить количество интересующего качества (меру интересующего признака), присущего каждому испытуемому и

25 Систему аксиом педагогических измерений см. в работах: *Аванесов В.С.* Проблема качества педагогических измерений // Педагогические измерения, № 2, с. 3–27, 2004 г.; *Аванесов В.С.* Основы теории педагогических заданий. С. 61–63. // Педагогические измерения, № 3, 2006 г. С. 47–66.

определить параметры заданий проектируемого теста.

Цель IRT — это проведение высококачественных педагогических измерений уровня подготовленности испытуемых и уровня трудности заданий, куда входит поиск подходящих прогностических моделей и проведение расчётов пригодности модели для имеющихся данных.

Задачи IRT вытекают из этой цели и сводятся к разработке таких методов измерения, которые позволяют получить наилучшие (оптимальные) параметрические оценки уровня подготовленности испытуемых и оценки трудности заданий на основе выборочных статистик и других эмпирических данных.

IRT позволяет решить три ключевые задачи педагогического измерения:

1) найти параметры заданий;

2) найти параметры испытуемых;

3) подобрать функцию $P_j(\theta) = f(\theta - \beta_j)$.

Применение логитов в моделях IRT как меры подготовленности испытуемых и меры трудности заданий даёт ряд преимуществ.

Первое преимущество вытекает из стандартизованного характера этой единицы. Как и всякая стандартная единица измерения, она представляет собой результат преобразования исходных данных, что даёт возможность объективно сравнить достижения разных студентов по разным учебным дисциплинам, суммировать эти достижения и проводить затем вполне объективированные рейтинги.

Второе преимущество вытекает из специфических особенностей модели Г. Раша. Получаемые с её помощью оценки уровня подготовленности знаний, в силу независимости от конкретного подбора заданий, приобретают признаки объективности измерения, что положительно отражается на качестве оценок, используемых в педагогическом контроле.

Третье преимущество связано с возможностями компьютеризации расчётов истинных (модельных) значений тестовых баллов, полученных с помощью этой модели. Относительная несложность и быстрота расчётов, выполняемых с помощью программы, одновременность получения параметров тестируемых по измеряемому свойству и параметров трудности заданий являются достаточно мотивирующим моментом для внедрения этой модели в практику.

Четвёртым преимуществом рассматриваемой модели является устойчивость рассчитываемых значений уровня знаний и трудности задания, их относительная инвариантность. Эта устойчивость позволяет утверждать, что подобные модели луч-

ше, чем какие-либо другие позволяют оценить интересующие качества личности, недоступные непосредственному измерению. Устойчивость статистических оценок требует, однако, большого числа испытуемых.

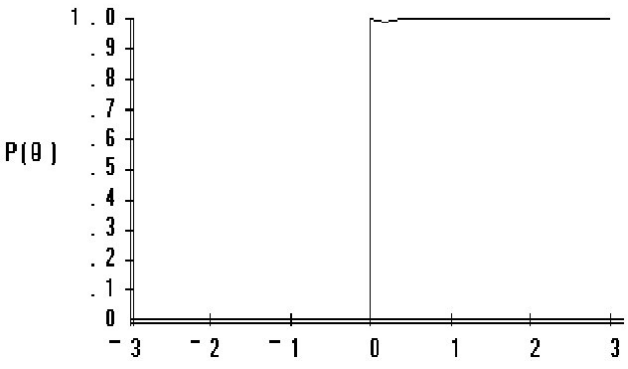

Рис. 4.

Чем выше крутизна кривой задания, тем уже интервал, на котором это задание дифференцирует испытуемых по уровню их подготовленности, тем выше дифференцирующая способность задания. Таким образом, возникла мысль об улучшении модели G. Rasch за счёт введения в выражение (5) второго, после параметра β_j, параметра a_j. Параметр a_j даёт информацию о задании с точки зрения оценки его дифференцирующей способности на заданном промежутке континуума измерения.

Дифференцирующей способностью задания (discriminant ability of the item) называется его свойство различать испытуемых по уровню подготовленности. Чем выше дифференцирующая способность задания, тем лучше деление испытуемых на подготовленных и неподготовленных. С ростом дифференцирующей способности графический образ задания стремится к вертикальному положению (рис. 4).

Три основные математические модели IRT

Математические модели педагогических измерений позволяют соединить интересующие латентные переменные величины с наблюдаемыми значениями этих величин у испытуемых и у заданий.

Известны три основные модели педагогических измерений.

Однопараметрическая модель педагогического измерения. Первая модель появилась в 1958 году, когда у Г. Раша возникла идея выразить вероятность правильного ответа на за-

дание j посредством функции вида[26]

$$P_j(\theta) = \{x_{ij} = 1 | \beta_j\} = \frac{\exp(\theta - \beta_j)}{1 + \exp(\theta - \beta_j)}, \quad (5)$$

где $P_j(\theta)$ — вероятность правильного ответа испытуемых любого уровня подготовленности на задание определённого уровня трудности под номером j.

x_{ij} = 1, если ответ испытуемого i на j-е задание правильный;

θ — уровень подготовленности (знаний), латентная переменная;

β_j — уровень трудности конкретного, j-го задания проектируемого теста, измеряемой на латентном континууме трудности заданий.

exp — константа e, иррациональное число, равное, округлённо, 2,71828.

Формулу (5) удобно представлять в строчной записи:

$$P_j(\theta) = \{x_{ij} = 1 | \beta_j\} = \exp(\theta - \beta_j) / (1 + \exp(\theta - \beta_j)) \quad (6)$$

В начале 50-х годов прошлого столетия датский математик G. Rasch стал рассматривать матрицу тестовых данных как результат взаимодействия множества испытуемых с множеством заданий. При этом естественным образом принималась *аксиома* — чем труднее задание для данного испытуемого, тем ниже вероятность правильного ответа. Из этой аксиомы неизбежно вытекало свойство функциональности модели: вероятность правильного ответа испытуемых на задание j есть функция от взаимодействия двух параметров — от уровня подготовленности испытуемых и от уровня трудности задания (β_j). Формально это условие можно записать $P_j(\theta) = f(\theta - \beta_j)$, что позволяет говорить, что эта функция от одной переменной величины, от разности значений $\theta - \beta_j$. Графический образ такой функции представлен на рис. 6.

Соответственно, вероятность неправильного ответа на задание j, обозначаемая (Q_j), равная, как принято в теории вероятностей, 1 – P, он выразил так (см. выражение 7)[27]:

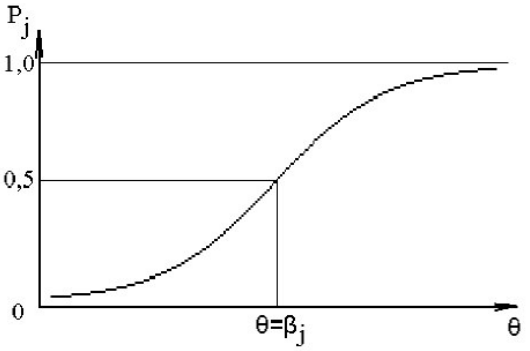

Рис. 5. Характеристическая кривая j-го задания текста

---

26 Приводимая формула излагались ранее в статье «Педагогическое измерение латентных качеств» в журнале «Педагогическая диагностика», № 4, 2003 г. Однако на стр. 72, обнаружились редакционные ошибки, появившиеся при электронной пересылке статьи. В этой публикации упомянутые ошибки устранены.

27 Мультипликативное отношение вида $\theta_I \cdot \beta_j$ или θ_I / β_j здесь не подходило из-за шкалы возможных значений θ_I и β_j. В шкалах логитов нулевые (т.е. средние) значения встречаются чаще других, а потому возможность обнуления оценок или деления на ноль делали мультипликативное отношение параметров в функциях $P_i = f(\theta_I \cdot \beta_j)$ и $P_i = f(\theta_I / \beta_j)$ абсолютно неприемлемым. Вот почему был избран аддитивный вариант $P_i = f(\theta_I - \beta_j)$. Теоретически обе оценки могли принимать значения: $-\infty < \theta_I < \infty$ и $-\infty < \beta_j < \infty$. Практически очень редки случаи, когда любая из этих двух оценок выходила бы за пределы значений от –5 до +5.

$$Q_j\{x_{ij}=0\,|\,\beta_j\}= \quad (7)$$
$$=1-\exp(\theta-\beta_j)/(1+\exp(\theta-\beta_j))$$

Если буквой L обозначить разность $(\theta - \beta_j)$, то тогда модель можно представить ещё удобнее, как $P_j(\theta)=\{x_{ij}=1\,|\,\beta_j\}=\frac{1}{1+e^{-L}}$, или $P_j(\theta)=\frac{1}{1+e^{-L}}$, а в строчной записи $P_j(\theta)=1/1+\exp(-L)$.

По логике G. Rasch

$$\frac{P_{ij}}{Q_{ij}}=\frac{\exp(\theta_i-\beta_j)/1+\exp(\theta_i-\beta_j)}{1-\exp(\theta_i-\beta_j)/1+\exp(\theta_i-\beta_j)}. \quad (8)$$

Формулу (8) можно упростить введением вспомогательной переменной

$$V=\exp(\theta_i-\beta_j), \quad (9)$$

откуда следует

$$\frac{P_{ij}}{Q_{ij}}=\frac{\frac{V}{1+V}}{1-\frac{V}{1+V}}. \quad (10)$$

После элементарных преобразований отношение $\frac{P_{ij}}{Q_{ij}}$ становится равным V, т.е., $\exp(\theta_i-\beta_j)$.

Пример практического применения однопараметрической модели для расчёта вероятностей правильного ответа испытуемых различного уровня подготовленности интересующийся читатель может найти в нашем журнале[28].

Двухпараметрическая модель измерения. Несколько позже возникла идея улучшения модели Г. Раша за счёт введения в формулу (5) параметра, a_j. Параметр a_j даёт информацию о задании с точки зрения оценки его дифференцирующей способности, на заданном промежутке континуума измерения. Графически значение параметра a_j выражается крутизной характеристической кривой задания, аналитически — значением производной функции в точке перегиба. После введения в выражение параметра a_j получается двухпараметрическая модель педагогического измерения.

$$P_j\{x_{ij}=1\,|\,\beta_j a_j\}=$$
$$=1-\exp a_j(\theta-\beta_j)/(1+\exp a_j(\theta-\beta_j)) \quad (11)$$

Или, короче,

$$P_j(\theta)=\frac{1}{1+e^{-L}}, \quad (12)$$

где L теперь представляет $a_j(\theta - \beta_j)$.

Чем выше крутизна кривой, тем уже интервал, на котором это задание дифференцирует испытуемых по уровню их подготовленности. Эмпирические пределы значений для параметра a_j — от минус 2,80 до плюс 2,80.

Рассмотрим пример применения двухпараметрической модели для расчёта вероятности правильного ответа испытуемо-

28

Аванесов В.С. Основы теории педагогических заданий. Педагогические измерения № 2, 2006 г. С. 26–62.

го i на задание с параметрами $\beta_j = 1$, $a_j = 0{,}5$. Для начала можно взять случай очень низкого уровня подготовленности испытуемого, $\theta_i = -3{,}0$.

Первый шаг. Находим значение L для двухпараметрической математической модели педагогического измерения $a_j(\theta - \beta_j)$.

Подставляя имеющиеся данные, получаем:

$$L = 0{,}5(-3{,}0 - 1{,}0) = -2{,}0\ ,$$

$$e^{-L} = 2{,}718^{-(-2{,}0)} = 7{,}389.$$

Второй шаг. Находится значение знаменателя формулы (12).

$$1 + 7{,}389 = 8{,}389.$$

Третий шаг. Находится вероятность правильного ответа

$$P_j(\theta) = \frac{1}{1+e^{-L}} = \frac{1}{8{,}389} = 0{,}119.$$

Интерпретация полученного результата: для испытуемых очень низкого уровня подготовленности, равного –3,0 логита, вероятность правильного ответа на задание уровня трудности 1,0 логита равна 0,119. Откуда следует, что правильный ответ малоподготовленного испытуемого на трудное задание маловероятен.

Теперь можно посмотреть, как меняется вероятность правильного ответа на то же задание с уровнем трудности $\beta_j = 1$ и параметром $a_j = 0{,}5$ в случаях, когда испытуемые имеют различный уровень подготовленности. Для этого достаточно провести небольшой вычислительный эксперимент, в котором надо последовательно брать разные уровни подготовленности, а полученные данные свести в табл. 1.

В этой таблице представлены результаты по определению вероятности правильного ответа испытуемых различного уровня подготовленности на задание с уровнем трудности $\beta_j = 1/0$ и со сравнительно низким уровнем дифференцирующей способности a_j=0,5.

Таблица 1

θ_i	$L = a_j(\theta - \beta_j)$	e^{-L}	$1 + e^{-L}$	$P_j(\theta)$
–3,0	0,5(–3 – 1) = –2	7,389	8,389	0,119
–2,0	0,5(–2 – 1) = –1,5	4,482	5,482	0,182
–1,0	0,5(–1 – 1) = –1	2,718	3,718	0,269
0	0,5(0 – 1) = –0,5	1, 649	2,649	0,377
1,0	0,5(1 – 1) = 0	1	2	0,500
2,0	0,5(2 – 1) = 0,5	0,607	1,607	0,622
3,0	0,5(3 – 1) = 1	0,368	1,368	0,731

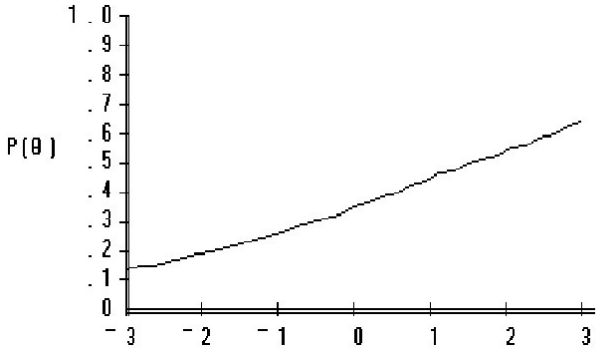

Рис. 6.

Результат вычислений представлен на графике (рис. 6).

Представленное на графике задание имеет очевидный дефект, выражающийся в том, что не все хорошо подготовленные испытуемые имеют шанс ответить правильно на данное задание. Вероятная причина такого дефекта — плохая формулировка содержания задания, при которой возможна другая интерпретация смысла задания.

Трёхпараметрическая модель педагогического измерения. Поскольку в заданиях с выбором одного правильного ответа всегда присутствует возможность угадывания, А.Бирнбаум посчитал необходимым учесть эту вероятность, а потому предложил добавить в двухпараметрическую модель третий параметр задания, c_j, со значением вероятности угадывания правильного ответа. Например, в заданиях с выбором одного ответа из пяти значение c_j принимается равным 0,2, что равно вероятности угадать правильный ответ теми испытуемыми, которые не знают правильный ответ. Это предложение привело к созданию трёхпараметрической математической модели педагогического измерения.

$$P_j\{x_{ij}=1|\beta_j,a_j,c_j\}=$$
$$=c_j+(1-c_j)\exp a_j(\theta-\beta_j)/(1+\exp a_j(\theta-\beta_j)) \quad (13)$$

В другом варианте,

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

где L по-прежнему представляет $a_j(\theta-\beta_j)$.

F. Lord называл c_j параметром псевдоугадывания. Хотя он сам не объяснил причину такого названия, можно думать, что этим было высказано, во-первых, сомнение, что это действительно параметр, а не предложение, основанное на утверждении о потенциальной, но не реальной вероятности угадывания правильного ответа. Во-вторых, в действительности, этот «параметр» применим не ко всем испытуемым, а только к тем, кто не имеет даже самых

туманных представлений о правильном ответе на задание. А таковых не так уж и много среди испытуемых. Следовательно, F. Lord прав: c_j является не «параметром», а лишь его тенью.

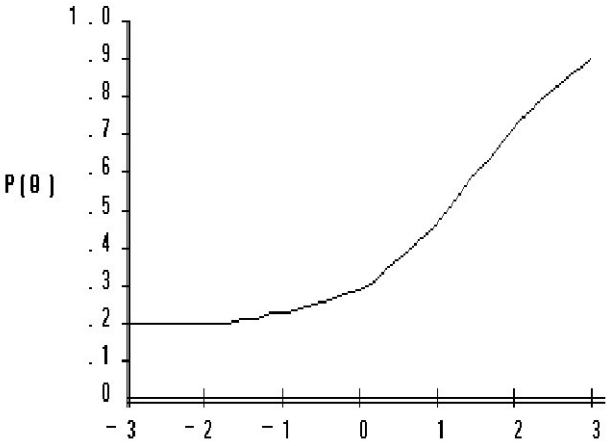

Рис. 7.

Для вычисления в трёхпараметрической модели можно провести некоторые вычисления. Обратимся к табл. 2 результатов по определению вероятности правильного ответа испытуемых различного уровня подготовленности на задание с параметрами трудности β = 1,5, с параметром крутизны a_j=1,3 и c_j = 0,2.

Графический образ задания с параметрами β = 1,5; a = 1,3 и c = 0,2 представлен на рис. 7.

Как отмечает D.L. McArthur, модели G. Rasch, A. Birnbaum, и F.M. Lord имеют много общего с математической точки зрения, однако они различаются концептуально[29]. Различны они и по генезису. F.M. Lord, его предшественники M.W. Richardson и D.N. Lawly начали с поисков

Таблица 2

θ_i	$L = a_j(\theta - \beta_j)$	e^{-L}	$1+e^{-L}$	$P_j(\theta)$
–3,0	1,3(–3 – 1,5) = –5,85	347,2340	348,2340	0, 202
–2,0	1,3(–2 – 1,5) = –4,550	94,6320	95,6320	0,208
–1,0	1,3(–1– 1,5) = –3,250	25,7900	26,7900	0, 230
0	1,3(0 – 1,5)= –1,950	7,0290	8,0290	0,300
1,0	1,3(1 – 1,5) = –0,650	1,9160	2,9160	0,474
2,0	1,3(2 – 1,5) = 0,650	0,5220	1,5220	0,726
3,0	1,3(3 – 1,5) = 1,950	0,1420	1,1420	0,900

29 *McArthur D.L.* Educational Assessment: A Brief History/ McArthur D.L. (Ed). Alternative Approaches to the Assessment of Achievement. Kluver Academic Publishers, Boston, 1987. 268 p.

математической модели презентации эмпирических данных. Подход же G. Rasch был априорным, теоретическим, направленным на создание математической модели измерения и получения такой единицы педагогического измерения, с помощью которой можно было бы корректно сравнить уровень знаний студента с уровнем трудности выполняемого им задания.

С самого начала своих поисков G. Rasch руководствовался идеей независимости оценки уровня знаний испытуемых от трудности заданий. Для реализации этой идеи ему понадобилось введение ещё одного понятия, распространённого до недавнего времени в азартных играх. Это понятие шанса на успех, равного отношению вероятности правильного ответа к вероятности неправильного ответа, p/q.

Смысл последнего становится ясным из сравнения нескольких, для примера, значений. При p= 0,5 и q = 0,5 отношение p/q = 1; тогда шансы на успех и неуспех равны. При p = 0,6 и q= 0,4 отношение p/q = 1,5; При p = 0,8 и q = 0,2, p/q= 4 и т.д., откуда видно, как меняются шансы на успех в зависимости от значения p и q. Отношение p_{ij} /q_{ij} даёт значение, равное шансу на успех испытуемого i по заданию j.

Итак, для ответа на общий вопрос об определении вероятности правильного ответа любого испытуемого (i), на любое задание (j) нужно выбрать одну из трёх моделей измерения:

Если на оси абсцисс отложить значения логитов уровня подготовленности, а по оси ординат — значения вероятности правильного ответа на задание j, то характеристические кривые с изменяющимися частными значениями параметра a_j, β_j и c_j, то можно получить различающиеся графические образы заданий, представленные на рис. 8.

Включение в тест заданий с очень пологими характеристическими кривыми порождает трудности интерпретации; эти задания часто измеряют, поми-

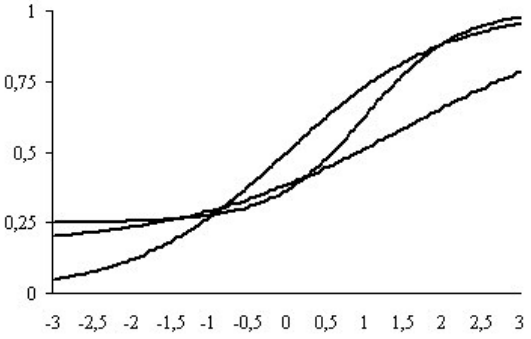

Рис. 8.

мо интересующего свойства, какое-то ещё и другое свойство.

С целью сопоставления кривых, полученных по другим математическим моделям, во все представленные формулы раньше вводилась константа 1,7. Модель G. Rasch, например, приобретает тогда вид:

$$P_j\{x_{ij}=1|\beta_j\}=$$
$$=\exp 1{,}7(\theta-\beta_j)/(1+\exp 1{,}7(\theta-\beta_j)) \quad (14)$$

Ф. Бейкер[30] справедливо полагает:

1. Если уровень дифференцирующей способности задания меньше среднего, то график задания больше похож на прямую линию. На рис. 9 представлен графический образ трудного задания с низкой дифференцирующей способностью.

Вероятность правильного ответа испытуемых на это задания принимает значения между 0,2 и 0,8 для всех испытуемых.

2. Если уровень дифференцирующей способности задания выше среднего, то кривая задания похожа на S — образную линию, и она довольно крутая в своей средней части. На рис. 10 представлено лёгкое задание с высокой дифференцирующей способностью.

3. Если уровень трудности задания меньше среднего, то вероятность правильного ответа у большинства испытуемых больше, чем 0,5. На рис. 11 приводится пример задания со средним уровнем дифференцирующей способности и с очень лёгким по уровню трудности. Оно выглядит так (см. рис. 11).

Из кривой видна довольно высокая вероятность правильных ответов у большинства испытуемых. Крутизна заметна только в левой части графика.

4. Если уровень трудности задания выше среднего, то вероятность правильного ответа у большинства испытуемых меньше, чем 0,5.

5. Задания располагаются на оси абсцисс в соответствии с уровнем их трудности, независимо от их уровня диффе-

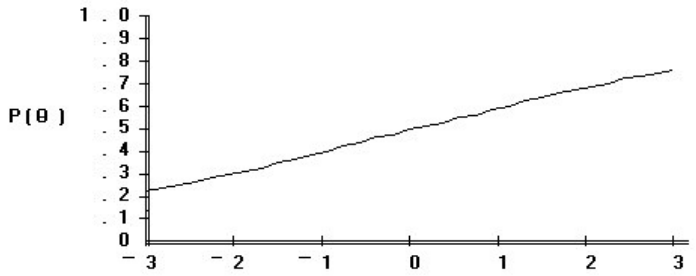

Рис. 9.

30 *Baker Frank.* The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD, 2001.

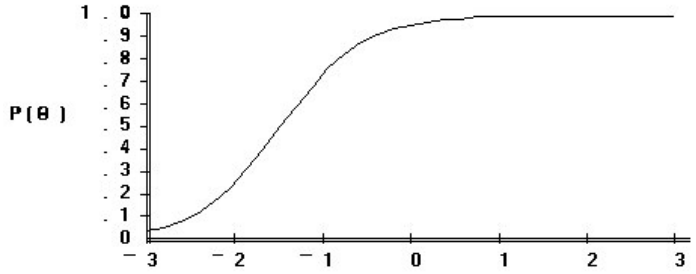

Рис. 10.

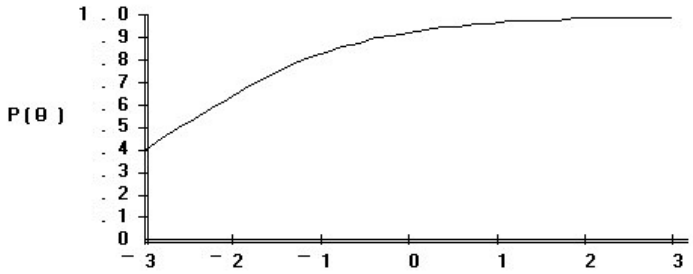

Рис. 11.

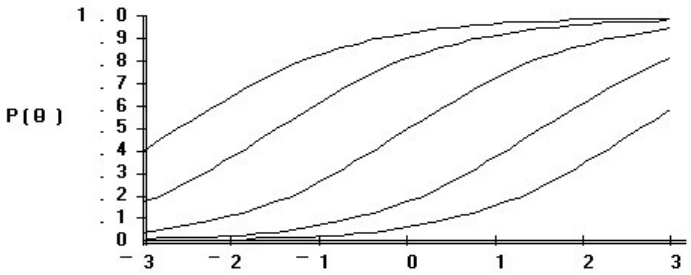

Рис. 12.

ренцирующей способности задания. Данный математический факт подтверждает независимость этих двух характеристик. На рис. 12. представлены задания с одинаковым уровнем дифференцирующей способности заданий, но с различным уровнем трудности[31].

Большинство содержательно осмысленных и хорошо оформленных заданий имеют график, направленный снизу-слева-верх-направо. Как это имеет место на рис. 13.

31

Frank Baker. The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD, 2001.

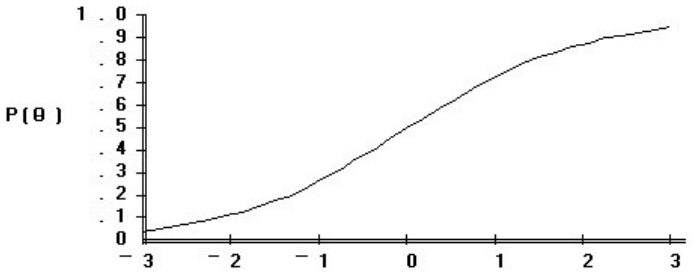

Рис. 13.

Нарушения содержательных, формальных, организационных и этических требований нередко порождают задания с графиками, направленными сверху-слева-вниз-направо. Что указывает на совершенно непедагогический случай — чем выше уровень подготовленности испытуемых, тем ниже оказывается вероятность правильного ответа. Ниже, на рис. 14 представлен случай абсолютно не тестового задания, имеющего параметры $\beta = 0, a = -0{,}75$. Слово «абсолютно» означает, что задание с таким графическим видом ни при каких обстоятельствах не может быть включено в тест. Такой график могут иметь задания, отрицательно коррелирующие с суммой баллов проектируемого теста.

Систематическое и наиболее полное описание возможностей IRT можно найти в работе Hambleton & Swaminathan[32]. Эти авторы справедливо отме-

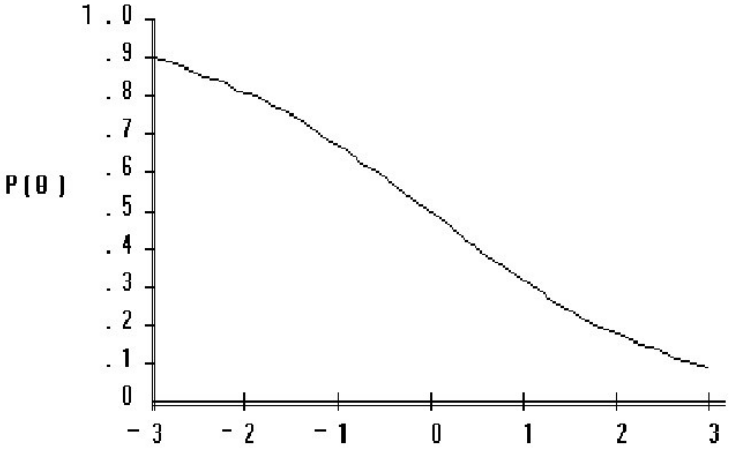

Рис. 14.

32

Hambleton R.K., Swaminathan H. Item Response Theory: Principles and Applications. Boston, 1985. 327 p.

чают позитивную роль классической теории измерений. Их суждения об ограничениях классической теории в переводе на русский языке оказались представлены как недостатки СТТ. Между тем, такие утверждения легко опровергаются множеством авторов, отмечающих высокий вклад СТТ в педагогические измерения и в практику тестирования. Поскольку применение СТТ в практике создания тестов не закончилось, то можно сказать, что каждая теория современна в той мере, в какой она используется для разработки тестов. А для этого применимы все известные теории, в большей или меньшей степени.

Педагогический тест как единство педагогической, статистической и метрической систем

В теории педагогических измерений статистические методы занимают настолько важную роль, что сама эта теория нередко понималась как статистическая теория педагогических измерений. Именно так случилось с самой первой теорией тестов, которая носила абсолютно статистический уклон. Что выражалось в понятиях теории и в методах обоснования качества тестовых результатов. Они были статистическими по своей сути. Начало статистической теории «тестов», как это тогда называлось, положил Чарльз Спирман[33].

Важно отметить, что педагогические вопросы содержания теста и тестовых заданий, вопросы научно педагогической терминологии и формы тестовых заданий статистическая теория тестов не рассматривала. Именно это обстоятельство способствовало попыткам автора этой статьи начать построение основ собственно педагогической теории педагогических измерений[34]. Аналогичный подход был осуществлён в стремлении построить основы теории педагогических заданий, опубликованных в предыдущих номерах нашего журнала.

Таким образом, только спустя столетие стало возможным рассматривать тест как статистическую систему заданий равномерно возрастающей трудности, имеющую в своей основе общий латентный фактор. А также и как содержательно-педагогическую систему заданий равномерно возрастающей трудности по той или иной учебной дисциплине[35]. Именно этот фактор выражает идею измеряемой переменной величины, показателем которой являются задания теста, выступающие в роли эмпирических индикаторов явно ненаблюдаемой величины. В педагогике чаще других в качестве таковой вы-

33

Spearman C. «General intelligence» objectively determined and measured. American Journal of Psychology, 15, 1904, Pp. 201–293.

34

Аванесов В.С. Методологические и теоретические основы тестового педагогического контроля. Дис. ... д-ра пед. наук, С-Пб госуниверситет, 1994 г.; *Аванесов В.С.* Основы педагогической теории измерений // ПИ № 1, 2004 г. С. 15–21 и другие работы автора.

35

На английском языке это выражено так: In test construction, the rule is «all items must be about the same thing, but then be as different as possible»! См.: http://winsteps.com/winman/advice.htm

ступает уровень подготовленности испытуемых. Таким образом, педагогический тест полезно представлять как единство двух систем — педагогической и статистической[36].

С появлением IRT появилась возможность создавать педагогический тест на основе математических моделей измерения. Этот факт открывает дорогу ещё одному возможному названию IRT. На русский язык IRT предлагается переводить как *математическая теория педагогических измерений*.

Таким образом, педагогический тест становится единством трёх, по меньшей мере, систем: педагогической, статистической и математической. Такого рода системное видение педагогического теста помогает лучше понять его состав, структуру, возможности улучшения качества измерений.

Исследование вклада каждой из этих систем в становление научно-обоснованного теста и теорий педагогических измерений — предмет других научных исследований.

36

В лекциях для продвинутого профессорско-преподавательского состава автор этой статьи развивает идеи теста как системы не только педагогической и статистической, но также логической, математической и метрической системы.