ПОСТРОЕНИЕ МОДЕЛЕЙ ПЕДАГОГИЧЕСКИХ ИЗМЕРЕНИЙ

Олег Деменчёнок
Восточно-Сибирский институт МВД России
AskSystem@yandex.ru

Рассмотрены вопросы построения математических моделей педагогических измерений: модели Раша, двух- и трёхпараметрической моделей Item Response Theory (IRT)[1], а также модели с фиксированными промежуточными категориями. Показано, что трёхпараметрическая модель не способна компенсировать влияние угадывания на результат педагогического измерения. Рассмотрены особенности интерпретация результатов тестирования.

Ключевые слова: тест, IRT, математические модели педагогических измерений, вероятность правильного ответа, уровень подготовленности, уровень трудности задания, стандартная ошибка.

Привычные всем оценки появились относительно недавно — в средние века. Оценочная балльная система возникла в немецких схоластических школах средневековья как уступка передовому общественному мнению, выступающему против телесных наказаний[2]. С тех пор оценочная деятельность педагога на удивление мало изменилась. Как и в средние века, оценка зачастую определяется субъективным мнением преподавателя, поскольку отсутствуют чёткие, однозначно определённые границы, разделяющие, например, «хорошо» и «удовлетворительно».

Уровень трудности, количество и последовательность заданий определяются преподавателем интуитивно. Положение усугубляет недостаток методического обеспечения оценочной деятельности. Разделы учебников по педагогике, посвящённые контролю знаний, непропорционально малы и не дают ответа на многие практически значимые вопросы. Например, в известном учебнике по педагогике[3] диагностике обученности и контролю успеваемости уделено всего 28 из 832 страниц (примерно 3,4%).

Однако ситуация меняется к лучшему благодаря тестовым технологиям. Растущая популярность тестов объясняется следующими факторами:

— повышенная точность и обоснованность тестовой оценки;

— снижение затрат учебного времени, особенно в случае компьютерного тестирования;

1 *Аванесов В.С.* переводит IRT на русский язык как «математическая теория измерений (МТИ). См.: Педагогические измерения, № 3, 2007. С. 3.

2 *Ксензова Г.Ю.* Оценочная деятельность учителя. Учебно-методическое пособие. М.: Педагогическое общество России, 1999. 121 с.

3 *Подласый И.П.* Педагогика. Новый курс: Учебник для студ. пед. вузов: В 2 кн. М.: ВЛАДОС, 1999. Кн. 1: Общие основы. Процесс обучения. 576 с. Кн. 2: Процесс воспитания. 256 с.

— исключение влияния субъективного мнения преподавателя;

— сопоставимость результатов освоения учебного материала (т.е. корректность сравнения данных успеваемости по годам, учебным группам и т.д.);

— снижение психологической нагрузки на преподавателей и обучаемых (особенно по сравнению с устной проверкой знаний);

— удобство самоконтроля.

Основные теории педагогических измерений

Повышенная точность и обоснованность тестовой оценки обусловлены тем, что в отличие от других форм контроля знаний тесты имеют научно обоснованную базу. Наиболее распространены две основные теории: статистическая (классическая) теория и Item Response Theory (IRT) — математическая теория измерений (МТИ).

Контроль знаний в рамках классической теории можно сравнить с определением площади некоторой фигуры по методу Монте-Карло[1]. Для применения этого метода фигуру вписывают в другую, известной площади (например, в квадрат), и случайным образом «бросают» точки, подсчитывая число попаданий в фигуру. При достаточно большом числе испытаний отношение числа точек, попавших внутрь фигуры, к общему числу точек, стремится к отношению их площадей. Тогда квадрат — это область, в которой проверяются знания; фигура неизвестной формы и площади — структура и глубина знаний тестируемого, а точки — тестовые задания.

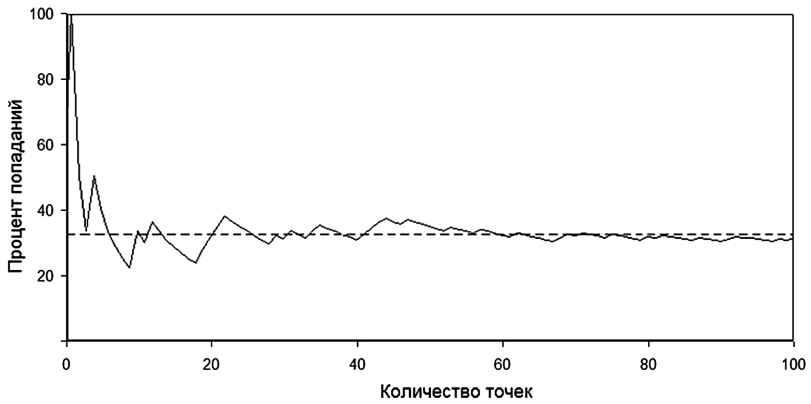

Рис.1. Определение площади фигуры по методу Монте-Карло

Классическая теория имеет неустранимый изъян: результат педагогического измерения зависит от трудности заданий. По любому учебному материалу можно составить задания разного уровня трудности. Если одному и тому же обучаемому предложить выполнить два теста: один — с набором простых заданий, второй —

1 *Гмурман В.Е.* Руководство к решению задач по теории вероятностей и математической статистике: Учеб. пособие для студ. втузов. 3-е изд., перераб. и доп. М.: Высш. школа, 1979. 400 с.

с набором более трудных заданий, то при использовании классической теории результат будет разным.

Item Response Theory (IRT) использует другой подход к измерению подготовленности обучаемых. Для облегчения понимания основной идеи IRT, можно сравнить тестирование с прыжками в высоту. Если планка требований низка, то результат почти всегда удачный; некоторая «средняя» высота преодолевается с переменным успехом, а попытки осилить более высокие уровни, как правило, неудачны.

Способность перешагнуть скамейку не означает готовности к высоким спортивным достижениям. Точно так же способность решать самые простые задачи не может служить подтверждением полного и глубокого усвоения учебного материала. Особенностью IRT является относительная независимость результата педагогического измерения от конкретного подбора заданий. Результаты выполнения тестов с разными по трудности заданиями одним и тем же студентом должны оказаться достаточно близкими.

Применение IRT для измерения знаний

В Item Response Theory принято, что уровень подготовленности обучаемого на момент выполнения теста — величина постоянная. Другими словами: уровень подготовленности обучаемого не должен, в принципе, и зависеть от того, какие именно задания включены в тест.

IRT позволяет найти некий средний уровень подготовленности студента, соответствующий результатам выполнения теста. Если целью педагогического измерения является изучение структуры знаний студентов, то можно провести отдельное тестирование по каждой учебной теме, считая в пределах темы уровень подготовленности студента постоянным.

Таким образом, цель педагогического измерения — это определение точки на шкале уровня подготовленности, соответствующей конкретному испытуемому. В IRT предполагается, что трудность тестовых заданий может быть объективно оценена. Используя специальный математический аппарат, можно совместить шкалы уровня подготовленности испытуемых и уровня трудности тестовых заданий, и измерять эти показатели по единой шкале.

На рис. 2 схематично изображена единая шкала измерения уровня подготовленности обучаемого и уровня трудности тестовых заданий. В порядке увеличения отмечены уровни трудности тестовых заданий β_1, β_2, β_3 и β_4. Пусть уровень подготовленности испытуемого равен θ (см. рис. 2). Тогда 1, 2 и 3 задание окажутся для тестируемого лёгкими, так как уровень подготовленности больше уровней трудности этих заданий. Можно ожидать, что эти задания будут успешно решены. Напротив, сложность четвёртого задания превышает уровень подготовленности испытуемого. Вероятно, испытуемый не сможет справиться с этим заданием. Мы

ожидаем, что он преуспеет на лёгких заданиях и потерпит неудачу на заданиях, которые слишком сложны.

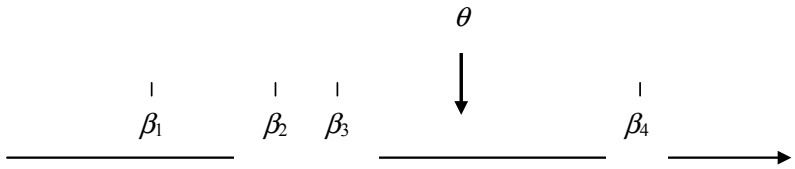

Рис. 2. Единая шкала измерения уровня подготовленности обучаемого и уровня трудности тестовых заданий

Такой подход позволяет выявить зависимость результатов тестирования от набора предложенных тестовых заданий. В известной работе Б.Д. Райта и М.Х. Стоуна[1] приведены удачные примеры, иллюстрирующие эту зависимость (рис. 3 и 4). На рис. 3 смоделирована ситуация, при которой студент с уровнем подготовленности θ выполняет пять разных тестов (каждый тест содержит 8 заданий)[2]:

• первый тест содержит очень лёгкие для этого студента задания, поэтому ожидаемый результат — правильные ответы на все 8 заданий;

• второй тест чересчур труден (уровни трудности всех заданий превосходят уровень подготовленности данного студента), правильных ответов, скорее всего, не будет;

• третий тест менее труден, решить первое тестовое задание студенту вполне по силам;

• четвёртый и пятый тесты легче, причём средняя сложность этих тестов одинакова (показана на рис. 3 пунктирной линией). Однако различается распределение тестовых заданий по уровню трудности. Вследствие этого различается ожидаемый результат тестирования — 7 правильных ответов для четвёртого теста и только 5 правильных ответов для пятого теста.

Получается, что в зависимости от трудности заданий один и тот же студент может набрать от 0 до 8 правильных ответов (доля правильных ответов от 0 до 100%). Хотя подготовленность испытуемого не меняется, при выполнении пяти разных тестов получено пять разных результатов. Очевидно, число правильных ответов зависит от меры трудности заданий теста.

При этом оцениваемый по единой шкале уровень подготовленности студента одинаков и никак не зависит от набора тестовых заданий:

• для третьего теста уровень подготовленности студента θ для четвёртого теста — между уровнями трудности заданий bдля пятого теста — между уровнями трудности β_1 и β_2;

• для четвёртого теста — между уровнями трудности заданий β_7 и β_8;

• для пятого теста — между уровнями трудности β_5 и β_6.

1

Wright B.D., Stone M.H. Best Test Design. Chicago: MESA PRESS. 1979.

2

В реальности редко бывает так, чтобы 8 заданий образовали тест. Для качественного педагогического измерения обычно требуется порядка тридцати заданий возрастающей трудности с общим временем выполнения около 30–40 мин.

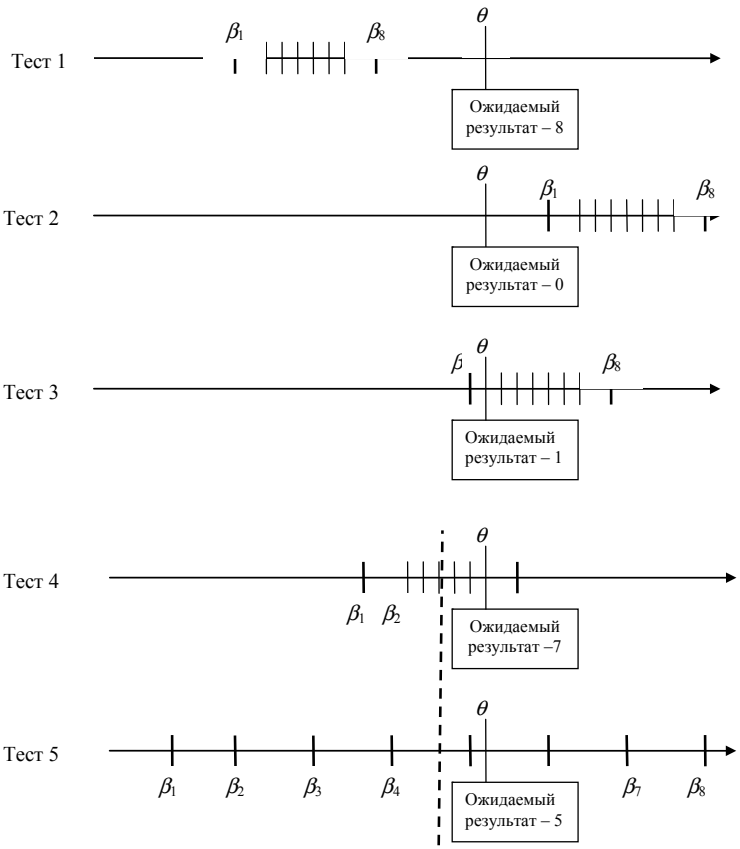

Рис. 3. Зависимость результатов тестирования от набора предложенных тестовых заданий

К сожалению, первый и второй тест малопригодны для измерения уровня подготовленности данного студента:

• все задания первого теста решены правильно, следовательно, уровень подготовленности студента больше трудности самого сложного задания q;

• все задания второго теста решены неверно, значит, уровень подготовленности студента меньше трудности самого лёгкого задания $\theta < \beta_1$. Т.е. уровень подготовленности на единой шкале расположен где-то правее β_8 (где именно — неизвестно);

• все задания второго теста решены неверно, значит, уровень подготовленности студента меньше трудности самого лёгкого задания $\theta < \beta 1$ (θ находится левее β_1).

В этих двух случаях точное значение уровня подготовленности определить невозможно. Всё, что мы можем сделать в таких ситуациях, — это заметить,

что студент, все ответы которого правильны (неправильны), имеет уровень подготовки существенно выше (ниже) уровня трудности теста. Если мы желаем оценить меру подготовленности такого испытуемого, то мы должны будем создать тест, соответствующий его уровню подготовленности.

Ещё одно важное замечание: если тест чрезвычайно лёгок, то высокий процент правильных ответов может быть достигнут при посредственных способностях.

На рис. 4 приведён пример, показывающий как разные тесты помогают различать уровни подготовленности двух обучаемых θ_A и θ_B. Оба студента проверяются пятью разными тестами, в каждом из которых по 8 заданий. Уровни подготовленности и, следовательно, расстояние между испытуемыми по единой шкале остаются неизменными.

Тем не менее ожидаемое различие результатов тестирования изменяется широко:

• Тест 1 составлен так, что уровни трудности всех заданий падают в интервал между уровнями подготовленности студента А и студента В. Можно предположить, что студент А не решит ни одно из заданий, а студент В решит все восемь заданий правильно. Ожидаемое различие результатов: 8 – 0 = 8;

• Тест 2 включает задания, уровни трудности которых значительно ниже подготовленности обоих студентов. Мы ожидаем, что оба студента справятся со всеми заданиями, потому что этот тест слишком лёгок для обоих. Ожидаемое различие результатов: 8 – 8 = 0;

• Тест 3 составлен из очень трудных заданий. Вероятно, оба студента получат 0, потому что этот тест слишком труден для них. Ожидаемое различие: 0 – 0 = 0;

• Тест 4 охватывает большую часть шкалы, но его задания столь разрежены, что уровни подготовленности обоих студентов попадают в промежуток между четвёртым и пятым заданием. Ожидаемое различие результатов: 4 – 4 = = 0;

• Тест 5 содержит два задания, которые разделяют уровни подготовленности студентов. Поэтому ожидаемое различие: 6 – 4 = 2.

Первый тест свидетельствует, что разница в уровнях подготовленности студентов существует, но не даёт возможности оценить величину этой разницы, так как точное значение уровней подготовленности определить невозможно. Непригодны для оценки различия подготовленности второй и третий тесты, поскольку также не могут измерить подготовленность каждого из студентов в отдельности. Четвёртый тест из-за неудачного распределения сложностей заданий даёт лишь грубую оценку подготовленности студентов — интервал между уровнями трудности четвёртого и пятого задания. Следствием недостаточной точности измерений является неразличимость уровней подготовленности студентов. И только пятый позволяет объективно оценить различие уровней подготовленности обучаемых θ_A и θ_B.

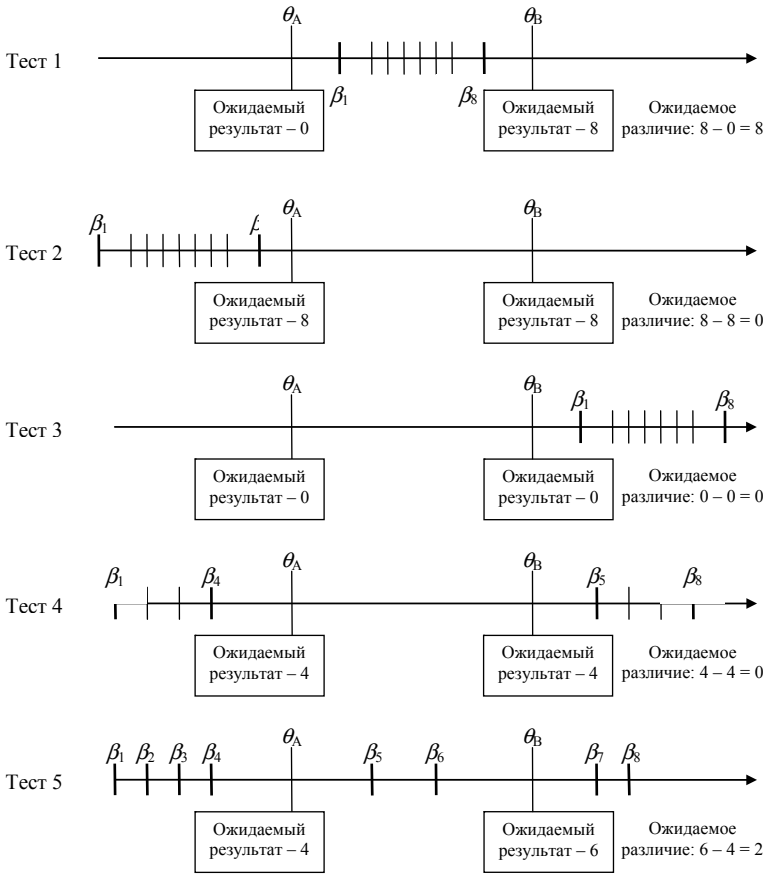

Рис. 4. Зависимость различия результатов тестирования двух студентов от набора предложенных тестовых заданий

Анализ примеров из классической книги[1] доказывает, что педагогические измерения для корректного определения подготовленности испытуемого должны оперировать не только количеством правильных ответов, но и параметрами тестовых заданий. Для этого требуется специализированный математический аппарат, основу которого составляет математическая модель[2]. Необходимо отметить ещё одну особенность IRT — вероятностный характер применяемых моделей. При рассмотрении двух последних примеров не случайно использовались слова «ожидаемый», «вероятно», «можно предположить» и т.п. Сильный студент может допустить случайную ошибку и дать неверный ответ на простое задание. А слабый студент — записать правильный ответ без знания предмета (угадать, списать). Поэтому жёстко определённая, детер-

1 *Wright B.D., Stone M.H.* Best Test Design. Chicago: MESA PRESS. 1979.

2 Математическая модель — приближённое описание какого-либо класса явлений внешнего мира, выраженное с помощью математической символики. Мощный метод познания внешнего мира, позволяет проникнуть в сущность изучаемых явлений (Большая советская энциклопедия, электронная версия. М.: Большая Российская энциклопедия, 2002).

минированная модель малопригодна. Модель должна описывать вероятность правильного ответа, причём вероятность правильного ответа сильного студента должна быть выше.

Модель Г. Раша

Умственная деятельность человека не изучена с такой полнотой, как, например, явления физики или механики. По этой причине нельзя составить модель из готовых формул, подобных закону Ома или Ньютона. Следовательно, модель педагогического измерения представляет собой эмпирическую зависимость. Разработчики эмпирических моделей сталкиваются с двумя важными проблемами:

• отбор параметров, оказывающих наибольшее влияние на рассматриваемый процесс;

• выбор математической формы, адекватно описывающей особенности процесса.

Георг Раш счёл возможным ограничить модель одним единственным параметром — разностью между уровнем подготовленности тестируемого θ и уровнем трудности задания β[1]. Поэтому его модель называют однопараметрической, хотя она оперирует двумя переменными.

Для выбора математической формы модели целесообразно исследовать характер взаимосвязи между параметром модели и вероятностью правильного ответа.

Рассмотрим случай, когда параметр положителен (уровень подготовленности тестируемого больше уровня трудности задания). Логично предположить, что в этом случае вероятность правильного ответа P больше вероятности ошибки, т.е. $P > 0{,}5$. Чем больше разность между уровнем подготовленности и уровнем трудности задания, тем легче испытуемому решить задание. Значит, с увеличением параметра вероятность правильного ответа должна возрастать. Поскольку вероятность не может быть больше единицы, вероятность правильного ответа будет неограниченно приближаться к ней (асимптотически стремиться к единице). Графически зависимость представлена на рис. 5.

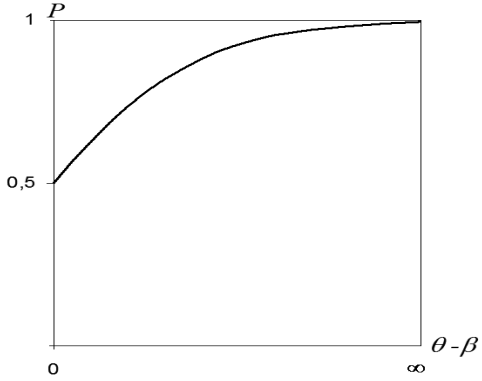

Рис. 5. Зависимость вероятности правильного ответа от разности уровней подготовленности испытуемого и трудности задания при $\theta > \beta$

1 *Rasch G.* Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research, 1960.

Аналогичный ход рассуждений приводит к выводу: если уровень подготовленности меньше уровня трудности задания, то вероятность правильного ответа $P < 0{,}5$; и с увеличением модуля разности P уменьшается, стремясь к нулю. В случае равенства уровня подготовленности и уровня трудности задания вероятность правильного ответа $P = 0{,}5$. Характер взаимосвязи между параметром модели и вероятностью правильного ответа приведён на рис. 6.

Далее нужно подобрать вид математической функции, график которой соответствует рис. 6. Любая математическая функция, график которой подобен кривой рис. 6, может стать основой модели измерения. Г. Раш выбрал для математической модели логистическую функцию:

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

где $e \approx 2{,}72$ — основание натурального логарифма.

Соответственно, модель Раша записывается в виде:

$$P = \frac{e^{\theta-\beta}}{1+e^{\theta-\beta}}. \qquad (2)$$

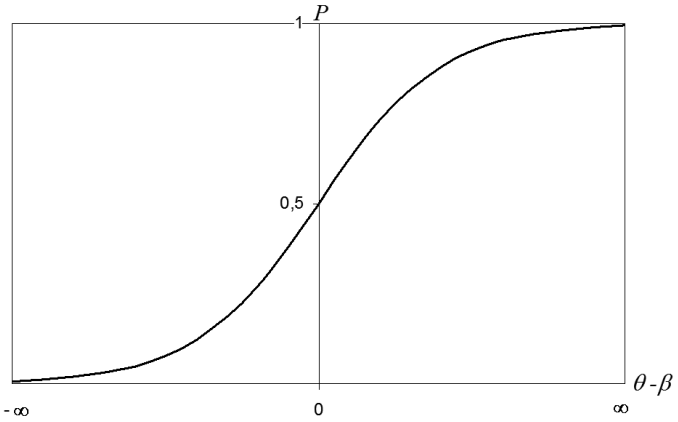

Рис. 6. Характер взаимосвязи между параметром модели Раша и вероятностью правильного ответа

Модель Раша также можно записать в алгебраически эквивалентной форме:

$$P = \frac{1}{1+e^{-(\theta-\beta)}}. \qquad (3)$$

Следует отметить, что логистическая функция — не единственный возможный выбор. Известны модели педагогических измерений, основанные на других функциях: на функции вероятности для нормального распределения[1] (с коэффициентом 1,7 эта зависимость практически совпадает с логистической функцией), на гиперболическом косинусе[2]. Теоретически возможны и другие варианты. Выбор математической формы модели измерения содержит элемент субъективизма. Во всяком случае, нет оснований утверждать, что логистическая

1 *Lord F.M.* A theory of test scores // Psychometric Society, Psychometric Monograph, 1952, № 7.

2 *Andrich D.*, Luo G. A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses // Applied Psychological Measurement, 1993, № 17. P. 253–276.

(или какая-либо другая) функция является единственно возможной или наилучшей.

Как известно, практика — критерий истины. Опыт подтвердил пригодность модели Раша для педагогических измерений. Эта модель с успехом используется уже несколько десятков лет.

Двухпараметрическая модель

Неоднократно предпринимались попытки улучшить модель Раша путём введения дополнительных параметров. Например, наклон кривой в центральной части графика для модели Раша постоянен. Подтверждается ли это опытными данными?

В работе Ф. Бейкера[1] изложен способ проверки уровня крутизны графика задания в центральной его части. Для этого надо:

• оценить уровень подготовленности группы испытуемых (как минимум, несколько десятков человек);

• разделить испытуемых на подгруппы так, чтобы в пределах одной подгруппы уровень подготовленности был примерно одинаков;

• предложить всем испытуемым выполнить одно тестовое задание;

• для каждой подгруппы рассчитать процент правильных ответов и нанести полученные точки на график задания теста (каждой подгруппе соответствует одна точка).

Примерный вид данных такой проверки приведён на рис. 7. Результаты свидетельствуют, что наклон кривой в центральной части графика модели Раша не всегда соответствует данным, полученным опытным путём. Повысить степень близости модели и экспериментальных данных можно за счёт введения коэффициента, регулирующего наклон кривой. Такая двухпараметрическая модель предложена А. Бирнбаумом[2]:

$$P = \frac{e^{\alpha(\theta-\beta)}}{1+e^{\alpha(\theta-\beta)}}, \qquad (4)$$

где a — второй параметр модели, называемый различающей способностью тестового задания.

При $a = 1$ двухпараметрическая модель полностью совпадает с моделью Раша.

Чем больше значение различающей способности a, тем ближе к вертикали центральная часть графика. Так, на рис. 7 сплошная линия соответствует $a = 1$ (модели Раша), а пунктирная линия представляет $a = 1{,}8$. По сравнению с моделью Раша двухпараметрическая модель является более гибкой (настраиваемой), что обеспечивает более высокое качество описания данных тестирования. Именно двухпараметрическая модель рассматривается в качестве основной модели измерения в классическом пособии по IRT Ф. Бейкера[3].

1 *Baker, F.B.* The Basics of Item Response Theory. 2 ed. Hieneman, Portsmouth, New Hempshire, 2001. 172 p.

2 *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.

3 *Baker, F.B.* The Basics of Item Response Theory. 2 ed. Hieneman, Portsmouth, New Hempshire, 2001. 172 p.

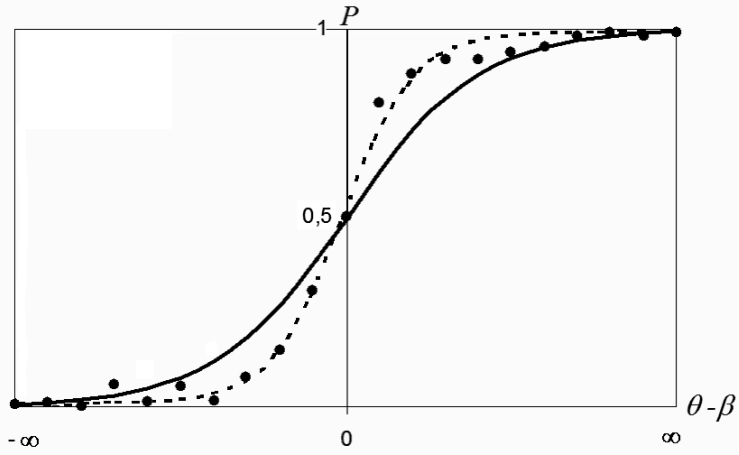

Рис. 7. Сравнение графика модели Раша (сплошная линия)

Анализируя графики на рис. 7, нетрудно заметить, что вероятность правильного ответа $P = 0{,}5$ для обеих моделей соответствует нулевой разнице между уровнем подготовленности и уровнем трудности задания. Поэтому за уровень подготовленности тестируемого принимается тот уровень трудности заданий, который преодолевается этим тестируемым с вероятностью 0,5.

Трёхпараметрическая модель

Широкое использование форм тестовых заданий, допускающих случайное угадывание правильного ответа, привело к идее учёта влияния угадывания.

Особенно высока вероятность случайного угадывания для задания с выбором одного правильного ответа. Например:

СТОЛИЦА БУРЯТИИ

1) Кызыл

2) Улан-Батор

3) Улан-Удэ

4) Усть-Орда

Для задания с выбором одного правильного ответа вероятность случайного угадывания обратно пропорциональна числу предложенных вариантов. При выборе одного из четырёх вариантов ответа вероятность угадывания равна 1/4 или 0,25. Отвечая случайным образом на тест, составленный из таких заданий, можно набрать примерно четверть правильных ответов. Очевидно, что в этом случае вероятность правильного ответа меняется не от нуля до единицы, а от 0,25 до единицы:

• 0,25 — вероятность случайного угадывания правильного ответа;

• в зависимости от параметров задания и уровня подготовленности, испытуемый может добавить к вероятности правильного ответа от нуля до 0,75.

Иллюстрация изложенного приведена на рис. 8.

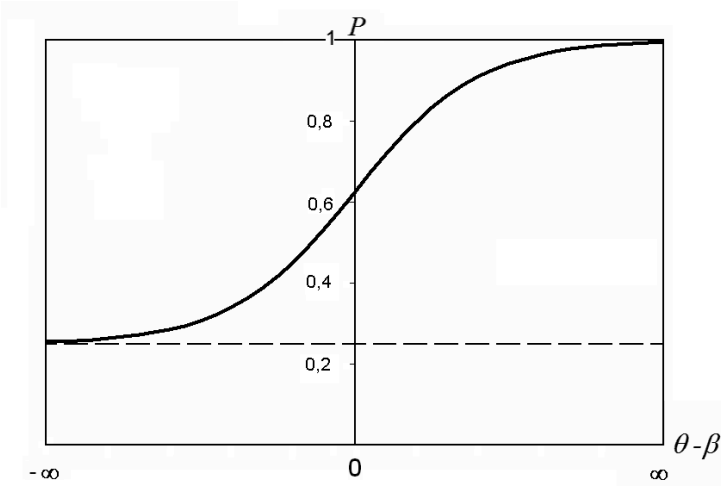

Рис. 8. График вероятности правильного ответа для трёхпараметрической модели

Параметр коррекции на угадывание правильного ответа обозначается символом с. При $c = 0$ график трёхпараметрической модели совпадает с графиком двухпараметрической модели. При увеличении *c* график сжимается так, что его нижняя граница поднимается до уровня $P = c$ (пунктирная линия на рис. 8). При этом смещается точка, соответствующая нулевой разнице между уровнем подготовленности и уровнем трудности задания (точка пересечения с вертикальной осью). Поскольку точка делит высоту графика пополам, то вероятность правильного ответа в этой точке равна:

$$P_0 = c + \frac{1-c}{2} = \frac{1+c}{2}. \qquad (5)$$

В точке P_0 уровень подготовленности равен уровню трудности задания. Поэтому за уровень подготовленности испытуемого принимается уровень трудности заданий, который преодолевается этим испытуемым с вероятностью $P = P_0$ (в других моделях — с вероятностью P = 0,5).

Математически трёхпараметрическая модель формулируется в виде[1]:

$$P = c + (1-c)\frac{e^{\alpha(\theta-\beta)}}{1+e^{\alpha(\theta-\beta)}}. \qquad (6)$$

Способна ли трёхпараметрическая модель компенсировать влияние угадывания на результат педагогического измерения?

Для проверки проведём небольшой эксперимент. Пусть 15 студентов выполнят тест из 16 заданий, в каждом из которых нужно выбрать один из четырёх ответов. Предположим, что 8 заданий каждый студент решил правильно, а ответы на остальные задания даёт случайным образом, пытаясь угадать. Наиболее вероятно угадывание двух ответов: $8 \times 0{,}25 = 2$. Но некоторые угадают больше двух правильных ответов, а некоторые — меньше.

1 *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.

Вероятность угадывания b из m правильных ответов можно найти по формуле Бернулли[1]:

$$P_m(b) = C_m^b p_1^b (1-p_1)^{m-b}. \quad (7)$$

где $C_m^b = \frac{m!}{b!(m-b)!} = \frac{m(m-1)...(m-(b-1))}{b!}$ — число сочетаний, p_1 — вероятность случайного угадывания правильного ответа.

Например, вероятность угадать ровно один ответ в восьми заданиях:

$$P_8(1) = C_8^1 p_1^1 (1-p_1)^{8-1} = \frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{1} 0,25(1-0,25)^7 \approx 0,27.$$

Результаты расчётов сведены в табл. 1. Расчёт по формуле Бернулли показывает, что двое из 15 студентов не угадают ни одного ответа; четверо угадают один ответ; пятеро — два ответа; трое — три ответа; один — четыре ответа. Пять или более ответов не угадает никто.

Таблица 1

Влияние случайного угадывания на результат педагогического измерения при использовании трёхпараметрической модели

Количество угаданных ответов, b	Вероятность угадывания, $P_8(b)$	Число студентов, угадавших b ответов, 15×	Уровень подготовленности студентов, q
0	0,100113	2	-2,6
1	0,266968	4	-0,7
2	0,311462	5	-0,2
3	0,207642	3	1,4
4	0,086517	1	5
5	0,023071	0	–
6	0,003845	0	–
7	0,000366	0	–
8	0,000015	0	–

Обработка результатов такого тестирования проведена с помощью компьютерной программы Estimate3PL (сайт www.asksystem.narod.ru). Уровень подготовленности студентов, решивших по 8 заданий, но угадавших разное количество правильных ответов, отличается разительно (см. табл. 1). Это доказывает, что трёхпараметрическая модель не может компенсировать влияние угадывания.

Полученные данные свидетельствуют, что угадывание правильных ответов способно внести существенные искажения в результат педагогического измерения. Для устранения этого негативного влияния рекомендуется использовать задания, которые благодаря своей форме устойчи-

1 *Вентцель Е.С.* Теория вероятностей. М.: Высшая школа, 2001. 576 с.

вы к угадыванию правильного ответа. В.С. Аванесов рекомендует переходить от заданий с выбором одного правильного ответа к заданиям с выбором нескольких правильных ответов[1], с числом ответов до 8–10. Для таких заданий вероятность случайного угадывания правильного ответа не превышает 0,001. В этом случае ни один из 15 студентов рассмотренного выше примера не смог бы угадать даже один ответ. Соответственно, отпадает необходимость учёта или компенсации влияния угадывания. Устойчивы к угадыванию задания на восстановление последовательности и соответствия, а также задания с кратким свободным ответом[2].

Модель с произвольными промежуточными категориями выполнения тестовых заданий

В рассмотренных моделях IRT принято считать, что результат выполнения задания x принимает только одно из двух значений: $x = 1$ при правильном ответе, $x = 0$ при неправильном ответе. Частично правильные ответы в основных моделях IRT не учитываются, т.е. отсутствует градация степени правильности ответа. Если результат выполнения задания может принимать больше двух значений (например, 0, 1 и 2), то обычно рекомендуют модели с произвольными и с фиксированными промежуточными категориями выполнения тестовых заданий[3].

Модель с фиксированными промежуточными категориями выполнения тестового задания подразумевает одинаковое количество категорий в ответах, а также неизменность трудности шагов в задании. Задания такой формы часто используются при проведении опросов, анкетировании, диагностировании различных проблем.

Более универсальна модификация модели Раша с произвольными промежуточными категориями выполнения тестового задания, известная в англоязычной литературе как Partial credit model (PCM). Именно эта модель рекомендуется для педагогических измерений[4]. Модель с произвольными промежуточными категориями описывает ситуацию, когда за выполнение тестового задания испытуемый может получить от 0 до x_{max} баллов. Эту модель можно получить математически из модели Раша. Приведём вывод этой формулы.

Чтобы достичь результата x, испытуемый должен последовательно преодолеть x шагов, за правильное выполнение каждого из которых он получает 1 балл; причём сложность выполнения каждого шага различна. Рассмотрим случай, когда за выполнение задания можно получить от 0 до 2 баллов. Очевидно, что набрать 2 балла можно только при условии, что первый шаг выполнен правильно. Тогда согласно теореме умножения вероятностей[5] вероятность одновременного появления двух событий (первое событие — успешное выполнение первого шага, второе — второго шага) равна произведению вероятности одного из них на условную вероятность другого

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

2 *Деменчёнок О.Г.* Влияние угадывания на значение тестового балла: корректировать или устранять? //Педагогические измерения, 2007, № 1. С. 56–70.

3 *Wright B.D., Mok M.M.C.* Introduction to Rasch Measurement. Chicago: University of Chicago, 2004.

4 *Wright B.D., Masters G.N.* Rating Scale Analysis: Rasch Measurement. Chicago: Mesa Press, 1982. 204 p.

5 *Вентцель Е.С.* Теория вероятностей. М.: Высшая школа, 2001. 576 с.

при условии, что первое событие произошло:

$$\pi_2 = P_1 \cdot P_2, \quad (8)$$

где P_1 и P_2 — вероятность успешного выполнения соответственно первого и второго шага; π_2 — вероятность получить 2 балла (во избежание путаницы, вероятность успешного выполнения i-го шага будем обозначать P_i, а вероятность получить i баллов — π_i)

Очевидно, что вероятность правильного выполнения первого шага равна сумме вероятностей получить 1 и 2 балла:

$$P_1 = \pi_1 + \pi_2. \quad (9)$$

Подставим (9) в выражение (8) и зададим P_2 по формуле (3):

$$\pi_2 = (\pi_1 + \pi_2),$$
$$\pi_2 - \pi_2 \cdot P_2 = \pi_2 \cdot P_2,$$
$$\pi_2 = \pi_2 \cdot P_2 :(1 - P_2),$$

$$\pi_2 = \frac{\pi_1}{1+e^{-(\theta-\beta_2)}} : \left(1 - \frac{1}{1+e^{-(\theta-\beta_2)}}\right) =$$

$$= \frac{\pi_1}{1+e^{-(\theta-\beta_2)}} : \left(\frac{1+e^{-(\theta-\beta_2)}}{1+e^{-(\theta-\beta_2)}} - \frac{1}{1+e^{-(\theta-\beta_2)}}\right),$$

$$\pi_2 = \frac{\pi_1}{1+e^{-(\theta-\beta_2)}} : \frac{e^{-(\theta-\beta_2)}}{1+e^{-(\theta-\beta_2)}} =$$

$$= \frac{\pi_1}{e^{-(\theta-\beta_2)}} = \pi_1 \cdot e^{\theta-\beta_2}, \quad (10)$$

где β_2 — уровень трудности второго шага.

Аналогично можно получить вероятность результата выполнения задания с одним баллом:

$$\pi_1 = \pi_0 \cdot e^{\theta-\beta_1}, \quad (11)$$

где π_0 — вероятность получить 0 баллов; β_2 — уровень трудности первого шага.

Так как получение 0, 1 и 2 баллов составляет полную группу возможных событий, то сумма их вероятностей равна единице:

$$\pi_0 + \pi_1 + \pi_2 = 1,$$
$$\pi_0 + \pi_0 \cdot e^{\theta-\beta_1} + \pi_0 \cdot e^{\theta-\beta_1} \cdot e^{\theta-\beta_2} = 1,$$
$$\pi_0\left(1 + e^{\theta-\beta_1} + e^{\theta-\beta_1} \cdot e^{\theta-\beta_2}\right) = 1,$$
$$\pi_0 = \frac{1}{1+e^{\theta-\beta_1} + e^{\theta-\beta_1} \cdot e^{\theta-\beta_2}}. \quad (12)$$

Подставив (12) в формулы (10) и (11), получим:

$$\pi_1 = \frac{e^{\theta-\beta_1}}{1+e^{\theta-\beta_1} + e^{\theta-\beta_1} \cdot e^{\theta-\beta_2}}, \quad (13)$$

$$\pi_2 = \frac{e^{\theta-\beta_1} \cdot e^{\theta-\beta_2}}{1+e^{\theta-\beta_1} + e^{\theta-\beta_1} \cdot e^{\theta-\beta_2}}. \quad (14)$$

Обобщая полученные выражения, найдём вероятность достижения тестируемым результата x_{ij} (т.е. того, что тестируемый i выполнит ровно x шагов в задании j)[1]:

$$\pi_{ijx} = \frac{\prod_{k=0}^{x} e^{\theta_i-\beta_{jk}}}{\sum_{l=0}^{x_{\max}} \prod_{k=0}^{l} e^{\theta_i-\beta_{jk}}}$$

или

$$\pi_{ijx} = \frac{e^{\sum_{k=0}^{x}(\theta_i-\beta_{jk})}}{\sum_{l=0}^{x_{\max}} e^{\sum_{k=0}^{l}(\theta_i-\beta_{jk})}}, \quad (15)$$

где k = 0, 1 ...x_{ij} ...x_{max} — количество шагов; xmax — максимально возможное количество баллов за задание;

$$\beta_{j0} = 0, \sum_{n=0}^{0}(\theta_i - \beta_{jk}) = 0$$

Уравнение (15) является математическим выражением модели

1 *Wright B.D., Masters G.N.* Rating Scale Analysis: Rasch Measurement. Chicago: Mesa Press, 1982. 204 p.

с произвольными промежуточными категориями выполнения тестового задания. По сути, эта модель представляет собой модификацию модели Раша для заданий с градацией степени правильности ответа.

Педагогические измерения на основе моделей IRT

Применение простых математических моделей для физических явлений, таких, например, как закон Ома, проблем не вызывает. Чтобы найти неизвестную величину (например, силу тока), достаточно подставить в формулу известные значения (скажем, напряжение и сопротивление).

Модели педагогических измерений, представляющие собой вероятностные модели, значительно сложнее. В идеальном случае математическая модель полностью соответствует экспериментальным данным. Однако в действительности результаты измерения всегда содержат погрешности. Это создаёт определённые трудности интерпретации результатов вычислений.

Задача педагогического измерения с помощью выбранной математической модели сводится к выбору параметров модели, при которых результаты конкретного тестирования и результаты применения математической модели наилучшим образом совпадают. Другими словами: нужно так подобрать значения уровней подготовленности испытуемых и параметры тестовых заданий, чтобы расчёт по выбранной модели оказался максимально близок к результатам выполнения теста. Решение этой задачи во многом зависит от того, что именно мы условимся считать наилучшим, т.е. от критерия оптимальности. Теоретическую базу всех методов и приёмов, положенных в основу построения эмпирических математических моделей, составляет метод максимального (наибольшего) правдоподобия. В общем виде метод максимума правдоподобия можно сформулировать так: наилучшее описание явления — то, которое даёт наибольшую вероятность получить в результате измерений именно те значения, которые и были фактически получены[1].

В формальной записи этот метод может быть представлен в виде максимума произведения вероятностей всех наблюдаемых независимых событий[2]:

$$p_1 \cdot p_2 \cdot \ldots \cdot p_N = \prod_{i=1}^{N} p_i \to \max, \quad (16)$$

По методу максимального правдоподобия наилучшим будет признан тот набор значений θ и β, при котором произведение расчётных вероятностей фактически полученных результатов максимально:

$$\prod_{i=1}^{n}\prod_{j=1}^{m} P_{ij}(x_j) \to \max, \quad (17)$$

где $P_{ij}(x_{ij})$ — вероятность результата x_{ij} при решении i-м тестируемым j-го задания, найденная по выбранной математической модели; n — количество испытуемых; m — число тестовых заданий.

Критерий оптимальности также может быть выражен по методу наименьших квадратов. Требо-

1 *Львовский Б.Н.* Статистические методы построения эмпирических формул. М.: Высшая школа, 1988.

2 *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998.

вание наилучшего согласования расчётных и экспериментальных данных сводится к тому, чтобы сумма квадратов отклонений между ними обращалась в минимум:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij}-P_j\left(x_j\right)\right)^2 \to \min . \quad (18)$$

Аналитически выражения (17) и (18) неразрешимы относительно искомых параметров подготовленности тестируемых и тестовых заданий, что затрудняет подбор параметров. Поэтому задача оценки параметров модели IRT сводится к математической задаче численного поиска экстремума (максимума или минимума) функции. Известны варианты упрощённого решения, например алгоритм PROX[1], базирующийся на предположении о нормальности распределения уровня подготовленности испытуемых и уровня трудности заданий.

Хотя упрощённое решение задачи определения уровня подготовленности тестируемого по процедуре PROX может быть найдено без использования компьютера, такой подход в настоящее время не рационален.

Во-первых, исчезла необходимость упрощения решения — компьютеры доступны, и использование в решении моделей IRT не создаёт никаких проблем современным компьютерам.

Во-вторых, при упрощённом решении снижается точность педагогических измерений — реальное распределение уровней подготовленности и уровней трудности всегда будет отличаться от идеального нормального распределения. А поправочные коэффициенты всегда будут недостаточно точны для полной коррекции результатов вычислений.

В-третьих, ручной счёт слишком трудоёмок и ненадёжен — требуется провести большой объём вычислений, при котором ошибки, просчёты, неверная запись полученных результатов практически неизбежны.

Применение IRT фактически невозможно без использования компьютера на этапе оценки параметров математических моделей. Какие программы можно использовать?

Во-первых, универсальные программы для обработки числовой информации: электронные таблицы (например, Microsoft Excel), а также математические пакеты MathCad, Mathematica, Maple, MathLab и др. Основное преимущество универсальных программ — гибкость (возможна тонкая настройка поиска решения, расчёт и визуализация любых дополнительных параметров и т.д.). Недостаток универсальных программ — существенная трудоёмкость подготовительных работ (нужно не только ввести данные, но и составить все необходимые формулы, правильно задать ограничения и параметры поиска решения) и сложность для неспециалистов. Поэтому их нельзя рекомендовать для широкого использования в практической оценочной деятельности.

Во-вторых, для нахождения параметров математических моделей можно использовать специализированные компьютерные программы, например:

2

Wright, B.D. and Douglas, G.A. Better procedures for sample-free item analysis. Research Memorandum № 20, Statistical Laboratory, Department of Education, University of Chicago, 1975.

1. Программы Winsteps и Facets Д.М. Линека (John M. Linacre), сайт www.winsteps.com. Эти программы могут проводить расчёты на основе модели Раша и модели с произвольными промежуточными категориями выполнения тестовых заданий (PCM)[3].

2. Программа RUMM2020, созданная в Rumm Laboratory Pty Ltd (сайт www.rummlab.com), обеспечивает расчёт параметров модели Раша, двух- и трёхпараметрической моделей, а также модели с произвольными промежуточными категориями.

3. Написанная автором этой статьи программа Estimate3PL (сайт www.asksystem.narod.ru) выполняет подбор параметров модели Раша, двух- и трёхпараметрической модели.

Для использования этих программ достаточно ввести результаты тестирования и указать параметры работы. Определение уровней подготовленности испытуемых и параметры тестовых заданий выполняется автоматически.

Интерпретация результатов тестирования

Интерпретация тестовых результатов требует понимания некоторых важных особенностей IRT.

Во-первых, необходимо учитывать точность данных педагогического измерения. Выданные компьютерной программой значения уровней подготовленности испытуемых и параметры заданий — это наиболее вероятные значения этих неизвестных величин. Истинные же значения этих величин неизвестны. Поэтому нельзя определить погрешность путём сопоставления истинного и расчётного значения уровня подготовленности конкретного испытуемого. Вероятные величины ошибок можно оценить по формулам, полученным А. Бирнбаумом[1] (чем меньше σ, тем тем выше точность определения искомого параметра):

$$\sigma_{\theta_i} = \frac{1}{\sqrt{\sum_{j=1}^{m} \alpha_j^2 \left[\left(\frac{1-P_{ij}}{P_{ij}} \right) \left(\frac{P_{ij}-c_j}{1-c_j} \right)^2 \right]}}, \quad (19)$$

$$\sigma_{\beta_j} = \frac{1}{\sqrt{\sum_{i=1}^{n} \alpha_j^2 \left[\left(\frac{1-P_{ij}}{P_{ij}} \right) \left(\frac{P_{ij}-c_j}{1-c_j} \right)^2 \right]}}, \quad (20)$$

где σ_{θ_i} — стандартная ошибка уровня подготовленности i-го испытуемого; σ_{β_j} — стандартная ошибка уровня трудности j-го задания.

Формулы (19)–(20) справедливы для трёхпараметрической модели. При $c_j = 0$ и $\alpha_j = 1$ — модели Раша.

Все рассмотренные выше специализированные программы обеспечивают расчёт стандартных ошибок. Стандартная ошибка связана с погрешностью измерения зависимостью[2]:

$$\Delta\theta = \varepsilon \cdot \sigma_\theta,$$

где ε — аргумент функции Лапласа, при котором она равна половине заданного значения вероятности α (табличная величина, например: α = 0,68 соответствует ε = 1,0; α = 0,90 соответствует ε = 1,65; α = 0,997 соответствует ε = 3,0 и т.д.).

1 *Partchev I.* A visual guide to item response theory. Jena: Friedrich-Schiller-Universitat, 2004. 61 p.

2 *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998.

Найденный уровень подготовленности испытуемого нужно рассматривать не как конкретное числовое значение, а как интервал вида $\theta \pm \Delta\theta$. Например уровень подготовленности тестируемого, равный единице, при $\sigma_\theta = 0{,}3$ может трактоваться так:

— с вероятностью 68% уровень подготовленности этого испытуемого находится в интервале $\theta = 1 \pm 1 \cdot 0{,}3$ (или 0,7 ... 1,3);

— с вероятностью 90% — $\theta = 1 \pm 1{,}65 \cdot 0{,}3$ (или 0,5 ... 1,5);

— с вероятностью 99,7% — $\theta = 1 \pm 3 \cdot 0{,}3$ (или 0,1 ... 1,9);

— с вероятностью 100% — $\theta = 1 \pm 5 \cdot 0{,}3$ (или –0,5 ... 2,5);

Во-вторых, сопоставимость результатов тестирования обеспечивается тем, что параметры тестовых заданий оцениваются совместно, по единой шкале. Например, уровень подготовленности $\theta = 1$ при выполнении испытуемым теста с набором простых заданий равноценен результату $\theta = 1$ при выполнении теста с набором более трудных заданий только в том случае, если параметры всех тестовых заданий оценены совместно. Т.е. предварительно нужно собрать статистику ответов испытуемых на задания обоих тестов, объединить результаты и с помощью компьютерной программы получить параметры заданий, выраженные в единой шкале.

Без выполнения этого условия результаты тестирования несопоставимы. Поэтому некорректно сравнивать результаты тестирования, например по математике, если использованы разные, никак не связанные между собой тесты. По этой же причине невозможно прямое сопоставление уровней подготовленности студентов по разным учебным дисциплинам или годам обучения.

Третья особенность — сложность перевода уровня подготовленности обучаемого в педагогическую оценку. Тестовый балл, полученный при обычном тестировании, обычно достаточно просто ассоциируется с оценкой. Например, 8 баллов из 10 возможных или доля правильных ответов — 80%, вероятнее всего соответствуют оценке «хорошо». В IRT оценки уровня подготовленности обучаемого могут принимать как положительные, так и отрицательные значения. Известно, что большее значение уровня подготовленности соответствует лучшему знанию предмета. Однако непонятно, с какой педагогической оценкой общепринятой шкалы «неудовлетворительно — удовлетворительно — хорошо — отлично» можно сопоставить уровень подготовленности, равный, например, — 0,5 или 2,6.

Поскольку однозначных критериев перевода количественного показателя — уровня подготовленности тестируемых в показатель качества — в педагогическую оценку нет, то целесообразно привлечение мнения квалифицированного специалиста — преподавателя. Для этого некоторые работы тестируемых оценивает преподаватель с выставлением педагогической оценки. Зная оценки и уровни подготовленно-

сти тестируемых, несложно найти интервалы уровня подготовленности, соответствующего каждой оценке.

Мировой и отечественный опыт убедительно доказывает, что острота большинства проблем контроля знаний существенно снижается при внедрении тестовой формы контроля. Поэтому разработка технологии тестового контроля, в частности IRT, адаптированной для практического применения в оценочной деятельности преподавателя, представляется важной задачей, имеющей как теоретическое, так и практическое значение. Вопрос о целесообразности применения тестовой формы контроля каждый преподаватель решает сам; «право на свободу выбора... методов оценки знаний обучающихся» ему предоставлено федеральным законом РФ «Об образовании»[1].

5 Закон РФ от 10 июля 1992 года № 3266-1 «Об образовании».