Теория

АНАЛИЗ КАЧЕСТВА ПЕДАГОГИЧЕСКОГО ИЗМЕРЕНИЯ

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

Рассмотрены критерии качества педагогического измерения. Показано, что необходимыми условиями для признания практической ценности результатов педагогического измерения является справедливость положенного в основу математической модели предположения о нормальном законе распределения ошибок и значимости коэффициента детерминации между модельными и реально полученными результатами. Предложенная технология анализа результатов педагогического измерения реализована в виде компьютерной программы.

Ключевые слова: тест, Item Response Theory (IRT)[1], адекватность модели, погрешность измерения, коэффициент детерминации, проверка нормальности распределения.

1 На русский язык IRT В.С.Аванесов переводит как математическую теорию измерений (МТИ). См.: Педагогические измерения. № 3. 2007. С. 3.

Математическая модель педагогического измерения

Современный этап развития педагогических измерений невозможно представить без математической модели процесса взаимодействия испытуемых различного уровня подготовленности с заданиями, различающимися уровнями трудности.

Математическая модель — это «эквивалент объекта, отражающий в математической форме важнейшие его свойства — законы, которым он подчиняется, связи, присущие составляющим его частям, и т.д.»[2]. Математическое моделирование — процесс построения и изучения математических моделей реальных процессов и явлений. Любые науки, использующие математический аппарат, по сути занимаются математическим моделированием: заменяют реальный объект его моделью и затем изучают последнюю. Педагогические измерения не являются исключением: создаётся математическая модель, описывающая уровень подготовленности испытуемых, а затем производится обработка результатов тестирования с использованием математического уравнения модели.

Математическая модель — это одна из форм идеализации изучаемого явления, основанная на некоторых предположениях. Результаты математического моделирования практически всегда отличаются от экспериментальных данных, что приводит к необходимости анализа качества полученных результатов и их пригодности для практического применения.

Проблема качества педагогического измерения

Степень несоответствия фактических и расчётных данных может быть различной. В одних случаях разница между ними не очень заметна и может оказаться чисто случайной. В других расхождения значительны. Отсюда возникает задача оценки качества педагогического измерения, установления того, в каких случаях и с какой степенью вероятности можно считать разницу между фактическими данными и теоретически ожидаемыми достоверной и, наоборот, когда её следует считать несущественной, находящейся в пределах случайности.

Адекватность модели (от лат. adaequatus — приравненный, вполне соответствующий, верный, точный) — это способность модели отражать заданные свойства изучаемого процесса с приемлемой точностью. Если фактически полученные данные совпадают с теоретически ожидаемыми, по модели,

2 *Самарский А.А., Михайлов А.П.* Математическое моделирование. Идеи. Методы. Примеры. 2-е изд., испр. М.: Физматлит, 2001. 320 с.

данными, то это может быть достаточным основанием для принятия предлагаемой модели и признания инструментальной валидности результатов педагогического измерения. Но если фактические данные не согласуются с теоретическими, то возникает большое сомнение в практической ценности полученных результатов.

Вместе с тем адекватность модели определяется не только степенью её соответствия реальному объекту, но также целям исследования и назначению модели[3]. Критерии адекватности служат мерилом того, насколько эффективно модель справляется с возложенными на неё задачами. Если, например, модель используется для прогнозирования, то главным критерием адекватности будет качество прогноза.

Для моделей педагогических измерений основным критерием качества является точность[4] и надёжность измерения уровня подготовленности испытуемых и параметров тестовых заданий.

Стандартный подход к оценке адекватности модели основан на сопоставлении результатов моделирования с экспериментальными данными (например, адекватность моделирования прочности каких-либо объектов можно проверить, сравнив расчётные значения показателей прочности с результатами натурных испытаний). Этот стандартный подход непригоден для педагогических измерений. Знания (учебные достижения, умения, навыки, профессиональная компетентность и т.п.) в общем случае не поддаются непосредственному измерению. Поэтому процесс измерения принципиально отличен от измерения наблюдаемых (например, физических) величин. Если наблюдаемые величины (например, длину объекта) находят путём прямого сопоставления с эталонными мерами, то уровень знаний определяется по результатам косвенных измерений.

Результатом моделирования являются не истинные значения уровней подготовленности испытуемых и параметры тестовых заданий, а их приближённые оценки. К сожалению, целый ряд факторов существенно осложняет анализ качества результатов педагогического измерения:

- погрешность измерения может быть оценена только косвенными методами, потому что для выявления ошибок аппроксимации расчётные значения не с чем сравнить;
- оценка стандартных ошибок измерения проводится с использованием математической модели. Эта оценка лишена смысла в том случае, когда положенные в основу модели предположения не выполняются;

3

Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998. 1022 с.

4

Точность измерения — характеристика измерения, отражающая степень близости его результатов к истинному значению измеряемой величины. Чем меньше результат измерения отклоняется от истинного значения величины, то есть чем меньше его погрешность, тем выше точность (Большая советская энциклопедия, электронная версия. М.: Большая Российская энциклопедия, 2002).

• перепроверка результатов педагогического измерения путём расчёта по другим математическим моделям невозможна, так как получаемый при этом новый набор приближённых оценок также будет отличаться от истинных значений, а величина отличия неизвестна;

• изменение исходных данных (например, исключение из рассмотрения ответов по отдельным тестовым заданиям) приводит к вариации результатов (оценок уровней подготовленности).

Этапы проверки пригодности математической модели педагогических измерений

С учётом изложенного, для педагогического измерения автор полагает целесообразным проведение полного анализа пригодности математической модели педагогического измерения, который должен дать ответ на следующие вопросы:

• достижима ли требуемая точность педагогического измерения?

• пригодна ли выбранная математическая модель для обработки результатов тестирования?

• пригодны ли для практического использования результаты педагогического измерения?

Первый этап анализа — оценка точности измерения

Для моделей педагогических измерений главным критерием является точность измерения уровня подготовленности испытуемых и точность оценки параметров тестовых заданий.

Незнание истинных значений измеряемых параметров не даёт возможность рассчитать погрешность измерения. Однако можно оценить стандартную ошибку, которая характеризует погрешность измерения, вызванную действием случайных факторов. Случайные факторы (фрагментарность знаний испытуемого, случайный выбор ответов, ошибки ввода данных; ошибки, вызванные неверным истолкованием условия задания и т.п.) в каждом из отдельных измерений действуют непредвиденным образом то в сторону уменьшения, то в сторону увеличения результатов. Чем сильнее действие случайных факторов, тем больше отклонение расчётного значения относительно ожидаемого, а точность измерения — ниже.

Для трёх базовых моделей IRT[5,6] предельные, т.е. максимально возможные значения стандартных ошибок нахождения уровней подготовленности обучаемых и трудности тестовых заданий[7] рассчитываются по формулам:

5 *Baker F.B.* The Basics of Item Response Theory. 2 ed., ERIC Clearinghouse on Assessment and Evaluation, Madison, Wisconsin, 2001. 172 p.

6 *Деменчёнок О.Г.* Компьютерная программа для подбора параметров основных моделей IRT // Педагогические измерения. № 2. 2008.

7 *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, MA: Addison-Wesley Publishing, 1968. pp. 397–472.

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

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

где $\sigma_{\theta i}$ — стандартная ошибка уровня подготовленности i-го испытуемого; $\sigma_{\beta j}$ — стандартная ошибка уровня трудности j-го задания; P_{ij} — вероятность правильного ответа i-го тестируемого на j-е задание; m — количество тестовых заданий; n — число испытуемых; c_j — параметр коррекции на угадывание правильного ответа в j-м задании.

Ошибку измерения $\Delta\theta$ с нужной вероятностью можно найти из уравнения[8]:

$$\Delta\theta = \varepsilon \cdot \sigma_\theta, \quad (3)$$

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

Полученные с помощью модели уровни подготовленности испытуемых (уровни трудности заданий) являются приближёнными оценками этих величин. Истинные значения неизвестны, но с вероятностью α находятся в пределах доверительного интервала $\theta \pm \Delta\theta$.

Например, уровень подготовленности испытуемого $\theta = 1$ при $\sigma_\theta = 0{,}5$ означает:

- с вероятностью 68% уровень подготовленности находится в интервале $q = 1 \pm 1 \cdot \sigma_\theta$ (или 0,5 ... 1,5);
- с вероятностью 90% $q = 1 \pm 1{,}65 \cdot \sigma_\theta$ (или 0,175 ... 1,825);
- с вероятностью 99,7% $q = 1 \pm 3 \cdot \sigma_\theta$ (или –0,5 ... 2,5);
- с вероятностью 99,99% $q = 1 \pm 4 \cdot \sigma_\theta$ (или –1... 3).

Поэтому первым этапом проверки адекватности модели должна быть оценка доверительных интервалов для результатов педагогического измерения. Если погрешности окажутся слишком велики, то цель педагогического измерения не может быть достигнута, а результаты измерения непригодны для практического использования.

Таким образом, цель этого этапа — установить возможность достижения нужной точности измерения. Если ответ положительный, то далее надо проверить пригодность и качество модели. Если нужная точность недостижима, то продолжать анализ не имеет смысла.

Второй этап анализа — проверка пригодности модели

Результаты педагогического измерения и их стандартные

8

Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998. 1022 с.

ошибки рассчитываются по выбранной модели IRT и не могут быть проверены другими способами. Очевидно, что ошибочная модель приводит к ошибочным результатам. Поэтому требуется проверять правомерность модели в каждом конкретном случае её применения.

Модели IRT основаны на предположениях (допущениях):
- вероятность правильного ответа определяется разностью между уровнем подготовленности испытуемого и уровнем трудности задания;
- действие неучтённых в модели факторов пренебрежимо мало или взаимно компенсируется;
- ошибки являются случайными величинами.

Если модель адекватно описывает экспериментальные данные, то ошибки должны не противоречить этим предположениям.

Следствием выполнения стандартных допущений является:
- независимость ошибок модели;
- постоянство дисперсии ошибок модели для всех интервалов наблюдаемых данных;
- нормальный закон распределения ошибок.

В работе G. Karabatsos, посвящённой анализу пригодности модели, перечислен ряд факторов, приводящих к нарушению стандартных допущений[9]:
- несанкционированный доступ испытуемых к правильным ответам на все или отдельные тестовые задания (списывание, использование запрещённых справочных материалов, подкуп должностных лиц и т.д.);
- попытки угадывания ответа;
- ошибки при вводе ответа (психологическое напряжение, усталость, потеря концентрации внимания могут привести к техническим ошибкам при вводе правильного ответа);
- случайный выбор ответов без попыток угадывания;
- творческое осмысление задания (например, в задаче по физике студент может учесть, что Земля имеет форму не шара, а эллипсоида).

Составленный G. Karabatsos список трудно назвать исчерпывающим. Автор этой статьи полагает, что причинами нарушения стандартных допущений также могут стать:
- некорректность тестовых заданий (технические ошибки при вводе текста задания, фактические ошибки, двусмысленность формулировки, неверное значение эталона правильного ответа и т.д.);
- фрагментарность знаний испытуемого (даже в рамках одной темы отдельные учебные вопросы студент может знать лучше других, что не учитывается моделью измерения и приводит к росту погрешности модели).

9

Karabatsos G. Comparing the Aberrant Response Detection Performance of Thirty-Six Person-Fit Statistics // Applied measurement in education, 2003, №16(4), 277–298 p.

Действие указанных факторов приводит к нарушению теоретических предположений. Эффективным средством обнаружения отклонений от стандартных предположений является анализ погрешностей, позволяющий выявить основные виды нарушений стандартных предположений:

1) отклонение распределения ошибок модели от нормального закона распределения. Выявление такого отклонения означает неслучайность ошибок модели, что ставит под сомнение её адекватность;

2) выбросы — экспериментальные данные, резко отличающиеся от расчётных, вероятность которых крайне низка. В этом случае обычно проводят повторный эксперимент. Если результат повторяется, то следует исследовать природу его существования и уточнить модель, в противном случае выброс не принимают во внимание.

Критерии оценки закона распределения ошибок являются мерой справедливости допущений, положенных в основу математической модели.

Для проверки гипотезы о законе распределения погрешностей обычно используют критерий согласия Пирсона, также называемый критерием χ^2 — хи-квадрат (χ — греческая буква «хи»), критерий Колмогорова-Смирнова или значения коэффициентов асимметрии и эксцесса распределения данных.

Известные теоретики IRT Wright B.D. и Stone M.H полагают, что полный анализ модели педагогического измерения должен включать оценку степени соответствия данных теоретическим предположениям[10]. Они считают, что анализ на основе критерия χ^2 должен быть проведён для каждого испытуемого и каждого задания.

Идея критерия χ^2 состоит в оценке отклонений распределения экспериментальных данных от нормального распределения.

Чтобы найти значение критерия χ^2, нужно сгруппировать теоретические данные в интервальный ряд (желательно не менее 7 интервалов), причём в каждом интервале должно оказаться не менее 5 значений. Затем подсчитать эмпирические и теоретические частоты и вычислить статистику χ^2 по формуле[11]:

$$\chi^2 = \sum_{i=1}^{c} \frac{(A_i - E_i)^2}{E_i}, \qquad (4)$$

где c — число интервалов; A_i — эмпирическая частота (отношение числа результатов тестирования, попавших в интервал i, к общему количеству результатов тестирования); E_i — теоретическая частота (относительное количество расчётных значений, попавших в тот же интервал).

Для нормального распределения вычисленное значение χ^2

10 *Wright B.D., Stone M.H.* Best Test Design. Chicago: Mesa Press. 1979.

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

не превышает критического значения $\chi^2_{крит}$, которое выбирается из соответствующей таблицы[12].

При проверке гипотез методами математической статистики необходимо иметь в виду уровень значимости[13], который обычно выбирается из ряда 0,05; 0,025; 0,01 и 0,001. Различные значения уровня значимости имеют свои достоинства и недостатки. Меньшие значения дают бóльшую уверенность в том, что нормальное распределение ошибок не соблюдается, но при этом есть больший риск необоснованно признать распределение ошибок нормальным (ошибка второго рода или ложноотрицательное решение).

Например, вывод о несоответствии распределения ошибок нормальному закону при уровне значимости 0,05 означает:

- с вероятностью не менее 95% отличие распределения ошибок от нормального распределения достоверно;
- с вероятностью, не превышающей 5%, распределение всё-таки может быть нормальным.

Если положенное в основу математической модели предположение о нормальном законе распределения ошибок не выполняется, то нет оснований признавать адекватность модели. Такую модель следует забраковать, а полученные с её помощью результаты педагогического измерения — признать не имеющими практической ценности.

Проверку по критерию χ^2 следует провести для каждого уровня подготовленности испытуемого и уровня трудности задания. Если пригодность модели не будет подтверждена, то следует признать, что оценивание данного студента (тестового задания) в рамках выбранной модели невозможно.

В случае непригодности модели для тестового задания можно рекомендовать:

1) проанализировать формулировку задания. Обнаруженные ошибки исправить, после чего задание вновь может быть включено в тест для апробации[14];

2) исключить из рассмотрения ответы испытуемых на это задание и повторить обработку результатов тестирования.

В случае непригодности модели для испытуемого можно рекомендовать:

1) если позволяют требования к результатам педагогического измерения, принять меньшее значение уровня значимости. Возможно, на меньшем уровне значимости модель может быть признана пригодной;

2) оценить уровень подготовленности этого испытуемого в индивидуальном порядке (например, путём устного или письменного опроса);

3) при компьютерном тестировании — выдать дополни-

12 *Гмурман В.Е.* Руководство к решению по теории вероятностей и математической статистике. Учебное пособие. М.: Высшее образование, 2009. 404 с.

13 Уровень значимости — степень риска, заключающаяся в том, что исследователь может сделать неправильный вывод об ошибочности статистической гипотезы на основе экспериментальных данных (ошибка первого рода или ложноположительное решение).

14 *Baker F.B.* The Basics of Item Response Theory. 2 ed., ERIC Clearinghouse on Assessment and Evaluation, Madison, Wisconsin, 2001. 172 p.

тельные тестовые задания (с увеличением числа выполненных заданий растёт статистическая значимость результатов, и повторная проверка по критерию χ^2 может показать пригодность модели).

Итак, цель этого этапа — установить пригодность модели измерения. Если модель пригодна, то далее проверяем её качество. В противном случае — признаём, что оценивание данного студента (тестового задания) в рамках выбранной модели IRT невозможно.

Третий этап анализа — проверка качества модели измерения

В математическом моделировании именно близость экспериментальных и расчётных данных является основным критерием качества модели. Как правило, адекватность практически обосновывается достаточной степенью совпадения значений параметров модели и объекта.

В IRT такие критерии, хотя и упоминаются в отдельных работах, широкого распространения не получили. Так, Wright B.D. и Masters G.N. используют коэффициент надёжности модели и на 113 странице своей книги[15] приводят формулу этого коэффициента, которая полностью совпадает с формулой известного в статистике коэффициента детерминации.

Коэффициент детерминации R^2 (квадрат множественного коэффициента корреляции R) является универсальным и общепризнанным показателем близости расчётных и экспериментальных данных[16]:

$$R^2 = 1 - \frac{\sigma_\Delta^{\,2}}{\sigma^2}, \qquad (5)$$

где σ_Δ — средняя квадратическая ошибка (т.е. стандартное отклонение между расчётом по модели и наблюдаемыми данными); σ — среднее квадратическое отклонение экспериментальных данных.

Коэффициент детерминации R^2 — это статистический показатель, отражающий объясняющую способность модели и представляющий собой ту долю дисперсии (вариации) результатов наблюдений, которая объясняется уравнением математической модели. Он может принимать только положительные значения от 0 до 1.

Если $R^2 = 0$, то связь между экспериментальными данными и результатами моделирования отсутствует, и вместо модели можно с таким же успехом использовать среднее арифметическое наблюдаемых значений. $R^2 = 1$ соответствует идеальному совпадению экспериментальных (наблюдаемых) и теоретических (расчётных) данных. Чем ближе значение коэф-

15 *Wright B.D., Masters G.N.* Rating scale analysis. Chicago: Mesa Press, 1982.

16 *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998. 1022 с.

фициента детерминации к единице, тем ближе модель к эмпирическим наблюдениям. Например, R^2 = 0,8 означает, что модель объясняет изменение экспериментальных данных на 80%, а оставшиеся 20% приходятся на случайные ошибки или неучтенные в модели факторы.

Коэффициент детерминации R^2 можно рассматривать и как показатель надёжности результатов педагогического измерения. Надёжность характеризуется долей устойчивой части дисперсии. Чем выше значение R^2, тем эта доля больше. Следовательно, с увеличением R^2 снижается влияние случайных факторов, а надёжность педагогического измерения возрастает.

В двумерной корреляционной модели коэффициент детерминации равен квадрату коэффициента корреляции r:

$$R^2 = r^2, \qquad (6)$$

где $r = M\left[\frac{F - M(F)}{\sigma_f} \cdot \frac{Y - M(Y)}{\sigma_y}\right]$, (7)

M — обозначение математического ожидания; σ_f и σ_y — средние квадратические отклонения модели и экспериментальных данных; Y — множество экспериментальных данных; F — соответствующее множество значений модели.

Для практических расчётов коэффициента корреляции используется формула[17]:

$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})\cdot(f_i - \overline{f})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(f_i - \overline{f})^2}}, \qquad (8)$$

где $\overline{y}$ и $\overline{f}$ — средние значения экспериментальных и расчётных данных.

Формула коэффициента корреляции для случая, когда результат выполнения тестового задания x оценивается дихотомически — 1 («правильно») или 0 («неправильно»), принимает вид:

$$r = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \overline{x})\cdot(P_{ij} - \overline{P})}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \overline{x})^2}\cdot\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(P_{ij} - \overline{P})^2}}, \qquad (9)$$

где x_{ij} — результат j-го задания выполнения задания i-м тестируемым; $\overline{P}$ — средняя вероятность правильного ответа.

В уравнении (9) расчётное значение заменено вероятностью правильного ответа. Обоснуем эту замену. Как известно, для дискретной случайной величины, заданной значениями $x_1, x_2, ..., x_n$ и соответствующими этим значениям вероятностями $P_1, P_2, ..., P_n$, среднее значение (математическое ожидание) определяется формулой[18]:

$$M(x) = x_1P_1 + x_2P_2 + \cdots + x_nP_n. \qquad (10)$$

17 Теория вероятностей и математическая статистика: Учебное пособие для вузов /Под ред. Мхитаряна В.С. М.: Маркет ДС, 2007. 240 с.

18 Там же: *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998. 1022 с.

При дихотомическом оценивании вероятность того, что $x = 1$ равна P, а вероятность $x = 0$ равна $1 - P$. Тогда по формуле (10):

$$M(x) = 1 \cdot P + 0 \cdot (1 - P) = P.$$

Значит, для моделей педагогического измерения, основанных на дихотомическом оценивании, расчётное значение равно вероятности правильного ответа.

Коэффициент корреляции или его квадрат — коэффициент детерминации R^2 могут быть рассчитаны для каждого испытуемого

$$r_{\theta i} = \frac{\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)\cdot(P_{ij} - \bar{P}_i)}{\sqrt{\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)^2}\cdot\sqrt{\sum_{j=1}^{m}(P_{ij} - \bar{P}_i)^2}}, \quad (11)$$

и для каждого тестового задания

$$r_{\beta j} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)\cdot(P_{ij} - \bar{P}_j)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\cdot\sqrt{\sum_{i=1}^{n}(P_{ij} - \bar{P}_j)^2}}, \quad (12)$$

где $\bar{x}_i$ — среднее значение результата выполнения тестовых заданий i-м испытуемым; $\bar{x}_j$ — среднее значение результата выполнения j-го задания всеми испытуемыми; $\bar{P}_i$ и $\bar{P}_j$ — соответствующие средние значения вероятности правильного ответа.

Полученные значения покажут, насколько полно модель измерения объясняет вариацию результатов тестирования отдельно для каждого испытуемого и для каждого тестового задания.

Формулы (9) и (11–12) достаточно универсальны и могут быть использованы для случая, когда результат выполнения задания оценивается несколькими баллами. При этом следует заменить P_{ij} значениями, найденными по формуле (10).

Важный этап анализа — проверка существенности отличия от нуля коэффициента детерминации. Этим проверяется значимость построенной модели. Если окажется, что коэффициент детерминации существенно не отличается от нуля, то можно сделать вывод о нулевой «объясняющей» способности модели (т.е. модель объясняет наблюдаемые данные не лучше их среднего арифметического), надёжность измерения недостаточна (вариация ответов полностью зависит от случайных факторов).

Статистически незначимое отклонение от нуля коэффициента является нарушением предположения о модели измерения, согласно которому вероятность правильного ответа определяется разностью между уровнем подготовленности испытуемого и уровнем трудности задания. В этом случае вероятность

правильного ответа определяется действием неучтённых в модели факторов. Автор полагает, что при отсутствии значимости коэффициента детерминации R^2 использование полученных с помощью модели результатов неправомерно.

Значимость коэффициента детерминации проверяется с помощью *F*-критерия Фишера:

$$F = \frac{R^2}{1-R^2} \cdot \frac{l-k-1}{k}, \quad (13)$$

где $l = m \cdot n$ — количество результатов выполнения тестовых заданий, k — число входящих в модель переменных (переменными величинами считаются уровни подготовленности испытуемых, уровни трудности заданий, а также уровни различающей способности заданий).

F-критерий Фишера показывает, во сколько раз математическая модель описывает фактические данные лучше, чем среднее арифметическое. Если полученное значение *F* окажется больше критического $F_{крит}$, то на принятом уровне значимости можно сделать вывод статистической значимости отличия R^2 от нуля и, следовательно, значимости полученной модели. Поэтому условием адекватности модели и надёжности полученных с её помощью результатов является подтверждение значимости коэффициента детерминации[19].

При статистически незначимом отклонении R^2 от нуля ($F \le F_{крит}$) не подтверждается одно из базовых предположений модели IRT о том, что вероятность правильного ответа определяется разностью между уровнем подготовленности испытуемого и уровнем трудности задания. $R^2 = 0$ означает, что разность между уровнем подготовленности испытуемого и уровнем трудности задания не влияет на вероятность правильного ответа. Результаты педагогического измерения в этом случае можно обосновать только влиянием случайных и неучтённых в модели факторов, что лишает их какой-либо практической ценности. Следует признать качество полученной модели и надёжность результатов измерения неудовлетворительными.

Качественную интерпретацию коэффициента детерминации и коэффициента корреляции можно ориентировочно дать по шкале Чеддока[20] (табл. 1).

Весьма высокая Цель третьего этапа анализа — проверка адекватности модели измерения. Если окажется, что отличие коэффициента детерминации от нуля статистически не значимо, то следует признать адекватность полученной модели и надёжность результатов измерения неудовлетворительными.

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

20 Электронный учебник по статистике. М.: StatSoft; http://www.statsoft.ru/home/portal/glossary/GlossaryTwo/M/MultipleR.htm

Таблица 1

Количественная мера тесноты связи		Качественная интерпретация адекватности модели
r	R^2	
0,1–0,3	0,01–0,09	Слабая
0,3–0,5	0,09–0,25	Умеренная
0,5–0,7	0,25–0,49	Заметная
0,7–0,9	0,49–0,81	Высокая
0,9–0,99	0,81–0,99	Весьма высокая

Возможности расчёта в Microsoft Excel

На сложность вычисления описанных показателей качества модели можно не обращать особого внимания — функции для их расчёта имеются в готовом виде во многих математических и статистических пакетах компьютерных программ.

Так, в приложении Excel из среды Microsoft Office для проверки закона распределения погрешностей по критерию согласия Пирсона (критерию χ^2) используется встроенная функция ХИ2ТЕСТ. Эта функция вычисляет вероятность совпадения наблюдаемых (фактических) значений и теоретических (гипотетических) значений. Если вычисленная вероятность ниже уровня значимости (например, 0,05), то нулевая гипотеза отвергается и утверждается, что отклонения (погрешности модели) не соответствуют нормальному закону распределения[21]. Если вычисленная вероятность близка к единице, то можно говорить о высокой степени соответствия погрешности модели нормальному закону распределения.

Функция имеет следующие параметры:

ХИ2ТЕСТ(Фактический_интервал; Ожидаемый_интервал)

где Фактический_интервал — интервал данных, который содержит результаты наблюдений (в нашем случае — результаты выполнения тестовых заданий); Ожидаемый_интервал — интервал данных, который содержит теоретические (расчётные) значения для соответствующих наблюдаемых.

Значение коэффициента корреляции легко вычисляется при помощи функции КОРРЕЛ. Синтаксис этой функции:

КОРРЕЛ(Массив1; Массив2),

где Массив1 — интервал ячеек со значениями результатов выполнения тестовых заданий, Массив2 — интервал ячеек с расчётными значениями.

21

Гельман В.Я. Решение математических задач средствами Excel. СПб.: Питер, 2003. 240 с.

Критическое значение *F*-критерия Фишера можно найти без специальных таблиц, воспользовавшись функцией FРАСПОБР:

FРАСПОБР (вероятность;степени_свободы1; степени_свободы2)

где вероятность — уровень значимости (например, 0,05 или 0,01); степени_свободы1 — количество входящих в модель переменных; степени_свободы2 — количество экспериментальных данных, уменьшенное на число переменных и единицу.

Пример анализа адекватности модели педагогического измерения

Исходные данные для анализа (рис. 1) предоставлены В.С. Аванесовым. Таблица (или матрица) организована так, что столбцы — это результаты выполнения заданий (всего 10 заданий), а строки — результаты тестируемых (13 студентов). Сразу уточним, что столь малый объём данных не характерен для реального тестирования. Зато такой объём удобен для рассмотрения технологии анализа адекватности модели.

Для проведения анализа модели педагогического измерения использовалась бесплатная компьютерная программа Estimate3PL автора (сайт www.asksystem.narod.ru). Программа реализована в среде Microsoft Excel, поэтому доступны все функции и возможности этой электронной таблицы. Дополнительно включена возможность обработки результатов тестирования в соответствии с базовыми моделями IRT и анализ адекватности модели. После ввода исходных данных нажимаем кнопку *Поиск решения* (рис. 1), выбираем режим работы (рис. 2) и получаем результаты (рис. 3).

Уровни подготовленности испытуемых θ и соответствующие стандартные ошибки σ записываются в столбцах справа от исходных данных (столбцы M и N на рис. 3). Уровни трудности заданий β, их стандартные ошибки σ, дифференцирующая способность заданий a записываются в строках ниже исходных данных (строки 15–17 на рис. 3). Самой большой оказалась стандартная ошибка $\sigma = 1{,}19$ для уровня трудности первого задания (ячейка B16 на рис. 3). Это означает, что с вероятностью 68% $\beta_1 = -3{,}52 \pm 1{,}19$ (т.е. находится в интервале от –4,71 до –2,33). Точность педагогического измерения невысока. Это можно объяснить малым количеством выполненных тестовых заданий. Однако будем считать точность проведённого

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		Задание 1	Задание 2	Задание 3	Задание 4	Задание 5	Задание 6	Задание 7	Задание 8	Задание 9	Задание 10				
2	Студент 1	1	1	1	0	1	1	1	1	1	1				
3	Студент 2	1	1	0	1	1	1	1	1	1	0				
4	Студент 3	1	1	1	1	0	1	1	0	1	0				
5	Студент 4	1	1	1	1	0	1	0	1	0	0				
6	Студент 5	1	1	1	1	1	1	0	0	0	0				
7	Студент 6	1	1	1	1	0	0	1	0	0	0				
8	Студент 7	1	1	0	1	1	0	1	0	0	0				
9	Студент 8	1	1	1	1	1	0	0	0	0	0			Поиск Решения	
10	Студент 9	1	0	1	0	1	1	0	0	0	0				
11	Студент 10	0	1	1	0	0	0	0	1	0	1				
12	Студент 11	1	1	1	0	0	0	0	0	0	0				
13	Студент 12	1	1	0	0	0	0	0	0	0	0				
14	Студент 13	1	0	0	0	0	0	0	0	0	0				

Рис. 1. Исходные данные на листе Microsoft Excel

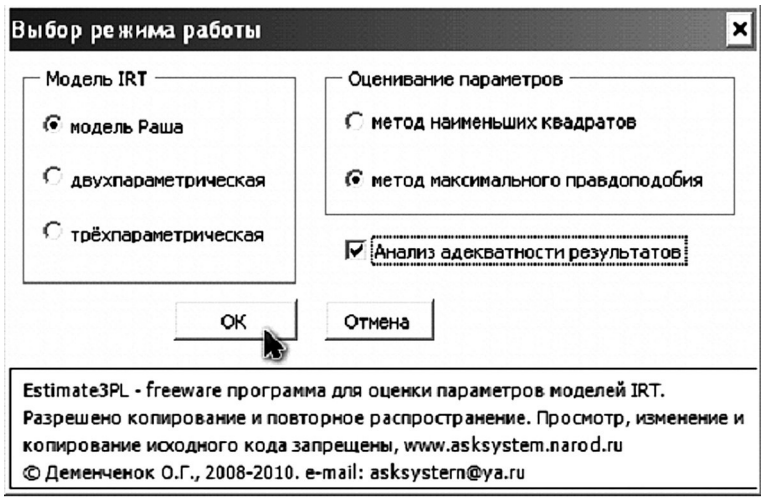

Рис. 2. Выбор режима работы

измерения удовлетворительной и продолжим рассмотрение результатов анализа модели.

Программа Estimate3PL выводит результаты анализа адекватности модели педагогического измерения в целом (т.е. для всей группы испытуемых):

- проверка закона распределения погрешностей по критерию согласия Пирсона (критерию хи-квадрат): вероятность случайности ошибок равна 0,999996. Вычисленная вероятность многократно выше уровня значимости 0,05, что

	A	B	C	D	E	F	G	H	I	J	K	M	N	O	P
1		Задание 1	Задание 2	Задание 3	Задание 4	Задание 5	Задание 6	Задание 7	Задание 8	Задание 9	Задание 10	θ	σ	R^2	P (хи-квадрат)
2	Студент 1	1	1	1	0	1	1	1	1	1	1	2,99	1,13	**0,03**	0,999
3	Студент 2	1	1	0	1	1	1	1	1	1	0	2,02	0,89	**0,11**	0,992
4	Студент 3	1	1	1	1	0	1	1	0	1	0	1,31	0,8	**0,29**	0,972
5	Студент 4	1	1	1	1	0	1	0	1	0	0	0,69	0,77	0,45	0,966
6	Студент 5	1	1	1	1	1	1	0	0	0	0	0,69	0,77	0,71	0,992
7	Студент 6	1	1	1	1	0	0	1	0	0	0	0,09	0,78	0,55	0,963
8	Студент 7	1	1	0	1	1	0	1	0	0	0	0,09	0,78	**0,28**	0,916
9	Студент 8	1	1	1	1	1	0	0	0	0	0	0,09	0,78	0,67	0,984
10	Студент 9	1	0	1	0	1	1	0	0	0	0	-0,54	0,82	**0,17**	0,825
11	Студент 10	0	1	1	0	0	0	0	1	0	1	-0,54	0,82	**0,01**	**0,002**
12	Студент 11	1	1	1	0	0	0	0	0	0	0	-1,28	0,9	0,84	0,997
13	Студент 12	1	1	0	0	0	0	0	0	0	0	-2,19	1,02	0,89	0,999
14	Студент 13	1	0	0	0	0	0	0	0	0	0	-3,44	1,24	0,77	1
15	β	-3,52	-2,45	-1,13	-0,17	0,27	0,27	0,72	1,2	1,74	2,41				
16	σ	1,19	0,92	0,73	0,67	0,66	0,66	0,68	0,71	0,77	0,88				
17	*a*	1	1	1	1	1	1	1	1	1	1				
18	R^2	**0,004**	0,34	**0,19**	0,35	**0,23**	0,52	0,45	0,33	0,76	**0,21**				
19	P(хи-квадрат)	1	1	0,993	0,991	0,938	0,981	0,968	0,788	0,998	0,071				
20	**Модель объясняет 48,9% вариации исходных данных. Время поиска решения 0 с**														

Рис. 3. Результаты расчёта параметров модели

подтверждает пригодность модели измерения. Ошибки с высокой вероятностью обусловлены случайными факторами;

- коэффициент детерминации равен 0,4887745 (модель на 48,9% объясняет вариацию ответов), что по шкале Чеддока соответствует заметной адекватности модели. Надёжность результатов не высока;
- *F*-критерий Фишера равен 4,4063, что больше критического значения при уровне значимости 0,05 ($F_{крит}$ = 1,631635). Адекватность модели подтверждена.

Рассматривая всю матрицу ответов, можно сказать, что достигнуто удовлетворительное качество модели (недостаток в том, что модель не объясняет 51,1% дисперсии исходных баллов). Однако конечными результатами моделирования являются оценки значений параметров испытуемых и тестовых заданий. Поэтому и для них следует проверить адекватность модели измерения.

Программа Estimate3PL рассчитывает коэффициенты детерминации R^2 для каждого испытуемого (столбец O) и для каждого тестового задания (строка 18). Также проверяется закон распределения погрешностей по критерию хи-квадрат и выводятся значения P (хи-квадрат) — вероятности того, что

расхождение между исходными данными и моделью вызвано случайными факторами (столбец P и строка 19). Так, для 11-го студента $R^2 = 0{,}84$ (модель на 84% объясняет вариацию ответов этого студента). По шкале Чеддока адекватности модели можно считать весьма высокой; проверка по *F*-критерию Фишера при уровне значимости 0,05 подтверждает значимость (пригодность) модели. P (хи-квадрат) = 0,997 (т.е. вероятность нормального распределения ошибок равна 99,7%), что подтверждает пригодность модели педагогического измерения для этого студента.

Для большей наглядности проведём графический анализ качества модели. На графике для одиннадцатого студента (рис. 4) видно, что результат тестирования и модель действительно хорошо согласованы. В полном соответствии с теорией, студент даёт правильные ответы на лёгкие задания (т.е. задания 1–3, уровень трудности которых меньше уровня подготовленности этого студента $\theta = -0{,}54$), а при решении сложных заданий терпит неудачу. Адекватность модели весьма высока.

Ответы четвёртого студента (рис. 5) меньше согласуются с теоретическими предположениями: он ошибается на двух лёгких заданиях и правильно решает более сложное. Однако эти задания близки по уровню трудности ($\beta_5 = \beta_6 = 0{,}27$, $\beta_7 = 0{,}72$) к уровню подготовленности этого студента

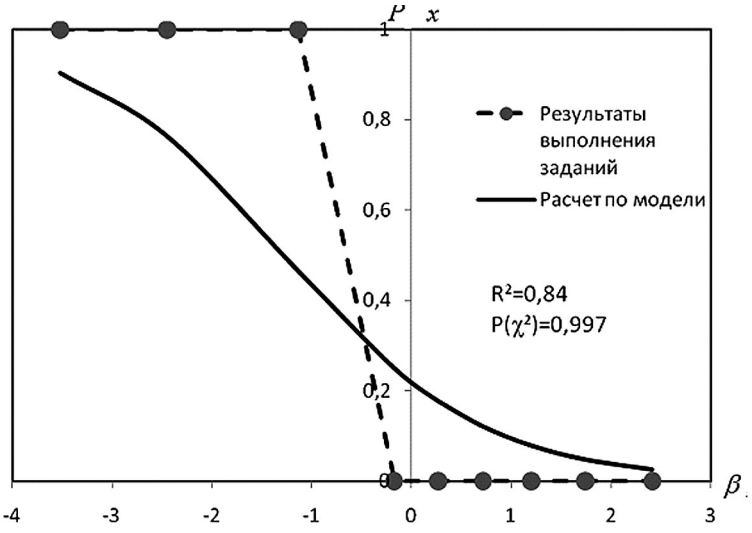

Рис. 4. Графический анализ модели для одиннадцатого студента

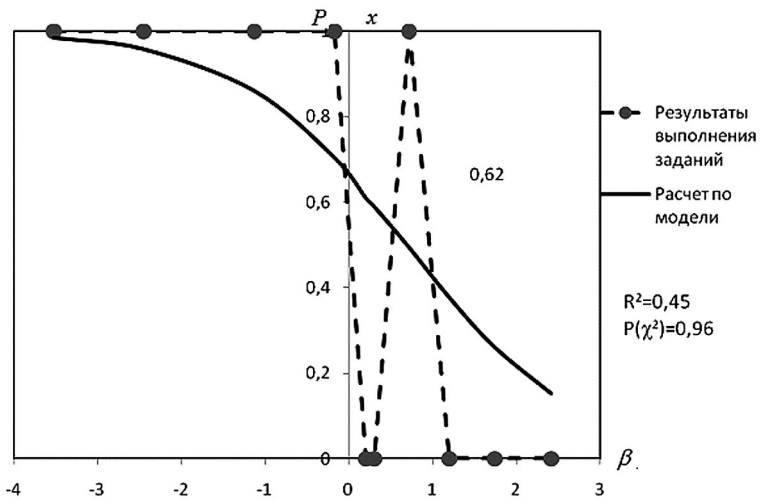

Рис. 5. Графический анализ модели для четвёртого студента

θ_4 = 0,45. Поэтому отклонения ответов от прогнозируемых с вероятностью 96% можно считать случайными (P(хи-квадрат) = 0,96). R^2 = 0,45 (модель объясняет 45% дисперсии ответов), адекватность модели достаточна.

Для удобства анализа программа Estimate3PL выделяет цветом значения R^2 и P(хи-квадрат), которые не подтверждают значимость и пригодность модели (на рис. 3 такие значения выделены жирным шрифтом). Наименее пригодной для практических целей оказалась модель измерений десятого студента: R^2 = 0,01 (модель объясняет всего 1% полученных ответов), P(хи-квадрат) = 0,002 (вероятность случайности ошибок модели 0,2%). И критерий Фишера, и критерий хи-квадрат при уровне значимости 0,05 опровергают пригодность модели. Действительно, на рис. 6 видно, что ответы хаотически расположены относительно графика модели измерения, причём отклонения достигают 0,95. Предположение о случайном характере ошибок, положенное в основу математической модели, в данном конкретном случае неправомерно: P(хи-квадрат) = 0,002, что много меньше стандартного уровня значимости 0,05.

Проверка по критерию Фишера не подтвердила адекватность моделей для студентов 1, 2, 3, 7 и 9 и заданий 1 и 3. Модель измерения в этих случаях не может объяснить ответы испытуемых, что лишает результаты педагогического измерения практической ценности.

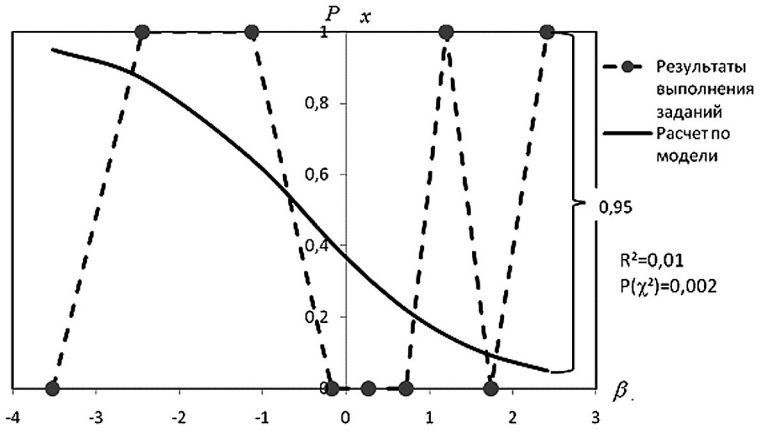

Рис. 6. Графический анализ модели для десятого студента

Очевидно, основная причина этого — малый объём данных (при большом количестве наблюдений даже весьма малые отклонения R^2 от нуля оказываются достаточными для признания значимости этого коэффициента и адекватности модели).

Далее проведён ряд пробных расчётов по тем же исходным данным с использованием других распространённых моделей и методов расчёта параметров модели, результаты сведены в табл. 2 (принят уровень значимости 0,05).

Анализ данных табл. 2 свидетельствует, что в данном педагогическом измерении:

- модель Раша в сочетании с методом максимального правдоподобия обеспечивает относительно небольшие стандартные ошибки. Проверка по критерию χ^2 подтверждает пригодность модели, критерий Фишера — значимость модели для всего набора исходных данных. Однако не удалось подтвердить возможность практического использования оценок уровней подготовленности 6 студентов и уровней трудности 4 заданий ввиду статистической незначимости (низкого качества) модели измерения (см. табл. 2);
- двухпараметрическая модель в сочетании с методом максимального правдоподобия дала большой разброс значений стандартных ошибок, которые для трёх студентов и одного задания превышают 1,8 (максимальная ошибка достигает 2,82). В целом модель измерения адекватна исходным данным, но не удалось подтвердить адекватность модели и результатов измерения для 4 студентов и 4 заданий;
- метод наименьших квадратов привёл к построению непригод-

Таблица 2

Модель, метод расчёта	Средняя стандартная ошибка, $\sigma_{ср}$	Максимальная стандартная ошибка, $\sigma_{мах}$	Вероятность случайности ошибок, $P(\chi^2)$	Коэффициент детерминации, R^2	Номера испытуемых и заданий, для которых $\chi^2>\chi^2_{крит}$	Номера испытуемых и заданий, для которых $F<F_{крит}$
Модель Раша, Метод максимального правдоподобия	0,84	1,24	0,999996	0,489	студент 10	студент 1 студент 2 студент 3 студент 7 студент 9 задание 1 задание 3 задание 4 задание 10
Модель Раша, Метод наименьших квадратов	1,17	3,79	0,000000	0,544	студент 10 задание 8 задание 10	студент 1 студент 2 студент 7 студент 9 студент 10
Двухпараметрическая модель, Метод максимального правдоподобия	0,94	2,82	0,9999999	0,562	-	студент 1 студент 7 студент 9 студент 10 задание 1 задание 3 задание 4 задание 5
Двухпараметрическая модель, Метод наименьших квадратов	0,52	2,47	0,000000	0,624	студент 9 студент 10 задание 6 задание 8 задание 10	студент 1 студент 9 студент 10 задание 1

ных для педагогического измерения моделей (вероятность случайности ошибок $P(\chi^2)$ много меньше любого стандартного значения уровня значимости). Это можно объяснить чувствительностью метода наименьших квадратов к выбросам (большим отклонениям между моделью и экспериментом) и малым объёмом данных.

Ни одна из рассмотренных моделей не смогла обеспечить приемлемое качество измерения. Очевидная причина этого — малый объём исходных данных, который казался недостаточным для надёжного отделения случайного и закономерного.

Тем не менее, пример позволил весьма подробно проиллюстрировать все основные этапы анализа адекватности модели измерения.

Выводы

1. Анализ адекватности модели является необходимым условием корректности педагогического измерения.

2. Проверку модели измерения целесообразно проводить в следующей последовательности:

а) оценка доверительных интервалов для результатов педагогического измерения. Если погрешности окажутся слишком велики, то результаты измерения непригодны для практического использования;

б) проверка закона распределения погрешностей для всей матрицы ответов и для каждого испытуемого и для каждого тестового задания в отдельности. Если нормальный закон распределения не будет подтверждён, то следует признать модель непригодной, а оценивание данного студента (тестового задания) в рамках выбранной модели невозможным;

в) оценка значимости коэффициента детерминации модели как показателя адекватности модели и надёжности результатов педагогического измерения.

3. При компьютерном тестировании в случае неподтверждения адекватности модели целесообразно предусмотреть автоматизацию выдачи дополнительных тестовых заданий. Критерием завершения тестирования может стать достижение адекватности модели педагогического измерения.