ПРОБЛЕМА АНАЛИЗА ПОГРЕШНОСТИ ПЕДАГОГИЧЕСКИХ ИЗМЕРЕНИЙ

Олег Деменченок

Восточно-Сибирский институт МВД России

AskSystem@yandex.ru

Опубликовано в ж. «Педагогические Измерения» №1 2009 г.

Средствами теории вероятностей показано, что известные формулы стандартных ошибок измерений с использованием базовых моделей IRT дают оценки предельных (максимально возможных) ошибок. Предложен способ минимизации эффекта завышения погрешностей для уровней подготовленности испытуемых. Проведён анализ информационной функции, даны практические рекомендации по формированию набора заданий.

Ключевые слова: тест, погрешность тестирования, информационная функция, адаптивное тестирование.

Постановка проблемы

Тестирование проводят для того, чтобы определить уровень подготовленности испытуемых. Чем быстрее и точнее будет получен результат педагогического измерения, тем эффективнее контроль знаний. Очевидно, что быстрота и точность – противоречивые требования. Чем больше затрачено времени, тем полнее может быть (правда, не всегда) представление об уровне подготовленности тестируемого. С другой стороны, контроль знаний отнимает значительную часть учебного времени в ущерб собственно обучению.

Для практического использования результатов педагогических измерений следует знать значение погрешности. Например, уровень подготовленности одного испытуемого оказался равен 3,0 логита; а другого – 3,1 логита. Если не учитывать погрешность, то можно сделать вывод о более высокой подготовленности второго студента. Однако, если, предположим, погрешность измерений равна 0,3 логита, то это означает, что уровень подготовленности первого студента оказался в пределах 3,0 ± 0,3= 2,7…3,3; а второго – 3,1 ± 0,3 = 2,8…3,4. По полученным значениям невозможно однозначно определить, какой из испытуемых подготовлен лучше.

Кроме того, без учёта погрешности представляется не вполне обоснованным и перевод результата тестирования в привычную пятибалльную педагогическую оценку. Пусть, например, границей между оценками «хорошо» и «отлично» принято значение 3,05. Учёт погрешности выявляет несовершенство метода, поскольку однозначность педагогической оценки в данном конкретном случае исчезает.

Следует отметить некоторые особенности педагогических измерений на основе

моделей IRT:

- косвенный характер измерения (уровни подготовленности обучаемых и параметры тестовых заданий определяются расчетным путём);
- непроверяемость полученных результатов практикой. Расчётные значения уровней подготовленности испытуемых не с чем сравнить для оценки погрешностей.

Из этого видно, что проблема анализа погрешностей и поиска возможностей их уменьшения до приемлемого уровня приобретает заметную актуальность.

Во всех известных автору публикациях по IRT приводятся формулы для расчёта стандартной ошибки нахождения уровней подготовленности обучаемых и трудности тестовых заданий[1].

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

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

где σ_{θ_i} – стандартная ошибка уровня подготовленности *i*-го испытуемого; σ_{β_j} – стандартная ошибка уровня трудности *j*-го задания; P_{ij} – вероятность правильного ответа *i*-го тестируемого на *j*-е задание; n – количество испытуемых; c_j –параметр коррекции на угадывание правильного ответа в *j*-ом задании.

В работе Бирнбаума указывается, что при оценке погрешности одного параметра другие параметры считаются найденными без ошибок; обоснование правомерности формул (1) и (2) отсутствует. К сожалению, такое обоснование не удалось найти и в последующих публикациях. Поскольку справедливость указанных формул не может быть проверена опытным путём, представляется важной теоретическая аргументация их справедливости.

Анализ ошибки измерения по модели Раша

Ошибка модели Δx определяется разностью фактического x и расчетного значения x_0:

[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, MA: Addison-Wesley Publishing, 1968. pp.397 - 472.

$$\Delta x = x - x_0, \qquad (3)$$

где x – результат выполнения тестового задания ($x = 1$ при правильном ответе, $x = 0$ при неправильном ответе).

Расчетное значение x_0 представляет собой теоретическое среднее значение случайной величины. 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[2]:

$$M(x_0) = x_1P_1 + x_2P_2 + \cdots + x_nP_n .$$

Вероятность того, что $x_0 = 1$ равна P, а вероятность $x_0 = 0$ равна $1 - P$. Тогда

$$M(x_0) = 0 \cdot (1-P) + 1 \cdot P = P .$$

Значит, для основных моделей IRT расчетное значение равно вероятности правильного ответа, и уравнение (3) можно переписать в виде:

$$\Delta x = x - P.$$

Уровень подготовленности испытуемого не поддаётся непосредственному измерению, он измеряется косвенно – расчетным путём по той или иной модели. Параметры модели независимы – уровни подготовленности испытуемых не зависят от уровней трудности тестовых заданий[3], т.е. уровень подготовленности обучаемого не должен зависеть от того, какие именно задания включены в тест.

Если параметры модели независимы, то для нахождения погрешности используется квадратичная сумма произведений частных производных и погрешностей измерения каждой используемой в расчете переменной величины[4]:

$$\Delta y = \sqrt{\left(\frac{\partial y}{\partial x_1}\Delta x_1\right)^2 + \left(\frac{\partial y}{\partial x_2}\Delta x_2\right)^2 + \ldots + \left(\frac{\partial y}{\partial x_n}\Delta x_n\right)^2} , \qquad (4)$$

где Δy – погрешность измерения величины y, рассчитываемой по значениям переменных величин x_1, $x_2 \ldots x_i \ldots x_n$; $\partial y / x_i$ – частная производная функции $y(x_1, x_2 \ldots x_i \ldots x_n)$ по переменной x_i; Δx_i – погрешность измерения переменной x_i.

Для модели Раша (5) уравнение (4) примет вид:

$$P = \frac{e^{\theta-\beta}}{1+e^{\theta-\beta}} = \frac{1}{1+e^{-(\theta-\beta)}} , \qquad (5)$$

где $e \approx 2{,}72$ – основание натурального логарифма;

---

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

[3] Wright B.D., Stone M.H. Measurement essentials, 2nd Edition. – Wilmington: Wide Range, Inc., 1999. - 204p.

[4] Тейлор Дж. Введение в теорию ошибок. Пер. с англ. – М.: Мир, 1985. - 272 с.

$$\Delta x = \sqrt{\left(\frac{\partial P}{\partial \theta}\Delta\theta\right)^2 + \left(\frac{\partial P}{\partial \beta}\Delta\beta\right)^2},$$

где $\Delta\theta$ – погрешность определения уровня подготовленности тестируемого, $\Delta\beta$ – погрешность определения уровня трудности задания;

$$\Delta x = \sqrt{\left(\frac{\partial\left(\frac{1}{1+e^{-(\theta-\beta)}}\right)}{\partial\theta}\Delta\theta\right)^2 + \left(\frac{\partial\left(\frac{1}{1+e^{-(\theta-\beta)}}\right)}{\partial\beta}\Delta\beta\right)^2},$$

$$\Delta x = \sqrt{\left(\left(\frac{e^{-(\theta-\beta)}}{\left(1+e^{-(\theta-\beta)}\right)^2}\right)\Delta\theta\right)^2 + \left(\left(\frac{e^{-(\theta-\beta)}(-1)}{\left(1+e^{-(\theta-\beta)}\right)^2}\right)\Delta\beta\right)^2},$$

$$\Delta x = \frac{e^{-(\theta-\beta)}}{\left(1+e^{-(\theta-\beta)}\right)^2}\sqrt{\Delta\theta^2+\Delta\beta^2},$$

$$\Delta x = \frac{1}{1+e^{-(\theta-\beta)}}\cdot\frac{1+e^{-(\theta-\beta)}-1}{1+e^{-(\theta-\beta)}}\sqrt{\Delta\theta^2+\Delta\beta^2},$$

$$\Delta x = P\cdot(1-P)\sqrt{\Delta\theta^2+\Delta\beta^2}. \qquad (6)$$

Уравнение (6) не даёт возможности найти погрешность определения уровня подготовленности тестируемого без знания погрешности уровня трудности задания. Поэтому необходимо рассмотреть дополнительные условия.

Анализ ошибки измерения при применении модели Г Раша, в случае $\Delta\theta >> \Delta\beta$

Рассмотрим случай, когда $\Delta\theta$ много больше $\Delta\beta$, например, $\Delta\theta$ =0,5 и $\Delta\beta$ =0,05. Тогда их квадратичная сумма будет мало отличиться от $\Delta\theta$:

$$\sqrt{0{,}5^2+0{,}05^2} = 0{,}502494 \approx 0{,}5.$$

Значит, в этом случае величиной $\Delta\beta$ в уравнении (6) можно пренебречь. Как известно, ошибка нахождения средней величины по результатам обработки статистического материала обратно пропорциональна квадратному корню из числа испытаний[5].

Если результаты тестирования в группе обрабатываются с учётом ранее полученных ответов, то количество ответов одного испытуемого может исчисляться десятками, в то время как по каждому тестовому заданию накоплено сотни и тысячи ответов. Практически это означает, что погрешность определения уровня подготовленности тес-

[5] Королюк В.С. и др. Справочник по теории вероятностей и математической статистике. – М.: Наука, 1985. - 640 с.

тируемого $\Delta\theta$ много больше погрешности уровня сложности задания $\Delta\beta$. Тогда уравнение (6) можно переписать в виде:

$$\Delta x \approx \frac{\partial P}{\partial \theta}\Delta\theta = P(1-P)\cdot\Delta\theta,$$

$$\Delta\theta = \frac{\Delta x}{P(1-P)}. \qquad (7)$$

Возведём в квадрат правую и левую часть равенства (7):

$$\Delta\theta^2 = \left(\frac{\Delta x}{P(1-P)}\right)^2 = \frac{\Delta x^2}{\left(P(1-P)\right)^2},$$

$$\frac{1}{\Delta\theta^2} = \frac{\left(P(1-P)\right)^2}{\Delta x^2}.$$

С вероятностью P дан правильный ответ (т.е. $x = 1$), вероятность $x = 0$ равна $1 - P$, следовательно:

$$\Delta x^2 = (x-P)^2 = (1-P)^2\cdot P + (0-P)^2\cdot(1-P) = (1-P)\cdot\left((1-P)\cdot P + P^2\right),$$

$$\Delta x^2 = (1-P)\cdot P.$$

Тогда

$$\frac{1}{\Delta\theta^2} = \frac{\left(P(1-P)\right)^2}{P(1-P)} = P(1-P). \qquad (8)$$

Просуммируем (8) по всем тестовым заданиям, учитывая, что уровень подготовленности данного испытуемого θ и погрешность его определения одинакова для каждого задания:

$$\sum_{j=1}^{m}\frac{1}{\Delta\theta^2} = \sum_{j=1}^{m}P_j(1-P_j),$$

где P_j – вероятность правильного ответа на j-е тестовое задание для данного испытуемого;

$$\frac{m}{\Delta\theta^2} = \sum_{j=1}^{m}P_j(1-P_j),$$

$$\Delta\theta^2 = \frac{m}{\sum_{j=1}^{m}P_j(1-P_j)}. \qquad (9)$$

Уравнение (9) служит отправной точкой для определения дисперсии ошибок. По определению дисперсия – это математическое ожидание квадрата отклонения от математического ожидания или средний квадрат отклонения от среднего:

$$D_\theta = M(\Delta\theta^2) - M(\overline{\Delta\theta})^2 .$$

Оценка параметров модели по методу максимального правдоподобия или наименьших квадратов является несмещенной. Несмещенная оценка θ – это оценка параметра, математическое ожидание которой равно значению оцениваемого параметра: $M(\theta) = \theta$. Несмещенная оценка лишена систематической ошибки, т.е. математическое ожидание ошибки равно нулю[6] $M(\Delta\theta) = \overline{\Delta\theta} = 0$. Тогда с учётом уравнения (9):

$$D_\theta = M(\Delta\theta^2) = \frac{m}{\sum_{j=1}^{m} P_j(1-P_j)} .$$

Дисперсия ошибки среднего связана с дисперсией параметра следующим соотношением[7]:

$$D_{\bar{x}} = \frac{D_x}{N},$$

где N – объём выборки или количество измерений.

Следовательно:

$$D_{\bar{\theta}} = \frac{D_\theta}{m} = \frac{1}{\sum_{j=1}^{m} P_j(1-P_j)} .$$

Стандартная ошибка среднего равна квадратному корню из дисперсии:

$$\sigma_{\bar{\theta}} = \frac{1}{\sqrt{\sum_{j=1}^{m} P_j(1-P_j)}} .$$

Анализ ошибки модели Раша в случае $\Delta\theta \approx \Delta\beta$

Если нет оснований полагать, что одна их двух указанных погрешностей много больше другой, то уравнение (6) можно использовать для оценки предельной погрешности. По аналогии с тем, что катет прямоугольного треугольника не может быть больше гипотенузы, вклад погрешностей определения любого из параметров модели не может быть больше суммарного влияния погрешностей всех параметров, т.е.:

$$\Delta x^2 = P\cdot(1-P)\Delta\theta^2 + P\cdot(1-P)\Delta\beta^2,$$

$$\Delta x \geq P\cdot(1-P)\Delta\theta,$$

$$\Delta x \geq P\cdot(1-P)\Delta\beta .$$

[6] Орлов А.И. Прикладная статистика. - М.: Экзамен, 2004. - 338 с.
[7] Тейлор Дж. Введение в теорию ошибок. Пер. с англ. – М.: Мир, 1985. - 272 с.

Описанный выше ход рассуждений приводит к следующим формулам для стандартных ошибок:

$$\sigma_{\bar{\theta}} \le \frac{1}{\sqrt{\sum_{j=1}^{m} P_j(1-P_j)}}, \tag{10}$$

$$\sigma_{\bar{\beta}} \le \frac{1}{\sqrt{\sum_{i=1}^{n} P_i(1-P_i)}}. \tag{11}$$

В англоязычной литературе ошибки средних значений обычно называют стандартными ошибками, символ среднего значения при этом опускают.

Для модели Раша a_j=1, c_j=0. Нетрудно заметить, что при этом формулы (10) и (11) соответствуют уравнениям (1) и (2).

Анализ ошибки двухпараметрической модели

Найдём частные производные для двухпараметрической модели (12):

$$P = \frac{e^{a(\theta-\beta)}}{1+e^{a(\theta-\beta)}} = \frac{1}{1+e^{-a(\theta-\beta)}}, \tag{12}$$

$$\frac{\partial P}{\partial \theta} = \frac{e^{-a(\theta-\beta)} \cdot a}{\left(1+e^{-a(\theta-\beta))}\right)^2} = a \cdot \frac{1}{1+e^{-a(\theta-\beta)}} \cdot \frac{1+e^{-a(\theta-\beta)}-1}{1+e^{-a(\theta-\beta)}} = a \cdot P \cdot (1-P),$$

$$\frac{\partial P}{\partial \beta} = \frac{e^{-a(\theta-\beta)} \cdot (-a)}{\left(1+e^{-a(\theta-\beta))}\right)^2} = -a \cdot \frac{1}{1+e^{-a(\theta-\beta)}} \cdot \frac{1+e^{-a(\theta-\beta)}-1}{1+e^{-a(\theta-\beta)}} = -a \cdot P \cdot (1-P),$$

$$\frac{\partial P}{\partial a} = \frac{e^{-a(\theta-\beta)}(\theta-\beta)}{\left(1+e^{-a(\theta-\beta))}\right)^2} = (\theta-\beta)\frac{1}{1+e^{-a(\theta-\beta)}} \cdot \frac{1+e^{-a(\theta-\beta)}-1}{1+e^{-a(\theta-\beta)}} = (\theta-\beta)P(1-P).$$

По аналогии с моделью Раша, стандартные ошибки уровня подготовленности испытуемого θ, уровня трудности β и дифференцирующей способности a тестового задания равны:

$$\Delta x \ge \frac{\partial P}{\partial \theta}\Delta\theta = a \cdot P(1-P) \cdot \Delta\theta,$$

$$\Delta\theta^2 \le \left(\frac{\Delta x}{a \cdot P(1-P)}\right)^2 = \frac{P(1-P)}{\left(a \cdot P(1-P)\right)^2} = \frac{1}{a^2 \cdot P(1-P)},$$

$$\sigma_{\bar{\theta}} \le \frac{1}{\sqrt{\sum_{j=1}^{m} a_j{}^2 P_j(1-P_j)}}, \tag{13}$$

$$\sigma_{\bar{\beta}} \le \frac{1}{\sqrt{a_j^{\,2}\sum_{i=1}^{n} P_i(1-P_i)}}, \tag{14}$$

$$\sigma_{\bar{a}} \le \frac{1}{\sqrt{\sum_{i=1}^{n} (\theta_i-\beta)^2 P_i(1-P_i)}}. \tag{15}$$

Анализ ошибки трёхпараметрической модели

По трёхпараметрической модели вероятность правильного ответа *i*-го тестируемого на *j*-е задание равна[8]:

$$P_{ij} = c_j + (1-c_j)\frac{e^{a_j(\theta_i-\beta_j)}}{1+e^{a_j(\theta_i-\beta_j)}} = c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i-\beta_j)}}, \tag{16}$$

где θ_i – уровень подготовленности *i*-го тестируемого; β_j – уровень трудности *j*-го задания.

Для оценки погрешностей найдём частные производные:

$$\frac{\partial P}{\partial \theta_i} = \frac{\partial\left(c_j + (1-c_j)\dfrac{1}{1+e^{-a_j(\theta_i-\beta_j)}}\right)}{\partial \theta_i} = (1-c_j)\frac{e^{-a_j(\theta_i-\beta_j)}\cdot a_j}{\left(1+e^{-a_j(\theta_i-\beta_j)}\right)^2},$$

$$\frac{\partial P}{\partial \beta_j} = (1-c_j)\frac{e^{-a_j(\theta_i-\beta_j)}\cdot(-a_j)}{\left(1+e^{-a_j(\theta_i-\beta_j)}\right)^2},$$

$$\frac{\partial P}{\partial a} = (1-c_j)\frac{e^{-a(\theta-\beta)}\cdot(\theta-\beta)}{\left(1+e^{-a(\theta-\beta))}\right)^2}.$$

Найдём квадрат погрешности измерения уровня подготовленности *i*-го тестируемого:

$$\Delta\theta_i^{\,2} \le \left(\frac{\Delta x}{\partial P_{ij}/\partial\theta_i}\right)^2 = \frac{P_{ij}(1-P_{ij})}{\left(a_j(1-c_j)\dfrac{e^{-a_j(\theta_i-\beta_j)}}{\left(1+e^{-a_j(\theta_i-\beta_j)}\right)^2}\right)^2},$$

$$\Delta\theta_i^{\,2} \le \frac{P_{ij}(1-P_{ij})}{a_j^{\,2}\left[(1-c_j)\dfrac{e^{-a_j(\theta_i-\beta_j)}}{1+e^{-a_j(\theta_i-\beta_j)}}\right]^2\left[\dfrac{1}{1+e^{-a_j(\theta_i-\beta_j)}}\right]^2}. \tag{17}$$

[8] Partchev I. A visual guide to item response theory. – Jena: Friedrich-Schiller-Universitat, 2004. – 61 p.

Первое выражение в квадратных скобках в неравенстве (17) представляет собой вероятность неправильного ответа:

$$1-P_{ij}=1-c_j-(1-c_j)\frac{1}{1+e^{-a_j(\theta_i-\beta_j)}}=(1-c_j)\left(1-\frac{1}{1+e^{-a_j(\theta_i-\beta_j)}}\right)=$$

$$=(1-c_j)\left(\frac{1+e^{-a_j(\theta_i-\beta_j)}}{1+e^{-a_j(\theta_i-\beta_j)}}-\frac{1}{1+e^{-a_j(\theta_i-\beta_j)}}\right),$$

$$1-P_{ij}=(1-c_j)\frac{e^{-a_j(\theta_i-\beta_j)}}{1+e^{-a_j(\theta_i-\beta_j)}}, \qquad (18)$$

Второе выражение в квадратных скобках можно преобразовать из аналитической формы трёхпараметрической модели (16) к виду:

$$P_{ij}=c_j+(1-c_j)\frac{e^{a_j(\theta_i-\beta_j)}}{1+e^{a_j(\theta_i-\beta_j)}},$$

$$\frac{P_{ij}-c_j}{1-c_j}=\frac{e^{a_j(\theta_i-\beta_j)}}{1+e^{a_j(\theta_i-\beta_j)}}. \qquad (19)$$

Подставим (18) и (19) в (17):

$$\Delta\theta_i^{\,2}\le\frac{P_{ij}(1-P_{ij})}{a_j^{\,2}(1-P_{ij})^2\left(\frac{P_{ij}-c_j}{1-c_j}\right)^2}=\frac{1}{a_j^{\,2}\left(\frac{1-P_{ij}}{P_{ij}}\right)\left(\frac{P_{ij}-c_j}{1-c_j}\right)^2}. \qquad (20)$$

После очевидных преобразований получим:

$$\sigma_{\theta_i}\le\frac{1}{\sqrt{\sum_{j=1}^{m}a_j^{\,2}\left[\left(\frac{1-P_{ij}}{P_{ij}}\right)\left(\frac{P_{ij}-c_j}{1-c_j}\right)^2\right]}}, \qquad (21)$$

$$\sigma_{\beta_j}\le\frac{1}{\sqrt{\sum_{i=1}^{n}a_j^{\,2}\left[\left(\frac{1-P_{ij}}{P_{ij}}\right)\left(\frac{P_{ij}-c_j}{1-c_j}\right)^2\right]}}. \qquad (22)$$

Таким образом, подтверждена справедливость формул (1) и (2) для определения предельных (т.е. максимально возможных) значений стандартных ошибок педагогических измерений с использованием базовых моделей IRT.

Определение величины погрешности

Стандартная ошибка связана с величиной погрешности соотношением[9]:

$$\Delta\theta = \varepsilon \cdot \sigma_\theta, \qquad (23)$$

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.).

При этом уровень подготовленности нужно рассматривать не как конкретное значение θ, а как значение в доверительном интервале $\theta \pm \Delta\theta$.

Например, уровень подготовленности испытуемого, равный единице, при σ_θ =0,3 может трактоваться так:

- с вероятностью 68% уровень подготовленности находится в интервале $\theta = 1 \pm 1 \cdot \sigma_\theta$ (или 0,7 … 1,3);
- с вероятностью 90% $\theta = 1 \pm 1{,}65 \cdot \sigma_\theta$ (или 0,505 … 1,495);
- с вероятностью 99,7% $\theta = 1 \pm 3 \cdot \sigma_\theta$ (или 0,1 … 1,9);
- с вероятностью 99,99% $\theta = 1 \pm 4 \cdot \sigma_\theta$ (или –0,2… 2,2).

Эффект завышения оценки погрешности

Использование формул (1) и (2) приводит к завышенной оценке значения погрешностей. Проиллюстрируем это на простом примере. Пусть вклад погрешностей определения уровня подготовленности испытуемого, уровня трудности и различающей (дифференцирующей) способности задания равен 0,3. Тогда суммарная погрешность:

$$\sqrt{0{,}3^2 + 0{,}3^2 + 0{,}3^2} \approx 0{,}52 .$$

Предельные оценки погрешности одного параметра исходят из того, что погрешности остальных параметров равны нулю. Это означает, что расчёт по формулам (1) и (2) даст существенно преувеличенную оценку погрешности: расчётное значение погрешности будет на 73% больше истинного (0,52/0,3≈1,73).

Наиболее практически значимой представляется точность расчёта погрешности уровней подготовленности обучаемых. Эффект завышения оценки погрешности для уровней подготовленности может быть минимизирован, если результаты тестирования в группе обрабатываются с учётом результатов выполнения того же теста в других группах. Например, одновременная обработка данных выполнения теста в 10 группах уменьшает погрешности определения параметров тестовых заданий в $\sqrt{10}$ раз:

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

$$\sqrt{0{,}3^2 + \frac{(0{,}3)^2}{10} + \frac{(0{,}3)^2}{10}} \approx 0{,}329 .$$

При этом предельная оценка погрешности уровней подготовленности испытуемых окажется ближе к истинному значению: расчётное значение погрешности на 9,5% больше истинного (0,329/0,3≈1,095). Если одновременно обрабатывать результаты 100 групп, то преувеличение практически исчезнет (менее 1%):

$$\sqrt{0{,}3^2 + \frac{(0{,}3)^2}{100} + \frac{(0{,}3)^2}{100}} \approx 0{,}30299 .$$

Анализ влияния модели измерения

Выражения (1) и (2) позволяют найти погрешность измерения для трёх базовых моделей IRT, причём величины погрешностей для каждой модели будут разными. Возникает естественный вопрос: какая из моделей обеспечивает большую точность измерений?

Анализ самих формул не даёт ответа на этот вопрос, т.к. модели отличаются не только значениями коэффициентов a_j и c_j, но и расчётными значениями вероятности правильного ответа. Поэтому были проведены пробные расчёты. Использовалась бесплатная компьютерная программа Estimate2PL (сайт www.asksystem.narod.ru), исходные данные – матрица результатов тестирования, опубликованная в известной работе Б.Д.Райта и М.Х.Стоуна[10] (эта матрица заложена в программу Estimate2PL в качестве примера).

Расчёты не выявили безусловного преимущества какой-либо из моделей. Так, например, стандартная ошибка уровня подготовленности испытуемых составила:

- для модели Раша: среднее значение ошибки 1,04; максимальная ошибка 1,15;
- для двухпараметрической модели: среднее значение – 0,91; максимальная ошибка 1,81;
- для трёхпараметрической модели: среднее значение – 3,28; максимальная ошибка 7,97.

Двухпараметрическая модель в целом оказалась точнее, но максимальная ошибка меньше у модели Раша (т.е. вариация значений ошибок измерения в модели Раша меньше). Ошибки определения уровней подготовленности по трёхпараметрической модели оказались настолько велики, что лишили результаты измерения практической ценности: стандартная ошибка ±3,28 перекрывает весь диапазон измерения –3…3.

То, что максимальная ошибка при использовании двухпараметрической модели превышает максимальную ошибку модели Раша, объясняется различием информационной ценности результата тестирования для этих моделей.

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

Анализ информационной функции

Чем больше информации, тем точнее наши сведения, т.е. меньше ошибка. В теории педагогических измерений Item Response Theory количеством информации[11] называют величину, обратную дисперсии ошибок, а информационной функцией – соответствующую аналитическую зависимость:

$$I = \frac{1}{D} = \frac{1}{\sqrt{\sigma}}.$$

В классическом пособии по IRT Ф.Бейкера[12] приведена формула, которая связывает стандартную ошибку уровня подготовленности тестируемого σ_θ и количество информации:

$$\sigma_\theta = \frac{1}{\sqrt{I}} = \frac{1}{\sqrt{\sum I_j}} = \frac{1}{\sqrt{\sum_{j=1}^{m} a_j{}^2 P_j (1 - P_j)}}, \qquad (24)$$

где $I = \Sigma I_j$ – общее количество информации, полученной при решении всех тестовых заданий (определяется простым суммированием количества информации по каждому выполненному заданию I_j).

Нетрудно убедиться, что при c_j =0 формула (24) соответствует уравнению (1).

Из формулы (24) следует, что количество информации, полученное в результате выполнения одного задания, для двухпараметрической модели описывается выражением:

$$I_j = a_j{}^2 \cdot P\,(1–P), \qquad (25)$$

Казалось бы, при $a_j > 1$ двухпараметрическая модель получает больше информации, чем модель Раша. Однако вносит свои коррективы и изменение расчётной вероятности правильного ответа.

Рассмотрим пример: пусть θ=1, β=1,2; a=3. Тогда количество информации, полученной при решении одного задания для модели Раша и двухпараметрической модели равно:

$$I_{Rash} = 1^2 \left(\frac{1}{1 + e^{-(1–1.2)}} \right) \left(1 - \frac{1}{1 + e^{-(1–1.2)}} \right) \approx 0{,}45 \cdot (1 - 0{,}45) = 0{,}248;$$

$$I_{2PL} = 3^2 \left(\frac{1}{1 + e^{-3(1–1.2)}} \right) \left(1 - \frac{1}{1 + e^{-3(1–1.2)}} \right) \approx 9 \cdot 0{,}35 \cdot (1 - 0{,}35) = 2{,}059.$$

[11] количество информации – показатель, характеризующий уменьшение неопределенности состояния системы.

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

Следовательно, при $a > 1$ и $\theta \approx \beta$ двухпараметрическая модель более эффективно использует результаты тестирования и работает точнее. Однако с увеличением разницы θ и β наблюдается обратная ситуация. Так, при θ=1и β=–2 получим:

$$I_{Rash} = 1^2\left(\frac{1}{1+e^{-(1+2)}}\right)\left(1-\frac{1}{1+e^{-(1+2)}}\right) \approx 0{,}95\cdot(1-0{,}95) = 0{,}045;$$

$$I_{2PL} = 3^2\left(\frac{1}{1+e^{-3(1+2)}}\right)\left(1-\frac{1}{1+e^{-3(1+2)}}\right) \approx 9\cdot 0{,}9999\cdot(1-0{,}9999) = 0{,}001.$$

Значит, с увеличением разницы θ и β двухпараметрическая модель получает значительно меньше информации и точность резко падает. У модели Раша точность снижается медленнее.

Из формулы (25) следует, что количество информации максимально при вероятности правильного ответа P= 0,5 (рис.1). Информативность заданий с вероятностью правильного ответа, близкой к нулю или единице, почти нулевая. Действительно: способность испытуемого решать очень простые задачи ($P \approx 1$) или неудачи в решений задач повышенной сложности ($P \approx 0$) мало информативны, так как не дают возможности уточнить уровень подготовленности тестируемого.

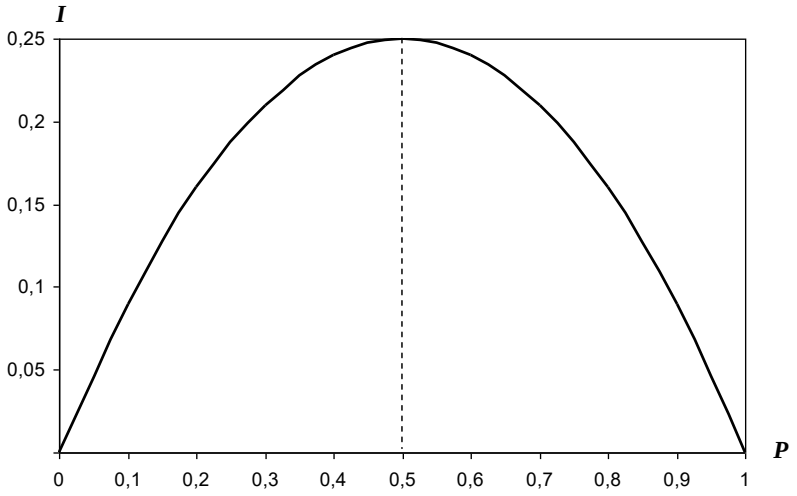

Рис.1. Зависимость количества информации от вероятности правильного ответа при a =1

Адаптивное тестирование

Вероятность правильного ответа в двухпараметрической модели описывается уравнением:

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

При a = 1 двухпараметрическая модель совпадает с известной моделью Г.Раша.

Легко заметить, что $P = 0{,}5$ в случае равенства уровней подготовленности испытуемого и трудности задания $\theta = \beta$:

$$P = \frac{1}{1+e^{-a(\theta-\beta)}} = \frac{1}{1+e^{-a\cdot 0}} = \frac{1}{1+1} = 0{,}5\,.$$

Именно на этом основан метод адаптивного тестирования: при правильном ответе следующее задание будет чуть более трудным, при неправильном ответе – более легким. Таким образом, поддерживается примерное равенство уровней подготовленности испытуемого и трудности заданий $\theta \approx \beta$, а средняя вероятность правильного ответа будет близка к 0,5[13]. Соответственно, количество информации (при $a = 1$):

$$I = \sum_{j=1}^{m} I_j = \sum_{j=1}^{m} a_j{}^2 P\,(1-P) \approx \sum_{j=1}^{m} 1^2 \cdot 0{,}5\,(1-0{,}5) = 0{,}25m\,.$$

Стандартная ошибка определения уровня подготовленности тестируемого:

$$\sigma_\theta = \frac{1}{\sqrt{I}} \approx \frac{1}{\sqrt{0{,}25m}} = \frac{2}{\sqrt{m}}\,. \qquad (26)$$

Например, для адаптивного теста из 30 заданий:

$$I \approx 0{,}25m = 0{,}25 \cdot 30 = 7{,}5$$

$$\sigma_\theta \approx \frac{2}{\sqrt{m}} = \frac{2}{\sqrt{30}} = 0{,}365\,.$$

Сравним точность адаптивного тестирования с точностью теста, в котором задания равномерно распределены по уровню сложности. Для этого по формуле (12) рассчитана вероятность правильного ответа для 30 заданий, равномерно распределенных по уровню сложности от –3 до 3, при дифференцирующей способности, равной единице. Расчет проведен для различных уровней подготовленности испытуемого в диапазоне от –3 до 3.

Затем для каждого значения θ найдено количество информации всего теста. Количество информации минимально при высоких и низких уровнях подготовленности (I = 2,53 при $\theta = 3$), максимум информации I = 4,42 соответствует нулевому уровню подготовленности (рис.2).

[13] Аванесов В.С. Применение тестовых форм в Rasch Measurement //ПИ, №4, 2006.

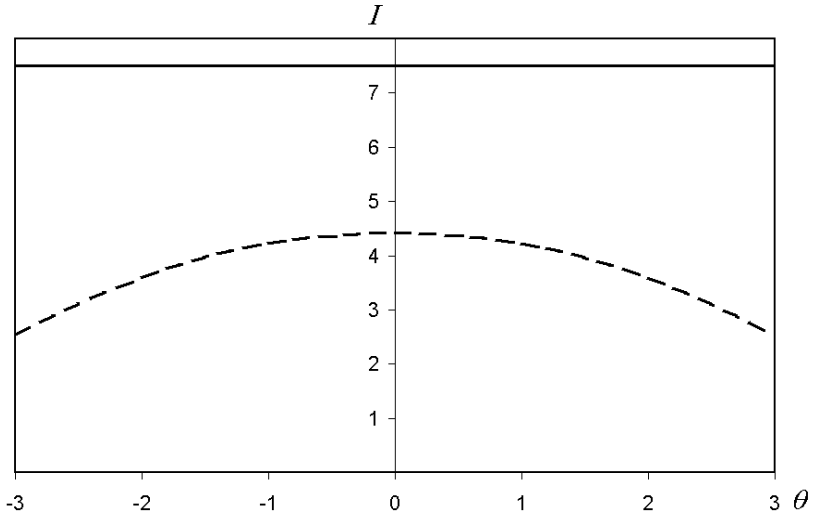

Рис.2. Зависимость количества информации теста с 30 заданиями от уровня подготовленности испытуемого: ———— - адаптивный тест, – – – – – - тест с равномерным распределением заданий по уровню сложности

Стандартная ошибка составила от 0,476 до 0,629 (рис.3). Точность адаптивного тестирования оказалась выше на 30-72% (0,629/0,365≈1,72). Ещё больше впечатляет разница в количестве заданий. Несложные расчеты показывают, что для обеспечения равной точности тест с равномерным распределением должен содержать от 50 до 90 (!) заданий (рис.4). Конечно, при адаптивном тестировании вероятность несколько отличается от 0,5 (например, при $P=0{,}47$ потребуется не 30, а 34 задания). Но даже в этом случае адаптивное тестирование даёт возможность в 1,5…2,5 раза сократить количество заданий, что убедительно доказывает его преимущество.

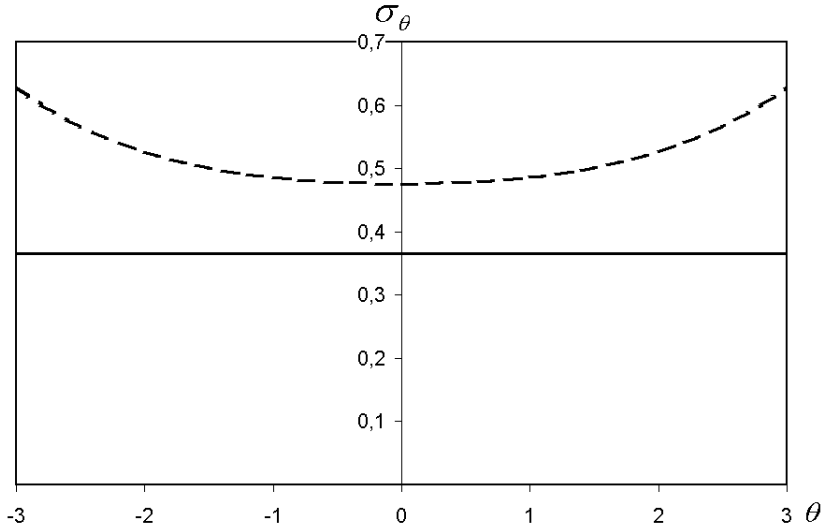

Рис.3. Стандартная ошибка уровня подготовленности испытуемого для теста с 30 заданиями: ———— - адаптивный тест, – – – – – - тест с равномерным распределением заданий по уровню сложности

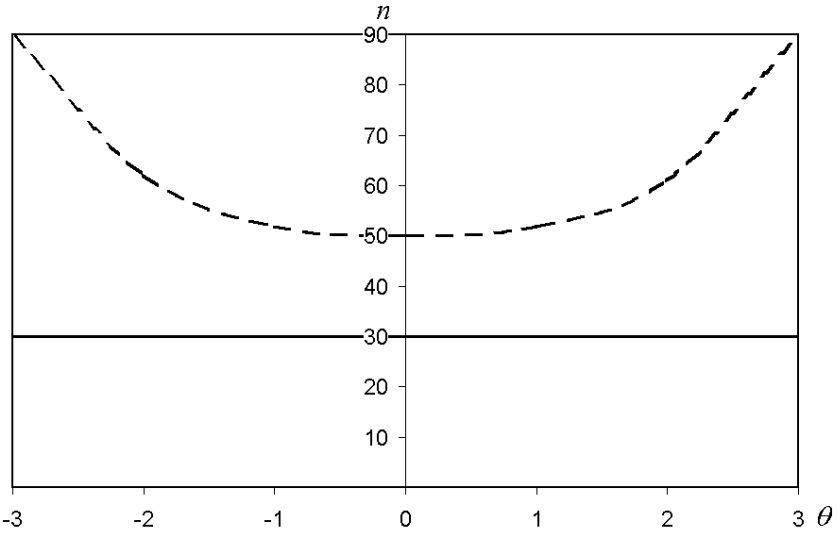

Рис.4. Количество заданий, обеспечивающих равную точность результата тестирования: ——— - адаптивный тест, – – – – – – - тест с равномерным распределением заданий по уровню сложности

Недостатки адаптивного тестирования

Во-первых, требуется большая база тестовых заданий, уровень сложности которых известен. Для каждого возможного уровня трудности нужно подобрать достаточное количество заданий (например, 30 заданий с $\beta = -3\ldots-2{,}8$; ещё 30 заданий с $\beta = -2{,}8\ldots-2{,}6$ и так далее). Если учесть, что уровни трудности заданий не могут быть заданы при составлении теста (они определяются статистически по результатам тестирования), то для адаптивного тестирования сложно сформировать подходящую базу заданий.

Во-вторых, невозможно адаптивное тестирование по бланкам (возможно только компьютерное тестирование). В англоязычной литературе используется специальный термин, подчеркивающий компьютерный характер такого тестирования – computerized adaptive testing (CAT)[14] – компьютеризированное адаптивное тестирование.

Практические рекомендации

Целенаправленно подбирая тестовые задания, можно влиять на форму информационной кривой теста. Так, в работе Н.Верхелста[15] для получения плоской информационной кривой теста с 18 заданиями в диапазоне уровней подготовленности $\theta = -2{,}5\ldots2{,}5$

[14] Wim J. van der Linden. A Formal Characterization of and Some Alternatives to Sympson-Hetter Item-Exposure Control in Computerized Adaptive Testing. – Law School Admission Council, 2006. 12p.

[15] Verhelst N.D. Item Response Theory. / Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment. - Strasbourg: Council of Europe, 2004. 42p.

предложен тест следующего состава (рис.5):

- 6 заданий с уровнем трудности β=–2;
- 1 задание с уровнем трудности β=–1,5;
- 4 задания с уровнем трудности β=0;
- 1 задание с уровнем трудности β=1,5;
- 6 заданий с уровнем трудности β=2.

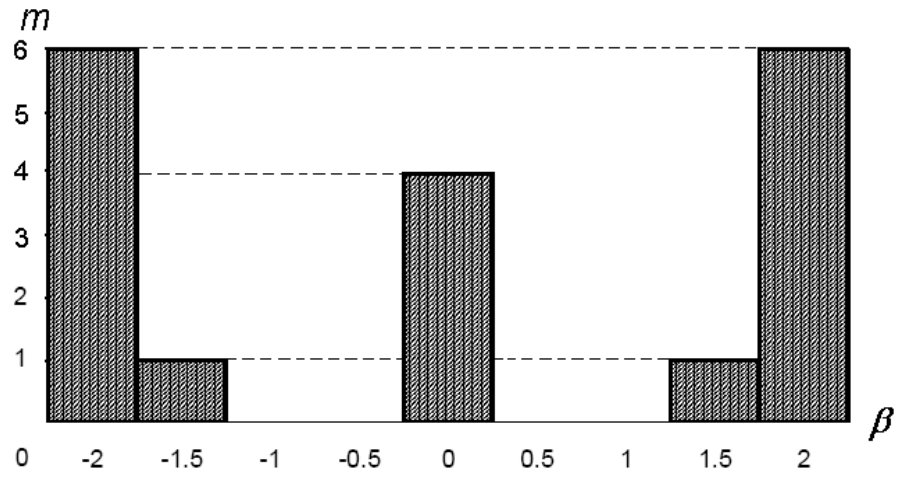

Рис.5. Предложенная Н.Верхелстом структура теста с 18 заданиями

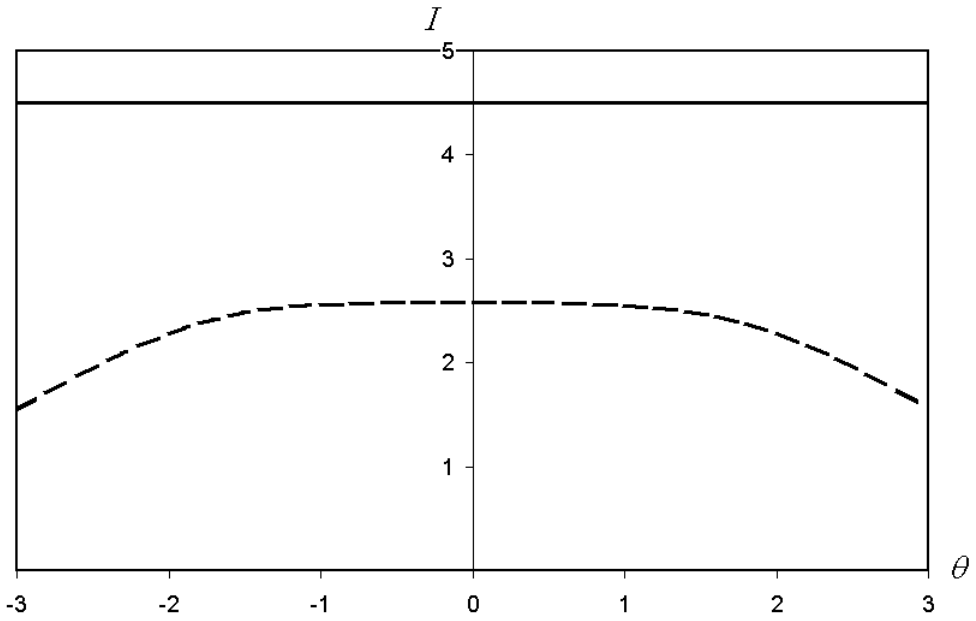

Рис.6. Зависимость количества информации теста с 18 заданиями от уровня подготовленности испытуемого: ———— - адаптивный тест; – – – – – - тест, предложенный Н.Верхелстом

При использовании модели Раша информационная кривая предложенного Н.Верхелстом теста имеет выровненную форму, а на участке θ=–1,8…1,8 практически горизонтальна. Это обеспечивает равную точность измерений в указанном диапазоне уровней подготовленности.

Автор полагает, что повысить точность педагогических измерений неадаптивного теста можно путём подбора заданий, которые по трудности соответствуют интересующему уровню подготовленности испытуемых (target group).

Пусть, например, оценка «зачтено» ставится при уровне подготовленности не меньшем $\theta \geq 1$. В этом случае не важно, насколько больше или меньше единицы уровень подготовленности конкретного испытуемого. Важно возможно более точно знать, преодолел ли испытуемый рубеж $\theta = 1$. Другими словами, важно обеспечить максимально возможную точность измерения уровней подготовленности, близких к единице. Подбирая задания с уровнем трудности, близким к θ (например, $\beta = \theta \pm 0{,}5$), при обычном тестировании получим точность, практически равную точности адаптивного тестирования (рис.7).

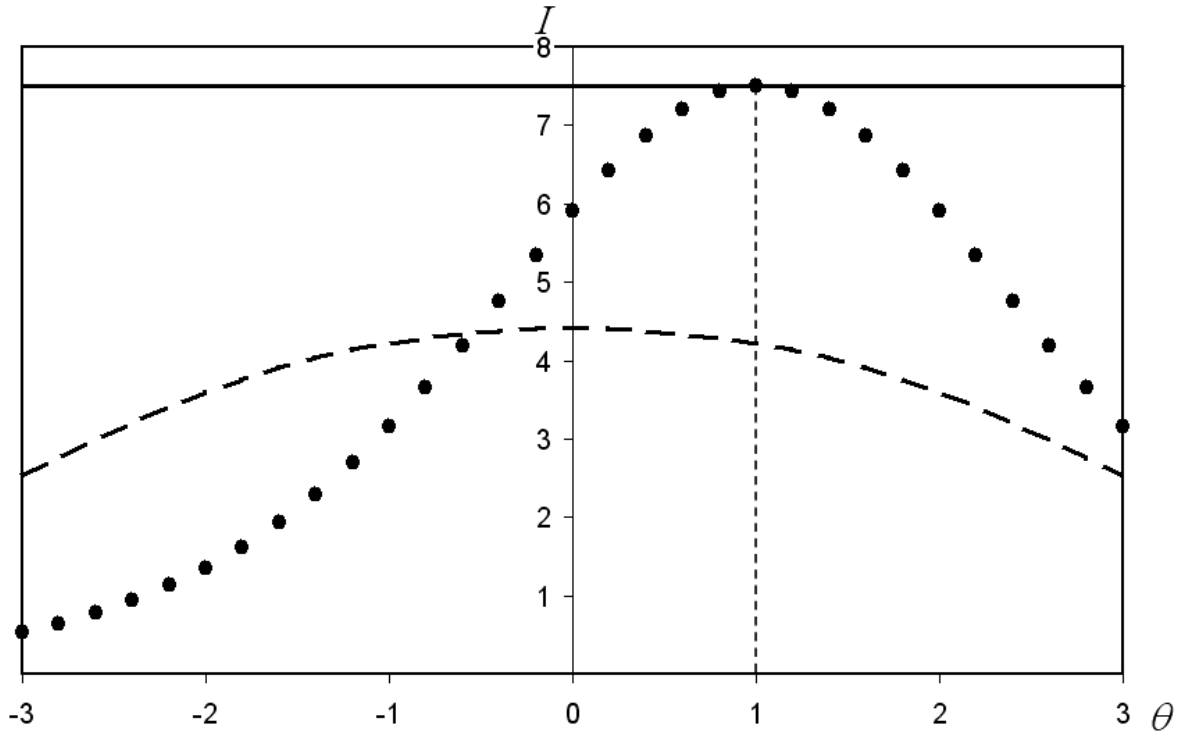

Рис.7. Зависимость количества информации теста с 30 заданиями от уровня подготовленности испытуемого: ———— - адаптивный тест, – – – – – – - тест с равномерным распределением заданий в диапазоне β=–3…3; • • • • • • • - тест с заданиями, уровень трудности которых равен единице

Так, для обычного теста с диапазоном изменения $\beta = 1 \pm 0{,}5$ требуется 31 задание, чтобы сравняться по точности с адаптивным тестом из 30 заданий при измерении уровня подготовленности $\theta \approx 1$.

Рассмотрим случай, когда требуется обеспечить высокую точность измерений в некотором диапазоне. Пусть, например, оценка «неудовлетворительно» ставится при уровне подготовленности $\theta < -1$, а оценка «отлично» – при $\theta > 2$. Следовательно, нужно

добиться высокой точности педагогических измерений в диапазоне θ =–1…2.

Если создать тест из заданий, равномерно распределенных в интервале β=–1…2, то можно увеличить информативность по сравнению с тестом из равномерно распределенных в интервале β =–3…3 заданий на 7-43% (рис.8). Это адекватно увеличению точности на 3,4…20%.

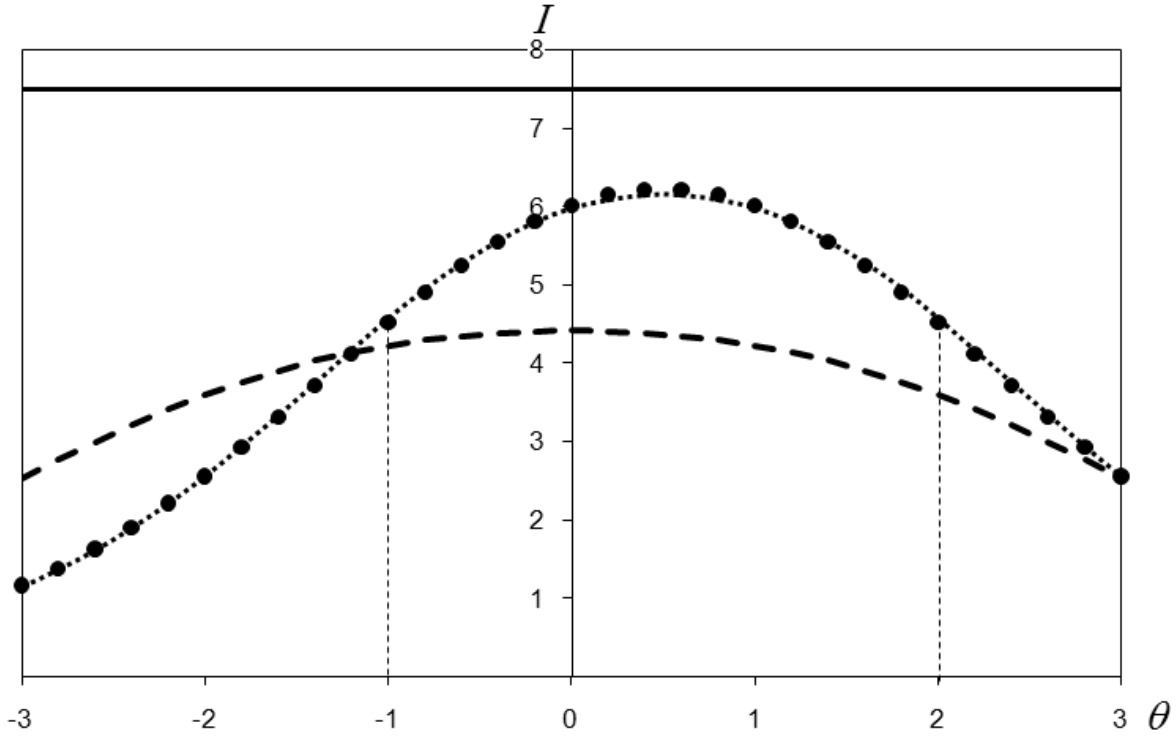

Рис.8. Количество информации теста с 30 заданиями:
———— - адаптивный тест,
– – – – – - тест с равномерным распределением заданий β=–3…3,
• • • • • • - тест с равномерным распределением заданий β=–1…2,
▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ - тест из 15 заданий β=–0,4 и 15 заданий β=1,4

Разумеется, сравняться по точности с адаптивным тестом в данном случае невозможно. Для достижения точности адаптивного теста требуется увеличить количество заданий. Так, для обеспечения точности, соответствующей адаптивному тесту с 30 заданиями тест с равномерным распределением заданий в диапазоне:

- β=–3…3 должен содержать от 51 до 63 заданий (51 задание при θ=0, 63 задания при θ=2);
- β=–1…2 должен содержать от 36 до 50 заданий (т.е. примерно на 30% меньше предыдущего случая, но на 40% больше в сравнении с адаптивным тестом).

Конечно, равномерное распределение в диапазоне, соответствующем измеряемым уровням подготовленности – это не единственный способ формирования набора тестовых заданий.

Поиск наилучшего решения проведён средствами Microsoft Excel по критерию максимизации минимального значения количества информации в указанном диапазоне:

$Min\, I(\beta) \to Max$.

Оказалось, что наилучшую точность педагогических измерений в диапазоне θ=–1…2 обеспечивает неадаптивный тест, состоящий из 15 заданий с уровнями трудности β=–0,4 и 15 заданий с уровнями трудности β=1,4. Однако точность этого теста практически совпала с точностью теста из 30-ти равномерно распределённых в диапазоне β=–1…2 заданий (рис.8). Увеличение информативности составило 1,3%; увеличение точности – 0,6%.

В целом, подбор задания, которые по сложности соответствуют интересующему диапазону уровней подготовленности, позволил снизить стандартную ошибку примерно на 15%.

Выводы

1. Подтверждена справедливость формул (1) и (2) для определения предельных (т.е. максимально возможных) значений погрешности педагогических измерений с использованием базовых моделей IRT.
2. Использование формул (1) и (2) приводит к завышенным оценкам погрешностей; расчётное значение погрешности может оказаться существенно больше истинного. Эффект завышения погрешностей для уровней подготовленности испытуемых может быть минимизирован, если результаты тестирования в группе обрабатывать с учётом ранее полученных результатов выполнения того же теста в других группах.
3. Пробные расчёты не выявили безусловного преимущества какой-либо из базовых моделей IRT в точности педагогических измерений.
4. Добиться максимальной точности педагогических измерений позволяет адаптивное тестирование. Для модели Раша при равной точности адаптивное тестирование даёт возможность в 1,5…2,5 раза сократить количество заданий по сравнению с тестом, задания которого равномерно распределены по уровню трудности.
5. Для повышения точности педагогических измерений при составлении неадаптивного теста целесообразно подбирать задания, которые по трудности соответствуют интересующему уровню подготовленности.