О многопараметрической оценке уровней подготовленности испытуемых и трудности заданий

Виктор Кромер
Новосибирский государственный педагогический университет
kromer@newmail.ru

Опубликовано в ж. «Педагогические Измерения» № 3, 2005 г.

Аннотация

Предложено обобщение двух известных вариантов модели тестирования Бирнбаума с учетом избирательности испытуемых и заданий. Основой обобщения является принятие модели Гуттмана с допущением случайных колебаний потенциалов испытуемых и заданий. Предложена концепция оценивания испытуемых с комплексным учетом их потенциалов и избирательности. Предложено также расщепление параметра, учитывающего вероятность угадывания в трехпараметрической модели Бирнбаума, на два отдельных параметра, характеризующие вероятность угадывания со стороны испытуемого и со стороны задания.

***Ключевые слова**: уровень подготовленности, трудность задания, дифференцирующая способность, модель Бирнбаума, модель Гуттмана*

Ввиду выявившихся недостатков классической модели педагогических измерений во 2-й половине XX в., появились модели совместного оценивания параметров испытуемого и заданий. В широко распространенных моделях семейства IRT (Item Response Theory) оценивание уровня подготовленности испытуемых и уровня трудностей заданий производится на единой интервальной шкале. Автором этой работы ранее предлагался термин *потенциал* как для уровня подготовки испытуемого, так и для уровня трудности задания.

В моделях IRT вероятность P_{ij} успешного выполнения i-м испытуемым j-го задания определяется разностью их потенциалов, т.е. нуль вводимой шкалы не определен и задается произвольно (обычно из условия равенства нулю усредненного потенциала всех испытуемых). В самой простой из моделей семейства IRT (модели Раша) $P_{ij} = (1 + \exp(\beta_j - \theta_i))^{-1}$, где θ_i и β_j – соответственно потенциалы i-го испытуемого и j-го задания. Модель принято считать однопараметрической, поскольку в качестве единственного параметра функции выступает разность потенциалов испытуемого и задания [4, с. 14].

Обобщением модели Раша является модель Бирнбаума, где для P_{ij} даются два выражения, в зависимости от того, определяем ли мы P_{ij} для испытуемого или для задания. По сути, речь идет о двух вариантах одной модели. В дальнейшем мы будем говорить о первом варианте модели

Бирнбаума, где $P_{ij}(\theta_i) = (1 + \exp(d_j(\beta_j - \theta_i)))^{-1}$, и втором варианте, где $P_{ij}(\beta_j) = (1 + \exp(d_i(\beta_j - \theta_i)))^{-1}$ [5, с. 13]. Дополнительные параметры d_i и d_j характеризуют избирательность, соответственно, задания и испытуемого. На практике за избирательностью задания закреплен термин *дифференцирующая способность*, и из двух описанных вариантов модели Бирнбаума применяется лишь первый, т.е. совместно с потенциалами испытуемых и заданий определяется дифференцирующая способность заданий, что позволяет отбраковывать задания с низкой избирательностью.

Модель Раша широко использовалась Центром тестирования Министерства образования. В 2001 г. этот центр использовал более прогрессивную модель Бирнбаума (первый ее вариант), что, тем не менее, вызвало нарекания со стороны испытуемых, их родителей, администраторов, журналистов и др. Причина заключается в том, что модель Раша обеспечивает монотонную (но, тем не менее, нелинейную) связь между исходным тестовым баллом и вычисленным в соответствии с моделью потенциалом испытуемого, переводимого затем посредством линейного преобразования в сертификационный балл. В модели Бирнбаума эта связь отсутствует, что снижает очевидную валидность тестового результата.

В первом варианте модели Бирнбаума потенциал *i*-го испытуемого однозначно задается не исходным тестовым баллом $\sum_{j=1}^{k} a_{ij}$, где k – количество тестовых заданий, а a_{ij} – элемент матрицы данных, а суммой $\sum_{j=1}^{k} d_j a_{ij}$. Таким образом, избирательность *j*-го задания d_j является его весом в конечном результате. Достоинством рассматриваемого варианта модели Бирнбаума является то, что по сравнению с моделью Раша снижаются требования к тестовым заданиям. Низкоэффективные задания с низкой избирательностью дают незначительный вклад в конечный результат, в отличие от модели Раша, где все задания равноправны по определению. Несправедливая критика использованной Центром тестирования в 2001 г. модели тестирования привела, в последующие годы, к возврату к модели Раша.

Второй вариант модели Бирнбаума практически не используется, поскольку в нем, как и в модели Раша, не учитывается избирательность заданий. Однако, в силу симметрии обоих вариантов, ему также присущи определенные достоинства. Все сказанное о первом варианте модели Бирнбаума применимо ко второму варианту при взаимной подмене испытуемых и заданий. Потенциал задания однозначно определяется суммой $\sum_{i=1}^{n} d_i a_{ij}$, где n – количество испытуемых. Избирательность *i*-го испытуемого d_i является его весом при определении потенциала задания. Таким образом,

снимается одна из проблем при шкалировании заданий и теста в целом – проблема неадекватных испытуемых.

Для правильного шкалирования заданий желательно удаление из матрицы данных всех неадекватных испытуемых. Однако возникающее при этом нарушение репрезентативности выборки ведет к непредсказуемости результатов. Возможно, разумным компромиссом является допустимость удаления из выборки до 5% (а лучше – не более 1%) от общего количества испытуемых (т.е. необходимо удалить самых неадекватных) [1, с. 106].

При обработке результатов по второму варианту модели Бирнбаума адекватность испытуемых, увязываемая в данном контексте с их избирательностью d_i, напрямую учитывается при вычислении потенциалов заданий, поскольку элементы i-й строки матрицы тестирования, соответствующей i-му испытуемому, учитываются с весом d_i. Отсутствует необходимость удаления неадекватных испытуемых, поскольку их влияние на конечный результат окажется ничтожным ввиду близости d_i к нулю.

Работа со вторым вариантом модели Бирнбаума требует тщательного предварительного отбора заданий. Возникает необходимость сведения двух вариантов модели Бирнбаума в одну, с учетом избирательности как испытуемых, так и заданий. Известно, что основное преимущество логистических функций, лежащих в основе модели Раша и ее обобщениях – аналитическая простота и вытекающие из этого вычислительные преимущества. Принято считать, что исход «противоборства[1]», испытуемого и задания вероятностен, и задается т.н. *характеристическими кривыми* (функциями) уровня подготовленности испытуемого или уровня трудности задания [4, с. 14–15].

Однако возможна и иная точка зрения. Исход встречи жестко детерминирован и определяется лишь знаком разности потенциалов испытуемого и задания. Иными словами, если потенциал испытуемого выше потенциала задания – соответствующий элемент матрицы равен единице, если ниже – нулю. Данная модель известна в педагогических и психологических измерениях как модель Л. Гуттмана [7, с. 86–88]. Принятие этой модели требует объяснения вероятностного характера исхода встречи испытуемого с заданием, который может быть объяснен вероятностным же характером потенциалов испытуемого и заданий.

Другими словами, все влияющие на текущий потенциал испытуемого и задания факторы ведут к тому, что к моменту встречи испытуемого с заданием потенциалы принимают случайные значения, распределенные по некоторому закону. В качестве такого закона (ввиду многочисленности факторов и малой степени влияния каждого из них) целесообразно принять нормальный закон распределения. При этом потенциал испытуемого ха-

[1] Это метафора принадлежит Г. Рашу.

рактеризуется математическим ожиданием θ_i и стандартным отклонением σ_i, а потенциал задания математическим ожиданием β_j и стандартным отклонением σ_j. При независимости изменений потенциалов испытуемого и задания результирующее стандартное отклонение разности потенциалов испытуемого и задания σ_r определится по известной формуле $\sigma_r = \sqrt{\sigma_i^2 + \sigma_j^2}$ [2, с. 26].

Как известно, логистическая функция является хорошей аппроксимацией нормального закона распределения. Расхождение при этом не превышает 1% [4, с. 17]. При описании характеристической функции трудности задания моделью нормальной огивы

$$P_{ij} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{d_j(\theta_i - \beta_j)} e^{-0,5x^2}\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(\theta_i - \beta_j)/\sigma_j} e^{-0,5x^2}\,dx$$

видно, что $d_j = \frac{1}{\sigma_j}$.

Обобщая для других моделей (второго варианта модели Бирнбаума и рассматриваемой трехпараметрической модели), получаем $\sigma_i = \frac{1}{d_i}$ и $\sigma_r = \frac{1}{d_r}$. Таким образом, $d_r = \frac{1}{\sigma_r} = \frac{1}{\sqrt{\sigma_i^2 + \sigma_j^2}} = \frac{1}{\sqrt{\frac{1}{d_i^2} + \frac{1}{d_j^2}}} = \frac{1}{\sqrt{\frac{d_i^2 + d_j^2}{d_i^2 d_j^2}}} = \frac{d_i d_j}{\sqrt{d_i^2 + d_j^2}}$, и вероятность успеха испытуемого определится как $P_{ij}(\theta_i, \beta_j) = \left(1 + \exp\left(\frac{d_i d_j}{\sqrt{d_i^2 + d_j^2}}(\beta_j - \theta_i)\right)\right)^{-1}$. Возможен и отказ от параметров d_i и d_j, с заменой их, соответственно, на более прозрачно интерпретируемые σ_i и σ_j. В этом случае $P_{ij}(\theta_i, \beta_j) = \left(1 + \exp\left(\frac{\beta_j - \theta_i}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right)\right)^{-1}$.

Предлагаемая модель полностью симметрична относительно испытуемых и заданий, чего нельзя сказать относительно моделей Бирнбаума. В то же время, введение еще одного параметра повышает вероятность ложного решения при вычислении параметров сторон тестирования, что связано с мультимодальностью функции правдоподобия при совместном определении потенциалов испытуемых и заданий [7, с. 120–121]. Для надежного определения параметров количество испытуемых и заданий должно возрастать с увеличением количества определяемых параметров. Для од-

нопараметрической модели Раша количество заданий не должно быть менее 20, а испытуемых – менее 200. Для двухпараметрической модели Бирнбаума минимальное количество составляет соответственно 30 и 500, а для 3-параметрической модели Бирнбаума 60 и 1000 [7, с. 129]. Предполагая, что предлагаемая модель при равенстве количества параметров количеству параметров в 3-параметрической модели Бирнбаума обладает аналогичной устойчивостью решений, приходим к выводу, что предлагаемая модель пригодна для интерпретации результатов широкомасштабных процедур тестирования.

Оценивание испытуемого двумя параметрами (θ_i и σ_i) вместо традиционного оценивания одним параметром θ_i требует новых подходов при критериальной оценке испытуемого. Критерий, как правило, устанавливается на одномерной линейной шкале. Необходим некий конструкт на базе θ_i и σ_i, значение которого и сравнивается с критерием. Поскольку θ_i является математическим ожиданием, а σ_i – стандартным отклонением потенциала испытуемого, конструкт $\theta_t = \theta_i + t\sigma_i$ при произвольном выборе t задает возможный текущий потенциал испытуемого при решении некоторой задачи (моделируемой в тесте заданиями), и вероятность P_t проявления испытуемым потенциала θ_t или меньшего задается известной функцией интегрального нормального распределения: $P_t = \Phi(t)$, где $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-0,5x^2} dx$ [2, с. 401]. При установлении порогового критериального значения K испытуемые со значениями $\theta_i \geq K + 1{,}64\sigma_i$ с вероятностью 95% будут проявлять при решении задач более высокий потенциал, чем K. Очевидно, именно конструкт $(\theta_i - 1{,}64\sigma_i)$ и должен служить для отбора испытуемых при предъявляемом требовании (почти) безусловного решения поставленной задачи. Данный критерий вполне применим в условиях учебного заведения при необходимости определения, удовлетворяет ли учащийся зачетным требованиям.

Однако, при необходимости решения нетривиальных задач повышенной сложности, и допуская, что испытуемый, потенциально способный справиться с подобными задачами хотя бы с 5%-й вероятностью, безусловно представляет интерес, критерием отбора должен служить конструкт $(\theta_i + 1{,}64\sigma_i)$. И, наконец, при требованиях самого общего характера (например, при переводе отметок интервальной шкалы потенциалов в шкалу школьных или вузовских оценок для аттестации) значение t полагается равным $t = 0$, т.е используется значение потенциала испытуемого θ_i без учета стандартного отклонения σ_i.

Предложенная концепция порывает с традицией трактовать вузовскую оценку «незачет» как оценку «2», а «зачет» как обезличенные «3», «4» или «5». Оценка «зачет», при учете не только среднего уровня подго-

товки испытуемого, но и возможных его колебаний, гарантирует высокую профессиональную пригодность испытуемого для решения вполне определенного (ограниченного) круга задач, что вполне отвечает содержанию данного понятия. Оценка же «3» свидетельствует, что испытуемый достиг определенного уровня развития измеряемого качества, но этот уровень является усредненным по всему спектру проявляемых в конкретных ситуациях уровней.

Модели Раша, двухпараметрическая модель Бирнбаума [4, с. 14–16] и предложенная нами трехпараметрическая модель адекватны реальности при использовании заданий открытой формы, заданий на установление соответствия или правильной последовательности, при условии дихотомичной оценки, что практически устраняет возможность угадывания правильного ответа. В заданиях с выбором одного правильного ответа, с ограниченным (как правило, не более шести) количеством вариантов ответа, высока вероятность успешного угадывания правильного ответа, что диктует применение моделей с дополнительными параметрами, учитывающими конечную вероятность успешного угадывания. Одна из таких моделей была предложена Бирнбаумом, и в литературе она известна как трехпараметрическая модель Бирнбаума [4, с. 17]. В общем случае вероятность успеха испытуемого при учете вероятности угадывания c составляет $P_{ij} = c + (1-c)P_{ij}^{*}$, где P_{ij}^{*} – функция успеха для модели без учета фактора угадывания (например, для модели Раша или двухпараметрической модели Бирнбаума). При незнании испытуемого, полной его готовности прибегнуть к угадыванию с целью повышения своего тестового результата (а такие рекомендации для процедуры ЕГЭ даются, например, в [6]) и равновероятности дистракторов (к чему следует стремиться), значение c будет составлять $c_j = \frac{1}{m_j}$, где m_j – предусмотренное количество вариантов ответа в j-м задании. Легко видеть, что c_j – это минимально возможное значение P_{ij} для j-го задания. В литературе отмечается, что, как правило, $c_j < \frac{1}{m_j}$ [7, с. 89]. Значение c_j в трехпараметрической модели Бирнбаума вычисляется наряду со значениями θ_i, β_j и d_j. Одним из распространенных методов оценки параметров является метод наибольшего правдоподобия Р. Фишера. В качестве точечных оценок латентных (скрытых) параметров принимаются такие их значения, при которых функция правдоподобия достигает глобального максимума [4, с. 51]. Теоретически для заданий с неравномерным распределением дистракторов значение c_j может превышать $\frac{1}{m_j}$, од-

нако значение c_j, меньшее $\frac{1}{m_j}$, может свидетельствовать как о добротности дистракторов, так и о том, что не все испытуемые прибегают к угадыванию. Причиной может быть как добросовестность испытуемых (следование инструкции, предписывающей выбирать ответ лишь в случае полной уверенности в его правильности), так и непонимание вытекающих из угадывания преимуществ. Тестовые результаты, полученные применением подобных тестов, измеряют другие концептуально независимые свойства, т.е. свойства, совершенно не относящиеся к модели-конструкту диагностируемого качества [3, с. 263].

Аналогично расщеплению единого параметра d_r на параметры d_i и d_j возникает задача расщепления параметра c_r на параметры c_i и c_j, характеризующие вероятность угадывания со стороны испытуемого и со стороны задания. Если дистракторы задания предполагают вероятность успешного угадывания c_j, а испытуемый прибегает к угадыванию с вероятностью c_i, искомая результирующая вероятность при условии независимости событий c_r находится по формуле умножения вероятностей $c_r = c_i c_j$, что позволяет выписать выражение для функции успеха пятипараметрической модели тестирования: $P_{ij} = c_i c_j + (1 - c_i c_j)\left(1 + \exp\left(\frac{\beta_j - \theta_i}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right)\right)^{-1}$.

Есть основания полагать, что данная модель наиболее полно позволяет охарактеризовать как испытуемых, так и задания. В то же время обострение проблемы мультимодальности функции правдоподобия [7, с. 120–121] заставляет искать новые алгоритмы параметризации. Возможно более точное задание нулевых приближений, оцениваемых методами классической теории тестов, частично снимает проблему. Возможно также упрощение модели путем жесткого задания $c_j = \frac{1}{m_j}$, что уменьшает количество параметров до четырех, и дискретное задание параметра c_i – либо 0, либо 1, что вполне отвечает природе исследуемых объектов (сторон тестирования). Подтвердить или опровергнуть правомерность данного упрощения может изучение характера распределения значений c_i и c_j при оценке параметров по полной пятипараметрической модели.

Литература

1. *Аванесов В.С.* Основы научной организации педагогического контроля в высшей школе. М.: МИСиС, 1989.

2. *Варден Б.Л., ван дер.* Математическая статистика. М.: Изд. иностр. лит., 1960.

3. Михайлычев Е.А**.** Дидактическая тестология. М.: Народное образование, 2001.

4. Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. М.: Прометей, 2000.

5. Челышкова М.Б. Разработка педагогических тестов на основе современных математических моделей: Уч. пособие. М.: Исследовательский центр проблем качества подготовки специалистов, 1995.

6. Шаповал В.В. О нравственной оценке одной тактики “высшего результата” при сдаче ЕГЭ // Вопросы тестирования в образовании. 2002. № 2. С. 252–260.

7. **Suen H.K.** Principles of Test Theories. Hillsdale, NJ: Erlbaum, 1990.