УДК 37.01

ВЫРАВНИВАНИЕ ПОКАЗАТЕЛЕЙ В СЛУЧАЕ ЭКСПЕРТНОГО ОЦЕНИВАНИЯ ЗАДАНИЙ

Е.Ю. Карданова

Новгородский государственный университет им. Ярослава Мудрого
E-mail: kar@novsu.ac.ru

Описана процедура выравнивания показателей, полученных при использовании различных вариантов одного и того же теста, в случае экспертного оценивания заданий. Данная процедура доступна в рамках моделей Г. Раша многопараметрического анализа.

Необходимость в выравнивании возникает, главным образом, при использовании различных вариантов одного и того же теста, при создании банка заданий, при использовании технологий адаптивного тестирования, а также при мониторинговых исследованиях достижений учащихся. В настоящей статье проблема выравнивания рассматривается относительно двух первых аспектов – использования различных вариантов теста и создания банка заданий.

Как правило, разработчики тестов создают несколько вариантов одного и того же теста, которые стараются сделать максимально параллельными. Однако на практике полного совпадения характеристик добиться невозможно, и мы не можем рассматривать такие варианты как эквивалентные и способные заменить друг друга. В связи с этим особую важность приобретает вопрос о равноценности баллов, выставленных участникам тестирования по результатам выполнения различных вариантов теста. Особенно это актуально при массовых процедурах тестирования (таких как ЕГЭ), когда создаются десятки различных вариантов одного и того же теста.

Для того, чтобы баллы участников по различным вариантам теста можно было сравнивать между собой, необходимо осуществить процедуру выравнивания показателей, которая позволяет установить связь между баллами испытуемых по различным вариантам теста и затем поместить их на одну общую шкалу.

Аналогичная проблема возникает при сравнении уровней трудности заданий. Для того, чтобы можно было сравнить по уровню трудности задания различных вариантов теста, необходимо поместить параметры заданий из различных вариантов на одну шкалу, т. е. осуществить процедуру выравнивания. В частности, это необходимо при создании банка заданий. Банк заданий – это множество откалиброванных тестовых заданий (т. е. заданий с известными параметрами), параметры которых помещены на общую шкалу.

Таким образом, необходимость в выравнивании возникает, во-первых, при использовании различных вариантов теста, чтобы поместить меры всех испытуемых на общую шкалу, а, во-вторых, при создании банка заданий, чтобы поместить на одну шкалу параметры заданий из различных вариантов теста.

Современная теория тестирования ТМПТ (теория моделирования и параметризации тестов [1]) позволяет выполнить процедуру выравнивания показателей различных вариантов и осуществить шкалирование на единой метрической шкале. Основы теории выравнивания с точки зрения ТМПТ изложены в работах [2, 3]. С различными методами и процедурами выравнивания в рамках моделей Г. Раша можно ознакомиться в работе [4]. Наиболее распространенный метод выравнивания – метод общих заданий, который подразумевает, что в процессе композиции вариантов теста разработчики включают в каждый вариант узловые задания, общие для нескольких вариантов. То есть, каждый вариант теста имеет подмножество заданий, перекрывающихся с другими вариантами теста. Параметры узловых заданий позволяют связать различные варианты теста и перенести оценки всех параметров на общую шкалу. Процедура выравнивания с помощью общих заданий подробно описана в [4].

Традиционно анализ результатов тестирования концентрируется на оценивании параметров заданий и мер испытуемых. Однако в случае, если выполнение заданий теста оценивается экспертами (примером таких заданий являются задания типа С в КИМах ЕГЭ), оценки испытуемых не могут рассматриваться как объективные в силу их зависимости от уровня строгости и других характеристик конкретного эксперта. Поэтому необходим дополнительный анализ, предполагающий использование математических моделей, параметрами которых, помимо характеристик участников тестирования и самого теста, являются и характеристики экспертов, оценивающих выполнение этих заданий. Многопараметрический подход в рамках ТМПТ позволяет оценить влияние различных аспектов деятельности экспертов на оценки испытуемых. Математические модели многопараметрического анализа в рамках моделей Г. Раша описаны в работе [5]. Применение этих моделей позволяет получить в конечном итоге объективные меры испытуемых, инвариантные относительно уровней трудности заданий конкретного теста и уровней строгости конкретной группы экспертов, оценивающих выполнение этих заданий.

Наличие общих экспертов, оценивающих выполнение заданий различных вариантов теста, позволяет осуществить процедуру выравнивания показателей, основанную на характеристиках этих экспертов. Таким образом, связывание двух вариантов теста может быть осуществлено на основе не только общих заданий, но и общих экспертов, обоснованию чего и посвящена настоящая статья.

Рассмотрим одну из моделей Раша многопараметрического анализа [5]:

$$\ln \frac{p_{nilk}}{p_{nil(k-1)}} = \theta_n - \delta_{ik} - \xi_l, \qquad (1)$$

где p_{nilk} – вероятность того, что испытуемый n получит k баллов за выполнение задания i при оценке экспертом l; $p_{nil(k-1)}$ – вероятность того, что испытуемый n получит k–1 баллов за выполнение задания i при оценке экспертом l; θ_n – уровень подготовленности испытуемого n; δ_{ik} – уровень трудности выполнения k-ого шага в задании i; ξ_l – уровень строгости эксперта l. Модель (1) представляет собой расширение модели Раша с произвольными промежуточными категориями выполнения заданий ([6]) на случай многопараметрического анализа, когда в расчет принимаются не только параметры заданий и испытуемых, но и параметры экспертов, оценивающих выполнении заданий испытуемыми.

Рассмотрим для простоты дихотомический случай (в случае политомического оценивания заданий все формулы аналогичны). Имеем:

$$\ln \frac{p_{nik}}{1 - p_{nik}} = \theta_n - \delta_i - \xi_k$$

или

$$p_{nik} = \frac{\exp(\theta_n - \delta_i - \xi_k)}{1 + \exp(\theta_n - \delta_i - \xi_k)}, \qquad (2)$$

где p_{nik} – вероятность того, что испытуемый n с уровнем подготовленности θ_n получит 1 балл за выполнение задания i с уровнем трудности δ_i при оценке экспертом k с уровнем строгости ξ_k. Эта модель представляет собой расширение однопараметрической дихотомической модели Раша на случай принятия в расчет дополнительного фактора – уровня строгости эксперта.

В результате применения процедур оценивания все параметры модели будут расположены на общей метрической шкале логитов с указанием точности оценки [1, 5]. Однако для каждого варианта теста шкала будет своей. Это объясняется тем, что шкала логитов, на которой находятся оценки испытуемых, параметры заданий и экспертов, является интервальной. Как известно [1], интервальные шкалы не имеют абсолютного нуля. С математической точки зрения это означает, что значения всех параметров модели определены с точностью до сдвига

$$\theta_n = \theta_n^{'} + \alpha + \beta, \quad \delta_i = \delta_i^{'} + \alpha, \quad \xi_k = \xi_k^{'} + \beta, \qquad (3)$$

т. к. соотношение (2), очевидно, инвариантно относительно преобразования (3).

Следовательно, на каждой из шкал существует неопределенность с выбором нуля (начала отсчета), которая устраняется в процессе калибровки. Существуют различные способы сделать это, но чаще всего при применении моделей Раша многопараметрического анализа для устранения указанной неопределенности каждая шкала центрируется таким образом, чтобы сумма уровней трудности всех заданий теста и сумма уровней строгости всех экспертов равнялись нулю. То есть шкала центрируется в среднем значении уровней трудностей всех заданий и в среднем значении уровней строгости всех экспертов. Другими словами, при оценивании параметров полагают, что

$$\bar{\delta} = \frac{\sum_{i=1}^{I} \delta_i}{I} = 0, \quad \bar{\xi} = \frac{\sum_{k=1}^{K} \xi_k}{K} = 0,$$

где I – число заданий в тесте, K – число экспертов.

Очевидно, что различные варианты теста имеют различные средние значения трудностей заданий и различные средние значения уровней строгости экспертов, вследствие чего возникает проблема несопоставимости баллов, полученных по различным вариантам теста.

Предположим, что нам известны результаты тестирования испытуемых по двум заданиям, принадлежащим различным вариантам теста, но одинакового уровня трудности, которые получены при оценивании двумя экспертами с одинаковым уровнем строгости. Пусть δ_1 и δ_2 – уровни трудности задания по 1-му и 2-му вариантам; ξ_1 и ξ_2 – уровни строгости эксперта 1-ого и 2-го вариантов. Положим

$$\alpha = \delta_1 - \delta_2, \quad \beta = \xi_1 - \xi_2.$$

Оценив разности α и β, можно перевести все значения параметров испытуемых, заданий и экспертов со шкалы одного варианта на шкалу другого. Таким образом, для выравнивания показателей по двум вариантам в случае экспертного оценивания заданий необходимо иметь общих экспертов и общие задания.

Существуют различные процедуры выравнивания, различающиеся по способу калибровки (отдельная или одновременная) и по использованию общих элементов (меняются или нет в процессе калибровки параметры связующих элементов). Рассмотрим процедуру выравнивания при отдельной калибровке всех вариантов с последующей трансформацией всех мер на общую шкалу.

Предварительно вычисляются константы смещения, которые используются для преобразования всех мер на общую шкалу.

1) Вычисление константы смещения по общим экспертам:

$$t^{(1)} = \frac{\sum_{k=1}^{n} (\xi_{k1} - \xi_{k2})}{n}, \qquad (4)$$

где $t^{(1)}$ – константа смещения по экспертам между вариантами 1 и 2, ξ_{k1} – оценка строгости k-ого эксперта по результатам оценивания 1-ого варианта, ξ_{k2} – оценка строгости k-ого эксперта по результатам оценивания 2-ого варианта, n – число общих экспертов в рассматриваемых вариантах. Так как в моделях Раша нам доступна информация о точности полученных оценок, то лучше пользоваться

взвешенной константой смещения, которая представляет собой взвешенное среднее разностей оценок строгости экспертов:

$$t^{(2)} = \frac{\sum_{k=1}^{n} (\xi_{k1} - \xi_{k2}) \cdot w_{k12}}{\sum_{k=1}^{n} w_{k12}}. \quad (5)$$

Здесь $t^{(2)}$ – взвешенная константа смещения между вариантами 1 и 2; $w_{k12}=1/\sigma_{k12}^2$ – вес k-ого эксперта, где σ_{k12} определяется формулой:

$$\sigma_{k12}^2 = \sigma_{k1}^2 + \sigma_{k2}^2, \quad (6)$$

σ_{k1}, σ_{k2} – средние квадратичные ошибки измерения k-ого эксперта при калибровке вариантов 1 и 2 соответственно (дисперсия единицы веса принята равной единице).

Средние квадратичные ошибки оценок (4) и (5) вычисляются соответственно по формулам:

$$\sigma(t^{(1)}) = \frac{1}{n} \cdot \sqrt{\sum_{k=1}^{n} \sigma_{k12}^2}, \quad \sigma(t^{(2)}) = \left(\sum_{k=1}^{n} w_{k12} \right)^{-1/2}.$$

2) Вычисление константы смещения по общим заданиям:

$$\tau^{(1)} = \frac{\sum_{i=1}^{m} (\delta_{i1} - \delta_{i2})}{m}, \quad (7)$$

где $\tau^{(1)}$ – константа смещения по заданиям между вариантами 1 и 2, δ_{i1} – оценка трудности i-ого задания 1-ого варианта, δ_{i2} – оценка трудности i-ого задания 2-ого варианта, – число общих заданий в рассматриваемых вариантах. Взвешенная версия константы смещения имеет вид:

$$\tau^{(2)} = \frac{\sum_{i=1}^{m} (\delta_{i1} - \delta_{i2}) \cdot w_{i12}}{\sum_{i=1}^{m} w_{i12}}. \quad (8)$$

Здесь $\tau^{(2)}$ – взвешенная константа смещения между вариантами 1 и 2; $w_{i12}=1/\sigma_{i12}^2$ – вес i-ого задания, где σ_{i12} определяется формулой:

$$\sigma_{i12}^2 = \sigma_{i1}^2 + \sigma_{i2}^2, \quad (9)$$

σ_{i1}, σ_{i2} – средние квадратичные ошибки измерения i-ого задания при калибровке вариантов 1 и 2 соответственно.

Средние квадратичные ошибки оценок (7) и (8) вычисляются соответственно по формулам:

$$\sigma(\tau^{(1)}) = \frac{1}{m} \cdot \sqrt{\sum_{i=1}^{m} \sigma_{i12}^2}, \quad \sigma(\tau^{(2)}) = \left(\sum_{i=1}^{m} w_{i12} \right)^{-1/2}.$$

Далее все параметры заданий, экспертов и испытуемых переводятся со шкалы 2-ого варианта на шкалу 1-ого. С этой целью положим:

$$\sigma'_{i2} = \delta_{i2} + \tau, \quad i = 1,...,I;$$
$$\xi'_{k2} = \xi_{k2} + t, \quad k = 1,...,K;$$
$$\theta'_{n2} = \theta_{n2} + \tau + t, \quad n = 1,...,N. \quad (10)$$

Здесь δ_{i2} – оценка трудности i-ого задания 2-го варианта, δ'_{i2} – оценка трудности этого же задания на шкале 1-ого варианта (которая выбрана в качестве единой шкалы); I – число заданий; ξ_{k2} – оценка строгости k-ого эксперта 2-ого варианта, ξ'_{k2} – оценка строгости этого же эксперта на единой шкале, K – число экспертов 2-ого варианта; θ_{n2} – оценка уровня подготовленности испытуемого n 2-ого варианта, θ'_{n2} – оценка уровня того же испытуемого на единой шкале, N – число испытуемых 2-ого варианта.

Сдвинутые указанным способом оценки параметров 2-ого варианта будут находиться на шкале 1-ого варианта. Если бы модель Раша была полностью адекватна результатам тестирования, контингенты участников тестирования, выполнявших различные варианты теста, были однородны, а ошибки измерения отсутствовали, то указанный сдвиг (10) привел бы к полному совпадению трудностей узловых заданий и уровней строгости общих экспертов в вариантах 1 и 2. Однако в реальных условиях такое совпадение практически невероятно. Вследствие этого необходимо оценить качество узловых заданий и экспертов, что осуществляется по двум направлениям:

1) оценивание узловых заданий и экспертов в смысле согласия реальных ответов испытуемых на эти задания и их ожидаемых значений, предсказанных используемой моделью [1];
2) оценивание узловых заданий и экспертов как элементов связующего звена между различными вариантами. Это подразумевает оценку устойчивости и инвариантности оценок параметров узловых заданий и экспертов в различных вариантах. С этой целью для всех узловых заданий вычисляется статистика

$$u_i = \frac{\delta_{i1} - \delta'_{i2}}{\sigma_{i12}}, \quad i = 1,...,m, \quad (11)$$

где δ'_{i2} и σ_{i12} определяются соответственно соотношениями (10) и (9). Аналогично для всех общих экспертов вычисляется статистика

$$v_k = \frac{\xi_{k1} - \xi'_{k2}}{\sigma_{k12}}, \quad k = 1,...,n, \quad (12)$$

где ξ'_{k2} и σ_{k12} определяются соответственно соотношениями (10) и (6).

Статистики (11) и (12) имеют асимптотически нормальное распределение с нулевым математическим ожиданием и единичной дисперсией. Поэтому значения этих статистик, по модулю большие 2

(то есть $|u_i|>2$ и $|v_k|>2$), указывают на слишком большие отклонения оценок трудности данного задания или оценок строгости данного эксперта соответственно. Такое задание или такой эксперт не демонстрируют достаточной устойчивости своих оценок в различных вариантах и их полезно удалить из процесса выравнивания.

Таким образом, рассматриваемая процедура выравнивания подразумевает отдельную калибровку всех вариантов и последующую трансформацию мер на общую шкалу. Процедура состоит из следующих этапов:

1. *Выбор общей шкалы.*

При выравнивании нескольких вариантов за общую шкалу, как правило, выбирается шкала одного из вариантов.

2. *Отбор узловых заданий и экспертов.*

Не рекомендуется выбирать в качестве узловых задания, которые имеют экстремальные уровни трудности (то есть очень легкие или очень трудные) и экспертов, которые имеют экстремальные уровни строгости (то есть очень строгие или очень снисходительные). Это связано в первую очередь с тем, что оценки параметров в центре распределения имеют меньшую ошибку измерения, чем на его краях. При априорном выборе количества узловых заданий и экспертов следует учитывать возможности удаления некоторых из них на последующих этапах.

3. *Калибровка всех вариантов.*

На этом этапе оцениваются меры испытуемых и параметры заданий и экспертов по каждому из вариантов. При этом необходимо проверить согласованность реальных ответов испытуемых на задания теста и их ожидаемых значений, предсказанных моделью измерения [1]. Такой анализ особо тщательно необходимо сделать по узловым заданиям и экспертам. Задания и эксперты, не демонстрирующие согласия с используемой моделью, желательно удалить из процесса выравнивания.

4. *Вычисление констант смещения.*

Могут быть использованы оценки (4) и (7) или их взвешенные версии (5) и (8). Одновременно оцениваются стандартные ошибки измерения полученных значений.

5. *Оценивание узловых заданий и экспертов.*

На этом этапе узловые задания и эксперты оцениваются как элементы связующего звена между вариантами. По формулам (10) вычисляются смещенные значения уровней трудности узловых заданий и уровней строгости экспертов *2-го* варианта и с помощью статистик (11) и (12) оценивается устойчивость оценок трудности каждого узлового задания и строгости каждого общего эксперта в различных вариантах. Узловые задания и эксперты, демонстрирующие неустойчивость мер, желательно удалить из процесса выравнивания. Однако необходимо выяснить причины этой неустойчивости. Например, к числу наиболее вероятных причин неустойчивости параметров заданий можно отнести следующие: различное положение узловых заданий в разных вариантах теста, влияющее на их трудность; незначительные с точки зрения разработчика изменения задания, повлекшие существенное изменение его трудности и невозможность использования в качестве узлового; завуалированная подсказка к узловому заданию, содержащаяся в других заданиях данного варианта и облегчающая выполнение этого задания. После удаления неподходящих для выравнивания заданий и экспертов необходимо заново откалибровать варианты и вычислить оценки смещения, то есть повторить этапы 3 и 4.

6. *Перевод всех параметров на единую шкалу.*

При создании банка заданий необходимо преобразовывать на единую шкалу только трудности заданий. При выравнивании различных вариантов теста необходимо преобразовывать на общую шкалу меры испытуемых. И то, и другое выполняется прибавлением соответствующих оценок смещения к оценкам соответствующих параметров по каждому из вариантов (формулы (10)).

С целью апробации описанной процедуры выравнивания был проделан численный эксперимент с реальными данными ЕГЭ по английскому языку 2004 г. Тесты ЕГЭ по английскому языку не содержали общих заданий. Поэтому в качестве «общего» предлагается выбрать одно из заданий части С, а именно написание открытки. Это задание оценивалось экспертами по 5 критериям: содержание, организация текста, лексика, грамматика, орфография и пунктуация. Таким образом, в качестве общих заданий можно (по согласованию с разработчиками теста) выбрать некоторые критерии оценивания этого задания, наименее зависящие от содержания открытки (например, критерии «Организация текста» и «Лексика»). Множества экспертов, оценивающих выполнение заданий различных вариантов ЕГЭ по английскому языку, частично перекрывались, что позволило осуществить процедуру выравнивания.

В заключении отметим, что провести процедуру выравнивания показателей можно только в рамках современной теории тестирования ТМПТ. Немногочисленные методы, которые могут быть использованы в классической теории тестирования, требуют специфических ограничений и невыполнимых предположений, и в любом случае не предполагают создания общей метрической шкалы. Более того, классическая теория тестирования не позволяет учитывать характеристики экспертов, оценивающих выполнение заданий теста.

СПИСОК ЛИТЕРАТУРЫ

1. Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. – М.: Прометей, 2000. – 169 с.
2. Wright B.D., Stone M.N. Best Test Design. Rasch Measurement. – Chicago: Mesa Press, 1979. – 223 p.
3. Kolen M.J., Brennan R.L. Test Equating: Methods and Practices. – N.Y.: Springer, 1995. – 334 p.
4. Карданова Е.Ю., Нейман Ю.М. Проблема выравнивания в современной теории тестирования // Вопросы тестирования в образовании. – 2003. – № 8. – С. 21–40.
5. Карданова Е.Ю. Математические модели многофасетного анализа // Вопросы тестирования в образовании. – 2004. – № 11. – С. 11–27.
6. Карданова Е.Ю., Нейман Ю.М. Основные модели современной теории тестирования // Вопросы тестирования в образовании. – 2003. – № 7. – С. 12–37.

Поступила 16.02.2006 г.

УДК 530.1(075.8)

МЕТОДИЧЕСКАЯ СИСТЕМА ОБУЧЕНИЯ ФИЗИКЕ В ТЕХНИЧЕСКОМ ВУЗЕ

Г.В. Ерофеева, Е.А. Склярова, Ю.Ю. Крючков

Томский политехнический университет
E-mail: skea@tpu.ru

Рассмотрены концепция, модель и методическая система обучения студентов физике в техническом университете. Концепция, модель и методическая система обучения разработаны с учетом специфики технического университета, а также особенностей студентов, выбравших техническое направление. Кроме того, учтены направления модернизации образования в России, современные методы, принципы и подходы в обучении физике и другим дисциплинам.

Модернизация российского образования направлена на дальнейшую информатизацию его, гуманизацию, гуманитаризацию и демократизацию, а также личностно-ориентированный подход к обучению [1].

В техническом университете новая парадигма образования не отменяет прежнюю (парадигму, ведущий лозунг которой были знания, умения, навыки, воспитание), ее важнейшим компонентом является концепция фундаментализации, которая трактует фундаментальность как категорию качества образования и образованности личности [2]. В современных условиях преобразований российского общества и направлений модернизации образования фундаментализация высшего образования рассматривается как системное и всеохватывающее обогащение учебного процесса фундаментальными знаниями и методами творческого мышления, выработанными фундаментальными науками [3].

Поскольку прикладные науки возникают и развиваются на основе постоянного использования фундаментальных законов природы, то общепрофессиональные и специальные дисциплины тоже становятся носителями фундаментальных знаний. Следовательно, в процесс фундаментализации высшего образования должны быть вовлечены наряду с естественнонаучными общепрофессиональные и специальные дисциплины.

В техническом университете такой подход обеспечит фундаментализацию обучения бакалавра, магистра, специалиста, аспиранта и докторанта на всех этапах, начиная с первого курса.

Из известных научных концепций усвоения социального опыта в России наибольшее распространение получили деятельностный подход, а именно его направления: теория содержательного обобщения Д.Б. Эльконина – В.В. Давыдова, теория поэтапного формирования умственных действий П.Я. Гальперина – Н.Ф. Талызиной [4].

Из принципов деятельности: индивидуализация, контекстность обучения (обучение ведется в контексте будущей профессиональной деятельности), проектно-организованное и проблемно-ориентированное обучения (развитие креативного мышления).

Дидактические принципы разработки методической системы: научность, наглядность, доступность, интерактивность, адаптивность, профессиональная направленность содержания, системность.

Таким образом, концепция содержательной и процессуальной частей образовательных программ бакалавров, магистров и др. в техническом университете содержит единство фундаментальной, гуманитарной и профессиональной составляющих, информатизацию обучения, учет и развитие подходов, принципов и требований к компетенциям выпускников согласно Государственному образовательному стандарту (ГОС).

Возможные пути реализации концепции обучения в университете представлены на рис. 1.

Специфика учебного процесса в техническом университете состоит в практической направленности изучаемых дисциплин, при этом физика представляет собой фундаментальную основу дисциплин технического направления (электротехника, микроэлектроника, материаловедение, сопро-