УДК 001.53:[530.1+371.263]

СТАТИСТИКО-ФИЗИЧЕСКАЯ МОДЕЛЬ ПЕДАГОГИЧЕСКОГО ИЗМЕРЕНИЯ

СВИРИДОВ Владимир Владимирович,
доктор физико-математических наук, профессор кафедры общей физики;
КОЧУКОВА Марина Викторовна,
ассистент кафедры нормальной физиологии,
Воронежский государственный медицинский университет
ХОНИК Виталий Александрович, доктор физико-математических наук, проректор по научной работе,
Воронежский государственный педагогический университет

АННОТАЦИЯ. *Рассматривается физическая модель педагогического измерения, воспроизводящая основные выводы известных моделей Раша, Бирнбаума и Дроздова, но при этом более интуитивная и эвристически ценная. В рамках модели показано, что для правильной интерпретации результатов реальных педагогических измерений следует определять не только уровень каждого испытуемого и задания, но и меру неопределенности этих уровней. Показано, что дифференцирующая способность задания в действительности зависит также от испытуемого, предложены процедуры учета этой зависимости.*
КЛЮЧЕВЫЕ СЛОВА: *педагогические измерения, латентные переменные, модель Раша, модель Бирнбаума, статистическая физика, распределение Ферми-Дирака, физическая модель.*

THE STATISTICAL AND PHYSICAL MODEL OF THE PEDAGOGICAL MEASUREMENT

SVIRIDOV V. V.,
Dr. Physics and Mathem. Sci., Professor of General Physics Department;
KOCHUKOVA M. V.,
Assistant of the Department of Normal Physiology,
Voronezh State Medical University named after N.N. Burdenko;
KHONIK V. A.
Dr. Physics and Mathem. Sci., Professor, Head of General Physics Department, Vice-Rector on Scientific Work,
Voronezh State Pedagogical University

ABSTRACT. *The article reviews the physical model of the pedagogical measurement. The model mainly reproduces the conclusions of the known models by Rasch, Birnbaum and Drozdov but appears to be more intuitive and of higher heuristic value. Within the framework of the model it is shown that to interpret properly the results of a real pedagogical measurement one should determine not only the levels of every testee and the task but also the degree of uncertainty of these levels. It is indicated that the discrimination power of the task also depends on the testee. The procedures to account for this dependence are proposed.*
KEY WORDS: *pedagogical measurements, latent variables, Rasch model, Birnbaum model, statistical physics, Fermi-Dirac distribution, physical model.*

1. Актуальность проблемы интерпретации результатов педагогических измерений

Развитие информационных технологий стимулирует широкое распространение формализованных педагогических измерений – прежде всего, в форме тестирования – как инструмента контроля состояния и деятельности образовательных систем. Однако интерпретация результатов педагогических измерений (далее – ИРПИ) сталкивается с проблемами, постоянно озвучиваемыми педагогической (и не только) общественностью, как-то:

– можно ли на основании успешного прохождения теста говорить о глубоком усвоении знаний или же только о хорошей механической памяти;

– не дискриминируем ли мы нестандартных личностей, заставляя их проходить стандартные тесты и т.д.

Фундаментальной основой этих проблем служит противоречие между жесткой формальной структурой тестирующей системы и глубоко неформальной, нередуцируемо сложной природой тестируемых индивидуумов.

Проблема ИРПИ усугубляется диалектически противоречивыми свойствами агента, осуществляющего взаимодействие между тестирующей системой и тестируемыми, – банка тестовых заданий (ТЗ). Заказчик обычно выставляет авторам ТЗ весьма жесткие и по возможности формализованные требования, призванные обеспечить желаемые свойства банка ТЗ уже на этапе его разработки. Однако контекстная зависимость ТЗ и влияние личности разработчика всё равно, как правило, настолько существенны, что точно соблюсти заданный уровень сложности и валидности банка ТЗ во время его разработки оказывается невозможно. Именно поэтому обязательной считается коррекция банка постфактум, по итогам его апробации [1].

2. Интерпретация результатов тестирования в рамках модели Раша

Один из наиболее популярных в настоящее время подходов к разрешению указанных выше противоречий основан на вероятностной модели ИРПИ, предложенной в 1960 г. Г. Рашем [2]. Впоследствии Раш распространил свой подход на более широкий круг задач, когда по результатам измерения наблюдаемых переменных необходимо делать максимально объективные (то есть, формализованные) выводы о значениях т.н. *латентных переменных* – параметров, которые непосредственному измерению не

Информация для связи с авторами: xelym@mail.ru

поддаются, но предположение о существовании которых позволяет объяснить эмпирически установленные закономерности [3]. К латентным переменным относятся многие важные, но плохо формализуемые характеристики личности, такие как «обученность», «толерантность», «честность» и даже весьма актуальная для современного российского образования «степень сформированности компетенций» [4].

Базой для Раш-анализа результатов педагогических измерений служит эмпирическая закономерность, выявленная Рашем первоначально в данных тестирования датских школьников на скорость и точность чтения. Для простейшей формы педагогических измерений с оценкой ТЗ по дихотомической шкале (правильно/неправильно) эту закономерность можно изложить следующим образом. Рассмотрим отношение k вероятностей правильного ответа, f, и ошибки, $1-f$, при выполнении испытуемым некоторого ТЗ: $k = f/(1-f)$. Тогда отношение величин k для разных ТЗ не зависит от того, *какого именно* испытуемого мы берем. Другими словами, это отношение зависит *только от самих ТЗ* и может служить мерой их относительной трудности. Аналогично, отношение величин k для разных испытуемых, выполняющих одно и то же ТЗ, не зависит от того, *какое именно* ТЗ использовано. Это отношение *зависит только от самих испытуемых* и может служить объективным мерилом различия их уровней подготовленности.

Из этого закона, установленного Рашем и подтвержденного на обширном статистическом материале, следует [5], что можно определить и выразить численно уровень готовности μ каждого испытуемого и уровень трудности ε каждого ТЗ[2] таким образом, что вероятность успешного выполнения i-го ТЗ n-м испытуемым $p_{ni} = f(\varepsilon_i, \mu_n)$, где f – некоторая универсальная функция, играющая роль метрики отображения шкалы уровней μ и ε на шкалу вероятностей p.

Явный вид метрической функции f не вытекает из априорных соображений, но он должен быть таким, чтобы обеспечивать выполнение сформулированного выше основного закона Раша и ряда очевидных требований:

1) $f(\varepsilon,\mu) \xrightarrow[\mu<<\varepsilon]{} 0$ (слабый испытуемый почти никогда не справляется с трудным заданием);

2) $f(\varepsilon,\mu) \xrightarrow[\mu>>\varepsilon]{} 1$ (сильный испытуемый почти всегда справляется с легким заданием);

3) $f(\varepsilon,\mu) \xrightarrow[\mu=\varepsilon]{} 1/2$ (если трудность задания соответствует уровню испытуемого, его шансы справиться с заданием – 50 на 50);

4) $\partial f/\partial\mu > 0$ (чем лучше подготовлен испытуемый, тем больше его шансы добиться успеха);

5) $\partial f/\partial\varepsilon < 0$ (чем труднее задание, тем меньше вероятность его успешного решения).

В модели, предложенной самим Рашем [2; 5; 6], в качестве метрики f используется простая гладкая функция, отвечающая сформулированным требованиям, – логистическая:

$$f(\varepsilon,\mu) = \frac{1}{1+e^{\varepsilon-\mu}}. \qquad (1)$$

Легко видеть, что она правильно воспроизводит основной закон Раша. Действительно, согласно (1), отношение вероятностей правильного ответа и ошибки $k = f/(1-f) = e^{\mu-\varepsilon}$. Для данного испытуемого (фиксированный уровень μ) и двух разных ТЗ отношение $k_1/k_2 = e^{\varepsilon_2-\varepsilon_1}$ зависит только от разности трудностей заданий, а для данного ТЗ (фиксированная трудность ε) – только от разности уровней готовности испытуемых. Существуют [5] и более содержательные соображения в пользу выбора метрической функции в виде (1).

Принимая модель Раша (1) и имея матрицу $\mathbf{A}_{ni}$ ответов испытуемых (например, $A_{ni} = 1$ может означать, что n-й испытуемый справился с i-м ТЗ, $A_{ni} = -1$ – не справился, $A_{ni} = 0$ – не выполнял это ТЗ), можно подобрать уровни готовности испытуемых μ_n и трудности ТЗ ε_i, которые минимизируют отклонения выборочных частот решаемости ТЗ от априорных вероятностей, определяемых формулой (1). Существуют доступные математические пакеты для решения этой задачи [7], которые, будучи интегрированы в тестирующую компьютерную систему, способны вычислять объективные оценки подготовленности тестируемых μ_n, трудности ε_i и валидности ТЗ практически в режиме реального времени.

3. Достоинства и ограничения модели Раша

В литературе по педагогическим измерениям и, шире, по теории латентных переменных (например, [5]), отмечается, что ИРПИ в рамках модели Раша имеет ряд очевидных преимуществ перед классической теорией педагогических измерений.

Уровни подготовленности испытуемых и трудности заданий определяются не по отдельности (что всегда порождает вопросы об эталонах трудности ТЗ и подготовки студента), а в рамках единой самосогласованной процедуры.

Шкала оценок параметров μ и ε линейна, в противоположность классической теории, в которой разница в 1 балл на краю диапазона набираемых баллов означает гораздо бо́льший разрыв в уровне подготовки, чем в середине диапазона (если трудное задание не решил никто, кроме одного студента, он явно на голову сильнее всех остальных).

Уровни готовности испытуемых и трудности заданий в модели Раша измеряются на одной и той же шкале (иначе их нельзя было бы вычитать друг из друга, как делается в (1)).

Наконец, теоретически в модели Раша оценка уровня испытуемого не зависит от использованного набора ТЗ, и наоборот.

Однако, при всей привлекательности и популярности Раш-анализа, с ним связан ряд нерешенных проблем принципиального характера. Рассмотрим две из них, наиболее актуальные в контексте данной работы.

Первая проблема – это сущность латентных переменных (в модели Раша (1) это μ и ε). Что они собой представляют? Какое свойство испытуемых и какое качество ТЗ в действительности выражают? Модель (1) позволяет рассчитать их значения, но не уточняет их смысл.

С той же проблемой в свое время столкнулся создатель системы «измерения интеллекта» (IQ) Г. Айзенк [8]. Говорят, что во время одной из дискуссий вокруг IQ, будучи доведен до отчаяния настойчивыми требованиями оппонентов дать конструктивное определение интеллекта, Айзенк заявил:

[2] Обычно в педагогической литературе уровни готовности испытуемого и трудности задания обозначаются иными символами [5; 6]. Однако для сопоставления идей теории педагогических измерений и статистической физики, которое является целью настоящей работы, удобнее использованные здесь обозначения.

«Интеллект – это то, что измеряют тесты IQ!». Определение операциональное и конструктивное, но не очень удовлетворительное, если наша цель шире, чем простое выяснение IQ данного субъекта, и заключается, скажем, в отборе претендентов на некоторый пост.

Поясним суть проблемы еще одним примером. Представим, что студенты-историки вместо теста по Древнему Китаю по ошибке выполнили тест по современной экономике Юго-Восточной Азии. Математический алгоритм сработает, и каждому студенту будет присвоено некоторое значение параметра μ. Однако вряд ли разумно было бы интерпретировать его как степень компетентности студента в азиатских экономических коллизиях. В сложившейся ситуации полезней было бы, скорее, знать какие-то величины, характеризующие степень растерянности и дезориентированности испытуемых. Но ни μ, ни тем более ε на эту роль, очевидно, не годятся.

Вторая проблема заключается в том, как определить единицы измерения латентных переменных μ и ε. Обычный ответ в рамках модели Раша гласит [5; 6], что μ и ε измеряются в *логитах* – таких единицах, что изменение μ или ε на 1 логит приводит к изменению отношения вероятности правильного ответа к вероятности ошибки, $k = f/(1-f)$, в e раз. Однако такой подход, не основанный на сравнении с универсальным эталоном, оставляет возможность того, что логиты, определенные на разных контингентах или на разных банках ТЗ, имеют разную величину. Практика педагогических измерений показывает, что эта возможность часто становится реальностью [9], в связи с чем возникает необходимость регулирования величины логита.

4. Модель Бирнбаума

Проблему унификации величины логита практики, использующие модель Раша для ИРПИ, пытаются решать путем введения т.н. якорных ТЗ [9], которые предлагаются всем испытуемым и призваны играть роль эталона при построении шкалы измерения μ и ε.

Более фундаментальный подход предложен А. Бирнбаумом [10]. Он основывается на усложнении модели Раша (1) путем введения нового параметра d:

$$f_B(\varepsilon,\mu,d) = \frac{1}{1+\mathrm{e}^{d(\varepsilon-\mu)}}. \qquad (2)$$

Параметр d принято интерпретировать как характеристику ТЗ [6; 10]. Чем больше d, тем быстрее меняется f_B с изменением μ вблизи точки $\mu = \varepsilon$, в связи с чем d называют дифференцирующей способностью ТЗ. Распределение (2) получило название «модель Бирнбаума» [6]. Оно позволяет регулировать эффективную величину логита через подгоночный параметр d.

К сожалению, добавление в теорию нового параметра, хотя и улучшает согласие между эмпирическими данными и теорией, не всегда гарантирует понимание смысла параметра и сути того, что призвана описывать теория. Скажем, если считать d характеристикой только ТЗ, то при $d >> 1$ испытуемый с уровнем готовности чуть ниже сложности задания ($\mu < \varepsilon - d^{-1}$) должен практически гарантированно ошибаться: $f_B \approx \mathrm{e}^{-d(\varepsilon-\mu)} << 1$. Однако если испытуемый подготовлен еще хуже (уровень μ гораздо ниже ε), то в реальности он начнет выбирать ответ случайным образом. При этом вероятность правильного ответа будет хотя и мала ($1/n$ для ТЗ на выбор единственного верного ответа, где n – количество предлагаемых вариантов ответа), но всё же мала не экспоненциально, как предсказывает модель (2). А это возможно, лишь если предположить, что для такого испытуемого d становится малой величиной: $d \sim (\varepsilon-\mu)^{-1} < 1$. Стало быть, дифференцирующая способность должна зависеть не только от задания, но и от испытуемого. Вопрос о характере этой зависимости требует дополнительного обоснования, которое нелегко найти, оставаясь в рамках достаточно абстрактной и малоинтуитивной модели (2).

5. Модель Бирнбаума и распределение Ферми-Дирака

Проблема сущности латентных переменных продолжает оставаться предметом дискуссий, затрагивающих широкую предметную область [11; 12]. Несмотря на быстрое развитие компьютерных технологий обработки результатов тестирования, их содержательно корректная интерпретация требует не столько изощренной математики, сколько увязки с педагогическим, психологическим, культурным, научным и философским контекстом. В настоящей работе предпринята попытка такой увязки с фундаментальными физическими представлениями. Оговоримся, что речь не идет о разработке физической теории, которая на принятом в физике уровне строгости описывала бы и предсказывала результаты педагогических измерений. Задачу мы видим, скорее, в том, чтобы построить качественную модель системы, подчиняющейся законам статистической физики и при этом ведущей себя как можно более подобно системе «испытуемые – ТЗ – банк ТЗ». Такая модель имеет прежде всего эвристическую ценность, позволяя получить или углубить интуитивное понимание своего психолого-педагогического прототипа – модели ИРПИ – и даже предложить некоторые усовершенствования последней.

Стартовая идея статистико-физической модели педагогического измерения (далее – СФ-модели) заключается в том, что модель Бирнбаума (2) в точности совпадает с известным в статистической термодинамике распределением Ферми-Дирака [13; 14], которое дает вероятность f того, что в системе фермионов с абсолютной температурой[3] $T = d^{-1}$ и химическим потенциалом (уровнем Ферми) μ в квантовом состоянии с энергией ε обнаружится фермион. Это совпадение не является чисто формальным сходством. Оно вытекает из важных смысловых параллелей между постулатами, на которых покоятся модель Бирнбаума и статистика Ферми-Дирака.

Во-первых, модели Раша и Бирнбаума являются принципиально вероятностными, что настойчиво подчеркивал сам Г. Раш [2]. Результат выполнения данного ТЗ данным испытуемым не может быть точно предсказан (даже если значения всех необходимых латентных переменных известны), а лишь охарактеризован вероятностью того или иного исхода. Совершенно аналогично, распределение Ферми-Дирака – одно из фундаментальных соотношений *статистической* физики – не позволяет точно сказать, будет ли данное квантовое состояние в мо-

[3] Чтобы избежать появления в (2) лишней размерной константы (постоянной Больцмана), температуру будем выражать в энергетических единицах

мент наблюдения занято фермионом, а лишь дает вероятность обнаружить его занятым.

Во-вторых, распределение Ферми-Дирака характеризует поведение частиц в условиях, когда существенна разница между фермионами и бозонами. С точки зрения статистической физики, фермионы – это частицы, подчиняющиеся *принципу Паули*: в одном квантовом состоянии не может находиться более одного фермиона[4]. Но если уподоблять успешное выполнение ТЗ заполнению квантового состояния, то очевидно, принцип Паули аналогичен утверждению, что ТЗ выполнено либо верно (~ состояние занято), либо неверно (~ состояние не занято), и никаких иных вариантов быть не может. Такая логика исключенного третьего – не что иное, как дихотомическая шкала оценивания [2; 5; 6], наиболее широко применяемая в практике реальных педагогических измерений. В принципе, предлагаемая нами СФ-модель позволяет путем некоторого усложнения описать и политомическую шкалу оценивания, но этот вопрос выходит за рамки данной работы.

В-третьих, в модели Бирнбаума отношение вероятности успеха к вероятности ошибки (с анализа которого начиналась вся теория Раша) $k = f/(1-f) = \mathrm{e}^{d(\mu-\varepsilon)}$. При фиксированном μ (один и тот же испытуемый) это отношение зависит только от характеристик задания: $k \sim \mathrm{e}^{-d\varepsilon}$, что соответствует основному закону Раша. В терминах статистики фермионов величина k имеет смысл отношения вероятности обнаружить систему с энергией ε (имеется одна частица в состоянии с такой энергией) к вероятности обнаружить систему в состоянии с нулевой энергией (частицы отсутствуют) и равна $k = \mathrm{e}^{(\mu-\varepsilon)/T}$. Но определенная таким образом величина – не что иное, как *фактор Гиббса*, относительно которого в статистической физике установлен чрезвычайно общий закон (*большое каноническое распределение* [13; 14]): вероятность того, что система находится в состоянии с заданной энергией при заданных значениях температуры и химического потенциала, пропорциональна фактору Гиббса[5]. Таким образом, и модель Бирнбаума, и распределение Ферми-Дирака базируются на фундаментальных постулатах, которые, хотя и имеют разное происхождение, математически выражаются одинаково.

Несмотря на всю глубину аналогий между моделью Бирнбаума и распределением Ферми-Дирака, между ними имеются и серьезные различия принципиального характера.

В статистической физике значения температуры $T\ (= d^{-1})$ и уровня Ферми μ одинаковы для всех подсистем данной системы. Это достигается само по себе, в ходе процесса релаксации, который приводит систему в термическое равновесие (одинаковая температура по всей системе) и химическое равновесие (одинаковый химический потенциал по всей системе). В моделях же Бирнбаума и Раша уровень подготовки μ – это индивидуальная характеристика испытуемого, которая, может быть, и стремится к некоему уравновешиванию с внешним миром, но вряд ли успевает заметно измениться в процессе тестирования. Аналогичным образом, как обсуждалось в п. 3, параметр d в модели Бирнбаума обычно рассматривается как индивидуальная характеристика отдельного ТЗ и потому не может играть столь же фундаментальной роли, как температура равновесной макросистемы. Поэтому применение аналогии между моделью Бирнбаума и распределением Ферми-Дирака требует разработки теоретической схемы, которая аккуратно учитывала бы эти различия. Ниже описывается возможный вариант такой схемы.

6. Статистико-физическая модель педагогического измерения (СФ-модель)

В общем случае процесс измерения заключается во взаимодействии измеряемого объекта с измерительным прибором. По результатам этого взаимодействия и судят о свойствах объекта. При этом, в отличие от эксперимента вообще, взаимодействие прибора с объектом носит ограниченный и односторонний характер. Односторонность означает, что объект измерения воздействует на измерительный прибор, а прибор, в идеале, не оказывает обратного воздействия. Ограниченность означает, что состояние прибора может изменяться лишь в пределах некоторой узкой области фазового пространства. Другими словами, прибор – это система, обладающая лишь небольшим количеством степеней свободы (в идеале – одной).

Мы предлагаем СФ-модель педагогического измерения, учитывающую эти соображения и воспроизводящую, по крайней мере в простейших случаях, модели Раша и Бирнбаума.

Объекты измерения (~ испытуемые) рассматриваются как большие фермионные системы, каждая из которых внутренне равновесна и потому характеризуется определенными уровнем Ферми μ и температурой T. Чтобы подчеркнуть макроскопический характер и внутреннюю сложность объектов измерения, будем далее называть их «индивидами». Уровень Ферми индивида, в нашей модели, – это аналог уровня подготовленности реального испытуемого. Задача измерения состоит в том, чтобы по возможности точнее установить значение μ. Вопрос о смысле параметра «температура индивида» и необходимости его измерения мы обсудим позднее.

Измерительный прибор в нашей СФ-модели – это измерительная микросистема (ИМС), приготовленная так, что она обладает небольшим количеством квантовых фермионных состояний с заданными энергиями ε_i. В простейшем случае ИМС имеет лишь одно состояние (унимодальная ИМС). В терминах педагогических измерений такой ИМС соответствует однобитное ТЗ, заключающееся в вынесении суждения об истинности или ложности предъявленного испытуемому утверждения. Более распространенные формы ТЗ для адекватного отражения в СФ-модели требуют рассматривать более сложные конструкции ИМС даже если выполнение ТЗ в целом оценивается по дихотомической шкале «успех / ошибка».

Рассмотрим, например, популярную форму ТЗ на выбор единственного верного ответа на фоне g неверных (дистракторов). Такое ТЗ в СФ-модели

---

[4] Альтернативным классом являются бозоны – частицы, которые могут накапливаться в одном и том же квантовом состоянии в любом количестве. Это свойство сложно отобразить на какую-то разумную логику оценивания результатов педагогических измерений, так что для СФ-модели интересна только фермионная статистика.

[5] Есть важное уточнение этого закона: если состояние *вырождено*, то есть фактически представляет собой группу разных квантовых состояний с одинаковой энергией, то вероятность обнаружить систему в таком состоянии пропорциональна еще и кратности вырождения g – количеству квантовых состояний в этой группе [14].

может имитироваться мультимодальной ИМС с $n = g + 1$ фермионными состояниями и правилом «ТЗ считается выполненным тогда и только тогда, когда заполнено выделенное состояние ИМС)».

Выделенное состояние мультимодальной ИМС выделено тем, что соответствует верному ответу. На внутреннем языке СФ-модели это должно быть состояние с наибольшей энергией ε. Действительно, согласно пятому требованию к метрической функции f (п. 2) заполнение состояния ИМС с наибольшей энергией наименее вероятно. Таким образом, энергия ε однофермионного состояния ИМС – это мера «цены» соответствующего ответа на ТЗ. Чем выше цена, тем меньше шансов выбрать данный ответ случайно, тем труднее выбрать его сознательно, тем больше от испытуемого требуется знаний и умения их применять и комбинировать.

Энергии остальных g состояний ИМС, соответствующих дистракторам, должны быть ниже, чем ε. В принципе, они могут быть разбросаны произвольным образом в некотором диапазоне и даже не слишком далеки от уровня верного ответа. Но в таком случае мы получаем СФ-модель ТЗ с дистракторами, которые неверны лишь в некоторой мере, большей или меньшей. Результат выполнения такого ТЗ явно требует уже политомической оценки, а описывающая его СФ-модель окажется слишком сложной, чтобы быть полезной. Поэтому мы примем упрощение, фактически эксплуатируемое в практике педагогических измерений, и будем считать все дистракторы абсолютно неверными ответами. На языке СФ-модели это означает, что все состояния ИМС, кроме выделенного, имеют нулевую энергию. Таким образом, простейшая ИМС, моделирующая ТЗ с одним верным ответом и g дистракторами является бимодальной, то есть обладает двумя энергетическими уровнями – ε и 0, причем нулевой уровень g-кратно вырожден.

Банку ТЗ в СФ-модели соответствует ансамбль ИМС. Если задача педагогического измерения заключается в диагностике достижения испытуемыми некоторого уровня обученности, то задания банка должны иметь одинаковую трудность, соответствующую этому уровню. На языке СФ-модели это звучит как требование ко всем ИМС ансамбля иметь одно и то же значение ε. Если же педагогическая задача состоит в ранжировании испытуемых, то трудности заданий банка должны равномерно покрывать весь ожидаемый диапазон обученности испытуемых (~ спектр энергий ИМС должен быть равномерным в диапазоне от минимально до максимально ожидаемых значений уровня Ферми индивидов). В обоих случаях значение энергии ИМС должно по возможности точно контролироваться на этапе ее создания (~ трудность ТЗ в идеале должна точно задаваться при разработке ТЗ).

Процесс измерения в нашей модели начинается с приведения одной из ИМС в контакт с индивидом (~ предъявление ТЗ испытуемому). При этом между ними становится возможен химический обмен (частицами) и сопряженный с ним энергетический обмен. Поскольку индивид – система макроскопическая, а ИМС – микроскопическая, то химический потенциал ИМС должен установиться на уровне μ индивида. Затем измерительная установка (~ система проведения тестирования) определяет, занято ли выделенное состояние ИМС (~ успешно ли выполнено ТЗ). Согласно статистической физике, вероятность этого для однобитного ТЗ (единственный выбор из двух) дается распределением Ферми-Дирака и в точности соответствует модели Бирнбаума (2). Для описанной выше более реалистичной бимодальной ИМС, которая моделирует ТЗ с единственным верным ответом, вероятность заполнения верхнего уровня (ассоциируемая с вероятностью успеха испытуемого в данном ТЗ) будет даваться формулой, учитывающей вырождение нулевого уровня ИМС и несколько отличающейся от модели Бирнбаума:

$$f_S(\varepsilon,\mu,T) = \left(1 + g \cdot \exp\left(\frac{\varepsilon - \mu}{T}\right)\right)^{-1}. \qquad (3)$$

Ранее эту метрику, из несколько других соображений, получил В.И. Дроздов и показал, что она реалистичнее, чем модель Бирнбаума, описывает влияние случайного угадывания ответов на результаты педагогических измерений [15]. Таким образом, описанная в данном разделе СФ-модель педагогического измерения математически вполне согласуется с известными апробированными моделями. Однако в ней смысл и свойства используемых переменных μ и ε оказываются понятнее и определеннее, что облегчает задачу ИРПИ и, как мы покажем далее, делает возможными некоторые важные уточнения традиционных подходов к этой задаче.

7. Уточнение понятия «дифференцирующая способность ТЗ» с помощью СФ-модели

Продемонстрируем эвристический потенциал СФ-модели педагогического измерения на примере проблемы интерпретации параметра d модели Бирнбаума. Как отмечалось, обычно d рассматривается как индивидуальный атрибут ТЗ. Но в СФ-модели педагогического измерения это означало бы, что в (3) температура $T = d^{-1}$ является атрибутом ИМС и никак не зависит от свойств индивида, находящегося в контакте с ней. Принять последнее не представляется возможным ввиду следующих соображений.

Температура, с физической точки зрения, есть мера интенсивности беспорядочного движения микросоставляющих макроскопической системы. Индивиды СФ-модели, как системы, макроскопические по определению, должны иметь собственные ненулевые температуры. Если химический потенциал μ индивида соответствует уровню подготовленности (грубо говоря, количеству знаний) соответствующего испытуемого, то температура T должна характеризовать степень беспорядочности, бессистемности этих знаний, отсутствия связей между их элементами.

Действительно, каждому практикующему преподавателю приходилось иметь дело с учениками, добросовестно вызубривающими материал (высокий уровень μ), но в целом бестолковыми, не способными применить большой объем своих знаний для выполнения даже несложных заданий (высокая интенсивность беспорядка, T). Модели (2) и (3) могут объяснить этот феномен, но лишь если считать параметр d (или, что удобнее в СФ-модели, $T = d^{-1}$) атрибутом не ТЗ, а испытуемого! В рассматриваемой ситуации температура испытуемого T оказывается высока настолько, что превышает значение разности $|\varepsilon - \mu|$, и вероятность успеха f, согласно (3), оказывается не сильно отличающейся от уровня случайного угадывания: $f \sim 1/(g+1)$.

Возможна, однако, и даже довольно распространена в практике педагогических измерений обрат-

ная ситуация, когда хорошо обученные и системно мыслящие испытуемые (~ высокий уровень μ и низкая температура индивида T) плохо справляются с не слишком сложным заданием (низкое проектное значение ε), хотя модели (2) и (3) в таких условиях предсказывают почти 100-процентную вероятность успеха. Причинами обычно оказываются неточная или неряшливая (содержательно или грамматически) формулировка задания, выход за пределы предметной области, ошибка или описка составителя ТЗ и т.п. В самых вопиющих случаях такие задания обычно выбраковываются из банка ТЗ по результатам апробации, но менее заметные аномалии в нем остаются. В общем, такие случаи можно рассматривать как результат случайных отклонений параметров задания от проектных значений, а размах отклонений количественно описывать температурой – только теперь уже температурой задания! Если она достаточно высока, то теперь уже она определяет значение параметров d в модели Бирнбаума (2) и T в СФ-модели (3) и опять сводит вероятность успеха к вероятности случайного угадывания.

Таким образом, дифференцирующая способность d в модели Бирнбаума на самом деле зависит как от задания, так и от испытуемого. На языке СФ-модели, температура T, которая фигурирует в (3), определяется как температурой индивида T_i, так и температурой ИМС T_m, причем оказывается ближе к более высокой из них.

Последний вывод с физической точки зрения понятен. В статистической физике температура выступает как мера неопределенности энергии. При конечной температуре T энергия системы из-за беспорядочного теплового движения может быть определена лишь с погрешностью порядка $\pm T$ (тепловое уширение энергетических уровней). Таким образом, неустранимая неопределенность уровня Ферми индивида μ примерно равна температуре индивида T_i, а неопределенность энергии ε измерительной микросистемы – температуре ИМС, T_m. При выполнении измерения неопределенности значений μ и ε накладываются и усиливают друг друга. Закон сложения температур при этом вытекает из теории вероятностей: если μ и ε – независимые случайные величины (в нашем случае это, очевидно, так), то при их вычитании (как в моделях (1)–(3)) складываются их дисперсии, то есть средние квадраты отклонений от средних значений [16]. Характерные отклонения μ и ε от средних (в случае ε – от проектных) значений пропорциональны температурам индивида и ИМС, а дисперсии – квадратам этих температур. Отсюда следует, что эффективная температура, определяющая вероятность того или иного результата измерения,

$$T = \sqrt{T_i^2 + T_m^2}\,. \qquad (4)$$

Согласно (4), если температуры T_i и T_m различаются хотя бы в три раза, эффективная температура T практически совпадает с большей из них, что полностью согласуется с качественным выводом, сделанным выше в этом разделе.

Таким образом, согласно уточнению модели Бирнбаума, выполненному с помощью СФ-модели педагогического измерения, вероятность успешного выполнения ТЗ испытуемым должна описываться выражением (3), в котором параметр T определяется формулой (4).

Теоретически величины, входящие в модель (3), могут устанавливаться обычным путем – минимизацией отклонений априорных вероятностей (3) от эмпирических частот решаемости заданий. Однако эта задача потребует слишком больших объемов выборки и объемов вычислений ввиду большого количества переменных – по две для каждого задания (ε и T_m) и каждого испытуемого (μ и T_i). Для преодоления этой трудности процедуру можно проводить в два этапа.

На первом этапе банк ТЗ апробируется на испытуемых, наверняка имеющих низкие собственные температуры – например, на специалистах в данной предметной области. При этом в модели (3) будут работать в основном температуры заданий, определение которых (наряду с измерением уровней сложности ε) и является основной целью первого этапа.

После завершения первого этапа банк ТЗ охлаждается: из него изымаются все задания с температурой, превышающей уровень отсечки, который должен быть ниже температуры большинства испытуемых из предполагаемого контингента. Только после этого банк ТЗ приобретает свойства валидного инструмента. При необходимости банк гомогенизируется по уровню трудности заданий.

На втором этапе с помощью охлажденного банка ТЗ выполняется основное тестирование, по результатам которого для испытуемых определяются не только их уровни μ, характеризующие в основном объем знаний, но и их температуры T_i, которые в большей степени чувствительны к упорядоченности и системности знаний. Одновременно проверяются полученные на первом этапе оценки трудности заданий.

Описанная процедура представляется вполне реализуемой на практике и сулит получение более глубокого понимания результатов педагогических измерений.

Таким образом, предложенная в данной работе статистико-физическая модель педагогических измерений позволяет выполнять глубокую интерпретацию их результатов, уточнять математическое описание и разрабатывать практические предложения по совершенствованию практики проведения педагогических измерений в образовательных учреждениях и организациях.

СПИСОК ЛИТЕРАТУРЫ:

1. Савельев, Б.А. Оценка уровня обученности студентов в целях аттестации образовательного учреждения профессионального образования: учеб. пособие [Текст] / Б.А. Савельев, А.С. Масленников. – Йошкар-Ола : Центр государственной аккредитации, 2004. – 84 с.

2. Rasch, G. Probabilistic models for some intelligence and attainment tests: Expanded ed. / G. Rasch. – Chicago : Univ. of Chicago Press, 1980. – 199 pp.

3. Rasch, G. On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements / G. Rasch // Danish Yearbook of Philosophy. – 1977. – V.14. – P.58–93. / [Электрон-

ный ресурс] : Institute of Objective Measurements, Inc. [сайт]. – Режим доступа: http://www.rasch.org/memo18.htm.

4. Свиридов, В.В. Измерение компетенций как латентных переменных в рамках системного подхода к разработке аттестационных педагогических материалов [Текст] / В.В. Свиридов, Е.И. Свиридова, М.В. Кочукова, Н.М. Ткачева // Теория и практика измерения латентных переменных в образовании и других социальных и экономических системах : материалы XI и XII Всероссийских науч.-практических конференций. – Славянск-на-Кубани : СГПИ, 2009. – С. 219–224.

5. Маслак, А.А. Измерение латентных переменных в социально-экономических системах: теория и практика [Текст] / А.А. Маслак. – Славянск-на-Кубани: СГПИ, 2007. – 424 с.

6. Нейман, Ю.М. Введение в теорию моделирования и параметризации педагогических тестов [Текст] / Ю.М. Нейман, В.А. Хлебников. – М. : Прометей, 2000. – 168 с.

7. Sick, J. Rasch analysis software programs / J. Sick // Shiken: JALT Testing & Evaluation SIG Newsletters. – 2009. – Vol.13. – N. 3. – P. 13–16 / [Электронный ресурс] : Shiken Research Bulletin [сайт]. – Режим доступа: http://jalt.org/test/sic_4.htm.

8. Eysenck, H.J. Check your own I.Q. / H.J. Eysenck. – London : Penguin Books, 1977. – 192 pp.

9. Белобородов, В.Н. Применение современной теории тестирования IRT в системе контроля измерительных свойств диагностических материалов [Текст] / В.Н. Белобородов, А.О. Татур // Педагогические измерения. – 2016. – № 2. – С. 85–97.

10. Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability / A. Birnbaum // Lord F.M., Novick M.R. Statistical theories of mental test scores.– Reading, Mass.: Addison-Wesley Publ. Co., 1968. – 568 pp. – P. 397–479.

11. Andrich, D. Controversy and the Rasch Model: A Characteristics of Incompatible Paradigms? / D. Andrich // Medical Care. – 2004. – V. 42. – N. 1 suppl. – P.7–16.

12. Measurements at the Crossroads : History, philosophy and sociology of measurement. Call for contributions : Quantification and measurement practices / 3rd Interdisciplinary conference, 2018, Paris (France) [сайт]. – Режим доступа: https://measurement2018.sciencesconf.org.

13. Ландау, Л.Д. Статистическая физика. Часть 1 [Текст] / Л.Д. Ландау, Е.М. Лифшиц. – М. : Наука, 1976. – 584 с.

14. Киттель, Ч. Статистическая термодинамика [Текст] / Ч. Киттель. – М.: Наука, 1977. – 336 с.

15. Дроздов, В.И. Обобщение модели Раша и Бирнбаума [Текст] / В.И. Дроздов // Известия Курского государственного технического университета, 2008. – №3 (24). – С. 77-80.

16. Гмурман, В.Е. Теория вероятностей и математическая статистика [Текст] / В.Е. Гмурман. – Изд. 4-е. – М.: Высшая школа, 1972. – 368 с.