

Теория

ТЕСТ КАК ПЕДАГОГИЧЕСКАЯ СИСТЕМА

Вадим Аванесов

testolog@mail.ru

Педагогические измерения и педагогическое тестирование имеют в России существенный потенциал развития. Однако этот потенциал остаётся почти неиспользованным. Вместо профессионально разработанных тестов страна получила бюрократизированный и некачественный единый государственный экзамен (ЕГЭ), одним из следствий которого является неизбежное ухудшение массового образования и невнимание к развитию подлинной тестовой культуры.

Успешному преодолению сложившейся ситуации может способствовать изменение образовательной политики нынешнего правительства РФ и консолидация педагогической общественности в деле разработки настоящих педагогических тестов.

1. Проблемная ситуация

В каждой науке есть вопросы, которые требуют к себе постоянного внимания. К таковым можно отнести вопросы развития понятийного аппарата, повышения качества методов исследования, адекватности методов формулируемым целям и интерпретации полу-

ПЕД
измерения

1

Спикер Совета Федерации Сергей Миронов считает «системной ошибкой» введение единого госэкзамена (ЕГЭ). Об этом он заявил в интервью журналистам. По словам спикера, члены Совета Федерации поддержали закон о ЕГЭ только потому, что он позволяет выпускникам школ из глубинки поступать в вузы наравне с горожанами. Однако, как считает С.М. Миронов, у новой системы оценки знаний один большой недостаток — она предполагает «натаскивание» учеников, а не выработку умения думать. «Давайте не забывать о том, что многие новшества советского образования в свое время взяли на вооружение США, Европа и Азия, мы же заимствуем у них далеко не лучшие образцы», — отметил спикер. Он подчеркнул, что голосовал против введения ЕГЭ. Совет Федерации одобрил закон о едином госэкзамене, который будет введен на всей территории России до 2009 года. Адрес публикации: <http://www.mironov.ru/Publications/23929.html>

чаемых результатов. Всё отмеченное относится и к педагогическим тестам.

Для исследования возникшей в России проблемной ситуации в области разработки тестов важно определиться с научным подходом. В истории исследования общественных явлений и наук успешно проявил себя историко-философский подход, который отчётливо проявился в трудах Гегеля. Этот философ подчеркивал, что всякое данное явление следует рассматривать как развившееся, как ставшее, и что высшая форма, как результат развития явлений, не существует без пути развития.

За последние годы в стране произошло множество событий, отрицательно повлиявшие на разработку тестов. Главное из них — внедрение Единого государственного экзамена, претендующего на роль последних. Однако нет пока весомых оснований, которые удовлетворяли бы данной претензии. Вместе с тем есть немало положений и фактов, свидетельствующих о ЕГЭ как о профессионально неприемлемом средстве педагогического измерения. Это попытка эклектического соединения объективного тестового метода с нетестовыми методами в одном государственном экзамене. А экзамен, очевидно, не тестовый метод. Также это продолжающаяся засекреченность ис-

ходных результатов, неприемлемое качество ЕГЭ с очень низкой точностью измерений. Что нарушает права учащихся на объективную оценку уровня их подготовленности. В отдельных публикациях стали известны заметные объёмы фабрикации результатов, особенно в субъективно оцениваемой третьей части ЕГЭ.

Отдельного внимания заслуживает использование данных ЕГЭ во вред и против воли личности, невозможность апелляции по содержанию т.н. КИМов, Сюда же можно отнести отсутствие независимой экспертизы и келейность в принятии решений. Не случайно Председатель Совета Федерации С.М. Миронов назвал узаконенное введение ЕГЭ в России системной ошибкой^{1,2}. О системной конфликтности ЕГЭ писали и другие авторы³.

В практике тестирования проблемную ситуацию порождает противоречие между возрастающими запросами общества в качественном образовании и качественных оценках, с одной стороны, и невозможностью реального достижения интересующего качества без изменения образовательной политики.

Главная политическая составляющая общего кризиса образования заключается в ускоренной коммерциализации образования в условиях сверх-

избыточного неравенства населения в доходах. Вместе с неэффективным накоплением нефтедолларов в стабилизационном фонде и Центробанке, такая политика продолжает отбрасывать Россию назад по индексу человеческого развития⁴, лишая страну возможности образовательного, технологического и культурного прогресса⁵.

Можно также отметить чрезмерное увлечение применением в практике заданий с выбором одного правильного ответа, как будто бы других форм оценки уровня подготовленности нет. Между тем эта форма более других критикуется педагогической общественностью, и уже давно. Ещё больше критикуется применение этой формы в ЕГЭ, что вполне справедливо. Уже давно отмечалось, что недостатки заданий с выбором одного правильного ответа снижают самостоятельность и творчество студентов, из-за того, что гораздо легче выбрать готовый ответ, чем самому его сформулировать. Поэтому метод выбора ответов может быть использован лишь при изучении описательного и ознакомительного материала⁶. Хотя с этим утверждением можно спорить, главное в другом — ошибочное применение таких заданий незаслуженно дискредитирует тестовый метод вообще.

2. Нерешённые вопросы теории

У многих авторов нет ясности относительно теоретической основы педагогических измерений, лежащих в основе практики разработки тестов. Раньше в качестве такой основы ошибочно рассматривались статистические теории⁷. Ситуация несколько выправилась после классических работ Лорда⁸, однако затем, в последние тридцать лет, она вновь ухудшилась под влиянием сторонников Item Response Theory (IRT). Последняя имеет на Западе и другое, более общее и точное название — Latent Trait Theory (LTT), переводимое как математическая теория измерения латентных качеств личности. По многолетней инерции в России IRT провозглашается как «современная» теория педагогических измерений, которая обеспечивает качество тестов.

Проблемную ситуацию определяет также тенденция роста различий во взглядах исследователей по такому ключевому вопросу, как определение теста и формирование понятийного аппарата педагогических измерений, широкое распространение устаревшей лексики. Растёт число пишущих по тестовой проблематике, и вместе с тем растёт число тех, кто рассматривает тест как элементарный перевод с английского языка слов

Теория

2

Введение ЕГЭ, в третьем чтении, поддержали 310 депутатов. Против ЕГЭ проголосовали 114, воздержался один.

3

Например, Артём Соловейчик. В чём сегодня опасность единого экзамена? // Первое сентября, № 50, от 10 июля 2001 года.

4

Г.Н. Соколова. Энциклопедия социологии: Индекс развития человеческого потенциала — система показателей, принятая Организацией Объединенных Наций (ООН) с целью интегративной оценки качества жизни населения страны — члена ООН. И.Р.Ч.П. определяется на основе следующих показателей: ожидаемая продолжительность жизни; уровень образования; реальный душевой валовой внутренний продукт. И.Р.Ч.П. отражает средний уровень обеспечения странной базового человеческого потенциала и свидетельствует о том, сколько еще предстоит сделать стране для достижения некоторых определенных целей: средней продолжительности жизни в 85 лет;

ПЕД
измерения

доступа к образованию для 100% населения; обеспечения населению достойного уровня доходов. При расчете И.Р.Ч.П. все три базовых показателя сводятся к единой системе измерения путем определения достижений по каждому показателю как относительной дистанции от желаемой цели. Максимальное и минимальное значение каждой переменной переносятся на шкалу от 0 до 1, причем каждая страна располагается в определенной точке этой шкалы. При расчете И.Р.Ч.П. определяется мера относительных достижений этой страны по каждой из трех базовых переменных и рассчитывается простая средняя величина всех трех показателей. Чем ближе страна к величине И.Р.Ч.П., равной 1, тем меньше тот путь, который еще предстоит пройти этой стране до уровня и качества жизни, принимаемых за должные.

5

Коммерциализация образования может быть полезной в элитарном образовании, но только при условии ускоренного роста заработной платы всех трудящихся. Но этого нет.

«проба, испытание, проверка». Без указания на существенные признаки, выводящие тест за пределы элементарного перевода, как того требует научная логика. Хотя тест как понятие подразумевает элемент испытания, он не может сводиться к нему, ибо это ещё и концепция, и определённое содержание, это также специфическая форма и качественные результаты с невысокой погрешностью, шкалирование заданий и испытуемых, а также умелая интерпретация данных — все, требующее научного обоснования.

Нет единства не только в отношении числа форм тестовых заданий, но и в отношении их названий. Слова «форма», «вид», «тип» теста (задания) повсеместно употребляются как равнозначные. Некоторые авторы продолжают называть задания с выбором одного правильного ответа по-прежнему, как и 20 лет назад, «закрытыми». Однако за прошедшие годы неоднократно приходилось размышлять в поисках ответа на вопросы — почему эти задания называются закрытыми, от кого они закрытые, кем, как и т.п. С момента возникновения действительно закрытого от педагогической общественности ЕГЭ стало ясно, что название «задание закрытой формы» нехорошее. Вот почему ему на смену пришлось предложить другое, более сущностное на-

звание — задание с выбором одного или нескольких правильных ответов.

Ещё сложнее обстоит дело с определением специфических свойств и требований, позволяющих чётко различать одну тестовую форму от другой. Отсутствие чётких признаков демаркации нередко приводит к смешению различных форм, препятствует созданию новых форм, мешает созданию общепринятой классификации уже имеющихся форм и видов заданий, разработке точных и понятных инструкций для испытуемых, созданию понятийного аппарата этой малоисследованной сферы педагогической науки.

Недостаточное внимание к формам тестовых заданий породило ситуацию, заключающуюся в том, что сборники упражнений и задач в различных методических пособиях содержат слишком мало таких заданий, которые можно было бы эффективно использовать в автоматизированных системах обучения и контроля. Между тем, как будет показано дальше, небольшие усилия по трансформации заданий позволяют сразу же разделить их тестовой формой и тем самым сделать задания технологическими.

Неоднократные призывы автора этой статьи обратить особое внимание на дальнейшее развитие правил и принципов

композиции всех реально существующих форм заданий⁹, а также на необходимость формотворчества в тестовой работе, похоже, не услышан. Подобная ситуация складывалась и в 30-х годах XX века. «Форма теста, — писал в 1928 году Г.С. Костюк, — является одной из тех проблем тестирования, научное изучение и разрешение которой чрезвычайно важно и неотложно»¹⁰. Этот призыв в те годы был проигнорирован. На важность внимания к проблеме формы тестовых заданий недавно вновь обратил внимание А.Г. Войтов¹¹.

Продолжающееся в массовом сознании ненаучное отношение к исследованию возможностей тестовых форм можно назвать главной причиной производства некачественных «материалов», называемых контрольно-измерительными материалами. Последние называть не то что тестами, но даже заданиями в тестовой форме невозможно.

Мало внимания уделяется вопросам формы и в зарубежной литературе.

3. Трудности с определением понятия «тест»

Определение теста представляет собой нелёгкую проблему. Одна из причин — кажущаяся очевидность сущности и видимости простоты. Главным, а нередко и

единственным признаком теста чаще других воспринимается возможность выбора одного ответа из нескольких, предлагаемых на выбор. В наши дни такое восприятие, как и простой перевод с английского слова «test» — проверка, проба, вряд ли можно считать признаком просвещённости в данном вопросе. При опоре на такого рода обыденный смысл почти полностью исчезает специфика теста как научно обоснованного метода исследования, обладающего совокупностью существенных признаков, отличающих тест от всех других форм и методов проверки знаний. Естественная при этом подмена научного понятия «тест» элементарным истолкованием смысла этого слова создаёт много трудностей. Но при таком толковании смысла слова «тест» под него можно подвести почти любое средство контроля, а потому, как писал Спиноза, всегда сохраняется опасность смешения определяемой вещи с какой-либо иной.

На самом деле смыслы слова «тест» и научного понятия «тест» сильно различаются. Игнорирование этого момента в теории приводит к засорению понятийного аппарата науки словами обыденной речи, не имеющими заметного научного содержания. А в практике это порождает тысячи разного рода публикаций под общим именем «тест». Однако это не тест.

Теория

6

Молибог А.Г.

Вопросы научной организации педагогического труда в высшей школе. Изд. 2, доп. Минск, Высшая школа, 1975.

7

Gulliksen H.

Theory of Mental Tests. N-Y. Wiley. 1950 — 486 p.; Cronbach L.J., Gleser G.C., Nanda H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. N-Y: John Wiley and Sons. 1972.

8

Что подчеркивается, например, точным названием классической монографии: Lord F.M. & Novick M.R. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley. 1968.

9

Аванесов В.С.

Композиция тестовых заданий. М. Центр тестирования, 3 изд. 2002. 239 с.

10

Костюк Г.С.

Избранные психологические труды. М. 1988. с. 73.

ПЕД
измерения

11

Войтов А.Г.
Форма и формы тестовых заданий. // Педагогические Измерения, №2, 2005 г. С. 25–34.

12

Cronbach L.J.
Essentials of Psychological testing. 4th ed., 1980. P. 26. Harper & Row, Publishers, N-Y, 630 pp.

13

Rights and Responsibilities of Test Takers: Guidelines and Expectations. 2007.
<http://www.apa.org/science/ttrr.html>

14

Дидактические тесты: технология проектирования. Методическое пособие для разработчиков тестов. Минск, 2004. С. 21.

15

Федотова Г.А. Игнатьева Е.Ю. Технология профессионального образования. Уч. пособие. Нов ГУ им. Ярослава Мудрого. Великий Новгород, 2006. 142 с.

16

Дидактические тесты: технология проектирования. Методическое пособие для разработчиков тестов. Минск, 2004. С. 21.

Ещё меньше уделяется внимание поиску подходящих определений педагогического теста. Ли Кронбах честно признал, что в тестовой литературе нет достаточно удовлетворительного определения теста, как педагогического, так и психологического. Психологический тест он определяет как метод систематического наблюдения за поведением личности и описания результатов с помощью числовых шкал и фиксированных оценочных градаций¹². В работе «Права и обязанности испытуемых» тест определён как психологический и педагогический метод, разрабатываемый и используемый профессионалами в образовательных и прочих учреждениях, используемый для оценки испытуемых¹³. Приведём пример и сравнительно нового определения: «под тестом обычно понимается достаточно краткое, строго стандартизованное испытание, которое позволяет количественно выразить результат и, следовательно, даёт возможность осуществить его тематическую обработку»¹⁴.

На протяжении многих лет оспаривается идея теста как системы заданий возрастающей трудности. Некоторые авторы вместо системы заданий используют словосочетание «совокупность» заданий. Например, в только что вышедшей работе¹⁵ педагогический тест оп-

ределяется как совокупность тестовых заданий, позволяющая по результатам их выполнения объективно измерить уровень подготовленности по конкретным разделам определённой области знаний (С.122).

Противопоставление суждений по данному вопросу можно встретить даже в одной и той же работе. В уже упоминавшейся работе белорусских коллег¹⁶ на стр. 54 написано, что «задания следует располагать в порядке возрастания предполагаемой трудности». На другой странице, 60, написано: «Сейчас всё больше подвергается сомнению точка зрения о том, что педагогический тест — это система заданий возрастающей трудности. В противовес ей выдвигается положение о том, что расположение заданий в тесте должно отражать логику той или иной предметной области, учитывать значимость заданий в структуре всего материала и «кривую утомляемости» учащихся»¹⁷.

В этот вопрос хотелось бы внести ясность. В текущем контроле, где трудно говорить о применении настоящего теста, речь может идти скорее о применении системы заданий в тестовой форме, а потому тематический подход полезен и неизбежен. Другое дело, если ставится вопрос об итоговом тестовом контроле подготовленности, по всей учебной дисципли-

не. Здесь главным становится требование теста как системы заданий равномерно возрастающей трудности. Именно трудность заданий становится одним из главных предметов шкалирования на заключительном, пятом этапе разработки теста.

4. Определение теста

Правильный путь исследования — писал Спиноза — состоит в том, чтобы формировать мысли из некоторой данной дефиниции: исследование будет продвигаться тем успешнее и легче, чем лучше мы определим какую-либо вещь¹⁸. Ключевым понятием теории и методики педагогических измерений является «тест».

Среди основных направлений улучшения качества педагогических тестов можно выделить усиление внимания к вопросам композиции заданий в тестовой форме и поиски наиболее подходящих определений педагогического теста. Ясно, что примитивным голосованием в пользу того или иного определения такой вопрос не решается. Это вопрос науки, где голосование бывает более вредным, чем полезным.

В то время как научно сформулированное понятие — это мысль, отражающая в обобщённой форме существенные признаки и связи. К элементам теста можно отнести инструк-

ции для испытуемых, задания, ответы к заданиям, ответы, система оценивания, матрицы тестовых результатов и интерпретации.

Предлагаемая ниже дефиниция педагогического теста для цели дальнейшего развития уже прошла долгий путь многолетнего уточнения. Этот факт непрерывного уточнения находится в некотором противоречии со стремлением иметь устойчивые определения, без которых наука невозможна. В философии сложилась традиция рассмотрения научных понятий с двух противоположных позиций. Как отмечал Б.М. Кедров, каждое понятие может быть зафиксировано, во-первых, как сложившееся, готовое на данный момент времени, а потому как неизменное, постоянное. Научные понятия могут и должны быть рассмотрены исходя из того, что они выражают и резюмируют собой на каждом этапе определённую ступень развития науки. В результате складываются две логические схемы: формально логическая, оперирующая относительно неизменными, готовыми понятиями, и диалектологическая, оперирующая изменчивыми, развивающимися понятиями¹⁹.

На текущий момент формулировка такова: педагогический тест — репрезентативная система параллельных заданий равномерно возрастающей

Там же.

Спиноза.

Трактат об усовершенствовании интеллекта.

Кедров Б.М.

Характер изменения объема и содержания развивающихся понятий. Анализ развивающегося понятия. М.: Наука, 1967. 439 с.

трудности, специфической формы, определённого содержания, позволяющая качественно оценить структуру и эффективно измерить уровень подготовленности испытуемых. По сравнению с данными ранее определениями здесь добавились два термина — «*репрезентативная система*» и «*равномерно возрастающей трудности*».

В основу этого определения положены три идеи.

Первая: тест рассматривается не как обычная совокупность или набор заданий, а в виде более богатой по научному содержанию категории «система». В самом общем виде системе принято понимать как комплекс взаимосвязанных элементов, образующих некоторую целостность²⁰.

Вторая идея, заложенная в нашем определении, состоит в том, что в нем совершен отход от укоренившейся традиции рассмотрения теста как простого средства проверки, пробы, испытания. В нем подчеркнута, что тест является качественным средством педагогического измерения, проводимого на латентной шкале, не существующей в явно заданном виде. Эта переменная является результатом концептуализации и операционализации. В соответствии с положением теории латентно-структурного анализа тестовые оценки не являются точными оценками

испытуемых на интересующей нас латентной переменной, такой, как, например, знание учебной дисциплины, учебная мотивация, интеллект и др.; они лишь репрезентируют эти значения с некоторой точностью. Вот почему в Item Response Theory и в латентно-структурном анализе (LSA) в качестве исходной аксиомы принимается положение о принципиальном несовпадении тестовых баллов испытуемых и их истинных баллов на латентной шкале. Если на оси абсцисс отложить значения латентного уровня подготовленности испытуемых (θ), а на оси ординат откладывать значения тестовых баллов (Y), то получится кривая, похожая на логистическую функцию.

Третья идея, выдвинутая в нашем определении теста — это включение нового критерия, эффективности теста, который ранее в литературе по тестам не рассматривался в качестве специально-научного критерия анализа и конструирования тестов.

Главный термин определения теста — «система» — означает, что в тесте собраны такие задания, которые обладают системообразующими свойствами. Хотя любой тест состоит из тестовых заданий, последние представляют не совокупность произвольно объединённых заданий, а именно систему. Среди си-

стемных признаков выделяется, в первую очередь, общая принадлежность заданий к одной учебной дисциплине, коррелируемость с внешним критерием, место задания на латентной переменной, относительно других заданий, точность измерения испытуемых данным заданием интересующего свойства. О точности измерения можно судить на основе соотношения уровня подготовленности испытуемых и уровня трудности задания.

Тест как система, обладает составом, целостностью и структурой. Тест состоит из заданий, правил их применения, оценок за выполнение каждого задания и рекомендаций по интерпретации тестовых результатов. Целостность означает взаимосвязь заданий, их принадлежность общему измеряемому фактору. Каждое задание теста выполняет отведённую ему роль, и потому ни одно из них не может быть изъято из теста без потери качества измерения. Структуру теста образует способ связи заданий между собой. В основном это так называемая факторная²¹ структура, в которой каждое задание связано с другими через общее содержание и общую часть вариации тестовых результатов.

Проявлению системного качества теста способствует и единая дисциплинарная общность заданий, реализующая идею измерения подготовлен-

ности студентов по одной какой-либо определённой учебной дисциплине. Совокупность таких заданий, отобранных в соответствии с требованиями теста, образует гомогенный тест, измеряющий одно какое-либо качество (свойство). В таком тесте не допускается использование заданий, выявляющих другие свойства. Наличие последних нарушает требование дисциплинарной чистоты педагогического теста. Это требование, естественное для гомогенного теста, очень часто нарушается из-за стремления увязать проверку знаний с проверкой чего-то ещё, что приводит к ухудшению качества измерения, к потере так называемого свойства одномерности теста или, короче, шкалы, если держать в уме ещё одно очень короткое определение: тест — это шкала.

Репрезентативная система заданий означает необходимость рассматривать тест как статистическую выборку из генеральной совокупности всех мыслимых заданий, которые потенциально можно было бы использовать для тестирования испытуемых. Эта идея лежит в основе двух статистических теорий — классической, известной по предыдущим публикациям, и расширенной, публикуемой в статье данного № 1, 2007 г. журнала ПИ «Методологический анализ теорий педаго-

Теория

ИЗДАНИЕ

21

Иберла К.
Факторный анализ.
Пер. с нем. М.: Статистика, 1980.

гических измерений». Реализация этой идеи в практической плоскости потребует создания принципиально нового научно-обоснованного, параметризованного и стратифицированного банка настоящих тестовых заданий.

Примером системных требований к заданиям теста является педагогическое измерение по теории G. Rasch. Даже самое отличное, отдельно взятое задание может не попасть в тест, если оно не обладает системообразующими признаками, отвечающими требованиям модели измерения. Не случайно измерение по G. Rasch называется «model based measurement».

Что касается второго нового термина определения теста, то идея *равномерно* возрастающей трудности задания вытекает из зависимости тестового балла испытуемого от числа заданий теста и вариации тестовых баллов по уровню трудности. Это идея хорошо выражена, графически и аналитически, в интересной работе Б. Райта и М. Стоуна²².

5. Вопросы интерпретации смысла терминов

Определённое содержание означает использование в тесте только такого контрольного материала, который соответствует содержанию учебного курса; ос-

тальное содержание в педагогический тест не включается ни под каким предлогом.

Содержание теста проходит экспертизу у опытных педагогов, которые призваны дать ответ на главный вопрос — можно ли с помощью предложенных заданий корректно оценить содержание, уровень и структуру знаний у данного контингента испытуемых? При оценке содержания теста всегда возникают вопросы о цели теста, его содержание и качестве. Анализ содержания заданий, а следовательно, и теста в целом, позволяет определить знания, умения, навыки и представления, требуемые для правильного выполнения задания. При применении заданий в тестовой форме для аттестации выпускников образовательных учреждений важно иметь такие задания, которые позволяют делать вывод о минимально допустимой компетентности выпускников²³. Принципы и другие вопросы содержания теста рассматривались в главе «Содержание теста и тестовых заданий».

Трудность теста определяется суммарной трудностью заданий, его образующих. В лёгком тесте у большинства испытуемых будут высокие баллы, но это тот самый случай, когда цифры становятся обманчивыми, если не знать, как они полу-

22

Wright B.D., Stone M.H.
Best Test Design.
Chicago, Ill.,
1979. 219 pp.

23

За рубежом их так и называют: Minimum Competency Tests.

чены. Здесь уместно напомнить самое короткое (и потому не точное, но удобное) определение педагогического теста — это система заданий возрастающей трудности.

В ряде образовательных учреждений из года в год снижается трудность экзаменационных заданий, потому что результаты работы педагогического коллектива нередко оцениваются по количеству отличных и прочих оценок. Но суть не в оценках, а в содержании и уровне трудности заданий, по которым эти оценки получены. Поэтому органы управления образованием пытаются установить образовательные стандарты и проводить уровневый контроль, ниже которых опускаться нельзя.

Качество педагогического и любого другого теста традиционно сводится к определению меры надёжности и валидности полученных результатов²⁴. Как и объективным, качественным можно назвать только тот метод измерения, который обоснован научно и способен дать требуемые результаты. В западной литературе традиционно рассматривается два основных критерия качества: валидность и надёжность.

Валидность означает пригодность тестовых результатов для той цели, ради чего проводилось тестирование. Самая главная угроза для снижения

валидности результатов — это формулирование двух и более целей применения какого-либо метода оценки уровня подготовленности испытуемых. Тезис ...никогда не следует использовать системы оценивания, разработанные в одних целях, для других целей; невозможно эффективно достичь сразу двух целей²⁵ прямо относится к анонимным авторам некачественного ЕГЭ²⁶.

Валидность зависит от качества заданий, их числа, от степени полноты и глубины охвата содержания учебной дисциплины (по темам) в заданиях теста. Кроме того, валидность результатов зависит также от баланса и распределения заданий по трудности, от метода отбора заданий в тест из общего банка заданий, от интерпретации тестовых результатов, от организации сбора данных, от подбора выборочной совокупности испытуемых.

Объективность педагогического измерения означает не столько абсолютное достижение этой цели, что желательно, но невозможно, сколько установку тестологов на максимальную объективность процесса создания и применения тестов. Объективности способствуют одинаковые инструкции для всех испытуемых, одинаковая система оценки результатов тестирования, автоматизированный подсчёт баллов испытуе-

Теория

24

24
Хороший анализ этих двух основных критериев качества педагогических тестов можно найти в работе Linn, R. (Ed.) Educational Measurement. 3-ed. Amer. Council on Education. 1980.

25

25
Новое в оценке образовательных результатов: межд. аспект. под ред А. Литтл, Э. Вульф. М. Просвещение, 2007. 367 с.

26

26
Не так давно был издан т.н. аналитический доклад «Единый государственный экзамен в общероссийской системе оценки качества образования», самые яркие стороны которого — это полное отсутствие настоящей аналитики и такое же отсутствие фамилий авторов. См. Школьные технологии, № 5, 2005 г.

мых и все остальное, повышающее качество тестирования.

Условиями повышения объективности тестирования является повышение уровня научной обоснованности как самих тестов, так и методов их разработки, использование современной техники хранения и передачи тестовых материалов, защита информации от нежелательного доступа, последовательная работа по устранению возможностей намеренного искажения результатов и других компонентов тестового процесса. Определение статистических характеристик является главным (после экспертной проверки содержания) средством диагностики качества теста по любой учебной дисциплине.

Принцип параллельности заданий. Параллельными называются задания, которые основаны на принципе вариативности и однородности, но, кроме того, имеют примерно одинаковую эмпирическую меру трудности всех вариантов задания и коррелируемые результаты. Из данного определения видно: второй принцип вытекает из принципа вариативности и однородности заданий, но по смыслу он шире, включает в себя формальные критерии и опирается на эмпирические данные результатов испытуемых. Различен и статус этих принципов. Первый и все остальные принципы являются по большей час-

ти логическими и содержательно-методическими, в то время как принцип параллельности заданий является общетеоретическим, опирающимся на статистику, а потому используется во всех формальных теориях педагогических измерений. Надо подчеркнуть, что параллельность по содержанию является только одним из необходимых условий параллельности тестовых заданий.

Идея разработки и существования параллельных вариантов тестовых заданий является фундаментальной для теоретически подготовленных разработчиков тестов, тех, кто знает так называемую классическую (статистическую) теорию тестов²⁷. Только непониманием важности такого подхода можно объяснить тексты некоторых авторов о преимуществах «современной» математической теории (Item Response Theory) перед другими теориями. На самом деле, ни статистическую, ни педагогическую теорию педагогических измерений²⁸, где изучаются вопросы содержания теста, не может заменить никакая математическая теория. *Для педагогики и практики образования адекватной может быть только педагогическая теория измерений. Все остальные могут быть только дополнительными и сопутствующими.* Отсюда легко понять неприемле-

27

Gulliksen, H. Gulliksen H. Theory of Mental Tests. N-Y. Wiley. 1950 — 486 p.

28

«Педагогическая теория педагогических измерений» было бы наиболее точным названием для излагаемой здесь теории, но необходимость избежать повтора одного и того же слова вынуждает писать о педагогической теории измерений. См., напр.: Аванесов В.С. Основы педагогических измерений // Педагогические измерения, № 1, 2004 г. С. 15–21.

мость претензий авторов математических теорий на лидерство в вопросах разработки педагогических тестов по различным учебным дисциплинам²⁹.

В западной литературе часто ошибочно пишут о «параллельных тестах»; так же ошибочно это переводится и на русский язык. На самом деле тест создаётся один, а к нему делаются параллельные варианты каждого задания. Создание параллельных вариантов создаёт предпосылки для применения теоретически более существенных методов проверки надёжности результатов теста. Надёжность результатов теста — это мера точности или непогрешимости педагогических измерений.

Не все вариативные задания могут соответствовать принципу параллельности содержания. Иначе говоря, вариативность есть необходимое и формальное, но не достаточное условие для возникновения феномена параллельности задания.

Понятие «педагогический тест» можно рассматривать в двух главных смыслах.

Во-первых, тест — это *метод* педагогического измерения; состоящий не из «некоторого», а точно определённого множества тестовых заданий возрастающей трудности, образующих *систему*. При этом сами задания теста должны

отвечать требованиям формы и содержания, иметь соответствующую меру трудности и вариации, корреляцию, подходящий геометрический образ, быть частью содержательной и формальной системы. Несистемным заданиям в тесте места нет.

В гомогенном педагогическом тесте не допускается использование заданий, выявляющих другие свойства. Наличие таких свойств нарушает требование предметной чистоты педагогического теста. Ведь каждый тест измеряет что-то заранее определённое. Например, тест по физике измеряет знания, умения, навыки и представления испытуемых в данной науке. Одна из трудностей такого измерения заключается в том, что физическое знание изрядно сопряжено с математическим. Поэтому в тесте по физике экспертно устанавливается уровень математических знаний, используемых при решении физических заданий. Превышение принятого уровня приводит к смещению результатов; по мере превышения последние все больше начинают зависеть не столько от знания физики, сколько от знания другой науки, математики. Другой важный аспект — стремление некоторых авторов включать в тесты не столько проверку знаний, сколько умение решать физические задачи, вовлекая

Для того, чтобы убедиться в непедагогичности, достаточно посмотреть на некоторые тексты — о чём там пишут, и каким языком.

ПЕД	
	измерения

тем самым интеллектуальный компонент в измерение физической подготовленности.

Уже отмечалось, что в гомогенном педагогическом тесте задания располагаются в порядке возрастающей трудности — от самого лёгкого до самого трудного. Иначе говоря, главным *формальным*³⁰ *системообразующим* признаком педагогического теста, после содержания и формы, является различие заданий по степени их трудности.

Время нередко называется в качестве другого системообразующего фактора. Действительно, одно из соображений, положенных в основу создания тестов — иметь инструмент быстрого и относительно точного оценивания больших контингентов испытуемых. Требование экономии времени становится естественным в массовых процессах, каковым и стало образование. Одно из актуальных направлений современной организации тестового педагогического контроля — это индивидуализация контроля, приводящая к значительной экономии времени тестирования. Контроль ведётся с помощью заранее шкалированных по трудности заданий³¹. Предельное время тестирования по одному на стоящему тесту не может превышать сорока минут.

Другая сторона вопроса заключается в том, что от вре-

мени тестирования существенно зависит качество результатов. Каждый тест имеет оптимальное время тестирования, уменьшение или превышение которого снижает качественные показатели теста. Оптимальное время тестирования определяется эмпирически, по показателю дисперсии тестовых данных. Если по оси абсцисс отложить время тестирования, а по оси ординат — значение дисперсии тестовых результатов, получаемое после каждого пробного контроля, то, соединив точки, получим представление об изменении дисперсии; максимум значения последней укажет на оптимум времени, необходимо для тестового контроля.

Со временем тестирования связан и вопрос о количестве заданий теста. В классической теории теста исходили из физической, по сути, идеи увеличения точности в зависимости от числа заданий: чем больше, тем точнее. Однако эта идея входила в противоречие с реальными возможностями учебного процесса, с естественным нежеланием студентов отвечать на большое количество заданий. Поэтому время тестирования ограничивается, исходя из фактических соображений удобства тестирования и идеи достаточной дифференциации студентов, показателем чего является дисперсия тестовых баллов.

30

Слово «формально» появилось здесь для того, чтобы оградить некоторых читателей от соблазна трактовать тест единственным образом — только как систему заданий возрастающей трудности, без учета их взаимосвязанного содержания.

31

Четыре часа, отводимого на проведение ЕГЭ выпускников школ можно считать самым ярким показателем бездарности самого ЕГЭ.

6. Соотношение оценок и отметок

В зарубежной педагогической литературе используется близкое к измерению понятие «evaluation», что можно перевести как «оценивание», имея в виду не только конечный результат, но и процесс формирования оценки. Соответственно выделяются два основных вида оценивания — формирующее и итоговое³². Цель первого — оказывать посредством оценивания формирующее влияние на текущий процесс обучения в смысле его улучшения, за счёт установления обратной связи от студента к преподавателю. Цель второго — получить итоговые результаты обучения.

Оценка как понятие имеет два основных смысла — как суждение о ценности (или значимости) и как приблизительная характеристика некоторой величины. Педагогические оценки делятся на две группы: абсолютные и относительные. В формулировании абсолютных оценок используются термины «хорошо», «плохо» и т.п. Оценки преподавателей вузов ближе к абсолютным оценкам, разделенным на четыре уровня (градации) — отлично, хорошо, удовлетворительно и неудовлетворительно.

Оценки нередко отождествляются с отметками. Последние являются численными

аналогами оценочных суждений. Отметки вошли в жизнь в царский период, затем на волне ложно понимаемой демократии были отменены. В советский период они были вновь введены директивным путем — Постановлением Совета Народных Комиссаров РСФСР от 10 января 1944 года. Постановление было принято, как писалось, «в целях более чёткой и точной оценки успеваемости» и замены «словесной системы» оценки успеваемости — отлично, хорошо, посредственно, плохо и очень плохо — цифровой пятибалльной системой отметок 5, 4, 3, 2, 1. Этим постановлением предписывалось ввести цифровую систему отметок на следующий день после публикации³³.

Пятибалльная шкала имеет своим главным преимуществом простоту и привычность, что объясняет её широкую и длительную, по времени, распространённость. Но она имеет и недостаток — уже отмечавшуюся субъективность, большую зависимость от того, кто оценивает, и кого оценивают. Каждый преподаватель даёт оценку на основе своего личного понимания. Другой недостаток пятибалльной шкалы оценок касается её слабой дифференцирующей способности. Она позволяет провести грубую классификацию только на четыре группы — отличников и тех, кто

32

Bloom B.S. a.o. Handbook on Formative and Summative Evaluation of Student Learning. N-Y., McGraw-Hill, 923 p. 1971.

33

Сборник постановлений РСФСР. М.: 1944, № 43.

учится хорошо, удовлетворительно и неудовлетворительно. Более тонкое различие, особенно нужное при приёме в вузы и для эффективного управления качеством обучения, пятибалльная шкала не даёт и дать не может. Отсюда понятна необходимость совершенствования педагогических оценок на основе измерений.

В других странах сложилась традиция отмечать единицей отличные знания, а пятеркой — плохие. Из этого видно, что одинаковым оценкам можно ставить в соответствие различные отметки. По мнению Ш.А. Амонашвили, в отметках мало педагогики, в них больше власти; отметки — это костыли хромой педагогики»³⁴.

В связи с тем, что самая сильная сторона тестового метода заключается в возможности проведения педагогического измерения, рассмотрим этот вопрос подробнее.

Длительный период недоверия общественности к результатам педагогического измерения сыграл, как ни странно, положительную роль в формировании своеобразного рефлексивного мышления исследователей, занимающихся измерениями: они вынуждены были обращать внимание не только на создание теста, но и на поиск доказательств того, что их тест свободен от существенных погрешностей. Это мы-

шление создавалось под заметным влиянием формировавшихся в те же годы философской, педагогической и психологической теорий измерений. Именно в этих теориях уделялось самое пристальное внимание вопросам надёжности.

Одна и та же оценка может быть дана при разной структуре знаний. Знания испытуемого, набравшего свои баллы за правильные ответы на первых, сравнительно лёгких заданиях, отличаются от знаний того, кто набрал те же десять баллов на последующих заданиях.

Если в физических измерениях обычно стоит задача измерения явно выделенных объектов, то в педагогических измерениях такие объекты не существуют. Естественно, что сомнения в правильности выделения последних всегда выше, откуда и возникли понятия «надёжность — ненадёжность», отражающие идею перманентного недоверия к результатам педагогических измерений.

Подлинные тесты от мнимых отличаются тем, что первые, как и всякий метод, эмпирически проверяются до начала практического применения, а информация о качестве теста доступна для всех, как сейчас называют, пользователей. Без этого тест не должен ни применяться, ни продаваться, ни покупаться. Наличие результатов проверки теста, высокое каче-

ство самой проверки, доступность результатов тестирования для независимой экспертизы — все это признаки, обеспечивающие объективность тестовой работы³⁵.

7. Интерпретация результатов тестирования³⁶

На уровне обыденного сознания любая интерпретация результатов тестирования представляется нетрудным делом. Полученная сумма баллов принимается за истинный уровень знаний, хотя каждый признает, что знание и баллы, полученные по показателям знаний — разные явления. Ведь можно придумать тысячи разных показателей и все они будут с разным количеством баллов показывать уровень знаний. Возникает естественный вопрос: какие из показателей правильно указывают на наличие знаний, а какие — ошибочно?

Взросшая роль правильной интерпретации данных даёт основание для выделения *трёх видов интерпретации результатов тестирования*.

Если главной задачей ставится стремление выяснить, какие элементы содержания учебной дисциплины усвоены тем или иным испытуемым, то мы имеем дело с *предметно-педагогическим подходом к интерпретации* результатов тестирования

При этом требуется большое число заданий, посредством которых определяется: что из генеральной совокупности заданий (Domain) испытуемый знает, и что он не знает.

Вывод выстраивается вдоль логической цепочки: содержание учебной дисциплины → генеральная совокупность задания для измерения знаний → тест, как выборка заданий из этой совокупности, ответы испытуемого → вероятностный вывод о его знании учебной дисциплины.

При ориентации на такие тесты требуется большое число заданий и достаточно полное определение содержания изучаемой дисциплины. Интерпретация результатов ведётся педагогами-предметниками, на языке учебной дисциплины.

Споры ведутся вокруг двух главных вопросов.

- О правильности содержания теста, что означает безошибочность формулировок его заданий, предметно-научную обоснованность, допустимость теста для проверки интересующих знаний в данной группе испытуемых. При аргументации в пользу того или иного теста педагоги-предметники опираются на понятийный аппарат, язык принципы и вообще на знания преподаваемой ими учебной дисциплины. В литературе в таких случаях говорят о тестах с *чисто содержательно-ориенти-*

Code of Fair Testing Practices in Education. Prepared by Joint Committee on Testing Practices, American Psychological Association, Washington, D.C., 4 pp. 1988.

При написании этого раздела использованы в переработанном виде материалы книги «Композиция тестовых заданий» М.: Центр тестирования, 2002.

рованной интерпретацией результатов³⁷. Это и есть так называемый случай Domain Referenced Testing, что можно перевести как соотнесение знаний по результатам теста со знаниями, полный перечень которых представлен в образовательной программе и в генеральной совокупности (domain) заданий, охватывающих всю эту программу.

- Об обоснованности оценки знаний по всему учебному предмету на основе результатов тестирования испытуемых по небольшой выборке заданий теста. Фактически это вопрос обоснования точности индуктивного вывода о знании большого числа вопросов на основе ответов по малому числу заданий теста.

Второй вид интерпретации результатов тестирования связан с ориентацией на такие конкретные цели и задачи, как, например, проверка уровня усвоения требуемых знаний, умений и навыков, выступающих в качестве заданного стандарта или критерия усвоения. Эта задача возникает при аттестации, приёме в вуз, профотборе и т.п., где важно определить соответствие испытуемых заранее заданному требованию, уровню подготовленности или компетенции — в общем, какому-либо уровню, который на Западе тоже называется критерием. Именно в этих

случаях используются тесты с критериально-ориентированной интерпретацией³⁸. Но здесь имеется в виду сравнение содержания аттестационных материалов с результатами тестирования, с заранее принятыми решающими правилами и с выводом — что реально из включённого содержания в тест усвоено испытуемыми и можно ли их аттестовать на основании полученных данных. Например, для аттестации выпускников образовательных учреждений важно иметь такие задания, которые позволяют делать вывод о минимально допустимой компетентности выпускников. За рубежом такие тесты называют: Minimum Competency Tests.

При разработке таких «тестов» особое внимание обращается на определение зачётного уровня трудности. При отборе с очень низкими и очень высокими значениями критерия (например, значения проходного балла) тестовые результаты сильно теряют в своём качестве. При проверке минимально допустимого уровня знаний содержание заданий носит принципиально облегчённый характер. Поскольку такие задания должны выполнять почти все выпускники, допущенные учебным заведением к аттестации, здесь трудно говорить о тесте с достаточной дисперсией резуль-

37

Hively W. (Ed).
Domain Referenced
Testing. Educational
Technology Publications.
Englewood Cliffs, N-J,
1974.

38

Berk R.A. (Ed).
A Guide to Criterion —
Referenced Test
Construction. The John
Hopkins Univ. Press,
Baltimore, 1984.

татов, равно как и утверждать о методе объективного и эффективного измерения испытуемых с разным уровнем подготовленности.

Такие «тесты» нередко применяют органы управления образованием, стоящие перед необходимостью за короткое время проверить состояние образования в большом количестве учебных заведений и не позволять последним опускаться ниже предельно допустимого уровня. В западной литературе в таких случаях говорится о тестировании с *критериально-ориентированной* интерпретацией результатов³⁹.

Вывод выстраивается вдоль логической цепочки: *задания* → *ответы* → *выводы о соответствии испытуемого заданному критерию*. Под критериально-ориентированной интерпретацией имеется в виду сравнение содержания аттестационных материалов с результатами тестирования и вывод — что из заданного стандарта, в смысле требований, и на каком уровне реально усвоено.

При критериально-ориентированной интерпретации требуется несколько меньшее число заданий, посредством которых определяется, что испытуемый знает и что не знает из заданного перечня заданий. Иначе говоря, здесь ответы оцениваются относительно не всей области (Domain) требуе-

мых знаний, а только из области, ограниченной конкретным набором заданий. Как и в случае с Domain Referenced Testing, интерпретация результатов ведётся на языке учебной дисциплины, но уже преимущественно работниками органов управления образованием и теми педагогами, на мнения которых управленцы опираются при аттестации. Используемые при этом «тесты» не соответствуют требованиям, предъявляемым к традиционным и адаптивным тестам. Поэтому для диагностики заранее заданного уровня подготовленности используются, по существу, не тесты, а совокупности заданий.

Третий вид интерпретации тестовых результатов, полученных в традиционных тестах, называется нормативно-ориентированным. В тестах с *нормативно ориентированной* интерпретацией результатов роль на первое место выходит задача определения не столько полноты содержания проверяемых знаний, сколько определения сравнительного места, или рейтинга, каждого из тестируемых. Для этого и нужен тест со сравнительно меньшим, но достаточным для данной задачи числом заданий.

Интерпретация результатов тестирования ведётся на языке педагогических измерений, с опорой на так называе-

ПЕД
измерения

мые процентильные нормы. Эти нормы показывают, сколько процентов испытуемых имеют тестовый результат хуже, чем у любого взятого для анализа испытуемого, с его (её) тестовым баллом⁴⁰.

Такая интерпретация называется *нормативно-ориентированной*. Здесь вывод выстраивается вдоль следующей логической цепочки. Содержание учебной дисциплины → генеральная совокупность заданий для измерения знаний → тест как небольшая⁴¹ выборка заданий из этой совокупности, ответы испытуемых → сравнение результатов, рейтинг (вывод о месте или ранге испытуемого, по сравнению с другими).

Главные вопросы такой интерпретации — не «кто знает что», а кто подготовлен лучше и кто по результатам тестирования оказался выше или ниже нормы, кто занимает какое место? В этом суть нормативно-ориентированной интерпретации тестовых результатов.

Для ответа на эти вопросы не требуется тест с очень большим числом заданий. Наоборот, стоит задача, которую можно назвать, для краткости, мини-максом: минимумом заданий получить достаточный максимум дисперсии истинных компонентов измерения. А на основе таких баллов далее считаются так называемые процентильные нормы и проводится рей-

тинг испытуемых посредством расчёта процентного ранга испытуемых.

Хотя при такой интерпретации всегда остаётся место для критики теста с позиции неполноты его содержания, в практике этот подход всегда оказывался лучшим из-за возможности сократить число задний и принимать решения на основе теста с меньшим числом заданий. По существу, это форма редукции, успех которой зависит от цели и качества разработки теста, а также от метода обоснования валидности редуцированных результатов относительно концепции подготовленности испытуемых.

Тестирование с *нормативно-ориентированной интерпретацией результатов* нередко противопоставляется тестированию с критериально-ориентированной интерпретацией, как якобы лучшее и более приближенное к содержанию учебного предмета. Но это, можно сказать, признак непонимания роли каждого вида интерпретации в общем тестовом процессе.

В заключение этого раздела остаётся добавить, что тестирование без указания предполагаемой интерпретации результатов не имеет большого научного смысла. Таким образом, понятие «тест» может иметь широкий смысл, включающий тест как метод, результаты и интерпретацию результатов.

40
Подробнее об определении процентилей см., напр., Дж. Гласс и Дж. Стенли. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. 495 с.

41
Существенно меньшая, чем в содержательно ориентированной интерпретации.

Качество тестовых результатов зависит от целей и задач, от видов, формы и содержания, от условий тестирования и методов проверки данных, от уровня знаний испытуемых и подготовленности персонала, проводящего тестирование, а также от качества интерпретации результатов измерения и от других факторов.

«Тест» надо понимать как *результаты* применения метода измерения, представленные обычно в форме матриц данных. Исходные матрицы подлежат математико-статистической обработке с целью определения тестовых свойств отдельных заданий и теста в целом. Удивительно, что тексты на русском языке тяготеют к определению теста как метода, в то время как в большинстве работ западных авторов понятие тест чаще рассматривается как данные, подлежащие правильной интерпретации.

Оба отмеченных смысла характеризуют тест с разных сторон. В действительности тест на Западе толкуется одновременно как метод и как результат педагогического измерения. Одно дополняет другое. К этому необходимо добавить и третий элемент — интерпретацию результатов. Результаты теста сильно зависят от избранной автором интерпретации результатов.

8. Три особенности композиции заданий в тестовой форме

Как определялось ранее, композиция *заданий в тестовой форме* определяется как педагогическая деятельность, объединяющая возможности тестовой формы и *содержание* учебных дисциплин в учебные и контрольные задания, такие, которые отвечают требованиям современных образовательных технологий. Главный предмет композиции заданий — творческое соединение их формы и содержания⁴².

Форма и содержание тестовых заданий — два ключевых направления развития практики тестирования. Умелое соединение формы и содержания автор этой статьи называет композицией тестовых заданий. Цель композиции тестовых заданий — создание такого педагогического задания, которое можно включить в тест и использовать как в традиционно организованном учебном процессе, так и в автоматизированных системах контроля знаний. Творческое соединение формы и содержания является сутью композиции. Одновременно это является и главным предметом педагогической методики разработки заданий в тестовой форме. Идеи композиции на сегодняшний день вос-

Теория

12/0000

42

Аванесов В.С.
Определение исходных понятий теории педагогических измерений // Педагогические измерения, № 2, 2005, С. 11.

приняты в Республиках Беларусь и Казахстан.

Композиция может быть успешной, если обратить внимание на три особенности.

Первая касается инвариантности (в смысле одинаково равной применимости) форм и принципов разработки заданий для многих, если не всех, учебных дисциплин и всех образовательных учреждений. Тестовые задания по русскому языку, теории музыки и философии, правильно подготовленные, часто выглядят не менее изящно, чем задания по математике и химии. Отсюда и общий, инвариантный характер форм и принципов разработки заданий, подходящих для преподавателей всех учебных дисциплин и во всех образовательных учреждениях. Да и изучать тестовую форму, как показывает многолетний опыт, лучше на примерах разных изучаемых наук, абстрагируясь на время от привычного содержательного анализа заданий.

В наши дни рекомендуется использовать задания с большим числом ответов⁴³, что позволяет, во-первых, фактически избавиться от возможности угадывания правильных ответов, потому что вероятность угадывания становится меньше одной тысячной. Задания с выбором нескольких правильных ответов из числа предлагаемых на выбор и системы заданий в тес-

товой форме придут на смену распространённым сейчас заданиям с выбором одного правильного ответа. Потому что имеют преимущества в смысле повышенной трудности, полноты, глубины и точности проверяемых знаний.

Вторая особенность касается числа форм. На Западе, а теперь и в России, считается, что имеется много форм тестовых заданий⁴⁴. Однако внимательный анализ ситуации позволяет думать, что реальное число тестовых форм всё-таки четыре. Пока. Остальные являются либо вариантами одной из них, либо нежелательным смешением двух и более форм. Есть и такие формы, которые нельзя называть тестовыми. Вероятно, здесь и лежит разница в счёте. Поскольку каждая из форм заданий помогает проверить тот или иной вид знаний, внимание к этому вопросу вполне оправдано.

Третья особенность — это стремление преодолеть распространённое заблуждение о связи заданий различных форм с уровнями знаний. Например, считается, что задания с выбором одного правильного ответа из числа предложенных позволяют проверить знания только низшего уровня (отличить правильный ответ от неправильного), в то время как знания более высоких уров-

43

Аванесов В.С.
Применение тестовых форм в Rasch Measurement // Педагогические Измерения № 4, 2005. С. 3–20.

Аванесов В.С.
Применение заданий в тестовой форме в новых образовательных технологиях // Педагогические измерения № 4, 2006.

44

Войтов А.Г.
Форма и формы тестовых заданий // Педагогические Измерения, № 2, 2005. С. 29.

ней надо проверять посредством заданий других форм. Однако измеряемый уровень подготовленности испытуемых зависит в значительной мере не от формы, а от содержания теста. С формой же больше связаны виды проверяемых знаний.

8. Этапы разработки педагогических тестов

Принцип системности предполагает последовательную организацию всей работы по созданию тестов. При этом может оказаться полезной уточнённая концепция автора поэтапного становления тестового педагогического процесса. В прежних вариантах было три⁴⁵ и четыре этапа⁴⁶: 1) формулирование целей и задач, разработка концепции теста, создание заданий в тестовой форме; 2) разработка тестовых заданий; 3) композиция теста, как системы; 4) использование математических моделей измерения.

Теперь число этапов представляется полезным расширить до пяти.

Ввиду особой важности композиции *заданий в тестовой форме* эту часть работы целесообразно выделить в качестве отдельного, второго этапа.

Разработка тестовых заданий переходит на третий этап, где главное — применение математико-статистических пакетов и моделей математического измерения.

На четвёртом этапе создаются тесты, определяется их качество и эффективность. Наличие достаточного числа тестовых заданий позволяет перейти к разработке *теста как системы*, обладающей целостностью, составом и структурой.

На пятом этапе проводится шкалирование заданий по уровню их трудности и дифференцирующей способности, а также шкалирование испытуемых по уровню их подготовленности. Для этой цели рекомендуется использование двух основных статистических пакетов: Winsteps и RUMM-2020. Оба они позволяют получить измерения на одной и той же интервальной шкале натуральных логарифмов.

Теория

45

45

Аванесов В.С.
Проблема качества педагогических измерений // Педагогические измерения, № 2, 2004. С. 27.

46

Аванесов В.С.
Применение тестовых форм в Rasch Measurement // Педагогические измерения. № 4, 2005. С. 8–9.