

ЯЗЫК ПЕДАГОГИЧЕСКИХ ИЗМЕРЕНИЙ

Вадим Сергеевич Аванесов,

профессор, доктор педагогических наук, testolog@mail.ru

КАЖДАЯ НАУКА СТРЕМИТСЯ ИМЕТЬ СВОЙ СОБСТВЕННЫЙ ЯЗЫК, ПОЗВОЛЯЮЩИЙ КОРРЕКТНО ХАРАКТЕРИЗОВАТЬ И ИССЛЕДОВАТЬ ИНТЕРЕСУЮЩИЕ ОБЪЕКТЫ. ПЕДАГОГИЧЕСКИЕ ИЗМЕРЕНИЯ В ЭТОМ ПРОЦЕССЕ НЕ ЯВЛЯЮТСЯ ИСКЛЮЧЕНИЕМ. ЭТО СРАВНИТЕЛЬНО МОЛОДАЯ НАУКА, ВОЗНИКШАЯ В САМОМ НАЧАЛЕ XX ВЕКА НА СТЫКЕ ПЕДАГОГИКИ, ПСИХОЛОГИИ, МАТЕМАТИКИ, СТАТИСТИКИ, ЛОГИКИ И ФИЛОСОФИИ.

Методы оценивания педагогических измерений

На сегодняшний день сложились три основных метода оценивания и педагогического измерения уровня подготовленности учащихся и студентов, используемых в национальных и международных системах.

Первый, самый распространённый — это метод *тестирования*. Он применяется в США, странах Европы, в Казахстане и странах Юго-Восточной Азии. Его актуальность объясняется следующими преимуществами перед другими методами:

1. *Высокая научная обоснованность* настоящего теста.
2. *Технологичность* тестовых методов.
3. *Объективно фиксируемая и открыто проверяемая точность измерений*.
4. *Одинаковость правил измерения и адекватной интерпретации* тестовых результатов.
5. *Сочетаемость тестовой технологии* с другими современными образовательными технологиями.

Но тестовый метод имеет и отрицательные последствия — в смысле развития речи учащихся и студентов, отрицательного влияния на содержание учебного процесса. Он также грешит социально-психологическими издержками, особенно в тех случаях, когда тестовая оценка становится решающей для формирования жизненной и профессиональной траектории развития молодёжи¹. Поэтому применение тестов должно обязательно сопровождаться использованием других методов.

Второй метод — традиционные экзамены, где иногда используются задачи повышенной трудности. Они применяются в большей или меньшей степени в программах международного исследования PISA, в России, на Тайване² и других стран Юго-Восточной Азии. Преимущество этого метода — положительное влияние на развитие речи учащихся, возможности индивидуализации контроля, углублённой проверки знаний по избранным темам. Кроме того, лучше обеспечивается связь между содержанием обучения и содержанием экзамена. Но для массового процесса измерения уровня подготовленности в генеральной совокупности испытуемых эти методы не годятся.

Третий метод — контрольно-измерительные материалы (КИМы). Применяются исключительно в рамках проведения российского ЕГЭ. По имеющимся данным, их качество ниже допустимого уровня. Возможно, что именно этим объясняется отсутствие внятных публикаций о качестве этих материалов.

Постановка проблемы

Для научного понимания феномена педагогических измерений надо, говоря словами А.А. Зиновьева, осуществить логическую

¹ **Avanesov, V.S.** Consequences of the EGE in RUSSIA. KEDI Journal of Educational Policy. Special Issue: pp. 89–100. Guest Editor — Hoi Suen. V. 3, № 1, 2006. <http://testolog.narod.ru>

² **Линь Ю-Сю.** Образовательная реформа на Тайване: От единого вступительного экзамена к многовариантным формам приёма в вузы // Педагогические Измерения. №1. 2007. С. 71–76.

обработку языка. Причём она должна охватывать не отдельно взятые понятия, а весь их комплекс. Результатом её должна явиться не сумма разрозненных фрагментов, а целостная теория (система, концепция), построенная в соответствии с правилами логики и методологии науки.

Ситуация с языком науки становится сложной в период, когда её статус не определён должным образом. Тогда в этой науке начинается употребление языков нескольких наук. В педагогических измерениях раннего периода вначале использовался язык статистики и математики. О том, что мы имеем дело именно с математико-статистическими теориями, было чётко сказано в классической работе³ конца шестидесятых годов прошлого века.

Принято полагать, что начало научному этапу становления педагогических измерений дал Чарльз Спирман⁴. Хотя он занимал должность профессора психологии, язык и методы этого автора были преимущественно статистическими. Не случайно и созданная им теория

позже стала называться статистической теорией педагогических и психологических измерений. Многие авторы называли её также классической. В историю науки эта теория так и вошла под двумя именами. Один из смыслов второго её названия заключался в том, что статистика — это и есть настоящая (классический) язык психолого-педагогических измерений.

Однако последующее развитие науки показало, что смысл названия «классическая теория» оказался гипертрофированным. В середине XX века появилась другая, Item Response Theory,

не менее классическая теория и более продвинутая в научном отношении. А потому сейчас самое подходящее время называть теорию, начало которой положили Ч. Спирман, Г. Галиксен⁵ и тысячи других авторов именно статистической теорией педагогических измерений.

До построения целостной и развитой теории педагогических измерений дело может идти не скоро, особенно если принять во внимание синтетический характер знаний этой области. В силу ряда причин в России долгое время не было условий для становления языка педагогических измерений. Нужных условий нет, правда, и сейчас.

Для решения проблемы становления приемлемого языка педагогических измерений требуется анализ причин неудовлетворительного состояния этого языка, проведение терминологической работы и создание первых языковых систем.

Проблемой создания педагогической теории педагогических измерений в России занимался автор этой статьи. Идея необходимости перехода от тестирования к системе педагогических измерений возникла как результат сравнительного исследования вопросов методологии педагогических измерений и тестирования⁶, научного анализа практики централизованного тестирования⁷, введения российского Единого государственного экзамена (ЕГЭ)⁸ и личного опыта взаимодействия с практиками — работниками центров тестирования.

В наше время существенный признак педагогических измерений — это педагогически обоснованный отбор содержания теста, точный подбор формы заданий для того или иного содержания, система основных педагогических понятий этой теории⁹, получение данных об испытуемых и заданиях на интервальной шкале натуральных логарифмов. Последнее достигается использованием соответствующих алгоритмов трансформации исходных результатов практического тестирования посредством компьютерных программ типа RUMM 2020, Winsteps и др.

Многолетнее отставание педагогических измерений от потребностей практики тестирования и фактическое отсутствие учёных в штате большинства центров тестирования на пространстве СНГ превращает практику тестирования в зону постоянного

³ Lord F.M. Lord F.M., Novick M. Statistical Theories of Mental Test Scores. Addison-Wesley Publ. Co. 1968, Reading, Mass. 560 pp.

⁴ Spearman, Ch. The proof and measurement of association between two things. American J. of Psychology. 15, 72–101.

⁵ Gulliksen H. Theory of Mental Test Scores. N-Y, 1950.

⁶ Аванесов В.С. Вопросы методологии педагогических измерений// Педагогические Измерения. № 1. 2005 г. С. 3–27. <http://testolog.narod.ru/Theory34.html>

⁷ Аванесов В.С. Централизованное тестирование лучше Единого Государственного Экзамена// «Развитие тестовых технологий в России. Тезисы докладов Всероссийской научно-методической конференции / Под ред. Л.С. Гребнева. М.: Центр тестирования Министерства образования РФ. 2002. <http://testolog.narod.ru>

⁸ Аванесов В.С. Единый Государственный Экзамен в фокусе научного исследования// Педагогические Измерения, № 1, 2006 г. С. 3–31. <http://testolog.narod.ru/Education39.html>

⁹ Аванесов В.С. Научные проблемы тестового контроля знаний. М.: Иссл. Центр, 1994.

риска получения ненадёжных и невалидных — практически непригодных для обозначенных целей — результатов. Этот вывод легко подтверждается независимой экспертизой работы любого центра тестирования. Вывести тестирование из зоны перманентного риска получения ошибочных оценок личности может только стабильная опора на научные педагогические измерения.

Причины отсутствия общепринятого языка педагогических измерений

Можно выделить четыре причины отсутствия требуемого языка педагогических измерений.

Первая причина — недостаточная исследованность вопроса сущности самой науки о педагогических измерениях. Для начала можно отметить, что эта наука синтетична. Синтетичность здесь понимается в смысле необходимости синтеза научных достижений и лексики ряда наук. В педагогических измерениях сейчас фактически используется несколько языков — педагогики, статистики, математики и, кроме того, в заметной мере, языки психологии, психометрики, педагогической технологии, информатики и других наук.

Педагогические измерения имеют одновременно свойства как фундаментальной, так и прикладной науки. Например, физические измерения основаны на фундаментальных понятиях пространства, времени, массы, на физических законах и свойствах физических тел, на соотношениях изучаемых параметров, а также на свойствах чисел и аксиомах собственно физической теории измерений. Педагогические измерения тоже опираются на педагогические концепции, понятия, свойства, индикаторы интересующего свойства, на свойства чисел и аксиомы.

Это несомненные признаки фундаментальности, присущие любому измерению. Но одновременно педагогические измерения призваны разрабатывать методы измерения, используемые в педагогической практике. А это можно отнести к приложениям педагогической науки. Хотя сама российская педагогика данное приложение не рассматривала своей частью. В тридцатые годы XX века в СССР педагогические

и психологические измерения назывались буржуазной тестологией.

Вторая причина — задержка с развитием собственно педагогической теории измерений. Целый век использовалась статистическая теория, а последние полвека — математическая теория (Item Response Theory, IRT). Они и представляли язык педагогических измерений.

Третья причина исключительно российская — именно в России возник официально принятый язык контрольных измерительных материалов (КИМов) единого государственного экзамена (ЕГЭ), измерительный статус которых научно не определён и не подтверждён каким-либо удовлетворительным образом.

Четвёртая причина отсутствия общего и устойчивого языка педагогических измерений заключается в том, что с течением времени начинают различать явления, которые ранее не различались¹⁰.

С научно-исторической точки зрения, ни КИМы ЕГЭ, ни исходные результаты даже настоящего тестирования не могут рассматриваться как педагогические измерения. Они оказываются всего лишь первым этапом сбора данных для последующего логарифмического шкалирования с целью получения трансформированных баллов на интервальной шкале. Тестирование — это только начало педагогического измерения, это незаконченное педагогическое измерение. В новейшей литературе только шкалированные результаты тестирования признаются полноценными педагогическими измерениями.

Основные дотеоретические понятия

Вопрос о дотеоретических понятиях имеет свою историю. Они возникли в естественном языке много лет назад, когда ещё не было заметных следов теоретической работы.

Экзамены сложились в период, превышающий четыре тысячи лет¹¹. Особенно важно проводить экзамены там, где нужно проверить качество речи учащихся. В наше время формированию речи учащихся в большинстве образовательных учреждений

¹⁰ Зиновьев А.А. На пути к сверхобществу. М.: ЗАО Изд-во Центрполиграф, 2000.

¹¹ Аванесов В.С. Из глубины веков. <http://testolog.narod.ru>

уделяют заметно меньше внимания, а потому это качество стало сильно ухудшаться. С точки зрения массового процесса педагогических измерений экзамены стали формой не технологичной, субъективной и экономически неэффективной.

Задание. Под этим названием можно представить многие педагогические формы, используемые для упражнений учащихся, проверки и самопроверки уровня их подготовленности. В результате теоретического анализа было определено, что *задание* — это педагогическая форма, нацеленная на проверку усвоения учащимся изучавшихся тем и разделов курса. Каждое задание имеет как бы свою миссию, создаётся для определённой цели и обладает своими характеристиками. Главная миссия заданий — побуждение учащихся к активной самообразовательной деятельности, развитию и овладению необходимыми знаниями, умениями и навыками до требуемого уровня компетенции¹².

Педагогические задания выполняют как обучающие, так и контролирующие функции. Невнимание к обучающим возможностям заданий в тестовой форме стало одной из причин отставания российского образования от положения дел с этим в других странах¹³.

Контроль. Это слово можно рассматривать как синоним слову проверка (педагогическая деятельность, позволяющая определить уровень учебных достижений). Контроль может быть субъективным или объективным.

Педагогический контроль. Данное словосочетание содержит в себе элементы дотеретического и теоретического знания, поскольку позволяет отделить традиционные

формы контроля от современных технологических форм. Сейчас основным предметом педагогического контроля рассматривается преимущественно подготовленность, включающая в себя знания, умения, навыки и компетенции. Попытка перейти от донаучных форм контроля к на-

учным была предпринята автором этой статьи в пособии для слушателей учебного центра¹⁴. В этом пособии было отмечено, что в педагогических измерениях пробивает дорогу традиция, давно утвердившаяся в психологическом тестировании. Это формулирование тестовых заданий не в виде традиционных задач, вопросов, указаний, требований, а в виде утверждений в повествовательной форме предложений, которые затем превращаются в высказывания, истинные или ложные.

Оценивание. В зарубежной педагогической литературе широко используется 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 итоговое¹⁵. 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.

Оценки нередко путают с отметками. Последние являются численными представителями оценочных суждений. Они вошли в жизнь в советский период, директивным путём — 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.

Три источника и три составные части теории педагогических измерений

Язык педагогических измерений формируют три основные теории — это статистическая, математическая и собственно педагогическая теории педагогических измерений. Ни одна из этих теорий не отменяет и не заменяет другие. Все они нужны и применяются в меру их потенциальных возможностей.

Статистическая (классическая) теория позволяет дать оценку заданиям и тесту в целом, провести предварительную оценку

¹² Аванесов В.С. Основы педагогической теории измерений // Педагогические измерения. № 1. 2004 г. С. 15–21. и другие, в №№ 1–2. 2004 г. и в №№ 1–4 2005 г.

¹³ Аванесов В.С. Основы теории педагогических заданий. ПИ, №№2–3, 2006.

¹⁴ Аванесов В.С. Основы научной организации педагогического контроля в высшей школе. М.: МИСиС, 1989.

¹⁵ Bloom B.S. a.o. Handbook on Formative and Summative Evaluation of Student Learning. N-Y., McGraw-Hill, 1971.

надёжности и валидности тестовых результатов. Математическая (Item Response Theory)¹⁶ нацелена на углублённый анализ характеристик каждого отдельного задания, формирования их в оптимальную систему измерения, с требуемыми характеристиками качества результатов. Педагогическая необходима для оценки качества содержания, формы и педагогического понятийного аппарата педагогических измерений.

Краткий анализ каждой из теорий даётся ниже. Но прежде обратимся к языку практики.

Язык практики педагогических измерений

Основным предметом практики педагогических измерений являются создание теста, проведения тестирования и шкалирования тестовых результатов.

Тестирование — это практическая деятельность, практика применения тестов — педагогических, психологических и др.

Педагогическое тестирование определяется как преимущественно практика применения тестов для оценки уровня и структуры подготовленности.

Система педагогического тестирования — организация сбора и обработки данных посредством применения тестов на множестве испытуемых. Эта система может оказаться научно обоснованной, не обоснованной и даже антинаучной. Научной основой системы тестирования должна стать теория педагогических измерений. Нынешняя практика тестирования даёт немало фактов для разных выводов — как против тестов, так и в их пользу. Важно понять простую истину: сами по себе тесты — полезный метод, но их применение в практике может оказаться и вредным. Многое зависит от цели, качества, соблюдения принципа добровольности, возможности потенциального или реального ущерба личности.

Язык статистической теории педагогических измерений

Как уже отмечалось, первым языком теории педагогических измерений стал язык статистики, созданный Чарльзом Спирманом. Это язык оказался единым для теории как

педагогических, так и психологических измерений. В начале XX века не делали заметных различий между этими двумя направлениями применения статистической теории. Соответственно, такую объединительную линию, ввиду общности статистического языка, много лет проводил и продолжает это делать научный журнал *Educational and Psychological Measurement*¹⁷.

Самые распространённые статистические характеристики — доля правильных и неправильных ответов, коэффициент корреляции между ответами испытуемых по заданию и исходными тестовыми баллами испытуемых, показатели средней тенденции (средняя арифметическая, мода и медиана), показатели вариации исходных тестовых результатов. При разработке теста вычисляются пять показателей вариации — размах данных, сумма квадратов отклонений от средней арифметической, дисперсия (варианса), стандартное отклонение, коэффициент вариации. Кроме того, для исходных тестовых баллов считается коэффициент асимметрии и мера эксцесса.

После вычисления и анализа данных по этим значениям вычисляют коэффициент надёжности тестовых результатов. Все перечисленные статистики удобно вычислять стандартными статистическими пакетами «Статистика», SPSS и им подобными. Формулы для вычисления статистических характеристик читатель найдёт в статьях журнала «Педагогические измерения», на сайте автора¹⁸, в учебниках по статистике¹⁹. Ряд статистических вопросов классической теории педагогических измерений связан с вероятным угадыванием правильных ответов.

Вероятность угадывания правильных ответов

С первых дней своего существования статистическая теория педагогических измерений

¹⁶ Аванесов В.С. Там же. См. две статьи автора по IRT в ПИИ № 2 и № 3, 2007г. и др.

¹⁷ <http://epm.sagepub.com>. Позже этот журнал стал публиковать работы и по математической теории измерений, применяемой также с равным успехом к педагогическим и к психологическим измерениям.

¹⁸ <http://testolog.narod.ru>

¹⁹ Глас Дж. Стэнли Дж. Статистические методы в педагогике и психологии: Пер. с англ. Л.И. Хайрусовой. Общ. ред. Ю.П. Адлера. М.: Прогресс, 1976.

Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных. Изд-во «Речь», 2007.

начала уделять внимание феномену возможности угадывания правильных ответов, при применении заданий с выбором одного правильного ответа. Математико-статистическая сторона этого явления уже была рассмотрена в работах авторов журнала²⁰. В КИМах ЕГЭ этот вопрос обходят стороной. Хотя угадывание является там особо существенным источником погрешностей измерения — тем большим, чем меньше число предлагаемых ответов в каждом задании с выбором одного правильного ответа.

При выборе из четырёх ответов могут быть угаданы правильные ответы примерно в четвёртой части от общего числа заданий. В итоге испытуемые получают баллы, которые они не заслужили. Это и есть одна из самых распространённых форм искажения результатов за счёт формы заданий с выбором одного правильного ответа.

С целью преодоления феномена угадывания в ЕГЭ было принято решение использовать задания открытой формы, где угадывания практически нет. Правильные ответы по ним оцениваются в два раза выше — в предположении, что задания открытой формы труднее для испытуемых, чем задания с выбором одного правильного ответа.

Но задания открытой формы для проведения массовых исследований не годятся из-за нетехнологичности. Ведь не случайно настоящие тестовые службы в массовых обследованиях используют только задания с выбором, применяя далее автоматизированную считку результатов с бланков посредством оптических сканеров очень высокой производительности, от 5 до 50 тысяч бланков в час. Хороший пример успешного применения такой техники даёт тестовая служба Республики Казахстан. Там результаты тестирования по всей республике становятся известными испытуемым в тот же день.

Трудность заданий зависит, главным образом, не от формы, а от содержания заданий. Не случайно даже в ЕГЭ ряд заданий с выбором одного правильного ответа оказывается труднее ряда заданий открытой фор-

мы. На этот очевидный результат ошибочного волевого «шкалирования» оценок за выполнения заданий разной формы в Кимах ЕГЭ не обращается внимания. Здесь скрываются и другие источники ошибок измерения, вызванные *неверным толкованием* тестовых форм и их возможностей, субъективным приписыванием весовых коэффициентов заданиям открытой формы, ошибками сложения баллов, не обладающих свойством, допускающим такое сложение, и т.д.

Основные критерии оценки результатов педагогических измерений

К настоящему времени исследованы четыре основных критерия качества и эффективности педагогических измерений.

Надёжность. Это был первый критерий, предложенный основателем статистической теории педагогических и психологических измерений Ч. Спирманом. Идея надёжности берёт начало в идее точности измерения. Зная погрешность, легко корректировать меру точности результатов. Язык и методы обоснования надёжности на русском языке представлены в ряде работ автора²¹.

Валидность (от англ. слова valid — действительный, пригодный, действенный) — один из основных критериев качества педагогических измерений. Обсуждение вопроса валидности предполагает поиск ответов, по меньшей мере, на три вопроса.

Первый — валидность чего? Раньше считалось — теста. В последние годы валидность стала соотноситься с тестовыми результатами. Валидность результатов — это характеристика их возможности измерять именно интересующее свойство, и никакое другое.

Второй вопрос — от чего зависит валидность тестовых результатов. От цели разработки и применения теста, от содержания и формы тестовых заданий, от качества теста, условий проведения измерений и от интерпретации результатов. Валидность результатов зависит также от подбора контингента испытуемых, от их физического и психического состояния.

Третий вопрос — о критериях оценки валидности результатов педагогических измерений. Используются формальный

²⁰ Ким В.А. Коррекция баллов на угадывание. ПИ. № 4. 2006. С. 47.

Деменченко О.Г. Влияние угадывания на значение тестового балла: корректировать или устранять? ПИ № 1. 2007. С. 56.
Ким В.С. Коррекция исходных баллов испытуемых. ПИ. № 3. 2007. С. 37. и др.

²¹ Аванесов В.С. Проблема качества педагогических измерений. ПИ. № 2. 2004. С. 3–27.

(правильность формы), содержательный, концептуальный и статистический критерии.

В отечественной и западной тестовой литературе много пишется не только о валидности тестов, но и, ошибочно, о валидности отдельных заданий. Научные разработки последних лет позволяют подойти по-новому к этому вопросу, а именно: вместо валидности оперировать другим понятием — эффективностью заданий. Естественным следствием такого подхода является обогащение определения теста: в этом случае он становится системой не просто заданий, а эффективных заданий. Понятно, что система неэффективных заданий порождает неэффективный тест, а в некотором пределе — совсем не тест²².

Объективность педагогических измерений — это такое отражение интересующего свойства личности на числовой шкале, которое адекватно действительному распределению испытуемых по латентному свойству. Можно сказать, что объективность является самым важным и, вместе с тем, наименее исследованным критерием качества педагогических измерений. Объективность может возникнуть как результат применения системы методов измерения²³.

Идея объективности результатов педагогических измерений была осознана вместе с возникновением первых тестов. По мнению J. Mac Keen Cattell, тест — это средство для получения объективных оценок интересующего свойства личности. Для организации тестирования он считал необходимым создание условий, приближённых к эксперименту. Требуются одинаковость инструкций, равное время на выполнение заданий каждому участнику тестового процесса, добровольность участия испытуемых в тестировании, статистическая обработка данных, ограничение времени тестирования — не более одного часа²⁴. Подробнее о критерии объективности — в статье автора²⁵.

Эффективность педагогических измерений — сравнительно новый критерий оценки теста и тестовых заданий. Проблема эффективности педагогических тестов является частью общей проблемы эффективности форм и методов педагогической деятельности. Естественно поставить вопрос — почему тестирование относится к эффективной форме организации кон-

троля знаний, а сам тест считается эффективным и объективным методом диагностики уровня и структуры знаний? Краткий ответ на этот вопрос заключается в том, что настоящий тест экономичен, потому что не требует тех больших затрат живого труда преподавателей, которые сейчас имеют место быть²⁶.

Язык математической теории педагогических измерений Item Response Theory (IRT)

В России название IRT переводили так: «теория латентных черт», «теория характеристических кривых заданий», «теория моделирования и параметризации педагогических тестов», «современная» теория тестов и т.д. Столь заметные различия в переводах одного только названия IRT уже само по себе являются свидетельством неблагополучия в понимании её сути. Не лучшим образом обстояло дело с переводом на русский язык исходных понятий и положений IRT. Дело в том, что английское название этой теории не точное, не полное, отчасти устаревшее и метафоричное. А потому неперебиваемое, в принципе.

IRT определяется как математическая теория педагогических измерений (МТПИ), основным предметом которой является анализ формальных, математико-статистических свойств отдельных заданий и теста в целом²⁷.

В своё время модели IRT назывались современными²⁸. В последовавших затем у нас публикациях других российских авторов современные модели превратились в «современную теорию», что сразу же высветило два существенных недостатка такого толкования. Выходило, что пользователи теории IRT — современные

²² Подробно о валидности на русском языке см.: **Аванесов В.С.** Проблема качества педагогических измерений. ПИ. № 2. 2004. С. 3–27.

²³ Objective Measurement. <http://www.meaningfulmeasurement.com/Objective%20Measurement.pdf>

²⁴ **Cattell J. McKeen.** Mental Tests and Measurements. Mind, 1890, v.15, p. 373–380.

²⁵ **Аванесов В.С.** Проблема объективности педагогических измерений. ПИ. № 3. 2008 г.

²⁶ Подробнее об этом критерии: см.: **Вадим Аванесов.** Проблема эффективности педагогических измерений. ПИ. № 4. 2008 г. С. 3–24.

²⁷ **Аванесов В.С.** Item Response Theory: основные понятия и положения. ПИ. № 2. 2008 г.

²⁸ **Аванесов В.С.** Основы научной организации педагогического контроля в высшей школе. М.: Иссл. Центр, 1989.

авторы, в то время как опирающиеся на другие теории, не современные. Между тем, два сравнительно новых варианта классической теории измерения — Random Sampling Theory, а также Theory of Generalizability²⁹ — существенно не менее современны, чем IRT.

В главном фокусе изучения IRT находятся отдельные задания, изучаемые посредством математических моделей, хотя и разработке теста уделено большое внимание. Модель измерения определяется как структурное построение, позволяющее соединить так называемые латентные переменные с одним или с большим числом эмпирически наблюдаемых переменных³⁰.

Основной язык математической теории педагогических измерений рассмотрен автором в двух статьях³¹. Важно отметить, что вопросами определения педагогических

понятий теории измерений, содержания педагогических тестов и формы тестовых заданий занимается не IRT (МТПИ), а ПТИ — педагогическая теория измерений.

Язык педагогической теории измерений: основные понятия

Примерно до середины 50-х годов XX века наука о разработке тестов называлась преимущественно «Теорией тестов»³². И лишь с начала XXI века в качестве научной основы педагогических измерений повсеместно утвердился термин «Теория педагогических измерений»³³. Так сложилось исторически,

таким был идеал научности тестовых результатов. Однако попытки построения собственно педагогической теории педагогических измерений, других идеалов научности не прекращались ни в прошлом, ни в наше время. Известны попытки создания педагогической теории педагогических измерений в США³⁴.

В качестве педагогической основы теории педагогических измерений может служить система основных понятий педагогических измерений³⁵. Перечислим их:

Задание в тестовой форме

Тестовые задания

Педагогический тест

Испытуемые

Тестовый процесс

Содержание теста

Форма тестовых заданий определялась как способ связи, упорядочения элементов задания. Содержание теста существует, сохраняется и передаётся в одной из четырёх основных форм заданий. По форме все известные в теории и практике тестовые задания можно разделить на четыре основные группы. Вот их точные названия: задания с выбором одного или нескольких правильных ответов, задания открытой формы, задания на установление соответствия и задания на установление правильной последовательности. Тысячи примеров заданий различных форм, по многим, если не всем учебным дисциплинам читатель найдёт в книге автора³⁶.

*Композиция тестовых заданий*³⁷ — это название ещё одной книги, посвящённой вопросам творческого соединения формы и содержания. Композицию не случайно считают делом сосредоточения многих проблем, связанных с замыслом и исполнением, содержанием и формой. Хорошо сделанное задание — это всегда лучшее, на текущий момент, содержание и наилучшая форма; то и другое — результат, появляющийся вследствие осознания важности их взаимосвязи. Разработка теста начинается с композиции заданий в тестовой форме.

Кроме перечисленных основных понятий, основу ПТИ могут составить:

- концепция пяти этапов тестового педагогического процесса³⁸;

²⁹ Точного перевода названий этих теорий на русский язык пока нет.

³⁰ Bollen K.A. Structural Equations with Latent Variables. N-Y, Wiley & Sons, 1989.

³¹ Аванесов В.С. Item Response Theory: основные понятия и положения. ПИ. №№ 2–3. 2008 г.

³² Например, название классического труда того времени Gulliksen, H. «The Theory of Mental Test Scores». N-Y, Wiley, 1950.

³³ В России автор этого доклада ранее использовал понятие «тестология», теперь применяется только «педагогические измерения».

³⁴ Например, в работе Ebel, R.L. Measuring Educational Achievement. Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1965.

³⁵ Аванесов В.С. Определение исходных понятий. 3 редакция, 7 апреля 2007. <http://testolog.narod.ru/Theory46.html>; Аванесов В.С. Понятийный аппарат теории педагогических измерений// <http://testolog.narod.ru/Theory26.html>

³⁶ Аванесов В.С. Форма тестовых заданий. М.: Центр тестирования, 2005 г.

³⁷ Аванесов В.С. Композиция тестовых заданий. М.: Центр тестирования, 2003 г.

³⁸ Аванесов В.С. Пять этапов педагогических измерений. <http://testolog.narod.ru/Theory58.html>

- формулировки целей педагогических измерений³⁹;
- результаты исследования содержания⁴⁰, форм⁴¹ и принципов композиции тестовых заданий⁴²;
- математические модели⁴³ и аксиоматика теории педагогических измерений⁴⁴.

Ключевое место в первом номере журнала «Педагогические измерения» заняла авторская концепция трёх главных понятий — задания в тестовой форме, тестового задания и педагогического теста⁴⁵, как системы заданий возрастающей трудности.

В этих работах были даны определения следующих основных понятий теории педагогических измерений:

Латентные качества личности. Латентными называются интересующие исследователей положительные и отрицательные качества личности, не поддающиеся непосредственному измерению. Примерами являются «подготовленность студентов», «знание учебной дисциплины», «способность понимать» «интеллектуальное развитие» и многое другое. Попытки измерения подобных качеств⁴⁶ на уровне обыденного сознания оканчиваются словесными или численными оценками, содержащими в себе немалые погрешности. Методом измерения латентных качеств (свойств) личности является тест.

Педагогическое задание в тестовой форме. Это технологичное средство интеллектуального развития, образования и обучения, способствующее активизации учения, повышению качества знаний, а также повышению эффективности педагогического труда. Логическое преимущество задания в тестовой форме заключается в возможности его превращения после ответа студента в форму истинного или ложного высказывания. Задания в тестовой форме выгодно отличаются свойствами эффективности, краткости, лучшей понимаемости смысла заданий, быстротой ответа учащихся и определения меры трудности каждого задания, технологичности.

Научно-методическая ориентация педагогов на широкое использование обучающего потенциала заданий в тестовой форме в системе e-Learning даёт начало новой образовательной технологии. В ней задания начинают выполнять функцию не только

самоконтроля, но и организации образовательной и самообразовательной деятельности, делая это самым технологичным и эффективным образом. С точки зрения теории педагогических измерений⁴⁷ главным средством обучения в e-learning могут и должны стать не тесты, а задания в тестовой форме.

Взаимосвязь содержания и формы. При рассмотрении этого общего диалектического принципа применительно к тестам невольно возникает ассоциация с искусством. Настоящее искусство представляет собой, по словам Гегеля, законченное внутри себя соединение содержания с вполне соответствующей ему формой⁴⁸. Точно так же и настоящий педагогический тест можно охарактеризовать как результат взаимовлияния содержания заданий с наиболее подходящей формой. При этом, в отличие от широко распространённого философского истолкования активности содержания и пассивности формы, тестовую форму надо рассматривать вместе с содержанием, как активную сторону взаимодействия.

Тестовое задание определяется как составная единица теста, отвечающая содержательно-педагогическим требованиям к заданиям в тестовой форме и, кроме того, статистическим требованиям: известной трудности, достаточной вариации тестовых баллов испытуемых по заданию, положительной корреляцией ответов по заданию с исходными тестовыми баллами испытуемых.

Не все задания в тестовой форме могут стать тестовыми заданиями. Это заметно различающиеся

³⁹ Аванесов В.С. Основы педагогической теории измерений // ПИ. №1. 2004 г. С. 15–21.

⁴⁰ Аванесов В.С. Форма тестовых заданий. М.: Центр тестирования, 2006.

⁴¹ Аванесов В.С. Форма тестовых заданий. М.: Центр тестирования, 2006.

⁴² Аванесов В.С. Теоретические основы разработки заданий в тестовой форме. Уч. пособие для профессорско-преподавательского состава высшей школы. М.: МГТА, 1995. Аванесов В.С. Композиция тестовых заданий. М.: Центр тестирования, 2003.

⁴³ Аванесов В.С. Математические модели педагогического измерения. Научное издание. М.: Иссл. центр проблем качества подготовки специалистов, 1994.

⁴⁴ Аванесов В.С. Педагогическое измерение латентных качеств // Педагогическая диагностика. № 4. 2003 г.

⁴⁵ Там же.

⁴⁶ Понятия «качество», «свойство», «признак» удобно рассматривать как обобщённый аналог английского понятия trait.

⁴⁷ Серия статей в журнале «Педагогических Измерений» а также на сайте <http://testolog.narod.ru>

⁴⁸ Гегель Г. Энциклопедия философских наук // Соч. Т. 1. Логика, 1929.

понятия. Задания имеют шанс стать тестовыми только после эмпирической проверки меры их трудности и других статистических свойств, на типичных группах испытуемых. Корреляция ($r > 0,3$) указывает на способность заданий различать хорошо подготовленных учащихся от тех, кто подготовлен плохо.

Чем больше значение коэффициента корреляции, тем выше *дифференцирующая способность задания*. В последние годы вместо понятия «дифференцирующая способность задания» автор этой статьи чаще применяет понятие «различающая способность задания». Это связано с тем, что слово «дифференцирующая» в зарубежной теории педагогических и психологических измерений используется для нескольких целей.

Для того чтобы задание можно было включить в тест, его проверяют по множеству и других формальных математико-статистических требований. Это обычно выполняется с помощью статистических пакетов RUMM-2020 и WINSTEPS, которые созданы специально для углублённого анализа формальных свойств тестовых заданий, а также для проведения дуального (conjoint) шкалирования уровней трудности заданий и уровней подготовленности испытуемых.

Именно этот смысл, а не «калибровка», как нередко пишут на русском языке, следует из английского словосочетания *test calibration*. Свойства тестовых заданий были рассмотрены в статьях и работах автора⁴⁹. Задания в тестовой форме имеют шанс стать тестовыми только после эмпирической проверки их статистических свойств, на типичных группах испытуемых.

• *Педагогический тест* — система заданий равномерно возрастающей трудности, позволяющая оценить структуру и качественно измерить уровень подготовленности испытуемых. Тесты применяются для объективизации итогового контроля результатов обучения.

При культурном и качественном тестировании каждое задание теста должно обязательно иметь параллельные варианты. Параллельными называются варианты задания, имеющие сходное содержание из одной и той же общей дидактической единицы, примерно совпадающие меры трудности и вариации результатов испытуемых.

Авторы очень известной на Западе книги *Fundamentals of Item Response Theory*⁵⁰ пишут, что очень трудно, или даже невозможно, создать параллельные варианты теста для проверки надёжности тестовых результатов. И действительно, все варианты к каждому заданию, по каждому предмету должны иметь общее содержание. Они должны иметь хотя бы примерное равенство средних арифметических показателей, а также показателей вариации по каждому заданию. Только тогда обеспечивается справедливость и объективность выставляемых оценок⁵¹ испытуемых.

Создание параллельных вариантов заданий заметно облегчается при овладении методикой композиции фасетных заданий⁵². В рамках одной только математической теории создать тест с качественными фасетными заданиями практически невозможно. Поэтому авторов упомянутой зарубежной книги по математической теории измерений понять можно.

Остаётся напомнить, что *фасет* — это форма записи возможных параллельных вариантов задания, что является обязательным требованием при разработке теста, имеющего на каждое задание набор параллельных заданий. Все элементы одного фасета априорно считаются одинаково трудными по содержанию, что требует последующего эмпирического подтверждения. Каждому испытуемому компьютерная программа подбирает один вариант из фасета. □

⁴⁹ Подробнее см. **Аванесов В.С.** Композиция тестовых заданий. М.: Центр тестирования, 2002. С.163.

⁵⁰ **Ronald K. Hambleton, H. Swaminathan, H. Jane Rogers.** *Fundamentals of Item Response Theory*. p.6., Sage publ., 1991.

⁵¹ **Аванесов В.С.** Проблема объективности педагогических измерений// Педагогические измерения. № 3. 2008 г.

⁵² **Аванесов В.С.** Форма тестовых заданий. М.: Центр тестирования, 2005.