

ПЕДАГОГИЧЕСКОЕ ИЗМЕРЕНИЕ ЛАТЕНТНЫХ КАЧЕСТВ¹

Вадим Аванесов

В статье представлена классификация теорий измерения и дано определение латентных качеств (свойств) личности. Поставлена проблема построения общей теории педагогических измерений. (Подробнее об этом см. в кластере из примерно 200 работ автора и более чем 2000 уже опубликованных статей научных коллег, на сайтах viperson.ru и edumetrica.ru). Сформулировано уточнённое определение педагогического теста. Статья заметно обновлена в августе 2023 года

1. Основные понятия

Латентные качества. Латентными называются положительные и отрицательные качества личности, не поддающиеся непосредственному измерению. Примерами являются «подготовленность студентов», «знание учебной дисциплины», «способность понимать» «интеллектуальное развитие» и многие другие. Попытки измерения подобных качеств² на уровне обыденного сознания оканчиваются словесными или численными оценками, содержащими в себе немалые погрешности.

Попытки судить о невидимых вещах по видимым проявлениям имели место еще в глубокой древности. Например, Платон утверждал, что истинная астрономия занимается не изучением движения наблюдаемых небесных тел, а изучением законов движения звезд по математическому небу. Несовершенным изображением последнего является видимое небо³. В эпоху Возрождения скрытые качества рассматривались как некие силы, присущие всем вещам, причем каждая вещь представлялась как нечто двойственное: чувственно воспринимаемый предмет и внутреннее "скрытое качество"⁴.

Идея и методы измерения латентных качеств реализуются в тесной зависимости от эмпирических результатов. Именно на основе реально наблюдаемых данных ставится задача воссоздания непосредственно ненаблюдаемого качества, измеряемого с помощью модели. В науке эта ситуация не является новой. С ней уже неоднократно сталкивались физики при изучении микромира.

2. Педагогические измерения

¹ Статья представляет расширенный вариант доклада автора на V Всероссийской научно-практической конференции «Оценка эффективности образовательных инноваций и технологий», 27-28 июня 2003г., Славянск-на-Кубани, Краснодарского края.

² Понятия «качество», «свойство», «признак» удобно рассматривать как обобщенный аналог английского понятия trait.

³ Клайн Б.Л. В поисках физики и квантовая теория / Пер. с англ. М.: Атомиздат, 1971. - 286с.

⁴ Суворов Л.Н. Материалистическая диалектика.- М.: Мысль, 1980.- 264с.

Опора на основные понятия делает понятным смысл короткого, а потому общего определения: педагогическое измерение – процесс отображения числами уровней проявления интересующих латентных качеств личности. F.M.Lord и M.Novick определяют измерение как такое присвоение чисел, которое верно отражает расположение испытуемых на числовой шкале в зависимости от выраженности измеряемого качества⁵.

В отличие от элементарных оценок и некоторых простых физических измерений, процесс научно-педагогического измерения латентных качеств личности требует теоретизации. В последнюю входят: определение ведущего понятия, проверка логической правильности имени измеряемого качества; определение предмета измерения, а также системы индикаторов, понятийных и эмпирических, указывающих на наличие или отсутствие интересующего качества⁶. Далее требуются аксиоматика и математические формализмы, выбор подходящей модели и стандартизация условий измерения. И, наконец, полученные результаты подлежат аргументированной интерпретации с точки зрения их точности и соответствия ясно сформулированной цели.

3. Краткий анализ теорий измерений

Latent Trait Theory. Педагогическое измерение латентных качеств в западной литературе ассоциируется с Latent Trait Theory (LTT)⁷, которую можно перевести как теорию измерения латентных качеств.

В начале 40-х годов D.N.Lawley, задался целью получить такие оценки качества тестовых заданий, которые не зависели бы от состава оцениваемой группы испытуемых. Было понятно, что если задание с номером j давать хорошо подготовленной группе, то доля правильных ответов в этой группе будет заметно выше, чем у слабо подготовленных испытуемых. Возникает вопрос - а какова действительная мера трудности задания j для генеральной совокупности испытуемых? Для этого надо было найти некое свойство, присущее заданию, которое оказалось бы инвариантным относительно уровня подготовленности группы испытуемых. Таким свойством, как выяснилось, обладают графики, или иначе, так называемые характеристические кривые тестовых заданий и теста в целом. В качестве меры работоспособности задания, т.е., способности дифференцировать студентов по уровню их подготовленности этот автор стал рассматривать параметр крутизны кривой. Таким образом, каждому заданию теста в LTT ему удалось поставить в соответствие два параметра - трудности и крутизны графического образа каждого задания⁸.

Позже стала очевидной неточность названия Latent Trait Theory, где утверждается специфическая направленность данной теории на измерение

⁵ Lord F.M. Lord F.M., Novick M. Statistical Theories of Mental Test Scores. Addison-Wesley Publ. Co. 1968, Reading, Mass. - 560 pp.

⁶ Об этом процессе подробнее см., например, на стр. 87-105 книги: Аванесов В.С. Тесты в социологическом исследовании. М.: Наука, 1982.- 199с.

⁷ Теория измерения латентных качеств.

⁸ Lawley D.N. On Problems Connected with Item Selection and Test Construction // Proceedings of the Royal Society of Edinburgh. Section A Mathematical and Physical Sciences. 43 v. LXI, part III, p. 273 - 287, 1943.

именно латентных качеств. Однако это неверно. На самом деле все, без исключения, теории педагогических измерений направлены на оценку истинных компонентов измерений латентных качеств, явно не наблюдаемых. И все теории так же исходят из аксиомы устойчивого существования истинных компонентов измерения. В силу приведенных аргументов название Latent Trait Theory нельзя признать абсолютно правильным.

Item Response Theory. Вероятно поэтому в последние годы стали применять второе название Latent Trait Theory – Item Response Theory (IRT). В своё время модели этой теории большинство авторов модели IRT называли современными⁹. В последовавших затем у нас публикациях современные модели превратились в «современную теорию», что сразу же высветило два существенных недостатка подобной трансформации.

Во-первых, выходило, что пользователи теории IRT - современные авторы, в то время как опирающиеся на другие теории авторы становились не современными. Похоже, что такой разворот кому-то был нужен. Между тем, два сравнительно новых варианта классической теории измерения - Strong True Score Theory и Random Sampling Theory, а также Theory of Generalizability¹⁰, не менее современны, чем IRT. Куда их отнести?

Во-вторых, в фокусе изучения IRT находятся, строго говоря, не тесты, как системы заданий, а отдельные задания. Вот почему это, скорее, по сути, не «современная» теория тестов, а математическая теория оценки свойств и качества тестовых заданий.

Не случайно главным в IRT является утверждение о зависимости вероятности правильного ответа испытуемого от уровня его подготовленности и от параметров задания. Эту зависимость удобнее представлять в виде логистической функции. Число рассматриваемых параметров ставятся в соответствие модели измерения. Модель измерения определяется как структурное построение, позволяющее соединить так называемые латентные переменные с одним или с большим числом эмпирически наблюдаемых переменных¹¹.

Идеи и методы датского математика Георга Раша

Датский математик G.Rasch рассматривал матрицу тестовых данных как результат сопряжения множества испытуемых с множеством заданий. Он попытался формализовать идею зависимости результатов такого сопряжения от соотношения уровня подготовленности каждого испытуемого с мерой трудности каждого задания. Позже, в 1976 г., эта идея была развита им в виде так называемого фрейма взаимодействия двух множеств¹².

$$\Phi = P(O, A)$$

⁹ Аванесов В.С. Основы научной организации педагогического контроля в высшей школе. М.: Иссл. Центр, 1989. –167с.

¹⁰ Точного перевода названий этих теорий на русский язык пока нет.

¹¹ Bollen K.A. Structural Equations with Latent Variables. N-Y, Wiley & Sons, 1989.- 514p.

¹² Rasch, G. On Specific Objectivity: An Attempt of Formalizing the Request for Generality and Validity of Scientific Statements / Danish Yearbook of Philosophy. 1977, v. 14, p. 58 - 94, Munksgaard, Copenhagen. - 216p.

где Φ - фрейм сопряжения (взаимодействия);

O - множество заданий;

A - множество испытуемых;

P - исход, результаты взаимодействия множеств O и A .

К 1958 году у G.Rasch возникла идея выразить вероятность правильного ответа на задание j посредством так называемой логистической функции вида

$$P_j \{ X_{ij} = 1 \mid \beta_j \} = \exp(\theta_i - \beta_j) / (1 + \exp(\theta_i - \beta_j))$$

где $X_{ij} = 1$, если ответ любого испытуемого (i) на j -ое задание правильный;

θ_i - уровень знаний, латентная переменная;

β_j - уровень трудности j -го задания теста, измеряемой на латентном континууме.

Соответственно, вероятность неправильного ответа на задание j , обозначаемая (Q_j) равная, как принято в теории вероятностей, $1 - P$, выразилась так:

$$Q_j \{ X_{ij} = 0 \mid \beta_j \} = 1 - \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}$$

где взаимодействие θ_i и β_j представлено не мультипликативно, а аддитивно¹³, что является главным позитивным элементом этой модели измерения. Элементарные преобразования формулы посредством приведения разности к общему знаменателю $1 + \exp(\theta_i - \beta_j)$ позволяют выразить её в более удобном виде.

$$Q_j = 1 / (1 + \exp(\theta_i - \beta_j))$$

Симметрично возникла и модель, описывающая вероятность правильного ответа студентов с уровнем знаний θ_i на задания различного уровня трудности.

$$P_i \{ X_{ij} = 1 \mid \theta_i \} = \exp(\theta_i - \beta_j) / (1 + \exp(\theta_i - \beta_j))$$

Значения θ_i и β_j могут быть аппроксимированы из матрицы эмпирических данных.

Модели G.Rasch, A.Birnbaum, и F.M.Lord имеют много общего с математической точки зрения, однако они различаются концептуально¹⁴. Различны они и по генезису. F.M.Lord, его предшественники M.W.Richardson и D.N. Lawley начинали с поисков математической модели репрезентации эмпирических данных. Подход же G.Rasch был априорным, теоретическим, направленным на создание математической модели измерения и получения

¹³ Rasch, G. On General Laws and the Meaning of Measurement in Psychology /In Proceedings of the Fourth Berkley Symposium on Mathematical Statistics and Probability. Berkley: Univ. of California Press, 1961.

¹⁴ McArthur D.L. Educational Assessment: A Brief History /McArthur D.L. (Ed). Alternative Approaches to the Assessment of Achievement. Kluwer Academic Publishers, Boston, 1987.- 268p.

такой единицы педагогического измерения, с помощью которой можно было бы, на одной шкале, сравнить уровень знаний студента с уровнем трудности задания.

Преимущества математических моделей Георга Раша

Практические преимущества этого достижения в области педагогических измерений последних лет трудно переоценить.

Первое преимущество вытекает из стандартизованного характера полученной единицы измерения. Как и всякая стандартная единица, она представляет собой результат преобразования исходных данных, что даёт возможность объективно сравнить достижения разных студентов, по разным учебным дисциплинам.

Второе преимущество вытекает из специфических особенностей модели G.Rasch. Получаемые с её помощью оценки знаний, в силу относительной независимости от конкретного подбора того или иного задания, приобретают характер достаточно обобщённых результатов, что так же положительно отражается на качестве оценок, используемых в педагогическом контроле.

Третье преимущество связано с возможностями получения интервальной шкалы. Недавние исследования G.Karabatsos подтвердили преимущество моделей G.Rasch для построения строго интервальной шкалы тестовых баллов испытуемых, по сравнению с 2-х и 3-х параметрическими моделями¹⁵.

Четвертым преимуществом рассматриваемой модели является сравнительная устойчивость рассчитываемых значений уровня знаний и трудности задания. Это позволяет утверждать: однопараметрические модели удачно оценивают интересующие качества личности, недоступные непосредственному измерению.

Другие модели педагогического измерения

Тем не менее, неоднократно возникали попытки улучшения моделей G.Rasch за счет введения второго (после β_j) параметра. Этот параметр даёт информацию о задании j с точки зрения крутизны его графика (т.н. большинством авторов характеристической кривой). Обозначив этот параметр графика буквой a_j , выражение можно теперь представить в виде двухпараметрической модели

$$P_j\{X_{ij} = 1 \mid \beta_j, a_j\} = \exp a_j(\theta - \beta_j) / (1 + \exp a_j(\theta - \beta_j))$$

Добавление третьего параметра c_j , ассоциируемого с вероятностью угадывания правильного ответа в заданиях с выбором одного правильного ответа даёт трехпараметрическую модель.

$$P_j\{X_{ij} = 1 \mid \beta_j, a_j, c_j\} =$$

¹⁵ Karabatsos G. Axiomatic measurement theory as a basis for model selection in item response theory. Paper presented at 32nd annual conference of the Society for Mathematical Psychology, Santa Cruz, CA. 1999, July.

$$= c_j + (1 - c_j) \exp a_j(\theta - \beta_j) / (1 + \exp a_j(\theta - \beta_j))$$

Даже отлично знающие студенты, имеющие самые высокие значения θ_i , могут иногда терпеть неудачу по причинам, отнюдь не связанным со знанием предмета, а с какими-то другими факторами, которые иногда трудно определить. На этот и другие случаи труднообъяснимых причин рассчитана малоизвестная за рубежом четырёхпараметрическая модель измерения латентных качеств личности. Эта модель представляет интерес не столько практический, сколько теоретический. Вот её запись:

$$P_j\{X_{ij} = 1 \mid \beta_j, a_j, c_j, \gamma_j\} = \\ = c_j + (\gamma_j - c_j) \exp a_j(\theta - \beta_j) / (1 + \exp a_j(\theta - \beta_j))$$

где γ_j принимает значения больше единицы.

В IRT ставится важный исследовательский вопрос - как соотносятся между собой эмпирически наблюдаемые результаты испытуемого (X_i) с соответствующими значениями того же испытуемого на прямой (ось абсцисс), изображающей латентную переменную величину θ ? F.M.Lord сделал вывод, что линия регрессии тестового балла на латентную переменную имеет, в общем случае, вид логистической функции, откуда можно понять, что тестовый и истинный баллы связаны нелинейно¹⁶. Всё отмеченное позволяет предложить новый вариант перевода названия IRT на русский язык как *математической теории тестовых заданий*. Что близко к смыслу названия в английском языке.

Классическая теория тестов

Начало этой теории положил Ch. Spearman в 1904 году. Существенный вклад в её развитие внесли H. Gulliksen и многие другие авторы. Они предполагали наличие истинного компонента измерения, в каждом результате. Они же заложили основы аксиоматики психолого-педагогических измерений.

В настоящее время можно выделить по меньшей мере семь основных аксиом¹⁷.

1. Устойчивого существования интересующего качества (признака, свойства, характеристики) личности.
2. Возможности представления интересующего качества как измеряемой величины.
3. О существовании истинного значения интересующего качества у каждого испытуемого. Измеряемая величина предполагается варьирующей на множестве испытуемых.

¹⁶ 307. Lord F.M. Application of Item Response Theory to Practical Testing Problems. Hillsdale N - J. Lawrence Erlbaum Ass., Publ. 1980, - 266 pp.

¹⁷ Переработано по книге Gulliksen H. Theory of Mental Tests. N - Y. Wiley. 1950 - 486 p и книги Аванесова В.С. «Научные проблемы тестового контроля знаний». М.: Иссл. Центр, 1994. - 135с.

4. О неизбежной погрешности измерения, в соответствии с которым измеряемое значение (X) не равно истинному (T): X предполагается отягощенным некоторой погрешностью. Таким образом,

$$X \neq T$$

Откуда следует, что у каждого испытуемого

$$X_i = T_i + E_i$$

где E_i - некоторая случайная ошибка измерения, состоящая из суммы возможных ошибок различного происхождения, значения и знака.

В реальном процессе педагогического измерения каждый испытуемый имеет свои значения X_i , T_i , и E_i . Значения X_i рассматриваются как функция от истинного уровня подготовленности (T_i)¹⁸ и от значений ошибок измерения (E_i). Каждый из компонентов последнего равенства варьирует на множестве испытуемых.

5. Аксиома о случайных погрешностях измерений, распределяемых по так называемому в статистике нормальному закону.

6. Аксиома некоррелируемости истинных и ошибочных компонентов измерений. Из четвертой и шестой аксиом вытекает седьмая аксиома.

7. Дисперсия тестовых баллов принимается состоящей из двух компонентов: $s_x^2 = s_t^2 + s_e^2$.

Почленное деление последнего выражения на s_x^2 даёт

$$1 = \frac{s_t^2}{s_x^2} + \frac{s_e^2}{s_x^2}$$

Перенесем $\frac{s_e^2}{s_x^2}$ влево; получим $1 - \frac{s_e^2}{s_x^2} = \frac{s_t^2}{s_x^2}$.

Концептуально надёжность тестовых результатов равна правой части данного равенства; операционально же это понятие выражается левой частью:

$$r_{tx} = 1 - \frac{s_e^2}{s_x^2}$$

Последняя формула в западной литературе названа классической, вероятно, ввиду её концептуальной простоты и широкой применимости в разработке тестов.

Классификация теорий

Латентные качества изучаются не одной, а всеми научными теориями педагогического измерения. Такая позиция позволяет выработать интегрированную и сбалансированную позицию по отношению к положительным сторонам и реальным возможностям каждой теории. Периодически появляющиеся теории обычно не отвергают, а дополняют одна другую; они различаются не по признаку «современности», а по этапам их

¹⁸ В Latent Trait Theory (IRT) вместо T_i чаще других используется символ θ_i .

применения, по лексике, моделям и методам измерения, а также по прогнозируемости индивидуальных результатов и релевантности поставленным целям.

Например, в начале 30-х годов, физик по образованию, классик Чикагской психометрической школы L.L.Thurstone подчеркивал важность индекса надежности результатов измерения (r_{xt}). Этот индекс является мерой связи наблюдаемых значений каждого варианта теста с истинными компонентами измерений¹⁹. Если взять за основу классификации, например, философские понятия общего, особенного и единичного, то можно предложить такую классификацию теорий:

Общая теория педагогических измерений

Контурсы общей теории педагогических измерений удачно просматривались в работах R.Ebel²⁰ и других педагогов-тестологов. Последовавшее затем увлечение большинства американских авторов математической стороной проблемы измерения отвлекло их от построения целостной общей педагогической теории измерения. Главный предмет такой теории - разработка и применение заданий в тестовой форме, тестовых заданий и тестов для активизации учебного процесса, научной организации тестовой формы самоконтроля в процессе самообразования.

Настоящее тестирование начинается, развивается и прекращается вместе с учебным процессом. Отстраненное от такого процесса тестирование неизбежно становится объектом бюрократических искажений, несовместимых с тестовой культурой.

Другие предметы общей теории - разработка целей и задач педагогических измерений; создание программ проведения массового тестирования, их положительных и отрицательных последствий на личность, общество и государство. К сугубо научным вопросам общей теории можно отнести анализ понятий, положенных в основу измерений; содержание заданий и теста в целом; разработка педагогических критериев эффективности и качества проводимых измерений. Попытки разработки общей теории педагогических измерений представлены в ряде публикаций²¹.

Частные педагогические теории.

С общей теорией соотносятся частные теории. В качестве одной из таковых могла бы стать логико-педагогическая *теория композиции заданий в*

¹⁹Thurstone L.L. The Reliability and Validity of Tests 1932. Ann Arbour, Michigan. Он отмечал: «The index of reliability ...serves to show us the order of magnitude of the discrepancy to be expected when we use an obtained score to represent the theoretical or ideal true scores» (p.13).

²¹Ebel, R. Measuring Educational Achievement. Prentice Hall, Inc. Englewood Cliffs, New Jersey. 1965. -481pp.

²¹Аванесов В.С. Форма тестовых заданий // Труды иссл. центра. М. 1991.-33с.

Аванесов В.С. Основы научной организации педагогического контроля в высшей школе. М.: Иссл. центр, 1989. -167с.

Аванесов В.С. Методологические и теоретические основы тестового контроля". Дисс... докт. пед. наук. СПб госуниверситет, 1994.-339с.;

тестовой форме. Учитывая непосредственную зависимость качества любого теста от правильной композиции содержания и формы каждого отдельного задания, вряд ли можно признать конструктивными известные попытки недооценки в потребности такого рода теории. Ведь каждое нарушение формы, содержания, того и другого – источники погрешностей измерения. Прообразы такой педагогической теории можно обнаружить в работах многих авторов второго поколения западных тестологов и в публикациях на русском языке. (См. напр.²²). Предмет теории – разработка принципов композиции заданий в тестовой форме, позволяющих творчески соединять содержание каждого задания с подходящей для него формой.

Математические и статистические теории педагогического измерения.

К числу таких теорий надо отнести известную на Западе, но неисследованную в России Random Sampling Theory. Известные варианты классической теории тестов полезно рассматривать в связке с положениями Random Sampling Theory, учитывая их общую опору на статистические теории индуктивного вывода. Более общей является другая теория, Theory of Generalizability (Dependability)²³ так же не исследованная у нас. Это самые общие теории, сосредоточившие в себе мощный статистический аппарат исследования точности результатов измерения, возможности использования дисперсионных комплексов для оценки достоверности влияния различных источников погрешностей при измерении латентных качеств.

Специальные математические теории педагогического измерения. К ним можно отнести частную математико-педагогическую *теорию тестовых заданий*, в которую целесообразно включить часть положений классической теории тестов, известной на Западе под названием «Item Analysis», а также постулаты, модели и вычислительные методы IRT. Только взятые вместе они позволяют проводить тщательный анализ эффективности каждого задания с последующим решением о включении их в тест, в банк заданий, или об исключении из разработки²⁴. По сравнению с общими теориями, предмет исследования IRT является теорией меньшего уровня общности.

Основное средство педагогического измерения латентных качеств.

Наиболее подходящим для измерения латентных качеств является тест. Самый удивительный факт - в российской и западной литературе уделяется

²²Аванесов В.С. Теоретические основы разработки заданий в тестовой форме. Пособие для профессорско-преподавательского состава высшей школы. М.: МГТА, 1995.-95с.

Аванесов В.С. Композиция тестовых заданий. 1,2 и 3 изд.; 3 изд. М.: Центр тестирования 2002г. – 217с.

²³ Перевод названий этих теорий на русский язык пока нет.

²⁴ Именно так построена 11 глава книги автора «Композиция тестовых заданий». 3 изд. М.: Центр тестирования 2002г. – -239с. Идейным толчком к возникновению такой композиции послужила работа Tatsuoaka, K.K. Item construction and psychometric models appropriate for constructed response. Princeton, N-J, 1993. -56 pp;

недостаточное внимание определению теста. *Педагогический тест* - система заданий возрастающей трудности, специфической формы, позволяющая эффективно измерить уровень и оценить структуру результатов подготовленности испытуемых. Это определение гомогенного (одномерного) теста, т.е. измеряющего только одно интересующее качество или свойство личности. Из данного определения вытекает, что тест, как система, обладает целостностью, эффективностью (включающей в себя критерии качества результатов измерений), составом и структурой.

Особого пристального внимания заслуживает анализ такой «новации», не известной в педагогической науке, как т.н. контрольно-измерительные материалы (КИМы). О сущности этих «нововведений» писалось немало²⁵, но сейчас сложилась практика, в которой критерии оценки и перевода баллов из одной некачественной шкалы в другую, довольно спорную и непонятную массу, шкалу Центра тестирования, переводятся распоряжениями Минобразования РФ²⁶. Данный факт убеждает, что это министерство (сейчас Просвещения) обременено функциями руководства т.н. «экспериментом по ЕГЭ» и «созданием КИМов²⁷», что мало согласуется с традициями научного руководства такого рода делами.

²⁵ Напр., «И не тест, и не экзамен..». Ж. «Университет и школа» №1-2, 2002.с. 26-27

²⁶ Например, распоряжения №№748-13 от 11.06.03; №779-13 от 19.06.03; №817-13 от 20.06.03 и десятки, если не сотни других подобных документов.

²⁷ Появились публикации о КИМах, открывающие возможность начала нормальной дискуссии. Например, в сборнике «Единый Государственный экзамен. Научные основы, методология и практика организации эксперимента»: Под ред В.А.Болотова – (М.Логос. - 2002, 208 стр). одни авторы называют ЕГЭ и КИМы тестами; другие заявляют, что «говорить о ЕГЭ, как тесте, основываясь на результатах первого года эксперимента, конечно, рано» (с.34); третьи полагают, что «контрольные материалы для единого экзамена не сводятся к тестам в узком смысле слова». (стр. 14).

Привлекает внимание работа А.О.Татура: «Контрольно-измерительные материалы для единого государственного экзамена \\\ Высшее образование сегодня». №7-8 с.28-33, 2002. Автор даёт такое определение (с.28):. «КИМы – это стандартизованная экзаменационная работа, создаваемая в соответствии с требованиями теории педагогических измерений, позволяющая с достаточной объективностью и надёжностью провести независимую государственную аттестацию выпускников образовательных учреждений и отбор абитуриентов вузов». Из чего видно, что: 1) КИМы – это не тесты, а *экзаменационная работа*; 2) И что это экзаменационная работа, «позволяющая с достаточной объективностью и надёжностью провести *независимую государственную* аттестацию выпускников образовательных учреждений и отбор абитуриентов вузов».

Оставим в стороне ставшие ритуальными упоминания об «объективности и надёжности». Потому что даже сильно препарированные результаты подготовленного для узкого круга лиц, сборника «Статистика основных результатов единого государственного экзамена в 2003 году» свидетельствуют о прямо противоположном – об отсутствии какой-либо надёжности и объективности. А значит, и о непригодности этих КИМов для озвученных целей. Интереснее другое: в статье КИМы считаются инструментом «независимой государственной оценки»; - независимой, вероятно, ни от общества; ни от граждан страны!

Любопытна заключительная фраза этой статьи: «Дальнейшее развитие теории и практики создания и использования КИМов позволит нашей системе образования в полном объеме освоить достижения в области педагогических измерений». Из чего

Возникла необходимость нового истолкования сущности теста. Во-первых, надо уходить от пересечения тестовой лексики с традиционной педагогической лексикой экзаменов и вопросов. Во-вторых, в определение теста на русском языке надо специально подчеркивать, что тест не только метод, но и результаты измерения. Только тогда мы избавимся от устаревшей лексики «надежность и валидность теста» вместо более точного понятия «тестовых результатов испытуемых. Которые зависят не только от качества тестов, но и от качества подготовленных испытуемых. . А также и от таких ошибочных утверждений, как «надежность и валидность т.н. «КИМов» и «педагогических тестовых материалов», которые, по мнению самих же создателей, не являются, «пока», ни тестами, ни каким-либо другим научно признанным методами или т.н. материалами.

Свойства метода могут переноситься на свойства результатов, а могут, в силу множества источников погрешностей, и не переноситься. К этому необходимо добавить ещё один элемент: педагогический тест – это метод и результаты интерпретация данных. Потому что тест может быть одним и тем же, но в зависимости от множества причин результаты и их интерпретации могут существенно отличаться.

Такое истолкование сути теста позволяет по-новому взглянуть на перспективы развития тестовой культуры, особенно в связи с попытками создания так называемых «критериально-ориентированных тестов». На самом деле вопрос лучше ставить не о «тестах», а об интерпретации данных тестирования, в зависимости от того или иного критерия. Тест может быть один, с параллельными вариантами его заданий возрастающей трудности. А интерпретации результатов могут оказаться заметно различающимися. Да и число основных критериев вместо двух привычных – надёжности и валидности результатов – лучше расширить до четырёх, добавив к только что названным ещё два важнейших критерия - эффективности и объективности. Но для этого надо к попыткам настоящего тестирования надо привлекать больше учёных и представителей научной общественности. Сейчас это если где-то и делается, то явно недостаточно.

можно понять, что вслед за ЕГЭведением надо ожидать возникновения ещё одной псевдонауки – т.н. КИМологии. Тогда естественно поставить вопрос – где данные, свидетельствующие о том, что КИМы - это научно обоснованные методы, и что огромные расходы на них, оторванные, кстати, от скудных образовательных бюджетов страны и регионов, оправданы? Ведь все матрицы исходных результатов в Центре тестирования МО РФ по-прежнему недоступны для независимой экспертизы, что противоречит Закону.

Принципиальная невозможность попыток создать один «КИМ» для выпускников и абитуриентов мною уже отмечалась (Аванесов В.С. Понятийный аппарат педагогической тестологии //Педагогическая диагностика №2, 2002г.;стр. 35-37) а также: Аванесов В.С. «Основные инструменты коррупции в образовании»//Материалы научно-практической конференции «Социология коррупции» (М.:ИНИОН РАН, 20 марта 2003г., стр. 201-205) и в статье «Нелегитимный эксперимент в образовании: Ещё раз о недостатках Единого государственного экзамена» (Независимая газета 10 июня 2003г.). Чем это не материал для научной дискуссии?