ПРОБЛЕМА ОБЪЕКТИВНОСТИ ПЕДАГОГИЧЕСКИХ ИЗМЕРЕНИЙ

Вадим Аванесов
testolog@mail.ru

Опубликовано в ж. «Педагогические Измерения №1 2013 года.

Аннотация

Проблема объективности педагогических измерений – малоисследованная и спорная. Некоторые полагают объективными результаты любого теста. Другие – только методы, отвечающие критериям надёжности и валидности. Третьи считают подлинную объективность недостижимой вообще. G.Rasch создал модель, позволяющую математически обеспечивать так называемую специфическую объективность[1] измерения. Его последователи создали своеобразную систему теорий, моделей, принципов, решающих правил и методов обоснования качества и упомянутой объективности получаемых результатов измерений. Всё это получило название Rasch Measurement (RM), которое полагают объективным вообще, хотя правильнее было бы говорить об объективированном педагогическом измерении, потому что в нём всегда имеют место субъективные элементы. Некоторые такие элементы не устранимы, или трудно устранимы. Целенаправленное уменьшение субъективных элементов есть процесс продвижения к объективности.

В настоящей статье в дополнение к двум известным критериям качества результатов – надёжности и валидности - предлагается ввести ещё один критерий – объективности результатов педагогических измерений. Взятые вместе, эти три критерия лучше обеспечивают процесс достижения объективности и эффективности педагогических измерений.

Ключевые слова: объективность, процесс объективации тестовых результатов, Rasch Measurement.

Вопросы истории проблемы

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, приближенных к эксперименту. Требуются одинаковость инструкций, равное время на выполнение заданий каждому участнику тестового процесса, добровольность участия испытуемых в тестировании, статистическая обработка

[1] 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. – 216 p.

данных, ограничение времени тестирования не более чем одним часом[2]. В послесловии к этой статье основатель тестовых методов Ф. Гальтон высказал идею научного обоснования качества получаемых результатов[3].

Ещё в начале 1904 года E.Thorndike обратил внимание на необходимость выделения стандартной единицы измерения, отсутствие которой обрекает само измерение на необъективность[4]. Вопросы объективности результатов тестирования были рассмотрены в ряде работ многих других авторов. Установка на достижение объективности получаемых тестовых результатов, а также на адаптивность была выдвинута одним из основателей и классиком психометрики L.L.Thurstone. Вот некоторые положения из его работ: “Результаты измерения свойства личности не должны зависеть от того или иного набора заданий, а сравнительная трудность заданий не должна зависеть от выборки испытуемых» - писал он. «Надо так строить процесс измерения, чтобы результаты не зависели и от пропуска испытуемым некоторых заданий во время тестирования. Не обязательно всем испытуемым отвечать на все задания. Тестирование нужно организовать так, чтобы каждый мог начать и закончить тестирование на подходящем для него уровне трудности»[5]. Это, по сути, идеи инвариантности получаемых тестовых результатов.

Впервые проблема объективности психолого-педагогических измерений была поставлена основателем статистической (классической) теории тестов[6] Ч. Спирманом. В исследовании 1904 года он назвал объективность главным свойством тестов[7]. В работе 1907 года Ч.Спирман писал: «Объективные тестовые результаты практически не достижимы, но приближения к ним вполне возможны[8]. В 1910 году он ввёл в научный оборот понятие «коэф-

---

[2] Cattell J. McKeen. Mental Tests and Measurements. – *Mind, 1890, v.15, p.373-380.*

[3] Galton F. Remarks. In: Cattell, J. McKeen. Op. Cit. p.380.

[4] Thorndike, E.L. (1904). An introduction to the theory of mental and social measurements. New York: Teacher's College.

[5] Thurstone, L. L. (1926) . The scoring of individual performance. *Journal of Educational Psychology, 17, 445-457.*

[6] Spearman, C. (1904a). “General intelligence,” objectively determined and measured. American *Journal of Psychology, 15,* 201-293. Сейчас чаще говорят не о теории тестов, а о теории педагогических измерений.

[7] Spearman, C. (1904b). The proof and measurement of association between two things. *American Journal of Psychology, 15, 72-101*;

Spearman, C. 1907. Demonstration of formulae for true measurement of correlation. Am. J. of Psychology. 18, 160-169.

фициент надёжности» и предложил формулу коррекции коэффициента корреляции между результатами по двум тестам, с учётом меры их ненадёжности[9].

Во времена Ч.Спирмана достижение объективности тестовых результатов считалось невозможным. Но эта точка зрения входила в противоречие с общей идеей теста как метода достижения именно объективных результатов. Отсюда вытекала главная проблема методологии педагогических и психологических измерений – достижение объективности. Объективность считает главным требованием к измерению также A.J.Stenner. Он считает, что результаты объективного измерения должны быть полностью независимы от используемого метода и конкретной ситуации измерения и выделяет два вида объективности – общую и локальную. Общая объективность характерна преимущественно для физических измерений.

Концентрация внимания тестологов к одному интересующему свойству личности является важным условием проведения объективированных измерений, что ведёт к созданию так называемой одномерной шкалы. L.L.Guttman определял такую шкалу как совокупность заданий общего предметного содержания, позволяющая подготовленному испытуемому иметь более высокий ранг, по сравнению с менее подготовленным испытуемым[10].

Основные понятия, используемые в статье

Понятие «проблема» предполагает наличие некоторых затруднений, противоречий или неопределённостей в состоянии какой-либо сферы деятельности.

Понятие «объективность» означает независимость суждений от сознания и чувств отдельного субъекта, соответствие мнений и результатов действительности. Объективность является понятием преимущественно философии, где объективному знанию противостоит знание субъективное, понимаемое обычно как односторонне, неполное и недостаточное, отягощённое личным отношением и, нередко, личной предубеждённостью. В объективном измерении результаты не должны зависеть от того, кто измеряет[11].

Объективность предполагает освобождение от всего субъективного, от субъективных влияний; реальность, нейтральность. Объективностью также называют способность что-либо наблюдать и излагать «строго объективно». Но такой способностью человек не обладает.

---

[8] Spearman, C. 1910. Correlation from faulty data, British J. of Psychology. 3,271-295.

[10] Guttman, L. L. (1950) . The basis for scalogram analysis . *In Stouffer et al. (Eds.), Measurement and Prediction . New York: Wiley.* - P.62.

[11] Bond T.G. & Fox C. M. Applying the Rasch Model. Fundamental Measurement in the Human Sciences. Cook University, University of Toledo, 2001.

Напротив, во всяком познании и высказываниях любого рода взаимодействует весь комплекс факторов, относящихся к телесному, душевному и духовному бытию индивида, включая и действующие в нем подсознательные силы и трансцендентные переживания. Поэтому подлинная объективность достигается лишь весьма приблизительно и остается для научного труда идеалом; 2) духовная тенденция совершать действие не ради личной выгоды, а во имя высшего порядка. Предпосылкой объективности является способность непредвзято и без предрассудков вникать в содержание дела, повиновение порядку вещей и преданность делу[12].

Измерением называют процесс представления свойства испытуемых в виде числовой переменной величины. Переменной величиной можно назвать все то, что может быть больше или меньше, что может быть присуще объекту в различной степени. Числовая переменная величина выражается числами. Таким образом, измерение есть установление числового соотношения между испытуемыми по интересующему свойству. Если свойство одно, то измерение называется одномерным, если несколько - то многомерным.

Измерение испытуемых производится в предположении, что у каждого из них есть измеряемое свойство, в каком-то количестве. Если выясняется, что у кого-то нет данного свойства, то это даёт основания для исключения данного испытуемого из предполагаемой выборки лиц, обладающих интересующим свойством. В профессиональной литературе на английском языке имеется специальное понятие suitable persons, что означает, измерение данным тестом проводится не всех испытуемых, вообще, а только тех, кто по уровню подготовленности соответствует уровню трудности заданий разрабатываемого теста. При проверке качества создаваемых тестов совокупность таких испытуемых образует так называемую по-английски target group.

Педагогические измерения основаны на теории и, одновременно, нацелены для применения в практике. Отсюда вытекает прикладной характер проблем педагогических измерений, по отношению к педагогике, как фундаментальной науке. Фундаментальной называется наука, основные положения которой не выводимы из других наук.

Модель измерения определяется как структурное построение, позволяющее соединить латентную переменную величину с наблюдаемыми значениями этой величины.[13].

Объективность педагогических измерений можно определить как такое отражение интересующего свойства личности на числовой шкале, которое адекватно действительному

---

[12] http://www.galactic.org.ua/clovo/p-o2.htm

[13] Bollen K.A. Structural Equations with Latent Variables. N-Y, Wiley & Sons, 1989.-514pp.

распределению испытуемых по данному 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 измерения [14].

Нередко считается, что абсолютно объективных педагогических измерений не бывает, что в них неизбежно присутствуют элементы субъективности. Среди таких элементов следует назвать субъективный выбор предмета измерения, выбор одномерной или многомерной моделей, мера вариации заданий теста по уровню трудности, мера вариации испытуемых по уровню подготовленности, средний уровень подготовленности групп испытуемых, подбор определенного содержания, формы и числа предъявляемых тестовых заданий. К субъективным можно также отнести и такие процессуальные элементы как время предъявления теста, дифференцированная оценка значимости (веса) каждого задания в общей системе тестовых баллов, форма учёта возможности угадывания правильных ответов, списывания и т.п. нарушений учебной этики.

В общем, в педагогическом измерении объективное нередко пересекается с субъективным, а потому пессимизм относительно возможности достижения полностью объективных педагогических измерений имеет основание. Деятельность по целенаправленному уменьшению влияния субъективных факторов на результаты тестирования и на достижение максимальной возможной объективности можно назвать процессом достижения объективности педагогических измерений.

Исходные тестовые баллы получаются из подсчёта числа правильных ответов испытуемых на задания теста. Конечные результаты измерения получаются в результате шкалирования.

Интересующее педагога свойство личности может быть выражено явно или не явно. Явно выраженное свойство является основой для формирования счётного показателя. Например, показатель посещаемости занятий может зависеть от нескольких причин, в том числе латентных, и этот показатель при должной постановке учёта является прямо наблюдаемым, счётным. Но это не измерение. Для педагогического измерения нужны также теории, модель измерения, принципы и методы трансформации счётных данных в интервальную шкалу. Теории позволяют выделить интересующее свойство, определить форму и содержание заданий, принципы и методы измерения данного свойства.

---

[14] Objective Measurement. http://www.meaningfulmeasurement.com/Objective%20Measurement.pdf

Неявно выраженные свойства личности называются *латентными.* Для измерения латентных свойств используются тесты, педагогические и психологические. Примером латентной величины является уровень подготовленности испытуемых. Концептуально этот уровень включает в себя владение знаниями, умениями, навыками и представлениями. В последнее время к этому списку добавляют и компетенции, но вряд ли это оправдано. Потому что признание испытуемого компетентным указывает на другой, прагматический подход к проверке соответствия достигнутого уровня его подготовленности выдвигаемым конкретным работодателем критериям. Эти критерии могут заметно различаться.

В педагогике нет показателей, абсолютно объективно указывающих на уровень подготовленности учащихся. Можно выделить лишь признаки и подобрать тестовые задания-индикаторы, эмпирически проверяющие наличие интересующего признака. В RM созданы методы, проверяющие меру соответствия трудности теста уровню подготовленности выборки испытуемых, методы соответствия заданий модели измерения, оценки пригодности используемых эмпирических индикаторов интересующего свойства (заданий) для создания теста как системы заданий возрастающей трудности.

Постановка проблемы

Никто не станет спорить с тем, что результаты педагогических измерений должны давать объективную информацию. Но достижение объективности оказалось проблемой высокого уровня сложности, требующей системного анализа ситуации, предшествующей измерению и системного подхода к проведению самого измерения. Но этого как раз и не хватает в педагогических измерениях. Зато субъективизм и предубеждённость при оценивании одного человека другим – традиционные спутники такого процесса. Отсюда проистекает актуальность проблемы получения таких оценок, которые были бы свободны от субъективности, обоснованы системой таких научных методов, которые способствуют достижению объективных результатов.

Абсолютно объективных педагогических измерений не бывает, в них неизбежно присутствуют элементы субъективности. Среди таких элементов следует назвать субъективный выбор предмета измерения, выбор одномерной или многомерной моделей, варьирование заданий теста по уровню трудности, вариация испытуемых по уровню подготовленности, средний уровень подготовленности групп испытуемых, подбор определенного содержания, формы и числа предъявляемых тестовых заданий. К субъективным факторам можно отнести

также и такие процессуальные элементы как время предъявления теста, дифференцированная оценка значимости (веса) каждого задания в общей системе тестовых баллов, форма учёта возможности угадывания правильных ответов, списывания и т.п. нарушений учебной этики.

В общем, в педагогическом измерении объективное нередко пересекается с субъективными элементами, а потому пессимизм относительно возможности достижения полностью объективных педагогических измерений имеет основание. Деятельность по целенаправленному уменьшению влияния субъективных факторов на результаты тестирования и на достижение максимально возможной объективности можно назвать процессом объективации педагогических измерений. Следовательно, правильнее было бы говорить не об объективных, а об объективированных измерениях.

Ввиду практической невозможности достижения абсолютной объективности тестовых результатов внимание исследователей стали больше занимать критерии надёжности и валидности результатов педагогических измерений. Не случайно понятие «объективность» иногда ассоциируется с понятиями «надёжность» и «валидность». И в этой логике есть смысл. Обоснование надёжности и валидности тестовых результатов – это важная, хотя и неполная часть общего процесса объективации результатов измерений.

Не бывает объективных измерений без достижения приемлемой надёжности, которая на русском языке часто ассоциируется с понятием точности. Вместе с тем, объективные результаты не обязательно должны быть абсолютно надёжными. В педагогических измерениях достаточно приемлемой меры надёжности, что приходится определять в каждом тестировании. Однако если надёжность результатов низкая или неизвестная, то вопрос об объективности педагогических измерений нельзя ставить вообще. Следовательно, надёжность – необходимая, но недостаточная часть требований к объективности тестовых результатов. С другой стороны ясно, что ненадёжные результаты однозначно ведут к необъективности. В западной литературе уже много лет существует традиция проверки и публикации мер надёжности и валидности тестовых результатов. Эти критерии уже рассматривались в публикациях нашего журнала[15]

Для достижения объективности опоры на критерии надёжности и валидности в наше время уже недостаточно. Их предстоит дополнить третьим критерием – объективность педагогических измерений. Все три критерия полезно рассматривать как систему ключевых ка-

[15] Аванесов В.С. Проблема качества педагогических измерений. ПИ, №2, 2005г.

тегорий педагогических измерений, помогающих поставить дело обоснования педагогических измерений на более обширную систему теорий, методов и технологий.

Хорошей мерой точности измерений – общей и дифференцированной - являются значения стандартных ошибок измерений, вычисление которых в тестовой культуре является обязательным. В отличие от многих стран мира, где приняты стандарты проведения педагогических измерений, в России требования вычисления ошибок педагогических измерений не выполняются. Например, это легко видеть на примере т.н. Единого Государственного Экзамена (ЕГЭ). Неизбежным следствием отсутствия информации о погрешности ЕГЭ является либо фактическое их отсутствие там измерений, либо недопустимо низкое качество таковых, если какие-то измерения там всё-таки проводятся. Науке это неизвестно.

Измерение может быть точным, но не вполне адекватным цели тестирования или научной концепции, положенной в основу построения интересующей переменной величины. Это случай невалидности тех тестовых результатов, которые не годятся для практического применения в силу несоответствия качества получаемых результатов поставленной цели или научной концепции, на основе которой проводятся измерения. Поэтому проверка валидности результатов тестирования приближает к объективности получаемые результаты. Но и это не гарантирует полной объективности, поскольку валидность результатов сильно зависит от субъективного выбора теоретической основы предмета измерения, подбора заданий-индикаторов интересующего свойства, от типа интерпретации результатов.

Вопрос определения меры объективности всегда оставался открытым ввиду зависимости от критерия объективности знаний. Высшим критерием объективности обычно считается соответствие знаний действительности, их подтверждение практикой. Хотя важную роль практики в обосновании объективной истинности знаний трудно отрицать, практика всё-таки не всегда является критерием истины.

Как отмечают Е.К.Войшвилло и М.Г.Дегтярёв, именно такое понимание роли практики верно в трёх отношениях.

Во-первых, пишут они, критерий практики не всегда применим – по крайней мере, для проверки истинности высказываний[16]. Во-вторых, практика может подтверждать и некоторые ложные высказывания. И в-третьих, практика лишь подтверждает, но не доказывает ис-

[16] Войшвилло Е.К., Дегтярёв, М.Г. Логика: Учебник для студентов высших учебных заведений. – М.: Изд-во ВЛАДОС-ПРЕСС, 2001.- 528с. (С.446).

тинность утверждений теории[17]. В тестовой технологии задания в тестовой форме, если он правильно сформулированы, являются высказываниями, пригодность которых для проведения измерения проверяется логически и педагогически, а в эмпирическом исследовании для этого используются системы математических и статистических методов обоснования. Только после такого рода объективированной проверки посредством нескольких методов, часть заданий, отвечающих критериям пригодности, могут называться тестовыми заданиями[18].

Существенный недостаток исследований по вопросам объективности результатов тестирования отметил в своей монографии H. Gulliksen[19]. Общую проблему обоснования объективности результатов эмпирических исследований рассмотрела H.Peak [20]. J.Loevinger выдвинула два главных требования к объективным психологическим и педагогическим измерениям: шкала должна быть гомогенной (иметь одно общее предметное содержание) и у шкалы должно быть свойство монотонности[21]: чем выше уровень подготовленности испытуемых, тем большим должно быть значение тестового балла. Свойство монотонности шкалы исходных тестовых баллов ранее рассматривалось как главный признак объективности теста. Теперь внимание переместилось с исходных баллов на шкалированные значения.

В педагогике измерению подвергается одно или несколько свойств, попадающие в фокус внимания исследователей. При этом широко используется логический приём абстрагирования от других свойств личности. Самым распространённым, но не единственным предметом педагогического измерения является уровень подготовленности испытуемых по какой-либо одной учебной дисциплине.

Попытка измерить одним тестом знания двух и более учебных дисциплин (например, знание русского языка и литературы) оканчивается неудачей в виду возникающей в подобных примерах двойственности содержания. Двойственнность содержания невозможно корректно позиционировать на одной числовой оси. L.L.Thurstone называл такую ось континуумом. Как отмечал J.P.Guilford, континуум не имеет ни начала, ни конца; это некоторая

---

[17] Войшвилло Е.К., Дегтярёв, М.Г. Ук. Соч. С. 446.

[18] Аванесов В.С. Композиция тестовых заданий. М.: Центр тестирования, 2002 г. –237с.

[19] Gulliksen, H. *Theory of Mental Tests*. New York: John Wiley & Sons, 1950. pp, 392-393).

[20] Peak, H. (1953) Problems of objective observation. In L.Festinger & D.Katz (Eds). Research Methods in the Behavioral Sciences. Pp.243- 299. N-Y: Holt, Rinehart and Winston.

[21] Loevinger, Jane. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs,* 1947, *61,* P.46.

протяжённость, которая представляет собой непрерывную числовую ось. Интересующие объекты могут получить на ней числовые значения[22].

. В начале 60-годов прошедшего столетия G.Rasch[23] хотел назвать объективностью открытое им свойство независимости параметров заданий от параметров испытуемых, и наоборот, свойство независимости параметров испытуемых от параметров заданий. Но поскольку смысл этого традиционного философского понятия был более широким, чем позволяло его математическое открытие, он решил добавить к этому ограничивающее определение; получилось «специфическая объективность»[24]. Имелась при этом в виду только одна, математическая сторона процесса обоснования объективности педагогических измерений. Объективировать другие стороны этого процесса было гораздо сложнее.

G.Rasch хорошо понимал значимость своей модели и понимал трудности, которые бы неизбежно возникли, если бы предложенный им подход был назван «объективным» Именно поэтому сам G.Rasch избегал слова «объективность». Один из вариантов предложенного им измерения имеет названия «sample-free measurement»[25]. При таком варианте параметры заданий, насколько можно, не зависят от выборок испытуемых и от вида статистического распределения результатов испытуемых по уровню подготовленности в выборках[26].

Последователи G.Rasch - B.D.Wright, M.H.Stone, G.Masters, D.Andrich, J.M.Linacre и другие - создали оригинальную теорию и методику достижения объективных, по их

---

[22] Guilford J.P. Psychometric Methods. N-Y. McGraw-Hill, 1936 1st ed. , 1954 – 2 ed.

[23] Stenner A. J. (1997) Objectivity, Units of Measurement and Zeroes. Rasch Measurement Transactions 11:2 p. 560-561. http://www.rasch.org/rmt/rmt112d.htm

[24] Rasch G. On specific objectivity: an attempt at formalising the request for generality and validity of scientific statements. *Danish Yearbook Philos.* 1977; 14:58–94. - P.58.

[25] Sample-free measurement means "item difficulty estimates are as independent as is statistically possible of whichever persons, and whatever distribution of person abilities, happen to be included in the sample. Rasch, G. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press, 1961, *4,* 321-333.

[26] As Wright (1968) explains, "Object-free instrument calibration and instrument- free object measurement are the conditions which make it possible to generalize measurement beyond the particular instrument used, to compare objects measured on similar but not identical instruments, and to combine or partition instruments to suit new measurement requirements . . ..When we compare one item with another in order to calibrate a test, it should not matter whose responses to these items we use for the comparison. Our method for test calibration should give us the same results regardless of whom we try the test on. This is the only way we will ever be able to construct tests which have uniform meaning regardless of whom we choose to measure with them." (p. 87-88).

мнению, измерений. D.Andrich назвал эту теорию и методику новой парадигмой[27], хотя сам G.Rasch называл свой подход методом достижения т.н. «специфической» объективности, обеспечиваемой исключительно математическими свойствами предложенной им модели. Отсюда и слово «специфическая».

Может показаться поразительным, но с точки зрения сторонников новой парадигмы исходные баллы даже качественных тестов не считаются ни объективными, ни педагогическими измерениями[28], [29]. Тем более таковыми не могут считаться баллы российского ЕГЭ. В новой парадигме педагогическими измерениями признаются только те, которые отвечают требованиям инвариантной интервальной шкалы натуральных логарифмов (логитов). Как уже упоминалось, при таком подходе баллы испытуемых не зависят, математически, от набора заданий теста, а меры трудности заданий не зависят от выборки испытуемых.

Для воплощения в жизнь упомянутой парадигмы пришлось создать систему математических моделей, расширять теорию и методы, создавать компьютерные технологии шкалирования исходных данных. Понадобились также новые критерии и методы объективации получаемых выводов. На Западе вся эта система знаний и технологий была названа Rasch Measurement (RM). На русский язык RM можно перевести как систему объективированных педагогических измерений по теории Г.Раша.

В основу теории специфической объективности легли три предположения:

1. Уровень трудности заданий и уровень подготовленности испытуемых можно измерить в одной шкале, с общей стандартной единицей измерения.

2. При наличии такой шкалы вероятность правильного ответа испытуемого может стать зависимой от разности между уровнем его подготовленности и уровнем трудности задания теста.

---

[27] Andrich D. Controversy and the Rasch Model: A Characteristic of Incompatible Paradigms? Med. Care. Volume 42, Number 1, suppl, January 2004.

[28] Wright B.D. Scores are not Linear Measures. Rasch Measurement Transactions, 1992, 6:1, 208 http://www.rasch.org/rmt/rmt61n.htm

[29] См. раздел Raw Scores are NOT measures. In: Measurement for Social Science and Education. A history of social science measurement. http://www.rasch.org/memo62.htm

3. Результат противоборства испытуемого с заданиями теста можно прогнозировать. Чем больше уровень подготовленности испытуемого, тем выше должна быть *вероятность* его правильного ответа на задание фиксированного уровня трудности[30].

Формой реализации первого положения стали матрица тестовых результатов и вычисления, которые G.Rasch провёл с данными

- идея латентной переменной величины θ, представляющая уровень подготовленности;

- функция, описывающая зависимость вероятности правильного ответа испытуемого от уровня подготовленности испытуемого и от меры трудности задания.

Из второго предположения вытекает необходимость найти параметры заданий и параметры испытуемых, а из третьего - найти лучшую форму соотношения параметров.

Функциональная и графическая интерпретация второго и третьего положений – вероятность правильного ответа зависит от расстояния между значениями θ_i и β_j.

$$Pr_{ij} = f(\theta_i - \beta_j) \qquad (1)$$

Эпизодическое применение элементов системы RM для объективации результатов имело место в практике бывшего централизованного тестирования. Компьютерные программы RUMM-2020 и Winsteps используются в работе отдельных авторов и организаций. В целом вопросы RM в трудах русскоязычных авторов представлены мало и эпизодически.

Впервые к исследованию вопросов объективности педагогических измерений автор этой статьи обратился в 1976 г.[31] Затем проблема объективности педагогических измерений была затронута в диссертационном исследовании[32]. Там она рассматривалась как результат и следствие специально организованного научного процесса обоснования тестовых результатов, в котором критерии надёжности и валидности играют важную роль для достижения объективности. Было показано, что много работ западных авторов было посвящено вопросам определения надёжности, заметно меньше - валидности. И почти совсем ускользнул из поля зрения исследователей критерий объективности тестовых результатов.

По мнению членов программного комитета Института Объективных Измерений при Чикагском университете, главное средство преодоления недостаточной объективности педа-

---

[30] Rasch, G. (1960/1980) . Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: University ofChicago Press. (Original work published 1960.). P. 117

[31] Аванесов В.С. Вопросы объективизации оценки результатов обучения. - М.:НИИВШ, Отдел научной информации. 1976.- 66с.

[32] Аванесов В.С. Методологические и теоретические основы тестового педагогического контроля. Дисс….доктора пед. наук. С-Петербургский госуниверситет, 1994г. -337с.

гогических измерений – овладение совокупностью необходимых для этого знаний, а также нужным набором подходов и методов[33].

Несмотря на внушительный состав авторитетных исследователей, считающих проблему объективности практически решённой в рамках теории и технологии G.Rasch, автор данной статьи склоняется к другому мнению. Как ни покажется странным, но к мнению, совпадающему с позицией самого G.Rasch. RM –это методология, теория и технология , ведущая не к той объективности, о которой мыслят в философской науке, а к объективированным результатам измерения. В педагогических измерениях полная объективность недостижима или трудно достижима.

Со временем, вероятно, будут созданы и другие технологии, превосходящие RM, а потому больше приближающие тестологов к идеалу достижения полной объективности. Поэтому было бы полезно ещё раз заметить, что вместо слов «объективные измерения» лучше использовать словосочетание «объективированные измерения», имея в виду признанную в философии неизбежную погрешность любых измерений и расширенную возможность возникновения субъективных элементов в педагогических измерениях.

В научно организованных педагогических измерениях нет актуальнее вопросов поиска подходящей системы идей, теорий, моделей и методов для приближения к решению проблемы объективности результатов. Такого рода системные знания в мировой науке начали появляться. Но реальная потребность в таких знаниях сейчас оттеснена в России на научную периферию. В советский период истории тестов подобное явление уже имело место. Тогда это делалось под предлогом борьбы с буржуазной тестологией. В те же годы имел место и разгром советской генетики. Сейчас мотивы вытеснения иные.

В распоряжении Правительства РФ №910-з объяснялось, что ЕГЭ вводится с целью обеспечения *объективности* и унификации итоговой аттестации и вступительных испытаний в системе профессионального образования[34]. А в статье 1 п. 4.1 новой версии «Закона об образовании» ЕГЭ без видимых оснований назван формой *объективной* оценки качества подготовки лиц, освоивших образовательные программы среднего (полного) общего образования, с использованием заданий стандартизированной формы (контрольных измерительных материалов), выполнение которых позволяет установить уровень освоения ими федерального компонента государственного образовательного стандарта среднего (полного) общего обра-

---

[33] Definition of Objective Measurement. Written by the Program Committee of the Institute for Objective Measurement. December 2000. http://www.rasch.org/define.htm

[34] http://www.r-komitet.ru/obraz/EGE.htm

зования…»[35]. Автор этой статьи уже отмечал, что объективность метода порождается не законом или мнением чиновников, а свойствами метода[36]. Необоснованное мнение об объективности ЕГЭ, поднятое на пьедестал Закона, уже нанесло образованию России существенный урон.

Счёт или измерение?

Как уже отмечалось, с точки зрения сторонников G.Rasch, исходные баллы даже качественных тестов не считаются педагогическими измерениями[37], а КИМы ЕГЭ объективными методами педагогических измерений не являются точно. Такие выводы могут показаться неожиданными или субъективными. Но к этому подводит более строгое истолкование понятия «педагогическое измерение», распространившееся в последние годы, особенно в RM. Вопрос этот - не новый; его история прослеживается со времён публикации работы Гельмгольца «Счёт и измерение». Если судить по способу получения исходного тестового балла испытуемых (Y_i) (табл.1), то, очевидно, здесь мы имеем дело исключительно со счётом – чем больше правильных ответов на задания теста, тем выше индивидуальный балл испытуемого. Это действительно результаты счёта, а не измерения. В RM полученные таким образом результаты счёта подвергаются далее трансформации, посредством процесса шкалирования. После чего данные приобретают свойства значений, полученных на интервальной шкале натуральных логарифмов.

Исходными тестовыми баллами испытуемых называются результаты счёта - элементарные суммы баллов, которые получают испытуемые при ответах на задания теста. Примеры расчёта таких баллов читатель найдёт в столбце Y_i ранее публиковавшейся табл. 1.

Таблица тестовых результатов Табл. 1

№№	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Y_i	p_i	Q_i	p_i/qi	ln p_i/q_i
1.	1	1	1	0	1	1	1	1	1	1	9	.90	.10	9	2.20
2.	1	1	0	1	1	1	1	1	1	0	8	.80	.20	4	1.39
3.	1	1	1	1	0	1	1	0	1	0	7	.70	.30	2.33	.85
4.	1	1	1	1	0	1	0	1	0	0	6	.60	.40	1.50	.40
5.	1	1	1	1	1	1	0	0	0	0	6	.60	.40	1.50	.40
6.	1	1	1	1	0	0	1	0	0	0	5	.50	.50	1.00	0

[35] http://www.akdi.ru/gd/proekt/099384GD.SHTM

[36] Аванесов В.С. Проблема становления системы педагогических измерений. ПИ, №1, 2008. С.18. http://viperson.ru/wind.php?ID=435316&soch=1

[37] Wright B.D. Scores are not Measures. Rasch Measurement Transactions, 1992, 6:1, 208. http://www.rasch.org/rmt/rmt61n.htm

7.	1	1	0	1	1	0	1	0	0	0	5	.50	.50	1.00	0
8.	1	1	1	1	1	0	0	0	0	0	5	.50	.50	1.00	0
9.	1	0	1	0	1	1	0	0	0	0	4	.40	.60	.66	-.42
10.	0	1	1	0	0	0	0	1	0	1	4	.40	.60	.66	-.42
11.	1	1	1	0	0	0	0	0	0	0	3	.30	.70	.43	-.84
12.	1	1	0	0	0	0	0	0	0	0	2	.20	.80	.25	-1.39
13.	1	0	0	0	0	0	0	0	0	0	1	.10	.90	.11	-2.21
R_j	12	11	9	7	6	6	5	4	3	2	65				
W_j	1	2	4	6	7	7	8	9	10	11					
p_j	.923	.846	.692	.538	.462	.462	.385	.308	.231	.154	5				
q_j	.077	.154	.308	.462	.538	.538	.615	.692	.769	.846					
p_jq_j	.071	.130	.213	.248	.248	.248	.236	.213	.178	.130					
q_j/p_j	.083	.182	.445	.859	1.164	1.164	1.597	2.246	3.329	5.493					
$\ln q_j/p_j$	-2.489	-1.704	-.810	-.152	.152	.152	.468	.809	1.202	1.703					

Для новых читателей журнала напомним, что в этой матрице рассчитывают:

p_j - долю правильных ответов испытуемого i, по всем заданиям теста;

q_i - доля неправильных ответов того же испытуемого i, по всем заданиям теста;

p_j / q_i - потенциал подготовленности испытуемого i;

$\ln p_j/q_i$ исходное значение логита уровня подготовленности испытуемых[38]. Это значение в процессе дальнейшего шкалирования корректируется и обозначается далее греческой буквой θ_i.

$\ln q_j/p_j$ исходное значение логита уровня трудности заданий. И это значение затем корректируется и обозначается далее греческой буквой β_j.

В процессе последующего шкалирования результатов средние арифметические шкал уровней подготовленности испытуемых θ_i и уровней трудности заданий β_j приравниваются нулю посредством линейного переноса данных. Эти шкалы выравниваются и по уровню вариации, посредством умножения на соответствующие взаимно корректирующие числа. После чего появляется одна шкала с общей единицей измерения, равной одному логиту.

Единицей измерения являются так называемые логиты. Исходное значение логитов уровня подготовленности испытуемых и логитов трудности заданий вычисляется по эмпирическим результатам тестирования. Следуя традиции, обозначим уровень подготовленности греческой буквой θ_I, где подстрочный индекс i обозначает номер испытуемого. Теоретически

---

[38] 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; 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.; Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. With a Foreword and Afteword by B.D. Wright. The Univ. of Chicago Press. Chicago & London, 1980. -199 pp.

допускается, что θ_i - любое действительное число, но практически большинство значений θ_I находятся в интервале от –5 до + 5, что объясняется свойствами шкалы θ. В этой шкале средняя арифметическая принимается равной нулю. Пример вычислений исходных значений логитов – в последней строке и в последнем столбце таблицы 1.

Математическая модель

Для получения специфически, т.е. математически объективированных результатов измерения G.Rasch предложил использовать формулу:

$$P_j\{X_{ij} = 1 \mid \beta_j\} = \frac{\exp(\theta - \beta_j)}{1 + \exp(\theta - \beta_j)} \qquad (2)$$

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

θ_i - уровень знаний, латентная переменная;

βj - уровень трудности j-го задания теста.

В формуле (2), оба параметра функции – уровень подготовленности испытуемых и уровень трудности заданий связаны между собой операцией вычитания $\theta_i - \beta_j$. Чем больше разность, тем большей становится вероятность правильного ответа испытуемого i на задание j. Как выразился образно G.Rasch, значение разности *управляет* значением вероятности правильного ответа. Эту идею управления вероятностью выразили Wright B.D. & Stone M. D. в своей работе[39] (рис.2).

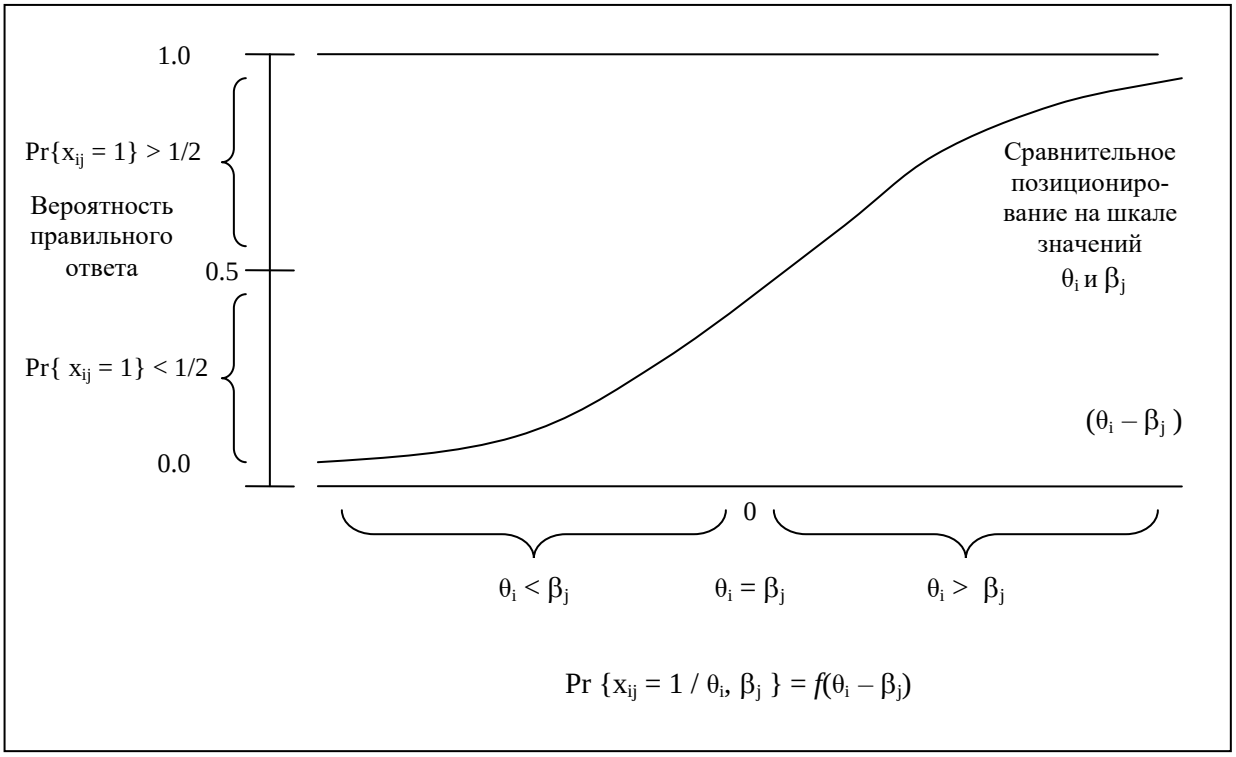
Из свойств функции, представленной на рис.2, можно вывести несколько полезных следствий:

Если $(\theta_i - \beta_j) = 0$, то вероятность правильного ответа испытуемого данного уровня подготовленности должна равняться ½. Формально можно записать так: $Pr_{ij}\{x_{ij} = 1\} = 0{,}5$. Если $(\theta_i - \beta_j) < 0$, то вероятность правильного ответа испытуемого данного уровня подготовленности должна быть меньше ½. $Pr_{ij}\{x_{ij} = 1\} < 0{,}5$. В таком случае, вероятность неправильного ответа возрастает.

Если $(\theta_i - \beta_j) > 0$, то вероятность правильного ответа испытуемого данного уровня подготовленности должна быть больше ½ . $Pr_{ij}\{x_{ij} = 1\} > 0{,}5$.

Рис.2.

[39] Wright, B. D. & Stone, M. H. (1979). Best Test Design. Chicago: Mesa Press.



Данная модель Г.Раша является одной из наиболее известных моделей педагогического измерения. Она применяется для оценки вероятности правильного ответа, для оценки параметров испытуемых и заданий, а также для оценки пригодности заданий для измерения по его модели. Это монотонно возрастающая функция, связывающая уровень подготовленности с вероятностью правильного ответа на тестовое задание. Из всех известных моделей измерения модель G.Rasch самая простая. Она требует информации о значениях только двух параметров модели: уровня подготовленности испытуемого, обозначаемого θ_i и уровня трудности задания, обозначаемого символом β_j. Значениях этих параметров указывают на положение испытуемых и заданий на одной и той же общей числовой оси, называемой латентной переменной; как видно из формулы (1) именно они нужны для определения вероятности успеха испытуемого при ответе испытуемого i на задание j.

Впервые идею представлять уровень подготовленности испытуемых и уровень трудности заданий на одной и той же числовой оси выразил L.R.Tucker[40]. Он же предложил определять меру трудности задания в той точке оси абсцисс, значение функции от которой равна 0,5. А это значение вероятности правильного ответа испытуемого на задание под номером j. Графически эта ситуация представлена на рис.1. Там приведен график зависимости вероятности правильного ответа от уровня подготовленности испытуемых.

[40] Tucker L.R. Maximum validity of a test with equivalent items. Psychometrika, 11 (1), 1-13.

Рис.1.

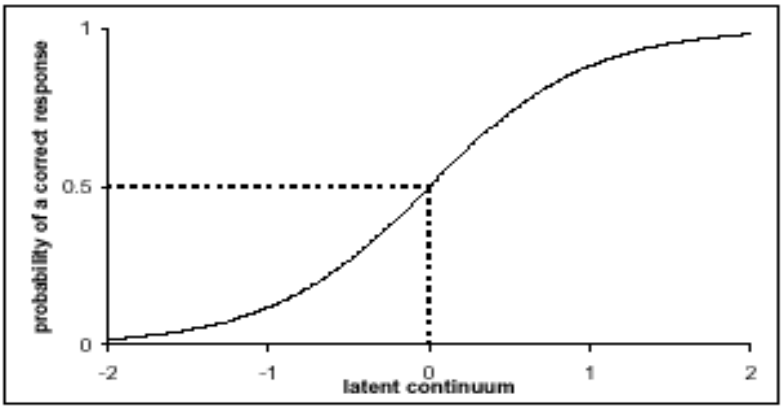


Рисунок полезно рассматривать как графический образ тестового задания, представленный в системе прямоугольных координат на плоскости. Трудное задание - и, соответственно, его график - позиционируется правее: соответственно правее перемещается и точка перегиба функции-задания. Это и есть формальный показатель меры трудности задания. График сравнительно лёгкого задания-функции располагается, естественно, левее. По оси абсцисс откладывается уровень подготовленности испытуемых (в стандартных единицах – логитах), по оси ординат – вероятность правильного ответа испытуемых.

Рассмотренные случаи подводят к мысли о возможности установить соответствие между значениями разности параметров и значениями вероятности правильных ответов: чем больше разность, тем выше вероятность правильного ответа. Именно эта идея составляет сердцевину математической, по сути, теории RM. Умозрительно разность (θ_i - β_j) может принимать любые значения: $-\infty \leq (\theta_i - \beta_j) \leq +\infty$. Однако 99 процентов тестовых результатов укладывается в диапазон $-5 < \theta_i - \beta_j < 5$.

В табл. 2 установлено важное соответствие между значениями разности параметров (3-ий столбец) и вероятностью правильного ответа испытуемого на задание уровня трудности j (5-ый столбец). Для лучшего понимания идеи зависимости вероятности правильного ответа испытуемых от разности двух основных параметров очень полезен числовой пример, подготовленный B.D.Wright[41] (см. табл. 2).

Зависимость вероятности правильного ответа от разности параметров. Табл.2.

Мера подготовленности испытуемого (θ_i)	Мера трудности задания (β)	Разность ($\theta_i - \beta$)	exp ($\theta_i - \beta$)	Вероятность правильного ответа исп. i на задание j	Мера информации

[41] Wright B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement 14 (2) pp. 97-116, Summer 1977. http://www.rasch.org/memo42.htm

5	0	5	148,4	.99	.01
4	0	4	54,6	.98	.02
3	0	3	23,1	.95	.05
2	0	2	7,39	.88	.11
1	0	1	2,72	.73	.20
0	0		1.0	.50	.25
0	1	-1	0,368	.27	.20
0	2	-2	0,135	.12	.11
0	3	-3	0,050	.05	.05
0	4	-4	0,018	.02	.02
0	5	-5	0,007	.01	.01

$$Pr_{ij} = \exp(\theta_i - \beta_l) / [1 + \exp(\theta_i - \beta_j)$$
$$I_{ij} = Pr_{ij}(1 - Pr_{ij})$$

В первом столбце B.Wright взял, для удобства, целые значения $\theta_{i,}$, представляющие пятерых испытуемых с различающейся подготовкой выше среднего уровня; подготовка остальных позиционирована на одном и том же одинаковом среднем уровне. В шкале логитов среднее значение равно нулю. Во втором столбце представлены шесть заданий среднего (нулевого) уровня трудности, а затем - возрастающей трудности. В третьем столбце представлены значения разности ($\theta_i - \beta$). В четвёртом столбце даны значения exp ($\theta_i - \beta$). В пятом столбце вычислены вероятности правильного ответа испытуемых в зависимости от уровня разности значений $\theta_i - \beta$ по модели G.Rasch. Например, испытуемый с подготовкой $\theta_i = 5$ имеет вероятность правильного ответа на задание средней трудности 0,99.

Применение информационной функции для повышения эффективности и объективности измерения

К настоящему времени разработан внушительный арсенал методов, позволяющих повысить объективность педагогических измерений. Это методы оценки соответствия заданий и испытуемых применяемой модели (Fit Analysis), сравнительный анализ двух гистограмм распределения результатов - испытуемых и мер трудности заданий теста (Item-Person Map), компьютерные программы, позволяющие вычислить объективированные значения параметров испытуемых и заданий, объективированное сравнение тестовых результатов (Test Equating), расчёт информационной функции и многие другие методы. Многие из предложенных

методов выходят за пределы проверки качества измерений по двум традиционным критериям - надёжности и валидности тестовых результатов. Но каждый из названных выше методов по-своему связан с объективностью, если объективность понимать как критерий более высокого уровня, вбирающий в своё содержание, по принципу включения, содержание критериев надёжности и валидности. Таким образом, предлагаемая в данной статье система критериев рассматривается как иерархическая, в той части, которая касается надёжности, валидности и объективности. Исследованный ранее автором этой статьи критерий эффективности – другого, прагматического толка.

В нижней части табл. 2 приведены две формулы. Первая из них уже известна: она позволяет вычислить вероятность правильного ответа испытуемых в зависимости от разности параметров уровня подготовленности и трудности заданий. Вторая формула позволяет посчитать т.н. количество информации, получаемой в процессе измерения испытуемого с уровнем подготовленности θ_i посредством заданий с уровнем трудности β_j. Информационная функция помогает повысить объективность через повышение точности и эффективности измерения.

В последнем столбце табл.2. приводятся значения информационной функции I, вычисляемой по формуле I = Pr (1 - Pr) где Pr - вероятность правильного ответа на задание, а 1 – Pr -это вероятность неправильного ответа. Последняя вычисляется из цепочки формул:

$$\Pr \{x_{ij} = 0 | \theta_i, \beta_j\} = 1 - \Pr \{ x_{ij} = 1 | \theta_i, \beta_j \}$$

$$= 1 - \{ \frac{\exp (\theta - \beta_j)}{1 + \exp (\theta - \beta_j)} \} = \qquad (3)$$

Значения информационной функции представлены в последней колонке табл. 2. В математической теории измерений (IRT) и в RM показатель количества получаемой информации является настолько важным показателем, что он оказывает влияние сразу на все базовые критерия оценки педагогических измерений.

Чем ближе значения θ_i и β_j, тем *точнее* оценивается уровень подготовленности данного испытуемого данным заданием.

Чем больше доля заданий теста, имеющих меру трудности β_{jj}, близких к значениям θ_i, тем *эффективнее* измерение испытуемых данного уровня подготовленности.

По значениям последнего столбца табл.2 Wright B.D. и Stone M.H. делают следующие выводы. Ответы испытуемых на задания, при разности $|\theta_i - \beta_j|$ в интервале менее одного ло-

гита информативнее для измерения примерно в два раза, чем ответы испытуемых на задания при разности $|\theta_i - \beta_j|$ в интервале больше двух логитов. И примерно в четыре раза информативнее ответов испытуемых на задания при разности $|\theta_i - \beta_j|$ в интервале больше трёх логитов[42].

Заметное и массовое несовпадение мер трудности заданий и уровней подготовленности испытуемых требует увеличения общего числа заданий для проведения измерений требуемого качества. Отсюда возникает снижение информативности и, кроме того, валидности и объективности педагогического измерения. В таких случаях западные тестологи говорят, что подбор заданий теста не адекватен уровню подготовленности испытуемых.

Пример ошибочного соотношения отобранных заданий и имеющихся испытуемых даёт применение части «С» в ЕГЭ. Там подбор сверх трудных заданий абсолютно неадекватен уровню подготовленности основной массы испытуемых, что приводит к нулевым значениям информационной функции и соответственно, к обнулению качества педагогических измерений. Не удивительно, что значения получаемого там низкого качества не публикуются. Обнародование таких данных означало бы признание краха КИМов ЕГЭ и прекращения финансирования их производства. Но правительство ещё готово к признанию такого рода дефолта.

Компьютерные программы повышения эффективности и объективности педагогических измерений

Становление идеи объективированного педагогического измерения совпало с периодом ускоренного развития компьютерной техники и программирования. Поскольку в процессе обоснования качества тестовых заданий и объективности педагогических измерений в целом используются математико-статистические вычисления, применение компьютеров и компьютерных программ сразу же получило широкое распространение.

Эффективность применения компьютерных программ очевидна, она проистекает из возможностей компьютеров точно обрабатывать огромные массивы информации, и за короткое время. Программы объединяют возможности компьютеров с содержательными и счётными задачами, и помогают решать такие задачи наилучшим образом.

Объективность педагогических измерений компьютерные программы решают посредством общих стандартных методов расчёта и единых решающих правил, применяемых при вычислениях параметров уровня подготовленности испытуемых и параметров заданий.

---

[42]Wright, B. D. & Stone, M. H. (1979). Best Test Design. Chicago: Mesa Press. P. 18.

В настоящее время в России применяются несколько зарубежных программ. Все они спроектированы и написаны для применения в системе Windows 95 и более продвинутых версий. В первую очередь можно отметить две наиболее распространенные программы - RUMM 2020 (прежняя версия RUMM 2010) и WINSTEPS. Обе программы предназначены для работы в интерактивном режиме в процессе разработки теста.

Из-за некоторых затруднений при вводе матриц исходных результатов и при обработке данных, оба пакета предполагают необходимость небольшого периода специального «фирменного» обучения пользователей пакета - от одного до пяти дней, соответственно в Австралии (RUMM 2020) и в США (WINSTEPS). Сроки обучения могут варьировать в зависимости от уровня статистической и компьютерной подготовленности пользователей. Желательны знания основ теории педагогических измерений, хотя бы на уровне одного-трёх недельных циклов занятий, проводимых автором этой статьи для начинающих тестологов[43], нужны также статистическая подготовка и навыки вычислительной работы на компьютере.

Оба пакета способны проводить шкалирование исходных тестовых баллов и мер трудности заданий на уровне требований интервальной шкалы. Пакеты выдают информацию о формальных свойствах каждого задания и теста в целом, об уровне и структуре подготовленности каждого испытуемого, о надёжности и валидности результатов, о качестве педагогических измерений и о степени соответствия результатов испытуемых уровню трудности заданий, а также информацию по другим критериям специфической объективности.

Название программы RUMM составлено из слов Rasch Unidimensional Measurement Model (RUMM). Что можно перевести как математико-статистический пакет для анализа заданий теста требованиям одномерной модели Г.Раша [44]. У этого пакета есть цветная графика, что позволяет легко и сразу дифференцировать свойства каждого варианта ответа на каждое задание[45].

Программа WINSTEPS является одним из вариантов трёх, по меньшей мере, программ, разработанных под руководством J.M.Linacre[46]. Два других варианта программы на-

---

[43] См. объявления на сайте http:testolog.narod.ru

[44] Содержательные свойства заданий и тестов - это предмет педагогической теории измерений. Подробнее см. например: Аванесов В.С. Основы педагогической теории измерений. ПИ, №1, 2004, с.15-22.

[45] Подробности о возможностях этого пакета, учебное пособие для пользователей и пробную демонстрационную версию RUMM, равно как и информацию о возможности приобретения можно найти по адресу http://www.rummlab.com.au/ or http://www.faroc.com.au/~rummlab/

[46] По-русски эта фамилия произносится как «Линека», с ударением на первом слоге.

зываются BIGSTEPS. Бесплатно предоставляемая демонстрационная версия называется MINISTEP[47].

Помимо этих двух упомянутых специализированных пакетов имеются и другие программы, позволяющих обрабатывать исходные результаты педагогических измерений и проводить шкалирование баллов. Среди таковых пакеты QUEST,WINMIRA, CONQUEST и RASCAL.

QUEST[48] представляет собой полезный статистический пакет. Он делает возможным анализ как тестовых заданий, так и вопросов социологических анкет. Обработка данных может проводиться с опорой на две наиболее популярные теории - классическую (статистическую) теорию педагогических измерений и на теорию RM. Использование этих теорий позволяет глубже понять сущность получаемых данных и отчленить те элементы результатов, которые возникают как следствие самой теории и применяемых в ней методов. По сути, здесь намечена и реализована идея достижения результатов, инвариантных (объективированных) по отношению к теории и методу[49]. Это замечательная возможность, открывающая возможности повышения объективности тестовых результатов, инвариантных по отношению к используемым теориям.

Программа WINMIRA практически неизвестна в России. У неё тоже есть интересные свойства гибридного пакета, а также широкий набор методов обработки данных на основе Item Response Theory. Сильной стороной данного пакета[50] является полная интеграция с бесчисленными возможностями другого замечательного и известного в России статистического пакета - SPSS.

CONQUEST интересен своими возможностями проведения не только одномерных, но и многомерных измерений. По мере повышения интереса к проведению многомерных измерений роль этого пакета будет неуклонно возрастать.

RASCAL позволяет шкалировать задания и испытуемых на основе одномерной модели G. Rasch, если данные представлены в дихотомической шкале (1/0). Примечательная особен-

---

[47] Консультацию по использованию и приобретению всех трёх вариантов можно получить по адресу http://www.winsteps.com/index.htm

[48] Разработали R.Adam and K. Siek Toon.

[49] Подробное описание возможностей статпакета интересующиеся читатели найдут по адресу http://www.scienceplus.nl/scienceplus/main/show_pakketten_categorie.jsp?id=38

[50] Информацию об этом статпакете и его демонстрационную версию можно найти по адресу http://www.scienceplus.nl/scienceplus/main/show_pakketten_categorie.jsp?id=38

ность этого пакета – вывод на печать таблицы перевода исходных тестовых баллов в шкалированные значения логитов испытуемых. Для хотя бы частичного устранения нынешнего произвола с баллами испытуемых ЕГЭ при приёме в различные вузы. Такой наглядно обозримый вывод данных был бы полезен для всех участников этого мероприятия. Это способствовало бы также и снижению чрезмерного психологического напряжения испытуемых и их родителей, суицида некоторых молодых людей, неудачно ответивших на ЕГЭ, а также закрыло бы возможности для манипуляций с баллами ЕГЭ.

Системный проект
продвижения к объективным результатам педагогических измерений

В книге B.D.Wright и M.H.Stone «Best Test Design» (BTD) 1979 года фактически впервые была изложена система методов, помогающих продвигаться к объективированным тестовым результатам в рамках теории и технологии RM. В те годы это был новый, нетрадиционный, и не совсем понятный проект для сложившейся к тому времени западной культуры педагогических измерений и для психометрики[51]. Создан был это проект как бы сторонними людьми.

Сам G.Rasch был не тестологом, а профессиональным математиком, вынужденно занявшимся вначале статистикой, а затем и разработкой теста для оценки уровня грамотности новобранцев датской армии. Когда в начале 50-х годов XX века стало известно о математическом решении им проблемы независимости определения параметров испытуемых и заданий, к нему был послан автор книги «Основы психологических измерений» Л.Кронбах[52]. Но последний не сумел разобраться (или не захотел) в сути работ Г.Раша, а потому дал по ним отрицательное заключение.

Несмотря на это, G.Rasch всё же был приглашён в США, но ассоциацией статистиков. Однако после первых же его лекций почти все слушатели разбежались – настолько содержание лекций и манеры лектора оказались непонятными и непривлекательными для собравшейся аудитории, далёкой от проблем педагогических измерений. Последним оставшимся слушателем оказался B.D.Wright, да и то только потому, что дал слово организаторам лекций выслушать всё, и до конца[53]. В итоге B.D.Wright оказался единственным в США, кто понял,

---

[51] Wright, B. D. & Stone, M. H. (1979). Best Test Design. Chicago: Mesa Press.

[52] Cronbach L.J. Essentials of Psychological Measurements. (1990). (5-ed.), New York: Harper Collins

[53] Rasch and Wright: the early years. Wright BD, Andrich DA. Rasch Measurement: *www.rasch.org/rmt/contents.htm*

что в лекциях датского профессора излагается действительно новое и нужное направление педагогических и психологических измерений.

Сам B.D.Wright получил вначале физическое образование, но затем стал искать себя в психологии, психоанализе, и, наконец, в психологических и педагогических измерениях. После ознакомления с идеями и методами G.Rasch он стал ярким пропагандистом нового направления, получившего позже название RM.

Название книги BTD можно перевести на русский язык, так, например, «проектирование лучшего теста». Но это не лучший перевод, а потому представляется, что лучше оставить привычное название BTD.

Во-первых, это название метафорично, а потому любой перевод такого названия обречён на неточность. И действительно, лучшего теста в реальности нет. Вряд ли можно найти такой тест, и обосновать, что он лучший.

Во-вторых, лучший тест трудно найти из-за того, многие тесты фактически не сравнимы между собой. Они имеют разное содержание, разные цели, несравнимые формы заданий, отличающиеся результаты и статистические характеристики.

В-третьих, критерии качества для разных тестов различны, хотя часть критериев может быть общей, например, теоретические показатели надёжности и валидности получаемых результатов. Последние, как известно, зависят не только от качества тестов, но и от цели, уровня подготовленности тестируемого контингента и от множества других условий.

Вслед за G.Rasch, авторы BTD выделили два условия достижения объективности измерений. Первое – шкалированные значения мер трудности заданий (β_j) не должны зависеть от уровня подготовленности испытуемых. Второе – шкалированные значения уровней подготовленности испытуемых (θ_i,) не должны зависеть от уровня трудности используемых в конкретном варианте теста заданий[54].

Позиционирование заданий и испытуемых в тесте

Можно определённо утверждать, что в объективно измеряющем педагогическом тесте задания должны располагаться в порядке возрастающей трудности. Также можно утверждать - чем выше уровень подготовленности испытуемых, тем больше вероятность правильного ответа на задания теста. При упорядочении испытуемых и заданий матрица исходных тестовых баллов принимает, скажем, треугольно подобный вид: в левом верхнем углу концентрируют-

[54] Wright, B. D. & Stone, M. H. (1979). Best Test Design. Chicago: Mesa Press. p. xii

ся единицы, в правом нижнем углу - преимущественно нули. (См. матрицу в табл. 1). Такое расположение не случайно, а статистически закономерно. Оно указывает на тенденцию зависимости результатов тестирования от уровня подготовленности испытуемых и от уровня трудности системы заданий теста.

Каждая строка матрицы образует профиль испытуемого, а каждый столбец - профиль ответов по заданию. Применительно к данным табл. 1 профилем испытуемого называется последовательность единичек и нулей, представленных в каждой строке. В правильном профиле знаний все нули следуют за всеми единицами. В каждом профиле подготовленности испытуемых можно увидеть зону устойчивых знаний, неустойчивых знаний и зону незнания.

При разработке теста полезно опереться также на понятие «подходящая мера трудности». Это означает, что в тесте должны быть задания переменной трудности, и если задания какого-то конкретного уровня трудности в тесте не хватает, то этот дефект сразу же становится заметным. Ухудшается дифференцирующая (различающая) способность теста на том участке шкалы, где не хватает заданий требуемого данного уровня трудности.

Посмотрим на пример позиционирования условных «тестов», с небольшим числом заданий. Напомним, что для качественного и объективного измерения желательно ориентироваться примерно на тридцать заданий теста. При очень качественных заданиях с выбором нескольких правильных ответов, подобранных как содержательная и формальная система, требуемое число заданий при создании теста может сократиться примерно до 20. Но для того, чтобы получить достаточное число качественных заданий для одного проектируемого теста, проверять эмпирически придётся примерно сто заданий.

На рис. 4 представлено пять примеров подбора заданий условного «теста» в BTD[55].

Первый пример – очень лёгкий «тест», с низкой вариацией заданий по уровню трудности. Не случайно все задания располагаются на левой, лёгкой части континуума. У испытуемого выше среднего уровня подготовленности ожидаемый исходный тестовый балл равняется максимуму числа всех имеющихся заданий.

Второй пример – очень трудный «тест», также с низкой вариацией заданий по уровню трудности. Все задания позиционированы на правой, трудной части континуума. У испытуемого выше среднего уровня подготовленности ожидаемый исходный тестовый балл может оказаться равным нулю.

---

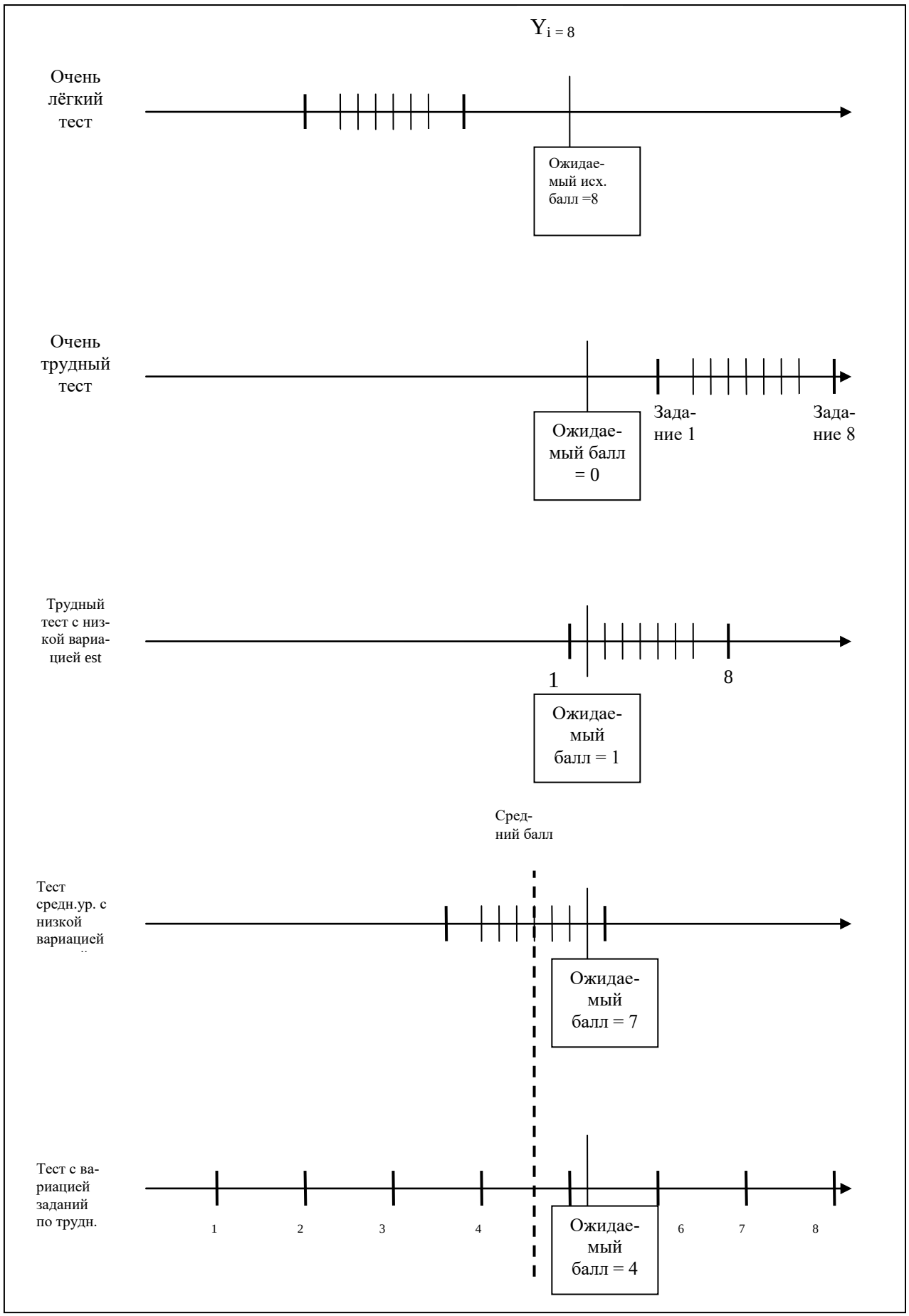
[55] Wright, B. D. & Stone, M. H. (1979). Best Test Design. Chicago: Mesa Press. . Переработано.

Третий пример – трудный «тест», с низкой вариацией заданий по уровню их трудности. У испытуемого выше среднего уровня подготовленности ожидаемый исходный тестовый балл может равняться одному.

Четвёртый пример – «тест» среднего уровня трудности, с низкой вариацией заданий по уровню их трудности. У испытуемого выше среднего уровня подготовленности ожидаемый исходный тестовый балл может равняться семи.

Пятый, последний пример на данной странице – «тест» с достаточной вариацией по мере трудности заданий. Ожидаемый балл испытуемого данного уровня подготовленности может равняться 4.

Трудно найти более подходящий пример для того, чтобы убедиться в том, что исходные баллы любого теста слишком зависимы от субъективных решений, а потому они не могут считаться педагогическими измерениями.


$Y_{i=8}$
Очень лёгкий тест
Ожидае-мый исх. балл =8
Очень трудный тест
Зада-ние 1
Зада-ние 8
Ожидае-мый балл = 0
Трудный тест с низ-кой вариа-цией est
1
8
Ожидае-мый балл = 1
Сред-ний балл
Тест средн.ур. с низкой вариацией
Ожидае-мый балл = 7
Тест с ва-риацией заданий по трудн.
1
2
3
4
6
7
8
Ожидае-мый балл = 4

Показатели объективированных измерений

В RM созданы десятки показателей, необходимые для проведения специфически объективного измерения. Они могут и должны стать предметом специального анализа. В данной статье приведены только несколько из них в качестве иллюстрации.

Первый показатель - мера относительной трудности теста. В старой психометрической литературе этот показатель относился к критерию валидности (пригодности) тестовых результатов для измерения испытуемых требуемого уровня подготовленности[56]. В случае заметного смещения среднего арифметического значения результаты могли быть интерпретированы как невалидные относительно нормативной группы испытуемых, если придерживаться лексики нормативно ориентированной интерпретации тестовых результатов. Напомним, что традиционно мера трудности теста определяется как сумма долей правильных ответов испытуемых на все задания используемого варианта теста ($M = \Sigma p_j$), или как средний арифметический балл исходных тестовых результатов испытуемых, полученных вследствие подсчёта числа правильных ответов на задания теста ($M = \Sigma X/N$). В таблице 1 применение обеих формул даёт, как и следовало ожидать, одно и то же значение среднего арифметического балла M = 5,0. В отличие от статистических исследований, в педагогических и психологических измерениях нередко используется деление не на N-1, а на N.

В RM мера трудности заданий стала теперь соотноситься с мерой подготовленности испытуемых численно, а не только интерпретационно. И если разность между этими двумя мерами превышает один логит, то измерение не соответствует принятому там канону объективности. Эта разность является одним из показателей продвижения к объективности тестовых результатов.

С понятием «мера трудности» связана давняя, и, казалось бы, спорная идея расположения заданий в тесте по мере возрастающей трудности. Автор этой статьи в самом начале свой научной работы познакомился с идеей А.Бине располагать задания теста именно так[57]. И вот около сорока лет эта идея входит в авторское определение теста как системы заданий равномерно возрастающей трудности; системы, позволяющей качественно оценить структуру

---

[56] Так называемая target group.

[57] «…it will be noticed that our tests are well arranged in a real order of increasing difficulty». Binet, A. & Simon, T. (1916). The development of intelligence in children. (Translations of articles in L'Annee Psychologique, 1905, 1908, and 1911). 1905, p. 185.Vineland, NJ: Vineland Training School .

и определить уровень подготовленности испытуемых[58]. За прошедшие годы неоднократно возникали попытки оспорить такое определение теста, вплоть до проведения экспериментальных проверок, как утверждалось в одном из сборников Центра тестирования бывшего министерства образования РФ. Там был сделан вывод, что теоретическое утверждение о тесте, как системе заданий возрастающей трудности противоречит результатам проведённого эксперимента.

В начале нашей статьи уже упоминалось, что практика, а равно и эксперимент, лишь подтверждают или не подтверждают отдельные положения, но не доказывают истинность утверждений теории. При расхождении теории и эксперимента требуется дополнительная проверка и более осторожная интерпретация. Ведь верной может быть теория, а неверной - избранная схема эксперимента. Или наоборот. В таких случаях нужна дополнительная и качественная проверка всей системы научного знания, включая теорию, эксперимент, выводы, что, конечно, не было сделано. В работах L.L.Guttman, B.D.Wright и M.H.Stone неоднократно подтверждалась идея именно упорядоченного расположения заданий педагогического теста, в зависимости от уровня возрастающей трудности. Примеры такого же расположения заданий мы встречаем в ряде работ D.Andrich.

Другой показатель, традиционный – это число заданий используемого варианта теста. Ещё на заре становления классической (статистической) теории педагогических измерений Spearman и Brown опубликовали статью, где показали зависимость надёжности тестовых результатов от числа заданий в тесте. Чем больше заданий, тем выше, при прочих равных, будет и надёжность педагогических измерений. Эту зависимость признана и в RM: большее число хороших заданий позволяют точнее позиционировать испытуемых на континууме, чем это позволяет делать меньшее число заданий.

А это означает, что для объективного измерения нужна достаточная система заданий равномерно возрастающей трудности. В ЕГЭ этого нет, там континуум трудности заданий неоднократно терпит существенный разрыв, и это очередной аргумент для обоснования непригодности КИМов для использования их в качестве метода педагогических измерений.

Среди других показателей можно отметить:

1. Средний уровень подготовленности испытуемых. По тестологическому канону, средний уровень подготовленности испытуемых должен соответствовать среднему уровню

---

[58] Аванесов В.С. Вопросы объективизации оценки результатов обучения. - М.:НИИВШ, Отдел научной информации. 1976.- 66с.

трудности заданий. Верно и обратное утверждение: средний уровень трудности заданий должен соответствовать среднему уровню подготовленности испытуемых. Как уже отмечалось, только при этом условии удаётся обеспечить максимум возможной информации. Чем больше несовпадение этих уровней, тем ниже качество педагогического измерения. Именно такой случай имеет место в результатах ЕГЭ.

Три уровни понимания проблемы объективности педагогических измерений

Можно выделить три уровня понимания проблемы.

Первый уровень имеет своей основой самое распространённое, обыденное мышление. Выставляемые при этом оценки упрощены до предела, а потому всем понятны. Примеры дихотомического оценивания: «зачёт-незачёт», аттестован - не аттестован». Иногда такие решения сопровождаются применением чисел, среди которых наиболее употребляемые 1 или О. Один балл ставится за правильный ответ на каждое задание, ноль баллов - за неправильный ответ. После чего используется удобное решающее правило. Например, если испытуемый набрал хотя бы четвёртую часть возможных баллов, считать его аттестованным. Здесь применяется принцип минимальной компетентности и подсчёт числа правильных ответов. Опять заметим, что это случай счёта, а не измерения. Счёт и сам по себе факт применения чисел не гарантирует проведения объективного педагогического измерения.

Любой педагог видит себя способным оценить уровень подготовленности учащихся своего класса посредством выставления привычной оценки. Совпадение значений у большого числа испытуемых является обычно признаком недостаточной точности измерений. Чем больше совпадающих значений, тем менее точно проводится измерение. Так, например, бывает при выставлении пятибалльной оценки.

Много совпадающих значений является также верным признаком не столько измерения, сколько счёта и педагогического оценивания. Как, например, в случае, когда половина учащихся класса получает по контрольной работе оценку «три». Мы не вправе считать, что оценочное мышление является ошибочным или некачественным, хотя бы потому, что лучше педагога оценить подготовку на данном уровне мышления и деятельности не сможет никто.

Второй уровень понимания проблемы предполагает знакомство с теориями педагогических измерений, с методикой разработки педагогических тестов. Требуется также понимание различий между педагогическими оценками и баллами педагогических тестов, умение применять статистические методы для оценки качества тестовых результатов. Преимущество тестовых баллов по сравнению с оценками педагогического наблюдения не вызывает сомне-

ния. До появления RM было принято исходные тестовые баллы испытуемых считать педагогическими измерениями. Теперь ситуация изменилась, требования усложнились, а потому исходные тестовые баллы уже больше не считаются результатами педагогических измерений.

Третий уровень понимания проблемы объективности педагогических измерений предполагает овладение теорией и методикой шкалирования исходных тестовых баллов. Шкалирование – это процесса трансформации результатов счёта в систему стандартных единиц измерения с общими средними и стандартными отклонениями на интервальной шкале. На такой трансформированной шкале становятся возможными линеаризация данных и допустимыми числовые операции, присущие интервальным шкалам. Там нет абсолютного нуля, но сохраняются отношения – на сколько больше или меньше стандартизованных баллов один испытуемых подготовлен лучше другого.

Измерение свойства испытуемых производится в предположении, что у каждого из них есть интересующее свойство, в каком-то количестве. Если выясняется, что у кого-то нет данного свойства, то это даёт основания для исключения данного испытуемого из предполагаемой выборки лиц, обладающих данным свойством. Считается, что такой испытуемый не входит в целевую группу испытуемых, для которых разрабатывается тест. Соответствие уровня трудности заданий теста уровню подготовленности целевой группы испытуемых - важно условие достижения объективности педагогического измерения. Отсутствие такого соответствия обрекает измерение на ненадёжность, невалидность и на необъективность результатов.

ЕГЭ – источник необъективных оценок

С появлением ЕГЭ часть образовательных ресурсов страны была брошена на создание т.н. КИМов ЕГЭ, научные свойства которых до сих пор не известны. Неофициальные источники информации, теоретическая немота, закрытость реальных результатов и массовые апелляции выпускников школ, ряд приведённых здесь аргументов прямо и косвенным образом свидетельствуют о непригодности и необъективности получаемых в ЕГЭ результатов, для тех амбивалентных целей и задач, которые сформулированы в официальных документах[59]. Известно, что погрешности оценивания в ЕГЭ превышают все мыслимые границы, а вопрос

[59] Аванесов В.С. Единый государственный экзамен в фокусе научного исследования. ПИ № 1, 2006.

объективности получаемых оценок там никогда и не ставился. Неизбежным следствием такого состояния таково: в настоящий момент основным источником массового появления необъективных оценок в России стал Единый Государственный Экзамен (ЕГЭ).

Казалось бы, проведение масштабного государственного экзамена должно было опираться на лучшие научные идеи и системы педагогических измерений. Но уже не раз приходилось писать[60], что в ЕГЭ нет ни тестов, ни педагогических измерений – в том смысле, как это принято в науке. В ЕГЭ есть лишь часть заданий, внешне похожих на тестовые, но которые по существу не тестовые.

Если прибегнуть к метафоре, то можно сказать, что исполнители ЕГЭ пытались создать ружьё, которое по просьбе заказчика должно было одной пулей попадать в несколько целей.

Первая цель – аттестация выпускников школ. Для оценки минимума знаний аттестуемых сейчас выпускников школ имеющаяся конструкция ЕГЭ избыточна, нетехнологична и расточительна. Аттестация стала очень дорогой[61] и абсолютно некачественной. Для нормальной аттестации хватило бы десятки три заданий в тестовой форме с выбором нескольких правильных ответов, что обеспечит радикальное снижение числа угадываемых правильных ответов, полную технологичность процесса аттестации и необходимый уровень качества и объективности.

Вторая цель – приём в вузы. Это требует заданий более высокого уровня трудности и более разнообразных тестов, с учётом профиля вузов и уровня требований к абитуриентам. Как уже отмечалось ранее, это проблема профессионального отбора, и она одним методом и одним уровнем трудности на все профессии не решается. Нужны кадры, обученные теории и методике профессионального отбора и нужны тесты, созданные для отбора на каждую профессию. Игнорирование этой рекомендации ведёт к несправедливости приёма в вузы и к профанации самих КИМов.

Третья цель ЕГЭ – это диагностика и учёт наиболее одарённой части выпускников школ. Специально для этой цели в ЕГЭ придумана часть «С». Материалы части «С» - одно-

---

[60] Аванесов В.С. Доживёт ли Единый Государственный Экзамен до 2009 года? http://www.prostranstvo.ru/news/news/show/1174318993.htm

[61] Точная стоимость проведения аттестации выпускников школ в Министерстве образования и науки не подсчитывается, а потому остаётся неизвестной. Озвучены лишь общие траты на проведение ЕГЭ, но при этом нет обнародованных финансовых отчётов по расходам. Образцовые отчёты министерства образования были лишь в царской России. Регресс в качестве управления и финансирования за прошедший век оказался разительным.

значно не тестовые, не технологичные, некачественные, не эффективные и не объективные. Формулировки допускают неоднозначные толкования, что само по себе неплохо с точки зрения преодоления распространённого догматизма в системе образования. Но при попытках объективированного педагогического измерения такая установка открывает дорогу фактически неконтролируемому субъективизму.

С точки зрения теории и технологии RM задания части «С» Единого государственного экзамена имеют запредельную меру трудности, что обнуляет метрическую полезность ЕГЭ К научно проводимым педагогическим измерениям часть «С» не имеет позитивного отношения. Погоня сразу за несколькими целями породила ошибочную конструкцию ЕГЭ. Таким образом, главные недостатки ЕГЭ – несовместимые цели и связанная с ними порочная конструкция. И это не единственная ахиллесова пята, делающая ЕГЭ некачественным.

Не случайно эта часть оказалась подверженной не просто искажениям, но и прямым фальсификациям. Повышенная фальсифицируемость именно этой части экзамена благоразумно удерживает Министерство образования и науки РФ от публикации рейтинга регионов по уровню подготовленности выпускников школ. Регионы- «передовики» частично известны, но не из официальной информации, а из средств массовой информации, в том числе, зарубежных. Статистика фальсификаций оценок за все годы проведения ЕГЭ тоже оказалась засекреченной федеральным органом управления образованием. Таковы традиционные издержки ведомственной образовательной политики, давно утвердившейся в России. И это вместо общественно-государственной образовательной политики[62], которая проводится в странах, имеющих качественное образование.

В ответ на критику сторонники ЕГЭ иногда отвечают, что там применяются не тесты, а так называемые контрольно-измерительные материалы (КИМы). И что к таковым материалам критерии оценки результатов педагогических измерений не всегда или вообще не применимы. Действительно, КИМЫ – явление, не известное педагогической науке, как неизвестны и критерии их качества. Итогом является отсутствие информации о погрешности баллов ЕГЭ, на основе которых принимаются решения о судьбах молодых людей. Массовые апелляции[63] испытуемых и их родителей, фактическая закрытая для граждан статистика ка-

---

Аванесов ВС. Приоритетный Национальный проект «Образование» как форма перехода к общественно государственному управлению образовательной сфере. http://viperson.ru/wind.php?ID=443355&soch=1

[63] По сообщениям прессы в 2008 году в ЕГЭ участвовало 1 млн. 96 тысяч человек. Только за один прошедший год год число россиян, не одобряющих ЕГЭ, выросло в 1,5 раза. 36% россиян неодобрительно относятся к введению в школах экзаменов в форме единого го-

чества результатов порождают в социуме сомнения в пользе проведения ЕГЭ. История повторяется – после введения ЕГЭ в России вновь появились противники тестов. Но тесты здесь абсолютно не при чём. Парадокс заключается в том, что одна из целей введения ЕГЭ было повышение *объективности* итоговой аттестации выпускников общеобразовательных учреждений [64].

Но в силу ошибочности конструкции нынешний ЕГЭ объективность аттестации в сущности не повысилась, а снизилась. Например, автором этой статьи уже не раз отмечалось, что качественное педагогическое измерение возникает в результате соответствия уровня трудности заданий уровню подготовленности испытуемых. Если бы авторы КИМов точно следовали бы исходным установкам на оценку испытуемых всех трёх целевых групп, то это должно было порождать мультимодальность в распределении результатов. Но, например, КИМ по математике, судя по косвенной информации, даёт иное распределение. Он оказался нацелен, по сути, на приём в вузы, т.е. на оценку наиболее подготовленных абитуриентов. Но именно из-за этого применять такой КИМ для аттестации выпускников массовых школ недопустимо. Эта пример рукотворной необъективности по отношению к слабым выпускникам, порождённой игнорированием основ теории педагогических измерений. Их знания измеряются некачественно.

Вследствие этого, чрезмерно высокий процент полученных двоек правильнее было бы воспринимать, скорее всего, как артефакт, порожденный односторонней целевой установкой авторов КИМов, принятой в реально противоречивой ситуации. Если бы они нацелились на аттестацию, то стало бы много высоких оценок. Тем самым, завалился бы тогда приём в вузы. Сейчас же в жертву необъективному методу принесена аттестация слабых выпускников школ. Они все получали в аттестат тройки, несмотря на то, что в действительности более 30 процентов выпускников получали на ЕГЭ двойки. Если умножить этот поразительный результат на восемь, по числу лет эксперимента, то факт масштабного засорения кадрового по-

---

сударственного экзамена (ЕГЭ), одобрительно же воспринимают его лишь 18%. К таким результатам привел опрос, проведенный фондом «Общественное мнение» в 100 населенных пунктах. Год назад распределение ответов было принципиально иным: 28% одобрительных и 23% неодобрительных суждений. Если бы социологи изучали мнение наиболее заинтересованной группы населения – экзаменуемых государством детей и их родителей, то статистика могла бы оказаться ещё более грустной. См. статью «Апелляционный бум». http://www.gzt.ru/education/2008/06/16/234500.html

[64] Решение коллегии Министерства образования и науки Российской Федерации «Об итогах проведения эксперимента по введению единого государственного экзамена в 2004 году и задачах эксперимента на 2005 год». http://ege.spb.ru/index.php?page=news003 23.03.2006 г.

тенциала России очевиден. Это и есть главный отрицательный результат «эксперимента» по ЕГЭ.

Программы RUMM 2020, WINSTEPS и другие выводят на печать две сопряжённые на одной оси гистограммы распределения уровней подготовленности испытуемых и мер трудности заданий, в соответствующих логитах (Item-person map). Чем ближе средние значения гистограмм одна к другой, тем меньше стандартная ошибка измерения, тем качественнее проводится измерение. Так распределены данные краткого примера данных «теста» в табл. 1, что является хорошим показателем качества и эффективности измерений.

Иное распределение баллов в КИМах ЕГЭ. Там часть «С» расположена сильно правее, что указывает на полное несоответствие (непригодность, неэффективность и необъективность) этой части для измерения уровня подготовленности основной массы испытуемых. А производимые при этом значительные денежные затраты являются чистой потерей образовательного бюджета.

На этом отрицательном фоне неожиданно и странно прозвучало заявление[65] о наличии в КИМах надёжности. В заявлении сказано, что по сравнению с 2005 г. повысилось качество КИМ. Средняя надёжность (коэффициент альфа, Кронбах) КИМ по всем предметам находится в пределах от 0,85 до 0,93. Это надо понимать как непрофессиональную информацию о средней температуре по больнице. Ещё с начала прошлого века (достаточно напомнить о работах Ч.Спирмана и последовавших за этим сотен тысяч других исследований) в литературе принято надежность не декларировать, а обосновывать конкретными результатами, аргументами научными методами. А этого как не было все годы т.н. «эксперимента ЕГЭ», так нет и сейчас.

Вот почему отмеченное заявление о высокой надёжности КИМов ЕГЭ, при полном отсутствии качественных научных отчётов и доказательств, воспринимается с обоснованным недоверием. Интересно было бы услышать – а в 2008 году, по сравнению с 2006 годом, надёжность увеличилась или уменьшилась? Вот вопрос, который следовало бы задать разработчикам КИМов. Нужны не заявления, а аргументированные отчётные данные. Но их не было все восемь лет проведения ЕГЭ. Остаётся не разгаданной тайной - почему Правительство РФ не требует и не публикует нормальный научный отчёт о действительных результатах ЕГЭ?

Теоретические вопросы достижения объективности педагогических измерений

[65] Ковалёва Г.В. Результаты ЕГЭ в 2006 году //Школьные Технологии. №6 2006г., С.146.

Становится понятно, что для достижения объективности педагогических измерений теперь уже недостаточно опираться только на одну теорию, метод или технологию. Применение нескольких теорий и сопутствующих им методов позволит получить результаты измерения, инвариантные относительно одной теории. А это может стать важной ступенькой в попытках достижения объективности получаемых результатов.

Можно выделить такие основные теории:

1. Классическая (статистическая) теория. Мнения о её недостатках или бесполезности для нынешней практики педагогических измерений оказались сильно преувеличенными. Хотя возможности этой теории в свете новых подходов оказались заметно ограниченными, ряд используемых там методов разработки тестов по-прежнему полезны. Особенно это касается методов обоснования надёжности и валидности результатов измерения. Критерии надёжности и валидности разрабатывались в рамках классической теории, важность этих критериев никто отрицать не сможет. Следовательно, и классическая теория будет применяться ещё очень долго, вероятно, вместе с другими теориями.

Например, в RM вместо одного классического коэффициента надёжности считается коэффициент надёжности заданий и коэффициент надёжности результатов испытуемых. Первый указывает на меру воспроизводимости порядка расположения заданий теста, второй – меру воспроизводимости порядка расположения результатов испытуемых. Эти два коэффициенты дают более объективную информацию о надёжности результатов, чем один прежний коэффициент.

2. Item Sampling Theory. В России практически неизвестная. Теория сфокусирована вокруг идеи случайного выбора заданий теста из хорошо определенной генеральной совокупности заданий. Что вообще говоря спорно. Создаваемый таким образом тест во всех своих вариантах вряд ли будет иметь сопоставимые результаты. Гораздо лучше идти по пути создания стратифицированной выборки заданий. В Item Sampling Theory не делается никаких предположений, касающихся процесса взаимодействия заданий и испытуемых.

3. Item Response Theory. Эта теория эффективна для проверки качества заданий и теста в целом. О ней уже опубликован ряд статей на русском языке, в том числе в нашем журнале[66]. Теория изучает процесс взаимодействия между испытуемыми разного уровня подготовленности и заданиями разного уровня трудности.

[66] Аванесов В.С. Item Response Theory. 2007. ПИ, №3 и №4.

4. Теория Generalisability . См. публикацию статьи Hoi K. Suen в нашем журнале[67]

5. Rasch Measurement (RM). В США некоторое время смотрели на неё как на однопараметрический вариант IRT. Но эта точка зрения себя уже исчерпала. С годами становились более понятными своеобразные и уникальные возможности RM для создания тестов с повышенным ресурсом объективности.

6. В качестве шестой по счёту, но не по значимости, автор предлагает рассмотреть возможности педагогической теории измерений, получившей оформленный вид и развитие на страницах нашего журнала[68] и в других публикациях автора этой статьи. Вклад этой теории в решение проблемы объективности может оказаться не малым. Это решение вопросов формы и содержания тестовых заданий, разработка понятийного педагогического аппарата, полезного для всех ранее перечисленных формальных теорий.

Каждая из упомянутых здесь теорий имеет свои варианты, поэтому существующий сейчас теоретический мир шире того, который затронут здесь. И он тоже заслуживает отдельной публикации.

Выводы

1. Объективное педагогическое измерение – идеал, к которому стремятся участники процесса педагогических измерений. Фактически получаются измерения разного уровня приближения к объективности - от нулевого до наиболее объективированного RM. Причины такого разнообразия – традиции монотеоретичности и монометодичности, недостаток исследований по проблеме объективности, отсутствие более широкой, чем RM, теории объективации результатов, недостаточная разработка методологии педагогических измерений - системы методов достижения объективированных результатов, независимых от теорий, методов, выборок испытуемых и от выборок заданий.

2. Специфически объективное измерение предлагается называть объективированным на основе теории и технологии G.Rasch. Процесс объективации результатов в RM осуществляется совокупностью разработанных там теорий, методов и технологий. Более развитой системы объективированных измерений сейчас, надо думать, нет.

3. В повышении объективности ключевую роль призвана сыграть педагогическая теория измерений. Очевидно, что без правильных, логически безупречных форм и педагогически обоснованного содержания заданий объективных тестовых результатов не бывает.

---

[67] Хои К.Suen, Пуи Ва Лей. Методологический анализ теорий педагогических измерений. ПИ №1, 2007. С. 3-20.

[68] Аванесов В.С. Основы педагогической теории измерений. ПИ №1, 2004г. С. 15-21.

4. Необходимо расширить число критериев оценки результатов педагогических измерений с традиционно известных двух - надёжности и валидности – до трёх, за счёт добавления критерия объективности. Это предложение позволяет полнее оценить получаемые результаты, Критерий объективности и связанные с ним технологии не отменяют традиционные критерии, а лишь дополняет и расширяет аргументационную базу достижения качества педагогических измерений.

5. КИМы ЕГЭ не являются педагогическими измерениями. Они не в состоянии обеспечить ни надёжность, ни валидность, ни эффективность и ни объективность получаемых там оценок. Восемь лет дорогостоящего «эксперимента» - достаточный срок для того, чтобы убедиться в его непригодности для решения реальных образовательных проблем. Вместо ЕГЭ надо организовать общественно-профессиональные системы итоговой аттестации выпускников школ и вузов, а также системы профессионального приёма в вузы.