КРИТЕРИИ КАЧЕСТВА ПЕДАГОГИЧЕСКИХ ИЗМЕРЕНИЙ

Вадим Аванесов
testolog@mail.ru

Каждая из наиболее известных основных теорий педагогических измерений — классическая (статистическая), математическая (Item response Theory, IRT) и Rasch Measurement (RM) — имеют свои критерии качества. В основном это критерии качества заданий, тестов и результатов испытуемых. Однако вне рассмотрения нередко оказываются такие вопросы качества, как приемлемость концепции измерения, обоснованность цели тестирования, выбор подходящего контингента испытуемых для разрабатываемого теста и отбор заданий подходящего уровня трудности для имеющейся совокупности испытуемых. В классической теории все эти вопросы относились к критерию валидности.

Измерения по модели Г. Раша (Rasch Measurement, RM) опираются на похожий, но другой критерий, называемый по-английски «Fit». Что означает соответствие. Авторы статей редко уточняют — соответствие чего и чему, каким требованиям, какому критерию? В RM чаще всего это соответствие заданий, тестов и результатов требованиям вероятностной модели Раша.

В статье даётся анализ критериев качества, используемых в трёх упомянутых теориях, а также в педагогической теории измерений.

Введение

В одной из своих работ Бенджамен Райт написал, что прогресс в науке возникает по мере создания простых методов решения сложных проблем[1]. В практике тестирования эта идея известного классика воспринимается двояко — буквально или диалектически.

При буквальной интерпретации метод кажется простым, если он позволяет быстро получить искомые результаты посредством

1 *Wright B.D.* (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97–116, p. 97.

кнопочного исполнения, с опорой на технологии и компьютерные программы. Простой метод понятен для большинства пользователей, но из этого свойства совсем не вытекает его качество. Например, имеющаяся сейчас практика создания так называемых КИМов ЕГЭ основана на простом подсчёте исходных баллов испытуемых. Вытекает ли из такого счёта какой-либо сущностный признак педагогических измерений? Нет, не вытекает.

Другое, диалектическое восприятие идеи Б. Райта основано на понимании, что сами идеи появляются в процессе развития теории, на которой далее основывается тот или иной метод измерения. Новые идеи развивают язык, формы и содержание методов педагогических измерений. Так появилась педагогическая теория измерений[2].

Общая история создания качественных методов показывает, что улучшение происходит обычно на основе усложняющейся, нередко, междисциплинарной теории, а также на совокупности используемых методов и технологий. То есть на пути, противоположном мысли упомянутого классика.

Анализ некоторых работ по RM на русском языке, печатаемых вне журнала «Педагогические Измерения», показывает на превалирование схематического подхода в отношении к вопросам обоснования критериев качества педагогических измерений. У большинства пользователей компьютерных программ обычный ход мысли сводится преимущественно к применению статистического критерия хи-квадрат, используемого для обоснования соответствия всего, что только можно проверить с помощью этого критерия.

Такого рода простой подход в RM не способствует утверждению и применению в практике другого подхода, основанного на педагогической теории педагогических измерений. Наблюдаемая сейчас недооценка этой теории не может быть долгой, по историческим меркам, и продуктивной для практики образовательной деятельности.

Все теории измерений накопили достаточно большой опыт обоснования качества тестовых результатов испытуемых, качества тестов и качества тестовых заданий. Среди критериев качества можно выделить: соответствие цели разработки теста той или иной образовательной политике, цели разработки теста — используемому контингенту испытуемых, отбору содержания каждого задания — цели теста в целом, уровня трудности заданий — уровню подготовленности каждого испытуемого.

Иным должен быть подход к исследованию качества т.н. КИМов ЕГЭ. Здесь для начала надо сделать доступными для анализа матрицы исходных тестовых результатов. Если бы качество т.н. КИМов ЕГЭ проверили на несоответствие уровню подготовленности испытуемых при аттестации и при приёме в вузы, то произ-

2

Аванесов В.С. Основы педагогической теории измерений / Педагогические Измерения, № 1. 2004. С.15–21.

водство контрольных материалов в России можно было бы прекратить. Названные материалы не содержат в себе ничего метрического, они не годятся ни для одной, ни для другой цели. Именно потому результаты засекречены. Этот вопрос уже рассматривался[3].

Цели педагогических измерений

Одним из главных вопросов является определение цели. Соответствие результатов цели — один из главных критериев качества педагогического измерения.

Цели педагогических измерений могут быть метрическими, педагогическими, психологическими и социально-политическими.

Пример *метрической* цели — построить шкалу уровня подготовленности испытуемых или шкалу уровня трудности заданий. Как это нередко бывает, далее возникает цепь уже неметрических вопросов — а для чего нужны эти шкалы, чем они полезны личности, науке, не приводят ли они к чрезмерному вторжению государства в жизнь каждого гражданина и т.п. На этом примере мы видим, что цели взаимосвязаны.

Пример *педагогической* цели — провести классификацию испытуемых по уровню их подготовленности, для комплектования уровневых классов в школах. Далее возникают вопросы о справедливости такой деятельности, не нарушает ли это принцип равенства граждан...

Важен и другой вопрос — для какой цели создаётся педагогический тест? Возможный ответ метрического толка — тест нужен для расположения испытуемых на шкале тестовых результатов, для проведения рейтинга, для получения объективной информации о состояния качества образования и для его улучшения. В наше время исходные баллы результатов тестирования требуется трансформировать в измерения уровня подготовленности испытуемых и уровня трудности заданий. Исследование свойств и качества тестовых заданий — это как бы другая сторона педагогических измерений[4].

Можно отметить различия между понятиями «тестовый балл испытуемого» и «измерение уровня подготовленности испытуемого», а также различия между понятиями «тестирование» и «педагогическое измерение». Как было уже отмечено, тестовые баллы — это результаты тестирования испытуемых, из которых могут получиться, а могут и не получиться результаты педагогических измерений[5].

Психологические цели педагогических измерений нередко формулируются в терминах оценки способностей личности учащихся, определения и стимулирования учебной мотивации, выявления ценностных ориентаций, использования тестовых результатов для выявле-

3

Аванесов В.С. Ошибочные цели — плачевные результаты. (Второй, расширенный вариант статьи). http://viperson.ru/wind.php?ID=632429&soch=1

4

de Boeck Paul, Wilson Mark. Explanatory item response model. A generalized linear and nonlinear approach. 2004. Springer-Verlag N-Y, LLC. P.13.

5

Wright B.D. and J M. Linacre Observations are Always Ordinal; Measurements, however, Must be Interval. *Archives of Physial Medicine and Rehabilitation* 70 (12) pp. 857–860, November 1989. http://www.rasch.org/memo44.htm

ния психологических затруднений в учебном процессе. Полезно при этом задаться вопросом о главных причинах возникновения учебных затруднений, об организации помощи учащимся, о формировании подходящей учебно-технологической среды...

И наконец, пример *социально-политической цели* — применение тестов для обеспечения равного доступа учащихся к качественному образованию, для противодействия коррупции. Здесь возникают совсем нетривиальные вопросы причин возникновения неравенства доступа граждан к качественному образованию, причин появившегося сверхизбыточного экономического неравенства в России, причём в масштабах, больших, чем в других странах мира.

А далее вопросы — насколько реальны возможности коррекции экономического неравенства посредством введения т.н. единого государственного экзамена, почему вообще вместо технологичных тестов в России применяются нетехнологичные самоделки вроде ЕГЭ?

Помимо формулирования целей, в концепции педагогических измерений явным образом указывается целевое свойство личности, измеряемое тестом и целевая группа испытуемых, для которой тест разрабатывается. В педагогических измерениях таковым свойством обычно является уровень подготовленности испытуемых.

Педагогическая переменная величина

С точки зрения математической теории педагогических измерений (IRT) и RM, идея измерения предполагает создание переменной величины. Графическим образом переменной величины является прямая линия. Результат измерения интересующего свойства выражается точкой на этой линии, представляющей меру интересующего свойства.

Педагогическая величина начинается с общей идеи измеряемого свойства. Затем следует подготовка заданий для выявления у испытуемых признаков интересующего свойства. Далее нужны основания, позволяющие считать, что переменная величина реализуется данным содержанием теста.

Возможность появления самого понятия «педагогическая переменная величина» связана с расширением смысла известного математического понятия «величина». Начало переменной величине даёт идея относительного уровня развития свойства, которое надо измерить у испытуемых. Эта же идея положена в основу содержания заданий.

Переменную величину образует система заданий равномерно возрастающей трудности, общего содержания[6]. Положение испытуемых на этой переменной величине определяется их ответами на систему заданий.

Наличие системы шкалированных заданий возрастающей

6

All items must be about the same thing, but then be as different as possible! http://winsteps.com/winman/advice/htm

трудности, имеющих содержание, адекватное названию измеряемой переменной величине, и форму, соответствующую требованиям к тестовым и образовательным технологиям, является важным условием педагогического измерения[7].

Термин «шкалированное задание» здесь означает тестовое задание с известной мерой трудности, выраженной на интервальной шкале.

Педагогическая теория измерений

Четвёртая по счёту теория педагогических измерений была кратко представлена в самом первом номере журнала «Педагогических Измерений»[8]. От классической (статистической) теории эта теория отличается педагогическим понятийным аппаратом, детальной разработкой проблем формы и содержания тестовых заданий. Отмеченная триада проблем образовала главный предмет этой теории.

Другими важными предметами педагогической теории измерений являются разработка и применение заданий в тестовой форме, тестовых заданий и тестов. Особенно высока роль заданий в тестовой форме для активизации учебного процесса, научной организации самообразования и самоконтроля.

Показано, что в педагогической теории измерений проблема обоснования качества результатов лежит преимущественно в плоскости, не двух, как это было прежде в статистической теории, а четырёх основных критериев — надёжности, валидности[9], объективности[10] и эффективности тестовых результатов[11].

В этой связи полезно дифференцировать основные понятия теории педагогических измерений от понятий, сложившихся в других теориях.

Тестирование — это эмпирический метод применения теста для сбора информации, педагогические измерения — это наука о разработке качественных тестов.

Важен вопрос — зачем нужен тест? Краткий ответ: тест нужен для расположения испытуемых на исходной шкале тестовых результатов. Разработка теста требует наличия метода трансформации результатов тестирования в измерения уровня подготовленности испытуемых.

Это утверждение основано на различиях между понятиями «тестовый балл испытуемого» и «измерение уровня подготовленности испытуемого», а также на различиях между понятиями «тестирование» и «педагогическое измерение». Как было уже отмечено, исходные тестовые баллы — это результаты тестирования испытуемых, а не результаты педагогических измерений[12]. Последними исходные баллы становятся после процесса шкалирования.

Измерение можно определить так же как процесс позиционирования (локализации) испытуемых на непрерывной числовой оси, в соответствии

Апрель

7

These calibrated items are operational definition of what the variable measures. *Wright, D. N. and M.H. Stone (1979). Best Test Design.* P. 3.

8

Аванесов В.С. Основы педагогической теории измерений / Педагогические Измерения. № 1, 2004. С. 15–21.

9

Аванесов В.С. Проблема качества педагогических измерений//Педагогические Измерения. № 2. 2004 г. С. 3–27.

10

Аванесов В.С. Проблема объективности педагогических измерений// Педагогические Измерения. № 3, 2008. С. 3–40.

11

Аванесов В.С. Проблема эффективности педагогических измерений. Педагогические Измерения. № 4, 2008. С. 3–24.

12

Wright B.D. and J M. Linacre Observations are Always Ordinal; Measurements, however, Must be Interval *Archives of Physical Medicine and Rehabilitation* 70 (12) pp. 857–860, November 1989. http://www.rasch.org/memo44.htm

с полученными ими баллами на интервальной шкале. Числовая ось представляется в виде горизонтальной прямой линии, отражающей различные уровни развития интересующего свойства личности. Эта ось позволяет представить интересующее свойство личности в виде числа. Само свойство называется педагогической переменной латентной величиной[13].

Измерения можно ещё определить как процесс перехода от интересующего свойства личности к переменной величине, являющейся операциональным представлением данного свойства.

Педагогическое измерение проводится посредством теста, образуемого системой заданий равномерно возрастающей трудности, отображающих единое содержание учебного курса. Отобранные для теста задания позволяют измерить уровень и структуру подготовленности испытуемых.

Для позиционирования испытуемого на шкале подготовленности его надо тестировать посредством заданий, определяющих, своим содержанием, переменную величину «уровень подготовленности испытуемых». А затем убедиться в том, что полученные тестовые результаты помогают определить место испытуемого на этой величине.

Для этого задания в тестовой форме приходится проверять на соответствие требованиям к тестовым заданиям[14], формировать тест как систему заданий равномерно возрастающей трудности, адекватного содержания, получить тестовые баллы испытуемых и затем трансформировать тестовые баллы испытуемых в измерения со свойствами интервальной шкалы.

Критерии качества заданий в педагогической и классической теориях

Основной «клеточкой» (единицей) теста является тестовое задание. Легко представить, что качество теста существенно зависит от качества каждого отбираемого для него задания. Поскольку не все создаваемые задания годятся для включения в тест, автор этой статьи их называет заданиями в тестовой форме. И только после статистического анализа часть заданий может называться тестовыми заданиями. Опыт показывает, что у специально обученного разработчика заданий успешными могут оказаться примерно 50% разработанных заданий, у менее подготовленных разработчиков — не более одной трети заданий.

Доля или процент успешных тестовых заданий от общего числа предложенных заданий в тестовой форме является хорошим критерием мастерства разработчика тестовых заданий.

В педагогической теории качество каждого задания рассматривается как зависимое от содержания и формы. В содержание заданий включаются только сущностные элементы изучаемого курса. Форма зада-

13

*Wright D. N. and M. H. Stone (1979). Best Test Design, P.*3. When we test a person, our purpose is to estimate their location on the line implied by the test. A measure is a location on a line. Measurement is the process of constructing lines and locating individuals on lines.

14

Эти требования представлены в книге *Аванесова В.С.* Форма тестовых заданий. М.: Центр тестирования, 2005.

ний подбирается в зависимости от вида проверяемых знаний. Сейчас чаще других начинают применяться задания с выбором нескольких правильных ответов из общего числа ответов, прилагаемых к каждому заданию. Они резко снижают вероятность угадывания правильных ответов и проверяют знания полнее, глубже и точнее.

Содержательная совместимость заданий теста определяется экспертно. Главным условием совместимости является принадлежность заданий преподаваемому курсу.

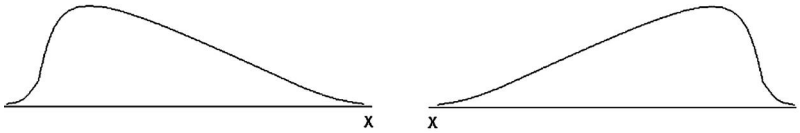
В классической теории вся работа по оценке качества тестовых результатов по критериям надёжности и валидности получила называние Item Analysis.

На ранних этапах работа по отбору заданий для теста проводится на основе следующих критериев:

— меры трудности каждого задания для испытуемых. Нет смысла включать в один и тот же тест два и большее число заданий одинакового уровня трудности. Наличие таких заданий указывает на признак избыточности;

— меры коррелируемости ответов испытуемых на каждое задание с суммами баллов тех же испытуемых по всем заданиям. Этим в классической теории определялась пригодность каждого задания теста. Используются там ещё методы факторного и множественного регрессионного анализа.

В зарубежных учебниках по классической теории психолого-педагогических измерений утверждалась необходимость добиваться нормального распределения результатов тестирования испытуемых. Это можно было достигнуть либо подбором заданий соответствующей трудности, либо трансформацией шкалы. При нарушении этого требования возникали асимметричные распределения (рис. 1).



Случай трудных заданий теста или слабый подготовленности испытуемых

Случай лёгких заданий теста или высокой подготовленности испытуемых

Рис. 1. Асимметричность распределения

Задания даются тем испытуемым, которые по своей подготовленности соответствуют целевой группе, после результатов статистической обработки данных делается вывод о тестовых свойствах заданий. При апробации теста в классической тестовой литературе обращается особое внимание на подбор групп испытуемых, наиболее подходящих по уровню своей подготовленности уровню трудности подобранных заданий теста[15].

Если целевых групп две и более, то качественный тест

15 Их называют suitable persons.

не получится. Поэтому для двух целевых групп российского ЕГЭ — выпускников школ и абитуриентов вузов — как это случилось в российском ЕГЭ — одни и те же т.н. КИМы ЕГЭ использовать нельзя[16]. Практика такого использования порождает слишком большие ошибки измерения, особенно для наиболее подготовленных испытуемых. Этот факт доказан по эмпирическим данным[17] и по результатам математического моделирования[18].

В России эти вопросы за прошедшие годы не получили заметного решения. Парадокс заключается в том, что сейчас вся страна под руководством Министерства образования и *науки* занимается ЕГЭ, не имеющим *никакого научного, даже хилого*, проекта. Не случайно, как было уже доказано, ЕГЭ вообще не является методом педагогических измерений[19].

Критерии качества заданий в математической теории измерений (IRT) и в RM

Дифференцирующей способностью задания (discriminant ability of the item) называется его свойство различать испытуемых по уровню подготовленности. С ростом дифференцирующей способности графический образ задания стремится к вертикальному положению (рис. 2). Точка деления на оси абсцисс равна средней арифметической. Этот график с исключительно высоким значением крутизны, а значит и дифференцирующей способности.

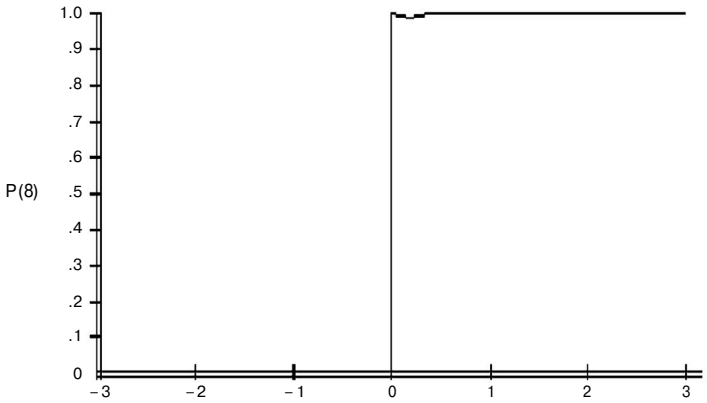


Рис. 2. Графический образ задания с очень высоким уровнем дифференцирующей способности

С появлением Item Response Theory появилась возможность определить графически, а также и по значениям параметров — как задание дифференцирует испытуемых различного уровня подготовленности? Пример трёх различающихся заданий

16 Российский опыт показывает, что «можно», но это опыт отрицательный, тиражируемый, к тому же, из года в год. А потому результаты оказываются некачественными. Неслучайно в Казахстане по предложению автора этой статьи 17 февраля 2012 г. было принято решение разделить единый метод Национального тестирования на две части — для аттестации и для приёма в вузы.

17 *Хлебников В.* Краткий анализ технологии и результатов единого государственного экзамена. Педагогические Измерения. № 4, 2008.

18 *Деменчёнок О.Г.* Погрешность баллов Единого государственного экзамена. Педагогические Измерения. № 4, 2011. С. 3–17.

19 *Аванесов В.С.* Являются ли КИМы ЕГЭ методом педагогических измерений? (Вторая редакция). http://viperson.ru/wind.php?ID=563869&soch=1

по уровню дифференцирующей способности представлен на рис. 3. Смысл значений параметров крутизны легко понять при сравнении трёх графиков заданий, имеющих одинаковый уровень трудности.

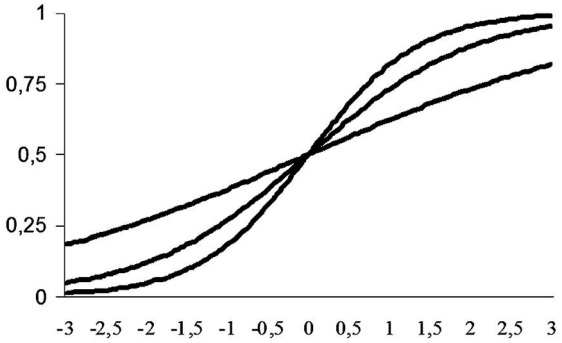


Рис. 3. По оси абсцисс откладываются значения логитов уровня подготовленности испытуемых. По оси ординат — значения вероятностей правильного ответа на задание, в зависимости от уровня подготовленности испытуемых и от значений параметров крутизны всех трёх графиков

На рис. 4 задание имеет низкую крутизну, а это признак очевидного дефекта, выражающийся в том, что не все хорошо подготовленные испытуемые имеют шанс ответить правильно на данное задание. Да и приращение вероятности правильного ответа на это задание в зависимости от уровня подготовленности испытуемых низкое.

Вероятная причина такого дефекта — плохая формулировка содержания задания, при которой возможна другая интерпретация смысла задания.

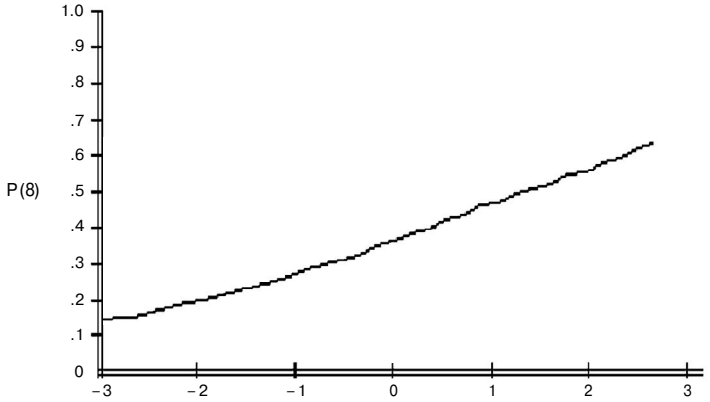


Рис. 4. Графический образ задания с низкой дифференцирующей способностью

Задание с высокой дифференцирующей способностью, но только для испытуемых с уровнем подготовки выше среднего уровня представлено на рис. 5.

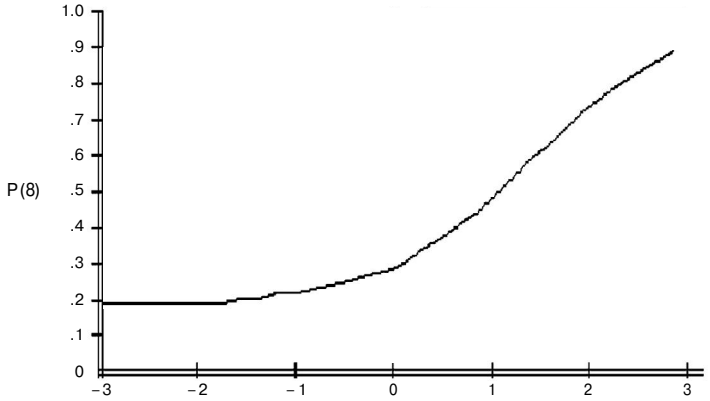


Рис. 5. Задание относительно трудное, дифференцирующее только хорошо подготовленных испытуемых

Критерии качества в RM

В RM анализ возможностей каждого задания ограничен тем, что крутизна каждого принимается одинаковой. Фактически она разная, но модель измерения такова, что параметр крутизны принимается один на все задания.

Можно озадачиться вопросом — хорошо это или плохо? С точки зрения оценки потенциальных возможностей каждого задания — это потеря ценной информации, однако с точки зрения создания системы заданий с непересекающимися графиками каждого задания — это хорошо. Меньше становится погрешностей измерения.

В RM одним из критериев качества заданий является уровень соответствия испытуемых уровню трудности содержания задания. Несоответствие этих уровней приводит к феномену т.н. *экстремальных* заданий. Это те, на которые либо все испытуемые отвечают правильно, либо те задания, на которые все испытуемые отвечают неправильно.

Аналогично в RM принят критерий соответствия среднего уровня трудности заданий теста среднему уровню подготовленности испытуемых. В качестве решающего правила принимается минимально допустимое расстояние этих средних между собой. Только при этих условиях можно говорить о соответствии уровня трудности заданий уровню подготовленности испытуемых.

Этот критерий соответствия («Fit») заданий теста уровню подготовленности испытуемых в процессе экспертной оценки результатов тестирования превращается в один из критериев

валидности результатов тестирования.

Хотя в литературе по RM термин «валидность» используется меньше, чем нужно. Это произошло, по мнению автора данной статьи, из-за психологической установки Б. Райта и его коллег провести как можно более глубокую демаркацию между разрабатываемыми ими методами, на основе теории Г. Раша, от остальных методов, основанных на предшествующих теориях.

Лучшим критерием качества измерений в RM являются совмещённые гистограммы — уровней трудности заданий, для соответствующих групп испытуемых (нижняя часть гистограммы, рис. 6) и уровней подготовленности испытуемых (верхняя часть гистограммы, рис. 6), в одной и той же стандартной шкале логитов.

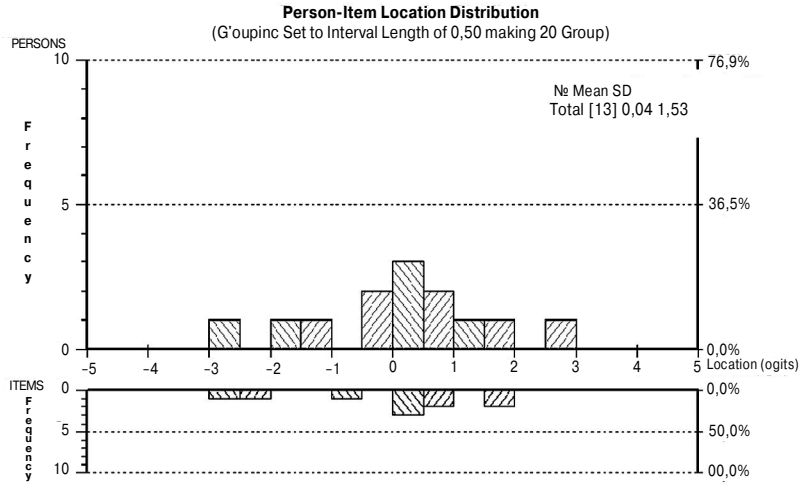


Рис. 6. Совмещённые гистограммы уровня подготовленности испытуемых и уровня трудности заданий

По оси ординат сверху и снизу откладываются соответствующие каждому уровню частоты. На таких гистограммах сразу же видны все недостатки проектируемого теста. В качестве статистического критерия принимается модель нормального распределения.

Из рисунка 6 можно вывести, что испытуемых очень мало (всего 13 человек), их результаты распределены с нежелательными для настоящего теста пробелами в исходных баллах, задания оказались неравномерно возрастающими по трудности, и их тоже мало. Нет задания, которое бы соответствовало уровню наиболее подготовленного испытуемого. В общем, результаты такого «измерения», оказываются неприемлемыми. Главная при-

чина — слишком мало испытуемых и мало заданий.

Несколько лучше обстоят дела в другом случае разработки теста (рис. 7), где много испытуемых и много заданий. Распределение результатов испытуемых в целом похоже на нормальное, хотя и с заметным нарушениями на шкале выше среднего уровня.

Есть претензии и к распределению заданий, к пробелам между ними по уровню трудности (слева и справа). В проектируемом тесте явно не хватает, как всегда, заданий повышенной трудности.

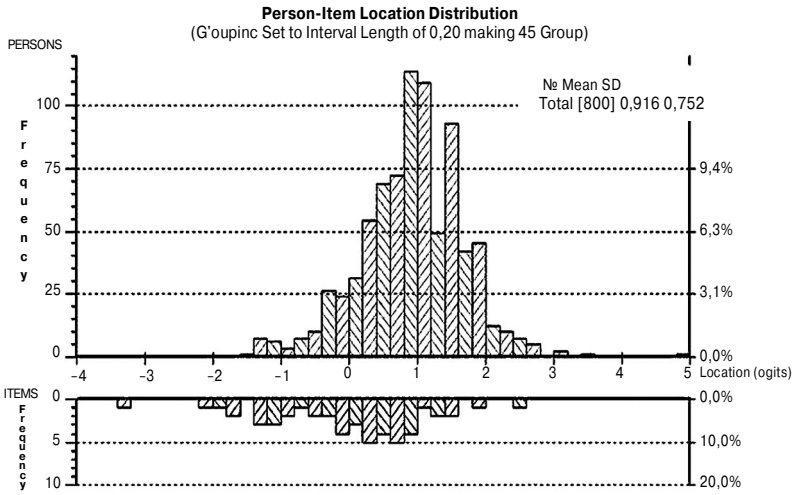


Рис. 7. Пример совмещённых гистограмм результатов проверки проектируемого теста

Приемлемыми результатами тестирования можно назвать такие, которые получены на одной целевой группе испытуемых, посредством теста, позволяющего получить статистически удостоверенные результаты тестирования, отвечающие известным критериям качества и эффективности. Для этого необходимо иметь:

— правильные, в основном, профили ответов испытуемых; это такие профили (вектор-строки испытуемых), где все нули испытуемых следуют за всеми единицами, полученными ими при ответах на задания;

— общая шкала уровня подготовленности испытуемых и меры трудности заданий. Такую возможность предоставляет RM;

— информацию о стандартных ошибках измерения для испытуемых различного уровня подготовленности;

— устойчивость параметров заданий в различных выборках испытуемых.

В RM наиболее часто применяемым критерием совмес-

тимости отдельного задания и общей совместимости всех заданий, образующих тест как систему заданий возрастающей трудности, является значение хи-квадрат. Чем больше значение хи-квадрат, делённое на число так называемых степеней свободы, тем лучше совместимость.

Хорошая совместимость появляется тогда, когда нет экстремальных и проблемных заданий, а также проблемных дистракторов к ним. Напомним, что дистракторами называются ответы неправильные, но правдоподобные.

Совместимость становится отличной, если все задания проектируемого теста задания не только свободны от дефектов, но и наилучшим образом соответствуют требованиям модели Г. Раша.