Теория

ОСНОВЫ ПРОВЕРКИ ДОСТИЖЕНИЙ УЧАЩИХСЯ: ПУТЕВОДИТЕЛЬ ДЛЯ ПРАКТИКОВ

В.А. Мясников, Н.Н. Найдёнова

Десятилетие 2003–2012 гг. объявлено десятилетием грамотности в мире по инициативе ООН. Буш объявил, что «в этой великой стране, называемой Америкой, не будет учащегося ниже среднего международного уровня». Поэтому основные исследования в мировом образовательном пространстве проводятся в русле этого направления. Каждая страна воплощает это по-своему, в том числе через участие в международных исследованиях. Сравнительные исследования учебных достижений в международном плане имеют богатую историю и хорошо проработанную методологию. Исследования внутри каждой страны по оценке учебных достижений имеют свою историю, которую необходимо встраивать в единое образовательное пространство, чтобы обеспечить своим ученикам и студентам одинаковые стартовые возможности во взрослой жизни. Сегодня широко проводятся исследования: PISA (грамотность 15-летних учащхся по математике, естествознанию и чтению), TIMSS (уровень знаний и умений 9-летних и 13-летних учащихся по математике и естествознанию), PIRLS (грамотность по чтению среди 9-летних учащихся) и другие.

Теоретическое развитие методы оценивания как учебных достижений конкретных учащихся, так и, прежде всего, средних учебных достижений, характеризующих разные страты: регион, район, город/село, школу получили в США и Австралии, некоторые вопросы — в Японии,

Канаде и Голландии. Прикладные аспекты развиваются в Великобритании, Ирандии, Скандинавии, Сингапуре. На всех специальностях высшего образования в социальных науках (психология, социология, педагогика) введён специальный курс по методологии социально-педагогических измерений.

В России под педагогическими измерениями понимаются разного рода проверки учебных достижений с помощью тестов, а под тестами понимается набор несвязанных заданий в закрытой форме. Такое толкование очень далеко от принятого в мировом образовательном пространстве.

В мире давно отказались от определения «педагогические», заменив его на более правильное — образовательные, т.е. вместо pedagogical употребляют educational. Хотя педагогические измерения существуют на практике, особенно в Европе, и состоят из текущего контроля за усвоением переданных знаний, не связанного с исследованием учебных достижений, способностей к чему-либо, учёта и выявления факторов и т.п. Итак, если проверяются умения применять знания, то говорят об образовательных измерениях.

Далее под термином «педагогические измерения» будем понимать «educational measurement», потому что в России это и имеется в виду: например, ЕГЭ — безусловно, педагогическое измерение в российском понимании (но, «educational measurement» в мировом понимании).

Сейчас также наблюдается путаница в определении: что значит специалист по педагогическим измерениям или тестолог? Сегодня под тестом понимается сбалансированное множество тестовых задач (заданий к ним), объединённых единой целью исследования. Причём испытуемые выполняют тесты, как правило, разного объёма и разные по содержанию, кроме того, задания предъявляются в трёх разных формах: открытой, краткой и закрытой. Общим является время выполнения и равные возможности у испытуемых выполнять тот или иной тест (у нас обычно это называется вариантом). Правда, в России на в ЕГЭ наблюдаются равные по структуре и длине варианты, временной фактор также равный, а вот право выбора варианта теста на ЕГЭ не реализовано. При этом известно, что наши варианты не равны друг другу, более того в разных регионах выполняются разные варианты, а Россия в образовательном плане представляет собой жёстко стратифицированную совокупность.

Итак, тестологом (специалистом по педагогическим измерениям) следует считать специалиста, умеющего проверять учебные достижения учающихся и знающего:

• как проводить разного рода проверки разными контрольно-измерительными средствами,

• как создавать надёжные контрольные измерительные материалы (КИМ),

• как оценивать результаты этого измерения в соответствии с принятыми в мире методами педагогических измерений.

Чаще под тестологом в России понимают предметника-разработчика тестовых заданий, не имеющего никакой квалификации в педагогических измерениях. Тестолог — это специалист в образовании, математической статистике, информационных технологиях, теории измерений в гуманитарных науках. На западе его прежде называли Jsychometrician», а теперь — «edumetrician». В принципе вся теория педагогических измерений выросла из теории психологических измерений (А. Анастази) и до сих пор очень тесно связана с психологическими аспектами ментальных измерений.

Что есть педагогические измерения? Измерение — это вес или вид деятельности, в которой участвуют:

• субъект, имеющий соответствующую квалификацию и осуществляющий само измерение;

• объект, который подвергается измерению собственных количественных и качественных свойств;

• инструмент, которым проводятся измерения.

Часто говорят, что измерение — это сравнение с эталоном. Это не совсем так: измеряя длину отрезка простой линейкой, не производится сравнения с эталоном метра. Но считается, что производители линеек соблюдали соответствие с эталоном в пределах разной величины ошибки. Ошибка измерения — это допустимое понятие в любом измерении. На этом примере видно, как разнесены во времени и пространстве производители измерительных материалов и средств и сам процесс измерения. Ясно, что производители и измерители владеют общей теорией измерения.

В педагогических измерениях производителями являются разработчики тестовых материалов. Под тестом понимается любой измерительный материал: стандартизированный тест, например, ЕГЭ в российском понимании, контрольная работа, диктант, сочинение, опросные листы, анкеты и другие формы, применяемые в образовании. Разработчик сдаёт свой тест, а собственно измерение проводится без его участия, включая оценивание результатов. Причём, разработчик обязательно сдаёт кроме самого теста и саму систему оценивания результатов, т.е. данный тест является той самой линейкой, причём уже проградуированной. Так принято в мире.

В России же принято после тестирования и первичной обработки результатов предоставить разработчикам таких тестов возможность определить нормы или критерии оценки, т.е. проградуировать линейку под того, кого измеряли, или под цели настоящего

времени. Нельзя измерять некалиброванным инструментом и калибровать его в процессе измерения. Именно так сейчас делается на ЕГЭ. Калибровка тестовых материалов проводится постоянно, независимо от процесса измерения, и ведут процесс калибровки сами производители под контролем независимой измерительной организации.

Итак, субъект в педагогических измерениях — это тестолог в мировом понимании. Им может быть учитель с квалификацией по педагогическим измерениям, работник органов управления или исследователь с той же квалификацией. Существуют в мире уже разработанные стандарты организации и функционирования органов, осуществляющих педагогические измерения системного характера. Понятно, что учитель проводит проверку самостоятельно, и его задача — собственно педагогические аспекты измерения без учёта системных факторов. Поэтому в банк заданий ЕГЭ попадают задания в основном текущего контроля приобретённых знаний на уроках. А учитель должен понимать, что при измерении такого уровня на результат влияют многие другие факторы и их надо учитывать при суждении о пригодности этого задания в качестве измерительного материала в целом: для других школ, областей, деревень, гимназий. Известен такой факт: что в школе лучше учатся девочки, но при окончании вуза уже превалируют мальчики, среди кандидатов и особенно докторов наук эта пропорция усугубляется, среди академиков подавно. Вот так простой фактор — гендерные отличия влияют на измерения уже в школе. Необходимо выделить психологический и педагогический аспект этого отличия. Но задания должны одинаково выполнять как мальчики, так и девочки. Поэтому анализ на гендерность присутствует всегда при педагогических измерениях.

Измерительный инструмент — тест, задания, диктант, устный опрос. Безусловно, имеется в виду калиброванный материал, прошедший репрезентативную апробацию и доработанный специалистами. Так, анализ тестовых заданий ЕГЭ показал, что часто работают всего два стоящих рядом варианта ответа: верный и один из трёх дистракторов. Если ученик выбрал один из двух оставшихся дистракторов, то его результат по тесту в целом будет очень низким. Все дистракторы должны быть равномерно привлекательны.

В мире сейчас строят тесты по так называемой юнитной системе: даётся задача и к ней набор заданий разного уровня сложности и разной формы. Обычно в тесте от четырёх до 15 юнитов, и примерно 50–70 заданий. Бывают юниты из одного задания, обычно из трёх-шести. Причём, ученик оценивается по пяти уровням: 1 — неудовлетворительный, т.е. не способный к квалифицированному труду

в будущем, 2 — базовый, имеет лишь базовые знания, но имеет и трудности с их применением, например, пропускает задания с открытым ответом, 3 — удовлетворительный, 4 — хороший и 5 — отличный. Существует ещё два подуровня: элита и маргиналы, имеются в виду лучшие среди пятого уровня и худшие среди неудовлетворительного.

Известно, что задание считается выполненным на неудовлетворительном уровне, если ученик набрал от 200 до 250 баллов, т.е. разработчик считал, что меньше 200 баллов никто не может набрать. Тогда при подсчёте баллов имеется в виду какое-то исключение: ученик с дефектом умственного развития или ошибка разработчика калибровщика данного задания; если же ученик нормальный и нормальная калибровка, то это недочёт самой системы образования. Также и отношение к трудным заданиям, за которые ставится «пятёрка» (550 до 700 баллов). При 800 баллах при нормальном уровне ученика и нормальной калибровке следует думать о дополнительном образовании помимо учебного заведения или гениальности.

Итак, существует единая шкала для конкретного задания и для конкретного ученика при индивидуальном оценивании. Также есть шкала для всех заданий и всех учащихся. Считается, что только результаты, оцениваемые на таких шкалах, можно сравнивать.

Что происходит при ЕГЭ: школа, в которой выполнялся сложный вариант, оценивается так же, как и школа, где дети выполняли простой вариант теста. Это позволяет уравнивать заведомо неравные результаты, например, давать в крупный город вариант более сложный, чем в село. И тогда не будет резкого отличия между сельскими и городскими учащимися. При выравнивании на общей шкале сельские учащиеся по объективным причинам всегда будут иметь более низкий результат на ЕГЭ.

Обычно в мире такая шкала дополняется рейтингом, в котором стратификация присутствует. Так 1% лучших учащихся Австралии принимается в любой вуз без экзаменов и бесплатно там учится. Но есть ещё отдельная шкала, например, если этот ученик входит в 1% лучших аборигенов, и входит в 1% лучших по Австралии — он имеет такие же права. Также в Америке, включая Гарвард.

Среди методов измерения, необходимых при анализе, всегда присутствуют нормализация, централизация и взвешивание, т.е. анализ данных, а особенно сравнение результатов, проводится всегда на таких триадно пребразованных данных. Нормализация — известная процедура в математической статистике, позволяющая учитывать дисперсию. Централизация бывает двух видов: вокруг заданий для оценки

учащихся и вокруг учащихся для оценки заданий.

Таким образом, при калибровке проводится централизация вокруг учащихся, а при шкалировании результатов ЕГЭ, например, вокруг заданий. Взвешивание необходимо для более точных оценок и учёта стратификационого и факторного влияния. Например, в широкомасштабном эксперименте по реформе содержания учитывались лишь численные характеристики. Так, в одном из регионов РФ в эксперименте участвовали 10 тыс. учащихся из разных школ — это достаточно представительно, но не было взвешивания. После взвешивания при полной репрезентативности получается примерно полный состав области по учащимся, но так как множество было сдвинуто в сторону городских школ и прежде всего гимназий, имеющих малый вес в этой области, представительность оказалась на уровне 50%, что не позволяет делать надёжные и достоверные выводы по эксперименту.

Объект педагогических измерений — ученик. Но в педагогических измерениях необходимо учитывать разные факторы, если необходимо выявить собственно системный результат; тогда исследуются и вовлечённые в учебный процесс объекты: учителя, администрация школ, районов и регионов, родители. Особенно это важно при калибровке заданий и тестов. Измеряем свойства ученика: например, способность к дальнейшему самостоятельному обучению или знание и умение хорошо производить тождественны преобразования и решать уравнения по математике: необходимо количественно выразить качество исследуемого свойства у объекта. На данный момент в мире практически не измеряют конкретные знания, оставляя это учителю, а измеряют конкретные способности, учебные достижения, готовность к социальной, трудовой и бытовой жизни. Считается, что школа должна подготовить ученика к жизни в современном мире, технически вооружённом, информационно насыщенном и постоянно изменяющемся социально и научно. Поэтому тесты составляются таким образом, чтобы ученика можно было оценить по этим позициям.

Наши тесты ЕГЭ совершенно на это не направлены. То, что ученик точно знает, когда было отменено крепостное право или прекрасно решает уравнения, но не может составить ни одного уравнения самостоятельно и правильно по текстовой задаче, мало что даёт будущему работодателю. В мире такие задания относятся к базовым, но их выполнение не позволяет получить даже удовлетворительную оценку. Базовые знания и умения всегда можно проверить и при заданиях другого уровня, создавая их по модели так называемой Пашиал кредит (Partial Credit). Как правило, школьники, легко решающие сте-

реометрические и текстовые задачи не испытывают трудностей с базовыми знаниями.

Методологические основы педагогических измерений состоят в целом из:

— общих вопросов проверки учебных достижений (методологии массовых сравнительных исследований; измерений в мониторинговых исследованиях; требований к сравнительному анализу);

— социально-педагогических проблем организационного характера (формирования и оценивания выборки, менеджмента; обработки данных и их анализа);

— теории и практики педагогических измерений (надёжности и валидности оценок учебных достижений; методов оценивания тестовых материалов; средств коррекции оценок с учётом угадывания, тенденциозности; теории измерения отношений; описательного шкалирования; измерения социальных факторов; аудиторных наблюдений; анкетирования; рейтингового шкалирования);

— социально-педагогического анализа данных (количественного и качественного анализа; анализа пропусков и стратегий выполнения теста; подготовки информации для анализа; измерения вариаций; фиксации единиц анализа для нормализации и взвешивания; многоуровневого анализа; многовариативного анализа; кластерного и факторного анализа; анализа данных из разных исследований; моделирования оценок; иерархических моделей влияния факторов на результат; анализа профилей оценивания).

Первичный анализ минимально должен включать следующие основные шаги:

1. Анализ тестов на нормальность по всем вариантам с указанием описательной статистики, гистограммы с наложенной кривой нормального распределения, карты распределения.

2. Разделение всей информации по варианту на страты: регион, район, центр региона, город разного типа, посёлок, село, школа.

3. Сводную статистику по стратам.

4. Расчёт суммы баллов для каждого ученика отдельно с выбором ответа и со свободным ответов, расчёт Z-балла.

5. Фит-анализ, карту распределения и коэффициент Альфа для варианта, страт.

6. Отдельный анализ якорных заданий в целом, по каждому варианту, по стратам.

7. Разметку карт распределения для якорных заданий.

8. Построение градиентов по Рашу (Rasch) для варианта и страт.

9. Определение дискриминантных заданий.

10. Итерационный фит-анализ.

11. Сжатие заданий методом главных компонент.

12. Многомерное шкалирование.

13. Анализ и объяснение дисперсионных отличий. Любое ис-

следование имеет свой ход развития, некоторые этапы должны следовать определённой последовательности. Например, нельзя выравнивать результаты учащихся без формирования нормальной группы сравнения. Единый цикл исследования состоит из семи шагов:

1. Задачи исследования;
2. Тематика исследования;
3. Планирование и выборка.
4. Сбор данных
5. Подготовка данных.
6. Анализ.
7. Выводы

Сделанные выводы могут привести к изменению самих задач исследования и к повтору цикла. Любая надёжная проверка средних учебных достижений по предмету, региону, стране, школе и т.п. состоит из нескольких циклов. Например, предпилотная проверка для уточнения измерительных инструментов; пилотная проверка для отладки инструментария оценивания, апробация для формирования трендовых моделей и основное исследование. Обычно все четыре стадии бывают в исследовании, минимум определяется двум стадиями.

Задачи исследования включают в себя описание целей исследования с детальной проработкой. Только после выполнения этого шага можно приступать к тематической разработке. При тестировании описание целей и задач должно включать:

- описание цели и задач тестирования;
- описание тестовых и анкетных материалов;
- определение оцениваемых признаков среди тестовых и анкетных материалов;
- описание объектов исследования;
- определение доступности и наличия поименованного списка объектов;
- перечень атрибутивных и количественных показателей;
- формулирование выводов, касающихся всей генеральной совокупности и имеющих отношение к тестовым и анкетным материалам, а прежде всего к объектам исследования;
- наличие сравнительного анализа данных разного типа;
- определение групп сравнения по объектам исследования или по тестовым и анкетным материалам;
- структуру планируемых итоговых выводов.

Этот этап планирования выборочного исследования очень важен, так как чёткое, полное и правильное описание целей и задач позволяет в дальнейшем правильно построить репрезентативную выборку, чтобы можно с делать достоверные и надёжные выводы по результатам следования с соотнесением этих выводов на всю генеральную совокупность.

Второй этап, состоящий из тематического планирования, состоит в детальной разработке спецификации контрольно-измерительных инструментов. Специфи-

кация обязательно включает таблицу верных ответов и систему кодировки открытых ответов, систему кодировки данных и их перекодировки для детального анализа. Например, оценивание успешности в зависимости от программы обучения.

Планирование исследования и формирование выборки — необходимое и достаточное условие любого массоового исследования. Более надёжные результаты можно получить при построении априорной выборки и дальнейшего взвешивания основных резуль-татов.

Итак, среди форм по организации и проведению следует выделить традиционно присутствующие во всех исследованиях:

• чёткое определение целей и задач исследования, представленное рядом специальных форм как международного типа, так и национального;

• согласование отчёта и форм представления информации на стадии организации исследования;

• обобщающая форма по странам, характеризующая страны по общим экономическим, социальным и культурным позициям;

• выделение обобщающего ядра тем по предметам (профили обучения);

• организационные формы: определение совокупностей, формирование представительных выборок, регламентация сбора данных, кодирование информации, ввод данных и верификация, анализ данных, документация инструментария.

Сбор данных включает сбор анкет и тестов, специальные сопроводительные процедуры, процедуры по качественному контролю за обработкой информации; документального сопровождения: сбора информации и подготовки данных; планов по подготовке и вводу информации; подготовке данных для компьютерного ввода и их проверке; документального сопровождения всех этапов ввода информации.

Аналитическая работа (шаг пятого цикла) с данными после тестирования, так называемый первичный анализ состоит из:

1. верификации ввода;

2. анализа пропусков;

3. корректировки информации через удаление элементов и переменных;

4. компенсации пропусков;

5. компенсации потерь информации с помощью взвешивания;

6. универсальной верификации по разным объектам:

• анкетам, идентификационным связям, включая перекодировку для анализа;

• вариантам теста, идентификационным связям, включая ротационную перекодировку для анализа;

• отношениям, семантическим дифференциальным измерениям о перекодировке данных;

• наблюдений в аудитории в связи с реализацией права испытуемого на получение любого варианта.

Мультивариативный анализ состоит из сжатия данных, удаления незначимых переменных, комбинации переменных в конструкты для вторичного анализа, анализа гипотез и моделей.

По итогам анализа на основании прогностических моделей делаются выводы о первом завершающем этапе исследования, о его продолжении и изменении задач по необходимости. Выводы о средней успешности обучения учащихся в стране по данному предмету формируются на многоуровневом, многовариативном, мультишкальном анализе данных. В выводах приводятся:

- анализ учебных достижений по предмету в форме единого теста, по отдельным частям и темам теста, по умениям и навыкам, по уровням компетентности и т.п.;
- достоверное и математически обоснованное выделение факторов, влияющих на уровень подготовки учащихся в стране, регионе и т.д.;
- определение индикаторов, характеризующих систему образования в стране, в мировом образовательном пространстве, в регионах, школах и т.д.;
- декларация доступности всех данных на электронных носителях для широкого использования через три года после основного исследования.

Среди российских исследований наиболее соответствует всем этапам цикла исследования единый экзамен (ЕГЭ). Цель ведения ЕГЭ в России очень благороден в своём декларативном обозначении: для большинства учащихся, удалённых от вузов и не имеющих средств на поездку для поступления в вуз, предоставляется возможность сдать экзамен рядом с домом. Кроме того, учащиеся избавляются от дополнительного стресса, связанного с повторным вступительным экзаменом после только что сданных выпускных испытаний в школе. ЕГЭ является тем инструментом оценивания уровня подготовки учащихся по отдельным предметам учебного плана, который достаточно объективен и информативен, так как независим от оценки учебных достижений школьными учителями. Таким образом, ЕГЭ — стандартная процедура оценки учебных достижений выпускников всех регионов РФ, во-первых, с целью выравнивания прав на высшее образование, во-вторых, для оценки учебных достижений в рамках школьной аттестации.

Ещё в 1978 г. Андерсон и Бол описали шесть способов оценки в любой сфере, определяющих последующую деятельность заинтересованных сторон в содействии успешности принимаемых решений. Эти оценки выносят достоверные суждения:

- о развитии и применении;
- о продолжении программ, расширении использования и сертификации;
- о модификации программ;
- о получении доказательств для сплочённой поддержки программы;

• о получении доказательств опровержений для оппозиции программы;

• в понимании базисных педагогических, психологических, социальных и других процессов.

Итак, наш единый экзамен пока использует лишь первых два способа. А успех определяется всеми шестью.

Шривен пишет о двух функциях оценки учебных достижений:

1) образующей или текущей;

2) итоговой или аттестующей.

Итоговое оценивание — это финальная стадия текущего оценивания. Они близки по целям и задачам, инструментарию и т.д.

Кронбах говорит о психологической или социополитической функции оценивания. Любое оценивание утилизирует программы, стиль поведения, отношения со средствами массовой информации и общественное мнение.

Необходимо сделать, по крайней мере, два шага в конструировании оценивания результатов ЕГЭ:

— определить функции оценки (суммарную или формулируемую);

— определить используемые приближения (объектно-ориентированные и т.п. или комбинации приближений).

Массовое обследование рассматривается обычно отдельно от оценивания (при оценке не учитываются организационные факторы), хотя проведение и организация обследования сильно влияют на качество оценивания. Массовое обследование — это обследование по классам, проверка и определение учебных достижений по предмету или измерение отношений в деятельности учащегося относительно исследуемых групп, типов. Например, при оценивании учебных достижений в освоении иностранного языка необходимо измерить результаты как письменного теста, так и устных навыков.

Для повышения информативности обследование должно быть систематическим, включать содержательную преемственность по предмету внутри самого измерения (в разные периоды обследование проводится тем же самым инструментом). Кроме того, через разные измерения должны быть похожие результаты, но достигаемые разными инструментами. Например, известны результаты учащихся РФ в международном исследовании PISA, которые объективно показали, что результаты учащихся Южного федерального округа в среднем ниже среднего результата по всей России, как в математике, так и в естествознании и чтении. ЕГЭ – другой инструмент измерения, но должны быть похожие результаты или должна существовать достаточно объективная интерпретация расхождений результатов. Например, работники сферы образования Южного округа имеют большой опыт в ЕГЭ. И поэтому научили своих школьников выполнять формальные тесты ЕГЭ достаточ-

но хорошо, а в тестах PISA, где почти отсутствует проверка воспроизведения знаний и почти все задания деятельностные (да и нет мотивации к натаскиванию для улучшения результатов), результат оказался ниже. Что объективнее? Какой инструмент лучше? Эти и прочие вопросы не ставятся в повестку дня. Объективное педагогическое измерение не зависит от инструмента измерения и квалификации персонала, проводящего измерения.

Бенсон и Михаэль в 1990 г. предложили два главных компонента в дизайне оценки: 1) критериальное определение со спецификацией необходимой информации, чтобы можно было объективно оценить эффективность самой программы; 2) выбор методов определения оптимальной стратегии или плана, которые дают возможность получить точную описательную и объясняемую информацию или информацию, позволяющую оценить погрешности между реальным воплощением стратегии и наблюдаемыми значениями.

Специалист по педагогическим измерениям (тестолог) должен быть проблемно-ориентированным, а не методистом, так как специалист будет стараться использовать в обследовании когда-то определенный метод, несмотря на сложившуюся ситуацию. А тестолог решает специфическую проблему в соответствии с целями исследования, применяя те методы, которые подходят в данный момент.

Обследования в их практическом воплощения можно разделить на четыре типа: эксперимент; квази-эксперимент (смоделированный эксперимент); массовое обследование на репрезентативной выборке; регулярное обследование.

Первые два типа относятся к так называемым предтестовым испытаниям. Третий тип часто называют апробацией. Четвёртый тип — обследование калиброванным инструментом.

Основное требование к первым трём типам — случайный выбор учащихся. Случайный выбор — это отбор учащихся по специальным процедурам формирования вероятностных репрезентативных выборок. Причём обязательно следить за репрезентативностью отбора на школьном уровне, как и за соблюдением репрезентативности во всех контрольных группах, определяющих достоверность результатов (региональная принадлежность, тип и вид учебного учреждения, род местности, профиль класса и т.п.).

Для сравнительных исследований выделяется нормальная группа сравнений, которая не означает множество реальных «живых» учащихся, участвующих в обследовании. Учащиеся, включённые в нормальную группу сравнений, представляют:

• нормальную группу, в которой учащиеся выполняли нормально-ориентированный тест, и распределение результатов подчиняется нормальному закону;

• тестовые баллы, оценивающие результаты учащихся этой группы, представляют средние результаты по стратификационным переменным (например, школьные и региональные результаты) в соответствии с нормальным распределением.

При апробационных испытаниях в нормальной группе сравнения обычно собирается дополнительная описательная статистика как: социально-экономический статус семьи, педагогический стаж учителя, состав педагогического учебного учреждения, практическая деятельность и т.п.

Необходимо при апробации выполнять следующие шаги: определять популяцию в соответствии с целью тестирования (например, в ЕГЭ это выпускники средних школ, их учителя и семьи); сбалансировать тестовые и анкетные материалы; если популяция велика, нужна идентификация репрезентативной выборки для анализа; формирование табличной и графической информации для представления результатов.

Представление информации ведётся и в рейтинговом режиме; низкий рейтинг затрудняет интерпретацию результатов. Апробация обычно — достаточно мощное по объёму исследование с соответствующей структуризацией информации. Однако она не даёт достаточной обобщённости выводов. При основном обследовании, кроме апробационной информации, представляется также информация о тех, кто проводил и организовал обследование, кто проверял задания в открытой форме, как проводилась верификация компьютеризованной информации и т.п.

Только результаты основного регулярного обследования позволяют собрать информацию, необходимую для планирования оценивания в системе образования, для мониторинга инновационных программ, для определения статистически значимой информации. Основные публикации в прессе, по крайней мере, с определёнными выводами должны базироваться на данных регулярных обследований.

Начиная с 1994 г. Пейном была предложена смешанная методология, совмещающая количественные и качественный оценки. При ней необходимо:

• участие тестологов разной ориентации (количественной и качественной);

• использование нескольких источников информации (стандартизированные тесты, альтернативные обследования и интервью);

• использование разных методов сбора информации (анализ данных по стратам, обследование учителей и т.п.);

• иримененение разных теоретических приближений (привлечение к анализу данных работников методических ассоциаций, специалистов в области программного обеспечения и т.п.).

К альтернативным исследованиям относятся обследования практической направленности, а

также критериально-ориентированные тесты. Организация информации для ученика называется портфолио. Информация в портфолио есть окончательное и наиболее удачное представление индивидуальной информации, взятой из результатов нормально-ориентированных тестов. Портфолио часто используется для экспертизы. Значимыми характеристиками серьёзного обследования являются надёжность и валидность.

Надёжность — стабильность и содержательная состоятельность обследования. Например, гарантируются одинаковые результаты тестирования участников в одно и то же время, но в двух разных обследованиях. Инструмент измерения должен быть надёжен в первую очередь. Чем больше заданий в тесте, тем он надёжнее. Например, тест из 50 заданий более надёжен, чем тест из 10 заданий. Но тест из 300 заданий будет уже достаточно ненадёжным (пропуски приобретают случайный характер и не позволяют надёжно оценить результаты). При тематическом разбиении теста по предмету следует выделять, по крайней мере, 10 заданий по каждой теме для надёжной оценки. Тест с 50–60 заданиями содержит задания из 5–6 тематических разделов. Следовательно, тест ЕГЭ по математике должен состоять всего из двух тем при длине теста в 25 заданий, чтобы измерение ЕГЭ было надёжным.

Внутрирейтерная надёжность — специфический тип надёжности. Когда респондент участвует и в альтернативном тестировании (например, результат по химии у выпускников следует оценивать не только по письменному тесту, но и по тесту на практическую деятельность), то объективная оценка учебных достижений по химии не может быть вне зависимости от результатов лабораторных и практических работ-тестов.

Также внутрирейтерная надёжность рассчитывается при оценке заданий открытого типа, так как, во-первых, необходимо учитывать влияние субъективного фактора, исходящего из опыта и пристрастий кодировщика. Во-вторых, необходимо определить число проверок (сколько раз будут проверять конкретное задание) разными рейтерами, как минимум двумя персонами. При этом около 80% оценок одинаковы.

Валидность — более трудная категория для описания. Новый взгляд на это понятие связан с тем, что надо оценить приближённость интерпретации, применений и акций, базирующихся на результатах обследования (Мессик, 1988). Валидность относится к приближениям, значимости и полезности специфического влияния, вытекающего из тестовых баллов. В России валидность — это достоверность теста.

Кроме измерительных важных оснований обследования следует отметить следующие факторы, во многом определяющие целесообразность самого исследования:

время (за траченное на проведение, сбор данных и оценивание); расходы на копирование, администрирование и оценивание); персонал (координация работ, специальные тренинги, оценки (тестовые баллы — интерпретация); оценивание (формирование агрегированных данных и значимое оценивание).

Количественная статистика бывает описательной и инфернальной. Вычисляется по итогам апробационных и предтестовых испытаний.

Качественная статистика вычисляется из регулярного обследования и смешанных дизайнов тестирования.

Количественная статистика — это частоты, проценты, средние значения, стандартные отклонения, процентили, статистики t и z, коэффициенты корреляции, ANOVA — анализ вариации.

Качественная статистика включает работу с категориальными и анкетными данными. Причём обязательно с подсчётом статистической значимости и ошибок измерения разного уровня. Устанавливаются связи с разными данными из разных обследований.

Пейн предложил четыре типа качественного анализа результатов обследования: феноменологический анализ — анализ данных анкет, интервью и заданий в открытой форме; содержательный анализ — анализ программных документов обследования (инструкций, писем, протоколов и пр.); аналитическая индукция — анализ разного рода новой информации, представляющей доказательства достоверности и надёжности; сравнительный анализ.

В заключение фиксируем следующее:

- цели и функции оценивания определяются внутри образовательных целевых установок;
- существует четыре типа дизайна оценивания — эксперимент, квази-эксперимент, апробация и основное исследование;
- определяется нормальная группа сравнения; формулируются основные определения и процедуры;
- определяются различия между количественным и качественным анализом.

Ребер в 1995 г. идентифицировал идеальное обследование как: конфиденциальность информации; классификацию информации; классификацию измерительных материалов, импульс к применению данного инструмента в образовательной системе; уровни применения (федеральный, региональный, местный, школьный и ученический).

В начале планирования обследования при разработке инструментария важно идентифицировать: цель обследования (прогностическая или итоговая аттестация); тематику обследования (предметное содержание и проверяемые навыки); язык, глоссарий обследования; тип обследования (нормативное, критериальное, альтернативное); типы тестовых оценок с детальным описанием оценивания заданий; группы сравнения.