

Основы проверки достижений учащихся: путеводитель для практиков

Владимир Афанасьевич Мясников, доктор педагогических наук, профессор, член-корреспондент РАО, директор Института теории и истории педагогики РАО, руководитель Центра сравнительной педагогики

Наталья Николаевна Найдёнова, национальный координатор РФ по выборке в IEA, член Научно-методического совета по тестированию Министерства образования и науки РФ, Институт теории и истории педагогики РАО

Сравнительные исследования учебных достижений в международном плане имеют богатую историю и хорошо проработанную методологию. Исследования внутри каждой страны по оценке учебных достижений имеют собственную историю, и её необходимо встраивать в единое образовательное пространство. Сегодня широко проводятся исследования: ПИЗА (грамотность 15-летних учащихся по математике, естествознанию и чтению), ТИМСС (уровень знаний и умений 9-летних и 13-летних учащихся по математике и естествознанию), ПИРЛС (грамотность по чтению среди 9-летних учащихся), СИВИК (граждановедческая грамотность учащихся 9-х и 11-х классов) и др.

Десятилетие 2003–2012 гг., объявленное ООН десятилетием грамотности, декларируется новой эрой в оценивании как учебных достижений конкретных учащихся, так и средних учебных достижений разных страт: региона, района, города/села, школы и др. Теоретическое развитие методы оценивания получили в США и Австралии, по некоторым аспектам — в Японии, Канаде и Голландии. Прикладные аспекты разрабатываются в Великобритании, Ирландии, Скандинавии, Сингапуре. На всех специальностях высшего образования в социальных науках (психология социология, педагогика и т.д.) введён специальный курс по методологии социально-педагогических измерений.

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

В мире давно отказались от определения «педагогические измерения», заменив его — «образовательными», т.е. вместо pedagogical употребляют educational. Хотя в Европе педагогические измерения существуют как средство текущего контроля за усвоением знаний, не связанного с исследованием учебных достижений, способностей к чему-либо, учёта и выявления факторов и т.п. И так, если проверяются умения применять знания, то говорят об образовательных измерениях.

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

Путаница наблюдается и в определении, кого считать специалистом по педагогическим измерениям, или тестологом. Под тестом понимается сбалансированное множество тестовых задач (заданий к ним), объединённых единой целью исследования. Причём испытуемые выполняют, как правило, тесты разной длины и различные по содержанию, кроме того, задания предъявляются в трёх формах: открытой, краткой и закрытой. Общим является время выполнения и равные возможности у испытуемых выполнять тот или иной вариант. Правда, в ЕГЭ наблюдаются равные по структуре и длине варианты, временной фактор также равный, а вот право выбора варианта теста в 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 жёстко стратифицирована.

Тестологом (специалистом по педагогическим измерениям) следует считать специалиста, умеющего проверять учебные достижения обучающихся и знающего:

- как делать разного рода проверки разными контрольно-измерительными средствами;
- как создавать надёжные контрольно-измерительные материалы (КИМ);
- как оценивать результаты этого измерения в соответствии с принятыми в мире методами педагогических измерений.

Тестологом у нас чаще считают предметника-разработчика тестовых заданий, не имеющего никакой квалификации в области педагогических измерений. Однако тестолог — это проблемно ориентированный специалист в образовании, математической статистике, информационных технологиях, теории измерений в гуманитарных науках, а не методист.

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

- субъект, имеющий соответствующую квалификацию и осуществляющий само измерение;
- объект, который подвергается измерению собственных количественных и качественных свойств;
- инструмент, с помощью которого проводятся измерения.

Часто говорят, что измерение — это сравнение с эталоном. Это не совсем так, ведь измеряя длину отрезка простой линейкой, мы не производим сравнение с эталоном метра. Но считается, что производители линеек соблюдают 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, включая оценивание результатов, про-

водится без его участия. Причём разработчик обязательно сдаёт и систему оценивания результатов. Иными словами, тест и есть та самая линейка, причём уже проградуированна. Так принято в мире. В России же после проведения тестирования и первичной обработки результатов разработчикам таких тестов предоставляется возможность определить нормы или критерии оценки. То есть, как бы заново проградуировать линейку под того, кого измеряли, или под цели настоящего времени. Нельзя измерять не калиброванным инструментом и калибровать его в процессе измерения. Именно так сейчас делается в ЕГЭ. Калибровка тестовых материалов проводится постоянно, независимо от процесса измерения, и ведут процесс калибровки сами производители под контролем независимой измерительной организации. И так, субъект в педагогических измерениях — это тестолог мирового понимания. Им может быть учитель с квалификацией по педагогическим измерениям, работник органов управления или исследователь с той же квалификацией. В мире существуют разработанные стандарты организации и функционирования органов, осуществляющих педагогические измерения системного характера. Понятно, что учитель организует проверку самостоятельно, и его задача — собственно педагогические аспекты измерения без учёта системных факторов. Поэтому, если привлекать к разработке КИМов лишь учителей, в банк заданий ЕГЭ попадают задания в основном текущего контроля приобретённых на уроках знаний. А при измерении такого уровня на результат влияют многие другие факторы и их надо учитывать, вынося суждение о пригодности этого задания в качестве измерительного материала в целом: для других школ, гимназий и пр. Известен такой факт: в школе лучше успевают девочки, но по окончании вуза доминируют мальчики, среди кандидатов и особенно докторов наук эта пропорция усугубляется, а среди академиков — тем более. Простой фактор — гендерное отличие влияет на измерения уже в школе. Необходимо выделить психологический и педагогический аспекты этого отличия. А собственно 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 и пр. Безусловно, имеется в виду

калиброванный материал, прошедший репрезентативную апробацию и доработанный специалистами. Анализ тестовых заданий ЕГЭ, показал, что часто работают всего два стоящих рядом варианта ответа: верный и один из трёх дистракторов. А если учащийся выбрал один из двух оставшихся дистракторов, то его результат по тесту в целом будет очень низким. А все дистракторы должны быть равномерно привлекательны. Или, к примеру, существует утверждение, что тройку можно получить за задания с выбором ответа, 4 — с выбором и кратким ответом и 5 — с выбором, кратким и полным ответом. То есть, если в мире принято считать, что для удовлетворительного образования человек должен уметь решить неформальную задачу, применяя приобретённые навыки в школе, то у нас эта задача — для отличников. Будущее страны оказывается таким образом в руках людей, умеющих выполнять элементарные задания, на которые их можно попросту натаскать.

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 4 до 15 юнитов, и примерно 50–70 заданий. Бывают юниты из одного задания, обычно 3–6. Причём учащийся оценивается по 5 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, имеется в виду лучшие среди 5-го уровня и худшие среди неудовлетворительного.

Известно, что задание считается выполненным на неудовлетворительном уровне, если учащийся набрал от 200 до 250 баллов, т.е. разработчик считал, что меньше 200 баллов никто не может набрать. Тогда имеется в виду какое-то исключение: ученик с дефектом умственного развития или ошибка разработчика-калибровщика этого задания; если же этих исключений нет, то причину нужно искать в несовершенстве самой системы образования. Так же и отношение к трудным заданиям, за которые ставится 5, если ученик набрал от 550 до 700 баллов. Если же есть 800 баллов при нормальном уровне ученика и нормальной калибровке, возникает мысль об особой одарённости учащегося.

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

Что происходит в ЕГЭ? Школа, в которой выполнялся сложный вариант, оценивается так же, как и школа, где ученики выполняли простой вариант теста. Это позволяет уравнивать заведомо неравные результаты, например, давать в крупный город более сложный вариант, 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 в ЕГЭ. А точнее сельский учащийся с высоким баллом, похоже, более успешен, чем учащийся городской гимназии с тем же баллом по ЕГЭ и прочими равными показателями социально-педагогического характера.

Обычно в мировой практике шкала дополняется рейтингом, в котором присутствует стратификация. Так один процент лучших учащихся Австралии принимаются в любой вуз без экзаменов и бесплатно там обучаются. Но есть ещё отдельная шкала: если ученик входит в один процент лучших аборигенов и входит в один процент лучших по Австралии, он имеет такие же права. Так же в Америке, включая Гарвард. Поэтому в ЕГЭ необходимо приводить рейтинг среди разных страт.

Среди методов измерения, необходимых при анализе, всегда присутствуют нормализация, централизация и взвешивание. То есть анализ данных, а особенно сравнение результатов проводится всегда на таких триадно преобразованных 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. При калибровке проводится централизация вокруг учащихся, а при шкалировании результатов ЕГЭ — вокруг заданий. Взвешивание необходимо для более точных оценок и учёта стратификационного и факторного влияния. Например, широкомасштабный эксперимент по реформе содержания. В учёт брались лишь численные характеристики. Так в одном из регионов РФ участвовало 10 тыс. учащихся из разных школ, — казалось бы, 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 — учащийся. Но в педагогических измерениях обязательно нужно учитывать разные факторы. Если необходимо выявить собственно системный результат, тогда проводится исследование и вовлечённых в учебный процесс объектов: учителей, администрации школ, районов и регионов, родителей. Особенно это важно при калибровке заданий и тестов. Объект измерения — учащийся, но мы измеряем его свойства: способность к дальнейшему самостоятельному обучению, или знание и умение хорошо производить тождественные преобразования и решать уравнения по математике, или что-то другое. То есть необходимо количественно выразить качество исследуемого свойства у объекта. В мире не измеряют конкретные знания, предоставляя это учителю, а измеряют способности, учебные достижения, готовность к социальной, трудовой и бытовой жизни и др. Считается, что школа должна подготовить ученика к жизни в современном мире — технически вооружённом, информационно насыщенном и постоянно изменяющемся. Поэтому тесты составляются таким образом, чтобы учащегося можно было бы оценить по этим позициям.

Отечественные тесты ЕГЭ на это не направлены. Какой прок работодателю или приёмной комиссии вуза от того, что ученик точно знает, когда было отменено крепостное право? Или если учащийся прекрасно решает уравнения, но ни одного уравнения не может составить самостоятельно и правильно по текстовой задаче. В мире такие задания относятся к базовым, но умение их выполнять не позволяет получить даже удовлетворительную оценку.

Методологические основы педагогических измерений

Эти основы включают:

— общие вопросы проверки учебных достижений (методологию массовых сравнительных исследований; измерения в мониторин-

говых исследованиях; требований к сравнительному анализу);

— социально-педагогические проблемы организационного характера (формирование и оценивание выборки, менеджмента; обработку данных и их анализа);

— теорию и практику педагогических измерений (надёжность и валидность оценок учебных достижений; методы оценивания тестовых материалов; средства коррекции оценок с учётом угадывания, тенденциозности и др; теорию измерения отношений; описательного шкалирования; измерения социальных факторов; аудиторные наблюдения; анкетирование; рейтинговое шкалирование);

• социально-педагогический анализ данных (количественный и качественный анализ; анализ пропусков и стратегий выполнения теста; подготовка информации для анализа; измерение вариаций; фиксация единиц анализа для нормализации и взвешивания; многоуровневый анализ; многовариативный анализ; кластерный и факторный анализ; анализ данных из разных исследований; моделирование оценок; иерархические модели влияния факторов на результат; анализ профилей оценивания).

Первичный анализ минимально должен включать следующие основные шаги:

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

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

Сводную статистику по стратам.

Расчёт суммы баллов для каждого учащегося отдельно с выбором ответа и со свободным ответом, расчёт Z-балла.

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

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

Разметку карт распределения для якорных заданий.

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

Определение дискриминантных заданий.

Итерационный фит-анализ.

Сжатие заданий методом главных компонент.
Многомерное шкалирование.

Анализ и объяснение дисперсионных отличий. Любое исследование 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.

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 с целью уточнения измерительных инструментов, пилотная проверка с целью отладки инструментария оценивания, апробация с целью формирования трендовых моделей и основное исследование. Обычно все эти четыре стадии имеются в исследовании.

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

- 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 и 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 — первый — этап планирования выборочного исследования очень важен, так как чёткое, полное и правильное описание целей и задач позволяет в дальнейшем грамотно построить репрезентативную выборку, чтобы можно было делать достоверные и надёжные выводы по результатам исследования с соотношением этих выводов на всю генеральную совокупность.

Второй этап, включающий тематическое планирование, состоит в детальной разработке спецификации контрольно-измерительных инструментов. Спецификация 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 массового исследования. Более надёжные результаты можно получить при построении априорной выборки и дальнейшего взвешивания основных результатов.

Итак, среди форм организации и проведения исследования 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 состоит не только из сбора анкет и тестов, но и включает специальные сопроводительные процедуры, утверждения 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 с данными после тестирования, так называемый первичный анализ, состоит из:

- верификации ввода;
- анализа пропусков;
- корректировки информации посредством удаления элементов и переменных;
- компенсации пропусков;
- компенсации потерь информации посредством взвешивания;
- 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 с перекодировке данных;

— наблюдениям в аудитории и связи с данными, ротациям и реализациям 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 формируются 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 Единый государственный экзамен (ЕГЭ). Цель ведения ЕГЭ в России очень благородна в своём декларативном обозначении: большинству учащихся, удалённых от вузов и не имеющих средств на поездку для поступления, 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. ЕГЭ — тот инструмент оценивания уровня подготовки учащихся по отдельным предметам учебного плана, который достаточно объективен и информативен, так как независим от оценки учебных достижений школьниками учителями. Таким образом, ЕГЭ есть 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.

Для повышения информативности обследования должно быть систематическим, включая содержательную преемственность по предмету внутри самого измерения (в разные периоды обследование проводится тем же самым инструментом). Кроме того, разные измерения должны приводить к похожим результатам, но достигаемым 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 учащихся выполнять формальные тесты ЕГЭ достаточно хорошо, а в тестах PISA, где почти отсутствует проверка воспроизведения знаний и почти все задания являются деятельностными (да и нет мотивации к натаскиванию), результат оказался ниже. Что объективнее? Какой инструмент лучше? Эти и прочие вопросы необходимо ставить в повестку дня. Объективное педагогическое измерение не зависит от инструмента измерения и квалификации персонала, проводящего измерения.

Обследования в их практическом воплощении 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 как за репрезентативностью отбора на школьном уровне, так 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 проведения 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 учащихся 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 до-

полнительная описательная статистика: социально-экономический статус семьи, педагогический стаж учителя, состав педагогического учебного учреждения, практическая деятельность и т.п.

При апробации необходимо выполнять следующие шаги: определить 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, идентифицировать репрезентативную выборку для анализа; сформировать табличную и графическую информацию для представления результатов.

Информация представляется и в рейтинговом режиме, то есть низкий рейтинг затрудняет интерпретацию результатов. Апробация обычно — достаточно мощное по объёму исследование с соответствующей структуризацией информации, но она не даёт обобщённых 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 компьютеризированной информации и т.п.

Следует заметить, что 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.

Значимые характеристиками обследования — надёжность и валидность.

Надёжность — показатель стабильности и содержательной состоятельности обследования. Например, гарантируются одинаковые результаты тестирования участников в одно и то же время, но в двух разных обследованиях. В первую очередь должен быть надёжен инструмент измерения. Чем больше заданий в тесте, тем он надёжнее. Например, тест из 50 заданий очевидно более надёжный, чем тест из 10 заданий. Но тест из 300 заданий будет уже ненадёжным (пропуски приобретают случайный характер и не позволяют надёжно оценить результаты). Для надёжной оценки при тематическом

разбиении теста по предмету следует выделять, по крайней мере, 10 заданий по каждой теме. То есть тест в 50–60 заданий должен содержать задания из 5–6 тематических разделов. Следовательно, тест ЕГЭ по математике должен состоять всего из 2 тем при длине теста в 25 заданий, чтобы измерение ЕГЭ было надёжным.

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

Кроме измерительных важных оснований обследования, необходимо отметить следующие факторы, определяющие во многом целесообразность самого исследования: время, затраченное на проведение, сбор данных и оценивание; цена (расходы на копирование, администрирование и оценивание); персонал (координация работ, специальные тренинги); оценки (тестовые баллы — интерпретация); оценивание (формирование агрегированных данных и значимое оценивание).

В заключение подведём итоги:

- 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; классификация информации; классификация 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. □