Мониторинг учебной эффективности

Майоров А.Н.

Почему тестирование? Какие тесты?

Поиск адекватного требованиям мониторинга инструмента для определения эффективности работы образовательных систем даёт единственный результат — это тесты и преимущественно тесты. Конечно, не все и не всякие, а в первую очередь тесты учебных достижений (другие синонимичные названия — тесты школьных достижений, педагогические, дидактические и т.д.).

Тест — это инструмент, состоящий из квалиметрически выверенной системы тестовых заданий, стандартизированной процедуры проведения и заранее спроектированной обработки и анализа результатов, предназначенный для измерения качеств и свойств личности, изменение которых возможно в процессе систематического обучения.

Наше утверждение базируется на достаточно значительном количестве доводов, основные из которых мы приведём.

Тесты как измерительный инструмент используются в большинстве стран мира. Их разработка и использование основаны на мощной теории и подтверждены многочисленными эмпирическими исследованиями. Тестология как теория и практика тестирования существует более 120 лет, и за это время накоплен громадный опыт использования тестов в различных сферах человеческой деятельности, включая образование. Тесты не являются универсальным средством, границы их использования хорошо известны. Всё это создаёт уверенность в том, что качественно подготовленный и использованный тестовый инструмент даст качественную и надёжную информацию, соответствующую реальному положению дел.

Тесты — не только более качественный, но и значительно более объективный способ оценивания. Объективность тестирования достигается путём стандартизации процедуры проведения (на всех этапах проведения тестирования невозможно внести субъективную составляющую в оценку) и путём стандартизации и проверки показателей качества отдельных заданий и тестов целиком.

Тесты дают столь необходимое при проведении мониторинга основание для корректных сравнений — их показатели ориентированы на измерение степени, определение уровня усвоения ключевых понятий, тем и разделов учебной программы, умений, навыков и пр., а не на констатацию наличия у учащихся определённой совокупности усвоенных знаний.

Тесты — широкий инструмент. Выполняя тестовую работу, каждый ученик выполняет задания, используя знания по всем темам программы. На устный экзамен обычно выносятся 2–4 темы, на письменный несколько больше (5–10). Тесты могут и, мало того, должны охватить учебный материал по всем темам.

Тест — широкий инструмент и с точки зрения интервала оценивания. Если провести аналогию с прыжками в высоту, то традиционная контрольная работа представит собой не линейку, а палочку, на которой нанесены две риски: 4 и 3, высотой 5. В случае выполнения всех заданий ученик получает отметку отлично. При этом совершенно не ясно, перепрыгнул он нашу стойку с двойным запасом или пролетел прямо над ней. То же можно сказать и про нижнюю отметку. Означает ли, что ученик ничего не знает, если он не выполнил ни одно задание? Скорее всего, нет. Давая широкие возможности для проявления достижений, тест представляет собой измерительный инструмент примерно трехметровой высоты (а вдруг зафиксируем рекорд мира?), риски на котором расположены практически от земли. В этом отношении тестирование приходит в противоречие с учительским стереотипом о том, что отличную оценку выставлять только тогда, когда все задания выполнены пра-

вильно.

Ещё одним отличием тестов от контрольных работ является их оснащение, то есть существование жёсткой процедуры проведения. Она регламентирует отношения между испытуемым и проводящим тестирование, поведение испытуемых во время тестирования, процедуры обработки результатов и интерпретации данных.

Отличие тестов от обычных экзаменов состоит в том, что тесты, используя единые критерии оценки, ставят всех учащихся в равные условия, а это приводит к снижению предэкзаменационных нервных напряжений.

Можно отметить и гуманизм тестирования. Всем предоставляются равные возможности, а широта теста позволяет ученику показать свои достижения на широком поле материала. Таким образом, ученик получает некоторое право на ошибку, которого он лишён при традиционном способе оценивания.

В основе тестирования лежат достаточно простые, логичные, не противоречащие здравому смыслу правила и законы. Они и позволяют найти полноценный ответ на те “почему?”, которые возникают у пользователей.

Привлекательными оказываются тесты и с точки зрения управления. Они дают широкую возможность для варьирования сложности тестового материала, широты охвата, целевой направленности, включения в тест нескольких компонентов структуры знаний, что позволяет создать инструмент, учитывающий самые взыскательные требования управленца. Система показателей качества теста даёт возможность оценить, насколько созданный инструмент реально соответствует этим требованиям, и использовать его в строгом соответствии с ними.

Кроме этого, тесты — эффективный инструмент с экономической точки зрения.

При тестировании основные затраты приходятся на составление качественного инструментария, то есть носят разовый характер. При увеличении количества аттестуемых эти затраты распределяются на них пропорционально, что приводит к снижению общих затрат.

Сравнительный анализ затрат на независимую городскую экзаменационную комиссию (ГЭК) в Санкт-Петербурге и экзамены с использованием тестов представлен на следующих диаграммах (см. диаграммы 1,2).

Pic.

Тесты многообразны, велики по номенклатуре и назначению. Какие из них подходят нам в большей степени?

Вероятно, в начале целесообразно отделить тесты от “не тестов”. Формально это сделать достаточно просто, используя приведённое определение: тест должен иметь в качестве составляющих, по крайней мере, три элемента — систему заданий, зафиксированную документально технологию предъявления и отработанную систему проверки обработки и анализа результатов, которые должны составлять единство.

В статье “Мониторинг социальной эффективности и условий деятельности образовательных систем” (Школьные технологии №5,99) мы обсуждали возможности использования психологического инструментария для нужд мониторинга. Основываясь на этих рассуждениях, разделим инструмент педагогический и психологический. Делая это с достаточной степенью условности, тем не менее можно считать, что педагогические тесты направлены на выявление тех личностных новообразований и приращений, которые получены в результате систематического обучения, а близкие им психологические — на выявление особенностей, полученных в результате всей жизнедеятельности.

Чтобы ответить на вопрос, какие из тестов учебной успешности целесообразно использовать для нужд мониторинга, рассмотрим некоторые основания для классификации этих тестов.

Тесты можно классифицировать по следующим основаниям:

1. По процедуре создания

могут быть выделены стандартизированные и нестандартизированные тесты.

Стандартизируются процедура и условия проведения тестирования, способы обработки и интерпретации результатов, которые должны привести к созданию равных условий для испытуемых и минимизировать случайные ошибки и погрешности как на этапе проведения, так и на этапе обработки результатов и интерпретации данных.

В образовании можно выделить ряд задач, которые могут быть решены нестандартизированными тестами. Однако для целей мониторинга необходимо использовать только стандартизированный тестовый инструмент.

2. По средствам предъявления существуют тесты:

— с использованием тестовых тетрадей, в которых находятся тестовые задания и в которых испытуемый фиксирует результаты;

— бланковые, когда испытуемые отмечают или вписывают правильные ответы (фиксируют ответы) на специальных бланках. Бланки предъявляются отдельно от заданий;

— компьютерные.

Каждый из способов предъявления имеет свои плюсы и минусы. Например, компьютерные очень быстрые, однако они провоцируют случайные ошибки и не оставляют исходных результатов на случай апелляции, что ограничивает их применение для аттестации учащихся. Бланковые позволяют экономить на бумаге, удобны для пересылки, дают возможность сканирования для обработки результатов, но также не оставляют исходных результатов. Наиболее качественные результаты могут быть получены с использованием тестовых тетрадей, но при этом возникают проблемы ввода результатов для обработки в компьютер и необходимости дополнительных расходов на печать самих тетрадей.

Для мониторинга подходит любой из этих способов, но нужно помнить об одном — предъявляя один и тот же тест в разных формах, мы получим разные результаты. Нельзя сравнивать результаты тестирования, полученные при разных способах предъявления.

3. По ведущей ориентации выделяются

— тесты скорости. Они содержат простые задачи, но время решения ограничено настолько, что ни один испытуемый не успевает решить за отведённое время все задачи;

— тесты мощности или результативности включают трудные задачи. Время их решения либо вовсе не ограничено, либо мягко лимитировано. Оценке подлежат успешность и способ решения задачи. Примером такого рода тестовых заданий могут быть задания для письменных итоговых экзаменов за курс основной школы;

— смешанные тесты, которые объединяют в себе черты двух вышеперечисленных. В них представлены задачи различного уровня сложности, от самых простых до очень сложных. Время испытания в данном случае ограничено, но достаточно для решения большинства предлагаемых задач большинством обследуемых. Оценкой в данном случае служат как скорость выполнения заданий (количество выполненных заданий), так и правильность решения.

Эти тесты наиболее часто применяются на практике и именно к ним относится большинство тестов учебных достижений, которые можно использовать для нужд мониторинга.

4. По виду нормирования

— тесты, ориентированные на статистические нормы. Основанием для сравнения в них служат соответствующим образом обоснованные, статистически полученные значения выполнения данного теста репрезентативной выборкой испытуемых;

— критериально-ориентированные тесты. Они предназначены для определения уровня индивидуальных достижений испытуемого относительно заданного критерия, существующего в реальной практике и заранее известного: уровня знаний, умений, навыков, необходимых для какого-либо вида деятельности;

— ненормированные.

Особенности этих тестов мы подробно обсудим в следующем разделе.

5. По целям использования выделяются следующие группы тестов (эта классификация и пояснения к ней приведены по соответствующему разделу книги Нормана Е. Гронлунда):

— знаний или поведения студента в начале обучения (определяющий тест);

— прогресса, достигнутого в процессе обучения (формирующий тест);

— трудностей обучения и их источников во время процесса обучения (диагностический тест);

— основных достижений в конце обучения (суммирующий тест).

Принципы и механизмы разработки одинаковы для всех видов этих тестов, но содержимое материала, включённого в тест, и степень сложности вопросов должны соответствовать целям тестирования.

Предварительный определяющий тест предназначен для оценки начальных способностей, обычно является несложным и охватывает очень небольшой диапазон знаний. Он может затрагивать минимум базовых знаний по теме обучения или другой ограниченный набор требуемых знаний. Он практически не отличается от суммирующего теста, даваемого в конце курса или раздела обучения.

Формирующий тест, используемый для контроля за прогрессом обучения, затрагивает ограниченный сегмент обучения, например, раздел или главу. С его помощью делается попытка оценить все важные результаты изучения данного сегмента. Акцент делается на оценке степени владения материалом и обеспечения обратной связи со студентом по корректировке отдельных ошибок в тех областях, где он не достиг успехов. Таким образом, формирующий тест состоит из серии отдельных тестовых вопросов, всесторонне охватывающих ограниченную область обучения. Он разрабатывается так, чтобы дать ученику конкретные инструкции для исправления обнаруженных в результате теста ошибок. Данные теста являются обучающими, они обычно менее сложны, чем суммирующие тесты, даваемые в конце процесса обучения.

Диагностический тест содержит относительно большое число вопросов, имеющих отношение к конкретной тестируемой области. Поскольку целью теста является определение трудностей обучения, внимание сосредоточивается на ответах учащихся на конкретный вопрос или группу вопросов, общий балл имеет второстепенное значение. Этот тест обычно больше фокусируется на распространённых ошибках учащихся, чем на попытке широкого отбора ожидаемых результатов обучения. В связи с тем, что тесты данного типа разработаны для тех учеников, у которых есть проблемы в обучении, они обычно имеют очень невысокий уровень сложности.

Суммирующий тест разрабатывается для оценки широкого диапазона результатов обучения, ожидаемого в конце учебного процесса. Сложность и представительность выборки являются важными аспектами данного теста, так как его результаты используются для простановки баллов и определения степени достижения задач курса обучения. Для того чтобы адекватно отобрать все ожидаемые результаты обучения, суммирующий тест обычно содержит вопросы более высокого уровня сложности, чем другие виды тестов.

Для нужд мониторинга можно использовать три из приведённых видов тестов. Если нас интересует динамика подготовленности учащихся на начальном этапе обучения, мы должны использовать определяющие тесты; если нас интересуют трудности в обучении, то мы будем использовать диагностические тесты; а если (и это наиболее вероятно и часто используемо) наш интерес состоит в оценке результатов обучения, мы применяем суммирующие тесты.

Необходимо подчеркнуть тот факт, что каждый из этих видов тестов имеет свои особенности. Использование одних тестов вместо других может привести к негативным последствиям.

Подходы и этапы разработки тестов учебных достижений

В этом разделе мы намерены обсудить зависимость разработки тестов от тех целей,

которые стоят перед разработчиками систем мониторинга.

Прежде всего нужно остановиться на двух подходах, которые в настоящее время сложились в тестировании — тестах, ориентированных на критерий (критериально-ориентированных), и тестах, ориентированных на норму (нормативно-ориентированных). Появившиеся как разные подходы к анализу результатов тестирования, отражающие разные основания для сравнения, сейчас эти два подхода определяют разницу на всех этапах создания теста.

В самом общем виде основанием для сравнения в тестах, ориентируемых на норму, являются результаты, полученные при предварительном тестировании группы учащихся, репрезентативной для какой-то общности. Например, предположим, что при итоговом тестировании по математике учащихся 8-го класса, углублённо изучающих историко-краеведческие дисциплины, был предложен тест из 70 заданий. Среднее количество заданий, с которыми справились учащиеся данной выборки, составило 33. Используя этот инструмент, мы провели тестирование учеников класса сходного профиля и выяснили, что ученик Петров справился с 33 заданиями. Оценивания этого ученика на основе нормы, мы можем сказать, что половина учеников справляется лучше, а половина — хуже Петрова. Аналогичную по подходу оценку можно дать и учащимся, которые выполнили другое количество заданий. В разделе “Нормирование” мы рассмотрим более подробно, каким образом даются эти оценки и как возможно выставление корректной оценки в школьных баллах. Сейчас же для нас важно, что оценка в рамках этого подхода даётся на основе предварительно полученных статистически обоснованных норм. Возможен и ещё один способ, когда оценка даётся относительно места ученика в группе (5-й из 40 или 27-й из 150 и т.д.). В этом случае нет необходимости получения предварительных норм, но отсутствует и возможность получения корректного сравнения для разных групп, что так важно для мониторинга.

Характеризуя подход, ориентированный на критерий, Гронлунд пишет: “Результаты второго типа тестов обрабатываются с точки зрения специальных знаний или навыков, которые студент может продемонстрировать (например, “он может определить все части микроскопа и продемонстрировать их правильное использование”). Он даёт возможность определить, что каждый студент может сделать с точки зрения конкретной задачи, не соотнося его действия с действиями других членов группы”. Критерий определяется на основе экспертного оценивания как по номенклатуре, так и по критическому уровню. Например, специалисты по русскому языку создают тест для оценки уровня владения определениями по теме “Части речи”. Они выясняют, какие определения включить в содержание теста, как оценивать ответы и тот уровень, превысив который можно считать, что ученик владеет определениями в достаточной степени. Включив в тест 18 определений, они говорят, что для получения положительной оценки достаточно дать правильный ответ на 12 из них. Этот подход даёт оценку только по дихотомической шкале: справился — не справился, прошёл — не прошёл, зачёт — незачёт и т.д. Однако он имеет широкие возможности для описания тех задач, с которыми ученик справляется индивидуально, тех задач, с которыми справляется меньше всего учащихся. Такая ориентация даёт возможность реализовать большие диагностические возможности этого подхода.

Сравнение подходов приведено в таблице:(См. табл.1)

Таким образом, приступая к построению системы мониторинга, мы должны определить цели и выбрать соответственно им подходы и область применения планируемого тестового инструментария.

Вариант и способ создания тестов будет зависеть и от того, как широко мы планируем использовать созданный тестовый инструмент.

Общий полный перечень этапов создания тестового инструментария представлен следующим списком:

1. Определение целей тестирования.
2. Определение ресурсных возможностей разработчиков.

3. Отбор содержания учебного материала.
4. Конструирование технологической матрицы.
5. Составление тестовых заданий.
6. Построение выборки для апробации заданий и тестов.
7. Компоновка заданий для апробации.
8. Апробация тестовых заданий.
9. Определение и расчёт показателей качества тестовых заданий.
10. Отбраковка заданий и составление теста.
11. Апробация теста.
12. Определение и расчёт показателей качества теста.
13. Составление окончательного варианта теста.
14. Стандартизация теста.
15. Нормирование теста.
16. Оснащение теста.

Этот список полный, поскольку в нём представлены все этапы создания тестов и общий, поскольку подходит для создания большинства видов тестов.

Создание теста начинается с определения целей тестирования. Конечно, для тестов, которые предполагается использовать для сравнения результатов между несколькими классами одного образовательного учреждения, и тестами, предназначенными для итоговой аттестации учащихся, существует значительная разница. Схема этапов составления тестов учебных достижений для разного применения представлена в таблице: (См.Табл. 2):

Требования к тестам

Попытаемся сформулировать достаточно обобщённые требования и правила составления тестов, использование которых возможно в рамках мониторинга. У нас нет намерения привести здесь полное руководство по созданию тестов, поэтому в тех разделах, которые требуют специальных знаний или информации, мы ограничимся качественными и надёжными ссылками. Здесь же представлен материал, который может быть полезен заказчикам систем мониторинга, руководителям образования и информационных служб, для того чтобы они могли иметь представление о возможностях тестирования и трезво оценивать его потенции. Конечно, при этом не удастся уйти и от некоторых достаточно специальных вопросов, которые руководителям могут показаться излишними, однако они будут чрезвычайно полезными разработчикам тестового инструмента для мониторинга.

Отбор содержания образования.Технологическая матрица

Выяснив цели составления тестов, уточнив подходы и выбрав уровень использования, разработчик определяет необходимые этапы создания тестового инструментария. После этого необходимо отобрать содержание образования, то есть составить модель объекта педагогического тестирования. Она может быть представлена в виде технологической матрицы. Технологическая матрица задаёт содержание, которое будет отобрано для проверки, и важность того или иного элемента содержания. Она может содержать уровни достижений, которые будут проверены, их соотношение, соответствие стандарту и некоторые другие компоненты.

В инструкции по составлению тестов NEAB (Northen Examinationes and Assesment Board) записано: “При имеющемся предмете тестирования разработчик обязан убедиться, что весь предмет охвачен предлагаемыми вопросами. Содержание предмета должно полностью покрываться матрицей по всем темам. Если же имеет место тестирование по отдельным подтемам, то и в этом случае необходимо, чтобы вся подтема была охвачена вопросами теста. В случае, если вопрос или часть вопроса не соответствует теме или не полностью ясна в рамках данной темы, от вопроса следует воздержаться”.

Таким образом фиксируется требование широты теста, полного учёта всех разделов предмета, который находит выражение в матрице.

Для тестов, ориентированных на критерии (критериально-ориентированных), отбор содержания теста является самым важным этапом создания, так как для принятия решения о достижении данной цели обучения, например стандарта, необходимо достаточно точно и полно описать содержание стандарта и выразить его представительной совокупностью заданий. Поэтому главной проблемой в разработке тестов, используемых для оценки достижения образовательных стандартов, является соотношение содержания стандарта и содержания теста.

В самом простом случае технологическая матрица может описывать только предметы, предметные области или отдельные темы разного уровня обобщения, которые должны войти в тест, и определять соотношение заданий в тесте. Например, (см. табл. 3).

Цифры представляют собой соотношение заданий в тесте. Они могут получаться, исходя либо из достаточно формального показателя, выделяемого на изучение той или иной темы времени, либо из важности той или иной темы. На этом этапе можно ощутить разницу между тестами, предназначенными для вступительных экзаменов в вузы, и выпускными экзаменами в образовательном учреждении. В первом случае материал будет отбираться по принципу его важности для продолжения образования в конкретном вузе или просто продолжения образования, во втором случае важность будет определяться в рамках всех тем, которые изучались в школе. Соответственно, между этими тестами будет существенная разница. В ряде стран такая разница отсутствует, но для этого им потребовались десятилетия согласований и уточнений. В наших современных условиях использовать результаты тестирования, полученные при вступительных экзаменах, в качестве выпускной оценки, по крайней мере, некорректно.

Представленная в качестве примера матрица не отображает ни уровней достижения учащихся, ни уровней овладения материалом.

Более сложные технологические матрицы содержат две шкалы и оформляются в виде таблицы. Например, в международном исследовании IAEP-II для сравнительной оценки естественнонаучной подготовки школьников использовалась двухмерная система заданий, представленная в таблице 4.

В соответствии с данной системой каждое задание теста предназначалось для проверки овладения учащимися определёнными умениями, характеризующими отдельные компоненты познавательной деятельности (воспроизведение, применение и интеграцию знаний) на материале различных разделов естествознания (биологии, физики и химии, наук о Земле и астрономии, методологии науки).

По сути, технологическая матрица представляет собой содержательно-деятельностную модель теста. Содержание горизонтальной строки матрицы, как уже было указано, не должно представлять трудностей для педагогов. Здесь, как правило, определяются предметы, разделы, учебные темы, разделы учебных тем. Выбор того или иного материала напрямую зависит от целей тестирования.

Сложнее обстоит дело с вертикальной составляющей. В западных или международных тестах она строится, как правило, на основе той или иной таксономии, представляющей собой некоторую реализацию идеи В.П. Беспалько о том, что цели должны быть сформулированы технологично.

Для реализации идеи второй шкалы необходимо выполнение простого правила: для отнесения задания к той или иной шкале необходимо взаимно однозначное соответствие конкретного тестового задания и уровня (или свойства, умения и пр.) той графе матрицы, к которой это задание отнесено.

К сожалению, в данной области на сегодня можно зафиксировать некоторый тупик. У нас нет отечественных разработок уровней обученности, которые достаточно однозначно могли бы восприниматься педагогическим сообществом, а использование западных разработок невозможно в силу различных подходов к оцениванию заданий и терминологиче-

ской неопределённости.

На уровне выделения групп педагогических целей ситуация достаточно однозначна: и наши, и зарубежные авторы в своих подходах достаточно единодушны (см. табл. 5).

Несмотря на терминологическую разницу, по содержанию области, выделяемые разными исследователями, близки между собой. К первой относят знание, различные уровни его усвоения. Ко второй — умения со своей иерархией подцелей. Наконец, к третьей — отношения, интересы, склонности, ориентации.

Проводя дальнейшую конкретизацию целей-результатов, многие исследователи выделяют уровни усвоения. Проведём анализ на примере первой области. Сравнительная таблица уровней усвоения, описанных разными авторами, представлена ниже (см.табл.6).

Если достаточно детально проанализировать эту таблицу, то становится ясно, что во всех работах речь идёт об одних уровнях, которые, вероятно, существуют реально (особенно хорошо это видно на примере трёх первых уровней).

В.П. Симонов, М.Н. Скаткин, В.В. Краевски и Б. Блум проводят дальнейшее уточнение и конкретизацию представленных в таблице уровней, правда, различные по качеству и объёму. В принципе можно было бы использовать любую из отечественных разработок, однако состояние педагогического сообщества таково, что согласовать использование того или иного уровня на сегодня не представляется возможным.

Использование таксономии Блума ограничено другим — в частности, разным подходом к анализу заданий. Для иллюстрации приведём пример, взятый из международного исследования IAEP-II: (См. рис.)

Попытка отнесения этого задания к одному из уровней таксономии Блума приводит в группе педагогов сначала к возникновению нескольких мнений. В процессе обсуждения вырабатывается согласованная позиция о том, что задание необходимо отнести к уровню “анализ”. Выбор мотивируется тем, что учащемуся для выполнения этого задания необходимо проанализировать строение тела птиц и сравнить их с требованиями условия. Анализ наших педагогов в любом случае проходит через ту деятельность, которую необходимо выполнить ученику в процессе решения той или иной задачи. Это в корне отличается от оценок западных специалистов, которые, разнося задания по уровням, ориентируются только на содержание самого задания. Представляется, что именно по причине такого несовпадения использование таксономии Блума в нашей стране крайне ограничено.

Для большинства тестов вполне достаточно одной качественно отработанной шкалы. Однако есть другие возможности использовать вторую шкалу технологической матрицы. Для более сложных тестов в качестве второй составляющей могут выступать:

— уровни овладения учебным материалом;

— специальные или общешкольные умения и навыки;

— уровень развития психических познавательных процессов и некоторые другие, в зависимости от целей тестирования.

Таким образом, технологическая матрица представляет собой достаточно универсальный инструмент отбора содержания и различных областей достижений учащихся для построения тестов учебных достижений.

Технология согласования

Подготовка инструментария требует привлечения значительного количества профессионалов, координацию между которыми необходимо наладить. В общем виде перед организаторами системы мониторинга стоит задача реализации идеи стандартизации в той форме, как её понимает международная организация по стандартам, — **нахождения способа решения повторяющихся задач в пользу и при участии всех заинтересованных сторон.**

Успех реализации мониторинга в значительной мере зависит от нахождения механизма, который может обеспечить эту формулу — “в пользу и при участии”. В этой связи требуется разработка организационного механизма (технологии) согласования.

Обычная административная система управления предусматривает наличие иерархической структуры соподчинения и выстраивается сверху вниз. На уровне Министерства образования (федеральной системы образования) разрабатывается образовательный стандарт, который доводится до региональных органов управления. Последние, снабдив его региональным компонентом, обращаются в одну или несколько подчинённых им учебно-методических организаций и поручают разработать систему контроля за выполнением этого стандарта. В учебно-методической организации выполняется некоторый комплекс работ (условимся, что выполняется качественно, хотя ответственности за качество работ никто не несёт). Результат этих работ утверждается региональным органом управления образованием и после этого становится нормативным документом, обязательным для выполнения всеми образовательными учреждениями и районными органами управления образованием. Хотя мы описали этот процесс начиная с министерства, он может инициироваться и на более низком уровне, причём суть его не меняется. Руководитель образовательного учреждения будет относиться к данной разработке с той степенью доверия, с которой он относится к органу управления, а педагог образовательного учреждения — с ещё меньшей.

Данная схема не удовлетворяет условиям, соответствующим характеру работ. Нам необходимо спроектировать комплекс социальной активности достаточно большого числа людей — от методологических оснований до конкретных работ, что невозможно сделать, оставаясь в рамках административного подчинения.

Для нас представляется несущественным (в отличие от административной схемы), в какой организации или каком учреждении работает тот или иной участник проекта. Внутри проекта он выступает как профессионал, ответственный за взятые на себя обязательства. Такой подход помогает преодолеть местнические интересы и неприятие, существующие между некоторыми педагогическими организациями разного подчинения, позволяет привлечь к выполнению работ профессионалов самого высокого уровня.

Реализация этих установок в проекте происходит через принципы добровольности участия в исследовании и выхода из эксперимента в том случае, если действия, которые реализуются в исследовании, противоречат нравственным установкам профессионалов. Поэтому первое спроектированное действие — приглашение профессионалов (напомним, вне всякой зависимости от их места работы и проживания).

Другой отличительной особенностью технологии согласования является то, что к обсуждению на различных этапах реализации проекта привлекаются практически все участники. Участие в работе над проектом даёт важное чувство причастности, авторства. Когда учитель получает в руки готовый документ, он относится к нему как к продукту собственной деятельности, а не как к навязанному, подозрительному инструменту.

Ещё одна особенность — процедура принятия решения. Решение могло быть принято только на основе полного согласия участников программы. Именно это обеспечивало реальное участие и заинтересованность всех участников, что так необходимо и важно для данной технологии.

Дальнейшие действия осуществлялись в соответствии с выбранной стратегией, за исключением редких случаев отступлений, продиктованных несоответствием реально сложившейся системы управления предполагаемому механизму реализации.

Следующим шагом проектирования было выделение субъектов взаимодействия внутри проекта:

Субъекты взаимодействия. Административный вариант

КО — Комитет по образованию

РОНО — Районный отдел народного образования

НМЦ (УМЦ) — Научно-методический центр

ОУ — Образовательное учреждение

УПМ — Университет педагогического мастерства

РГПУ — Российский государственный педагогический университет

ИОВ — Институт образования взрослых

Вузы

Административный вариант предусматривает выделение субъектов с точки зрения существующих административных и научно-методических структур. Оказалось, что такой вариант не соответствует целям и задачам создания тестов в силу неоднородности профессионального состава и отсутствия позитивных связей между некоторыми организациями. Поэтому была предложена другая схема определения субъектов взаимодействия, в основе которой лежат те или иные профессиональные позиции. Именно этот подход оказался реально подходящим и соответствующим установкам мониторинга.

Субъекты взаимодействия. Окончательный вариант

• педагоги
• методисты
• тестологи
• руководители (управленцы)
• родители
• учащиеся
• информатики
• издатели

Схематично модель согласования материалов и процедур эксперимента можно представить следующим образом: (См. рис.)

На основе модели согласования был создан механизм координации взаимодействия. Его суть показал пример разработки тестового инструментария, представленный в матричной форме (См. табл.).

Заполнение этой формы позволило наглядно представить всю структуру взаимодействия и рационально описать модель субъектов взаимодействия, создав схему работ.

Тестовые задания

В данном разделе представлены требования к основным видам тестовых заданий, которые применяются в тестах учебных достижений и носят “базовый” характер.

В самом общем виде тестовые задания должны:

— быть составлены с учётом соответствующих правил;
— соответствовать содержанию учебного материала;
— быть проверены на практике (апробированы);
— иметь рассчитанные показатели качества — сложность и дискриминативность;
— быть максимально разнообразными по форме.

Несмотря на многообразие видов заданий, все они могут быть сведены к нескольким типам или их сочетанию. Например, задача с переструктурированием данных может быть представлена как совокупность задач последовательности и соответствия; задания на нахождение ошибок — как частный случай заданий на исключение лишнего и так далее.

С точки зрения разработчика, минимальные требования к тестовому заданию заключаются в наличии трёх частей:

1. Инструкции.
2. Текста задания (вопроса).
3. Правильного ответа.

Дадим краткое описание этих частей.

1. Инструкция должна указывать на то, что должен сделать испытуемый, каким образом выполнять задание, где и как делать пометки и записи, описывать то, что ученик должен сделать руками.

Например:

— “ответ запишите в рамку, которая находится ниже задания. Для промежуточных вычислений используйте место слева от вопроса…”;

— “в третьем столбце над строчками впишите цифры, соответствующие понятиям, обо-

значенным буквами в этой же строке…”;

— “используя калькулятор, проведите вычисления, ответ запишите в бланке в строке 4…” и т.д.

В тестах допускается делать одну инструкцию для группы однотипных заданий и помещать её в начале данной группы заданий. Для проверки того, как испытуемые поняли инструкцию, желательно снабдить её несколькими примерами, которые разбираются вместе с проводящим тестирование.

2. Текст задания или вопроса представляет собой содержательное наполнение задания.

3. Правильный ответ — обязательный атрибут любого тестового задания. Без него задание, за исключением, пожалуй, самых тривиальных, теряет смысл, поскольку не может быть точно проанализировано и оценено с учётом авторского замысла.

Перечисленные части тестового задания являются минимально необходимыми для составления тестов.

Кроме того, составителям тестовых заданий целесообразно указывать:

— возраст (класс), на который рассчитано задание;

— тему (предмет или предметную область, в соответствии с технологической матрицей);

— предполагаемое автором время выполнения задания;

— сроки предъявления (календарные сроки, поскольку одно и то же задание, будучи предъявленным, например, в октябре и феврале, даст разные результаты и соответственно должно иметь разные характеристики);

— предполагаемую статистическую сложность;

— уровень, который соответствует данному заданию, или умения, которые оно выясняет;

— соответствие стандарту или программному материалу;

— данные составителя;

— возможные варианты невербальной поддержки;

— некоторые другие сведения, содержание которых определяется, как правило, специфическими целями создания данного инструмента. Как правило, для составителей заданий готовятся специальные бланки, в которых формализуется требуемая информация.

Существует основное требование к тестовым заданиям:

Тестовое задание должно иметь однозначный правильный ответ.

Данное правило требует пояснения. Часто понятие однозначности ответа трактуется как требование единственности или наличия предполагаемого образца. В данном случае однозначность понимается как возможность любого пользователя на основе сравнения ответа учащегося и правильного ответа, предложенного разработчиком, сделать однозначный вывод о том, выполнил данный ученик это задание верно или нет. Поэтому правильный ответ разработчика может заключаться не только в эталонном ответе, но и в описании схемы анализа, содержать конструкции “и, … и”, “…или…”, описывать вариант неправильного ответа, считая все остальные правильным. В инструкции NEAB записано: “Ясная схема оценки должна обеспечить пользователя теста аппаратом оценивания именно в рамках заложенной в тест оценки разработчика. Многие вопросы толкования могут быть сняты при разработке ясной и недвусмысленной схемы оценивания, которая содержит наиболее возможные варианты ответов, которые можно принять к рассмотрению и оценить как зачётные. Схема оценивания должна полностью соответствовать конкретному вопросу. Все формулировки ожидаемых ответов должны быть предельно ясными и недвусмысленными, чтобы при оценивании у проверяющего не могло возникнуть сомнения в правильности засчитываемого ответа. Единство требований к тестируемым может толковаться специалистами по-разному”.

Рассмотрим, типо-видовую классификацию тестовых заданий и выделим требования к ним. Существует два типа заданий, которые объединяют пять видов. К этим видам может

быть сведено всё многообразие существующих заданий без ущерба для их качества. Типы и виды тестовых заданий представлены на схеме: (См. рис.)

К заданиям закрытого типа относят задания трёх видов: альтернативных ответов (на схеме АО), множественного выбора и восстановления соответствия. Тестовые задания закрытого типа предусматривают различные варианты ответа на поставленный вопрос: из ряда предлагаемых выбирается один (или несколько) правильный ответ, выбираются правильные (или неправильные) элементы списка и др. Это задания с предписанными ответами, что предполагает наличие предварительно разработанных вариантов ответа на заданный вопрос.

1. Задания альтернативных ответов (АО).

К каждой задаче этого вида даётся только два варианта ответов. Испытуемый должен выбрать один из них — “да” или “нет”, “правильно” или “неправильно” и пр.

Форма задания:

утверждение 1 да нет
утверждение 2 да нет
утверждение 3 да нет
утверждение 4 да нет
и т.д.

Инструкция для заданий альтернативных ответов: **обведите кружком вариант ответа “да” или “нет”, который вы считаете правильным.**

Задания АО являются самыми простыми, но не самыми распространёнными при составлении тестов. Это связано в основном со специфичностью того материала, которому в большей степени соответствует форма заданий. Задания альтернативных ответов применяются для оценки одного элемента знаний. Их использование в виде отдельного вопроса приводит, как правило, к тривиальному тестированию и применяется достаточно редко. Эта форма целесообразна для использования в серии, когда для одного элемента знания задаётся несколько вопросов. В такой форме задания альтернативных ответов в большей степени подходят для выявления уровня овладения сложными определениями, знания достаточно сложных графиков, диаграмм, схем и т.д.

2. Задания множественного выбора. Это основной вид заданий, применяемый в тестах достижений.

Задачи со множественным выбором предполагают наличие вариативности в выборе. Они состоят из двух частей: формулировки задания и вариантов ответов. Испытуемый должен выбрать один из предложенных вариантов, среди которых чаще всего только один правильный. Однако задачи формулируются так, чтобы в них было не менее 3 правдоподобных, похожих на правильные ответов. Именно этим обеспечивается независимость результатов от случайного выбора.

Форма представления заданий:

Вопрос (утверждение):
A. Вариант ответа 1.
B. Вариант ответа 2.
C. Вариант ответа 3.
D. Вариант ответа 4.
E. Вариант ответа 5.

Инструкция для заданий множественного выбора: **обведите кружком букву, соответствующую варианту правильного ответа.**

3. Задания на восстановление соответствия.

В задачах соответствия (восстановления соответствия) необходимо найти или приравнять части, элементы, понятия конструкциям, фигурам, утверждениям; восстановить соответствие между элементами двух списков. К этому же типу следует отнести и задания, в которых требуется восстановить порядок ряда, упорядочить. Предложенные элементы задания могут рассматриваться как частный случай заданий на восстановление соответ-

ствия, в которых только один ряд.

Данный вид заданий имеет достаточно много модификаций, от которых зависят инструкции. Наиболее распространённой формой ответа, которая реально применяется педагогами, особенно в начальной школе, является вариант с использованием стрелочек: нарисуйте стрелочки от элементов первого списка ко второму, соедините стрелками соответствующие понятия и т.д. Сам по себе способ с использованием стрелочек вполне правомерен, однако он имеет два существенных недостатка: первый — сложность проверки, особенно когда необходимо проверить большое количество работ; и второй — опасность того, что ученики, привыкнув к рассматриваемому способу и встретив в дальнейшем классическую форму задания, воспримут её как неизвестную, что может снизить их результаты.

Форма представления:

Ряд 1	**Ряд 2**	**Место для ответов**
1.	**A**	_____
2.	**B**	_____
3.	**C**	_____
4.	**D**	_____

Инструкция: **“в графу для ответов впишите цифры ряда 1, соответствующие ряду 2”.**

Задания открытого типа. К ним относятся задания двух видов:

— свободного изложения (свободного конструирования). Они предполагают свободные ответы испытуемых по сути задания. Ограничения на ответы не накладываются, однако формулировки заданий должны обеспечивать наличие только одного правильного ответа.

— дополнения (задачи с ограничением на ответы). В этих заданиях испытуемые также должны самостоятельно давать ответы на вопросы, однако их возможности ограничены. Ограничения обеспечивают объективность оценивания результата выполнения задания, а формулировка ответа должна дать возможность однозначного оценивания.

Инструкция для заданий дополнения: **вместо каждого многоточия впишите только одно слово (символ, знак и т.д.); вместо многоточия впишите нужное слово; выпишите на бланк слова, которые пропущены в тексте; вместо многоточия впишите нужный символ; запишите ответ в отведённое место,** то есть один пропуск подразумевает одно слово (знак, символ, выражение).

Инструкция для заданий свободного изложения: **закончите предложение (фразу), впишите вместо многоточия правильный ответ; дополните определение, записывая ответ в бланке, и т.д.**, то есть вместо многоточия можно вписать словосочетание, фразу, предложение или даже несколько предложений.

Выполнение основного требования для заданий дополнения не представляется сложным, правильным ответом будет то самое выражение, слово и т.д., которое необходимо вписать испытуемому.

Для заданий свободного изложения выполнение основного требования к тестовым заданиям сложнее. Для этого необходимо формализовать сам ответ. В том случае, когда результатом выполнения задания служат цифровые выражения, структура фразы подразумевает два-три однозначных слова — это не сложно.

Для других случаев возможно:

а) выделить в ответе ключевое слово или фразу и в зависимости от их наличия оценивать ответ как правильный или неправильный;

б) выделить несколько смысловых, фиксируемых элементов и ранжировать на их основе правильные ответы. Например, “ответ на данный вопрос оценивается двумя баллами, если он содержит слова “геологической” и “организмами”; если он содержит одно из этих слов, то — одним баллом; во всех остальных случаях задание считается невыполненным”;

в) если предложенные варианты на подходят, использовать для задания иную форму.

Специфические требования к заданиям открытого типа:

— использовать не более трёх пропусков подряд, лучше один-два;

— дополнять наиболее важное понятие, определение, знание которого нужно проверить;

— дополнения лучше ставить в конце предложения.

Достаточно часто в тестах достижений можно найти попытки использования специфичных заданий, специально разработанных психологами для тестов интеллекта. Это в основном три вида заданий: *аналогии, классификации и исключения лишнего*. Их особенность заключается в том, что результат выполнения зависит не только от знания предметного содержания задания, но и от той интеллектуальной операции, выполнение которой предполагает данное задание. Как говорят психологи, эти задания нагружены разными факторами, один из которых — собственно результаты обучения, а другой отражает личностные особенности испытуемого. Поэтому использовать их в тестах нужно очень аккуратно, а лучше вообще отказаться в пользу нейтральных форм заданий.

По форме это могут быть задания как открытого, так и закрытого типа.

Дадим краткую характеристику каждому из этих трёх видов заданий.

Задания "аналогии". Форма представления: **А так относится к В, как С относится к …?** Задания аналогии имеют сокращённую форму записи, которая применяется тогда, когда задания представлены серией и нет необходимости повторять инструкции для каждого задания: А:В=С:?

Задания исключения лишнего (отношения и связей; "встретил лишнее — убери"). В таких заданиях испытуемому предъявляется список объектов, слов, фигур, чисел и т.д. — всего, что только может придумать разработчик тестов. Испытуемый должен найти общие закономерности отношения между элементами списка и на их основании сделать заключение о подобии или различии предложенных объектов. Очевидно, что при этом необходимо выявление отношений и связей.

Задания "последовательности". В них от учащегося требуется продолжить ряд, добавить элемент ряда в начало или середину. Наиболее известные задания этого типа — числовые последовательности.

Для правильного составления тестовых заданий важно соблюдать следующие требования:

— вопрос должен содержать одну законченную мысль;

— при составлении вопросов следует особенно внимательно использовать слова "иногда", "часто", "всегда", "все", "никогда", которые, с одной стороны, сами по себе содержат неопределённость и могут пониматься субъективно, что может приводить к ошибочным ответам, а с другой стороны, дают возможность учащимся догадываться о правильном ответе;

— вопрос должен формулироваться чётко, избегая слов "большой", "небольшой", "малый", "много", "мало", "меньше", "больше" и т.д.;

— чаще использовать количественные термины;

— избегать вводных фраз или предложений, имеющих мало связи с основной мыслью; не следует прибегать к пространным утверждениям, так как они приводят к правильному ответу, даже если учащийся его не знает;

— число ответов "да" и "нет" в тесте должно быть приблизительно равным, что исключает тенденцию отвечать одинаково на все вопросы;

— не следует задавать вопросы с подвохом (скорее всего, в заблуждение будут введены наиболее способные или осведомлённые учащиеся, которые знают достаточно много для того, чтобы попасться в ловушку. Кроме того, это противоречит цели — определению уровня знаний и понимания);

— лучше использовать один вариант правильного ответа, и если инструкция требует выбрать правильный ответ, то таковым должен быть только один. В противном случае в инструкции необходимо указать, что правильных ответов несколько;

— все варианты ответов должны быть грамматически согласованы с основной ча-

стью задания;

— неправильные ответы должны быть разумны, умело подобраны, не должно быть явных неточностей;

— как можно реже использовать отрицание в основной части, особенно — многократно в одном предложении. С одной стороны, это приводит к противоречиям при чтении задания, с другой — отрицательные знания не так важны, как позитивные;

— ответ на поставленный вопрос не должен зависеть от предыдущих ответов;

— место правильного ответа должно быть определено таким образом, чтобы оно не повторялось от вопроса к вопросу, не было закономерным, а давалось в случайном порядке;

— правильные и неправильные ответы должны быть однозначны по содержанию, структуре и общему количеству слов;

— если ставится вопрос количественного характера, то ответы к нему должны располагаться упорядоченно от меньшего к большему или наоборот;

— лучше не использовать варианты ответов "ни на один из перечисленных" и "все перечисленные". Применение первого целесообразно, когда существует недвусмысленный правильный ответ. Второй приводит к допустимости подбора вариантов с низкой дискриминативностью, поскольку разработчик знает, что все ответы правильные;

— лучше использовать длинный вопрос и короткий ответ. В противоположной ситуации на прочтение ответов уходит больше времени и больше сил тратится на анализ высказываний. Это противоречит поставленной в данном случае цели — выявлению усвоенных учащимся заданий.

Национальные системы ряда стран ставят специфические требования к тестовым заданиям, на которые у нас пока обращается недостаточно внимания. В качестве примера приведём требование инструкции NEAB по составлению тестов: *"Необходимо избегать вопросов, которые в каком-либо виде дают превосходство тестируемому определённого пола. Половой ориентации вопроса необходимо избегать в любом случае. Нельзя считать, что формулировка вопроса в мужском роде подразумевает лёгкость ответа в женском роде. Использование формулы "он/она" также нежелательно при формулировании вопросов. Лучше пользоваться неродовыми формулировками, типа "учащиеся", "школьники", а не "школьница", "учащийся". Лучше обращаться к группе, а не к отдельному учащемуся. Необходимо избегать и половых стереотипов типа: "Доктор — очень уважаемая профессия, он …". Следует предлагать нейтральную формулировку: "Врачи — люди уважаемой профессии, они …" ...Необходимо избегать в вопросе любой возможности его культурного толкования. Вопрос должен легко восприниматься человеком любого культурного слоя".*

После того как задания будут составлены, необходимо их упорядочить. Для этого существуют составленные на основе работы П. Клайна правила:

1. Составьте базу данных для заданий, при этом представьте каждую задачу на отдельном листе, предусмотрев место для занесения экспертных оценок, времени, необходимого для её выполнения, уровня сложности и прочих характеризующих её данных.

2. Проверьте содержание и формулировку задач во взаимосвязи друг с другом.

3. Располагайте задания каждого типа вместе. Инструкцию и пояснения необходимо давать один раз для каждой группы заданий. Это даёт возможность испытуемым приспособиться к данному типу заданий.

4. Располагайте задания в порядке возрастания предполагаемой трудности. Это предотвратит случаи, когда слишком старательный испытуемый тратит всё своё время (или большую его часть) на задания, которые он не может решить, и таким образом лишает себя возможности выполнить другие, по которым он мог бы получить баллы, а в результате все формы анализа заданий будут неточными. При апробации теста бывает полезно включить в инструкцию пункт о том, что если испытуемому не удаётся справиться с заданием, его

необходимо пропустить, а после окончания работы, если останется время, вернуться к вызвавшему трудность заданию.

5. Не комплектуйте вместе такое количество заданий, для выполнения которых среднему испытуемому потребуется более получаса — для детей начальной школы; для старшеклассников — более часа. (Примерно столько длится период сосредоточения у детей.) Если существует необходимость выполнения заданий большей продолжительности, технология проведения должна предусматривать перерыв. Необходимо отметить, что время появления утомления во многом зависит от мотивации (при этом слишком высокая и слишком низкая мотивация быстрее вызывают утомление), разнообразия материалов тестирования, способа проведения, эмоциональной подготовленности учеников.

6. При конструировании бланковых тестов лучше размещать задачи на листах брошюры так, чтобы они были пространственно разнесены и легко воспринимались. Задания и варианты ответов к ним должны располагаться на одной странице.

7. Важные части инструкции должны быть подчёркнуты или выделены особым шрифтом. Сделайте бланки ответов. Размножьте брошюры и бланки. Можно считать, что для проведения апробации всё готово, за исключением одного — необходимо выбрать контингент, на котором будет испытан тест. Правила построения выборки мы обсуждали в предыдущей главе. Для профессиональной работы необходимо обратиться к одному из изданий, приведённых в конце книги.

Проверка трудности задач. Определение места задачи в тесте.

Важным шагом в конструировании теста является проверка трудности предложенных задач. Для этого необходимо провести предварительное тестирование экспериментальной группы (выборки). После того, как определён состав выборки, испытуемым предлагается решить составленные задачи. Полученные ответы анализируются с целью установления трудности, обоснованности и дискриминативности каждого вопроса, пригодности каждого варианта ответов. Результатом анализа становится отбор и корректировка задач, а также их перераспределение внутри теста.

Трудность задачи является важнейшей характеристикой, определяющей её место в тесте. Трудность может быть субъективной и статистической.

Субъективная трудность задачи связана с индивидуально-психологическим барьером учащихся. В психологии величина этого барьера определяется различными факторами, в том числе:

1. Условиями решения задачи (временем, отведённым на решение, понятностью инструкции т.п.).
2. Уровнем формирования необходимых для решения знаний, умений и навыков.
3. Состояние испытуемого и т.д.

Для снижения влияния перечисленных факторов определяется стандартная форма процедуры тестирования.

В большинстве случаев для тестов достижений достаточно учитывать только правильность решения задач и меньше внимания уделять способу решения, характеру затруднений, энергетическим затратам испытуемого. В связи с этим определяется и используется статистическая трудность задач.

Статистическая трудность определяется долями решивших и не решивших задачу в выборке. Например, если задачу решили только 20% участников тестирования, то её можно оценить как трудную для данной выборки, если 80% — как лёгкую. При этом значимым является только факт выполнения или невыполнения задания, причины неудач не рассматриваются.

Статистическая трудность позволяет определить место задачи в тесте. Так, если задачу решает большинство испытуемых, то её, как лёгкую, помещают в начале; в том случае, когда с задачей справляется незначительный процент испытуемых, то её, как трудную, помещают в конце теста. **Самые лёгкие задачи (одну-две) выносят перед основными задачами теста и используют в качестве примеров.** Итогом распределения задач по

степени их трудности должна стать “лестница” усложняющихся задач, каждая ступень которой представлена процентом испытуемых, решивших соответствующую задачу.

Подчеркнём, что в тестах достижений трудности задач лучше всего определять в условиях “мягкого” лимита времени или совсем без его ограничения, фиксируя правильность и время решения.

Если трудность задания нормативно-ориентированного теста составляет меньше 20 или больше 80%, то его необходимо переработать или отбраковать. Для тестов, ориентированных на критерий, значение трудности не так существенно.

О статистической трудности заданий критериально-ориентированных тестов Гронлунд говорит: “В случае, когда целью тестирования является определение того, какие из учебных задач студент может осуществить, а не распределение студентов по результатам обучения, диапазон тестовых баллов не является значительным. В этом случае степень сложности вопроса определяется, исходя из сложности учебной задачи, степень освоения которой тестируется, и не делается попытка манипулирования степенью сложности вопроса с целью получения широкого разброса оценочных баллов”.

Определение дискриминативности (дифференцирующей способности) заданий.

Дискриминативность задач определяется как способность отделять испытуемых с высоким общим баллом по тесту от тех, кто получил низкий балл, или испытуемых с высокой продуктивностью учебной деятельности от испытуемых с низкой продуктивностью. Дискриминативность обозначает различительную способность задачи. Для её определения могут применяться коэффициент и индекс дискриминации, формула Фергюссона.

Самый простой и наглядный способ вычисления дискриминативности — применение метода крайних групп, когда при расчёте учитываются результаты учащихся, наиболее и наименее успешно справившихся со всем тестом. Как правило, берут по 27% лучших и худших по результатам выполнения всего теста. Индекс дискриминации вычисляется как разность долей испытуемых из высокопродуктивной и низкопродуктивной групп, правильно решивших данное задание.

$D = (Np_{верх}/N_{верх}) — (Np_{нижн}/N_{нижн})$, где:

$Np_{верх}$ — количество учащихся, верно выполнивших данное задание в группе лучших;

$Np_{нижн}$ — количество учащихся, верно выполнивших данное задание в группе худших;

$N_{верх}$ — общее количество испытуемых в группе лучших;

$N_{нижн}$ — общее количество испытуемых в группе худших.

Если ученики, лучше справившиеся со всем тестом, задание выполняют хуже или так же, как ученики, справившиеся со всем тестом плохо, дискриминативность признаётся неудовлетворительной. Это означает, что задание имеет существенные изъяны.

Типичными недостатками задач, оказывавшихся непригодными, являются:

1. Излишняя сложность, запутанность формулировки.
2. Неоднозначность условия.
3. Очевидность решения.
4. Зависимость результата от памяти или от других индивидуальных особенностей испытуемого, а не от уровня развития тех умений и навыков, для оценки которых разрабатывается тест (кроме заданий, где необходима именно работа памяти).
5. Абсурдность, нереальность вариантов ответов.
6. Появление двух и более правильных ответов, не оговорённое в условии.

Таким образом, дискриминативность ставит заслон некачественным заданиям.

Определение дискриминативности обязательно для тестов, использующихся для отбора учащихся, вступительных экзаменов, итоговой аттестации.

Показатели качества тестов

Требования к тестам как измерительному инструменту содержат требования к расчёту

показателей качеств тестов и требования к их оснащению.

Надёжность

Надёжность теста является одним из критериев его качества и показывает, насколько точно измеряет данный тест изучаемое явление, его “помехоустойчивость”. Она, как правило, определяется после проведения анализа задач и составления окончательной формы теста.

Надёжность характеризует точность теста как измерительного инструмента, его устойчивость к действию помех (состояния испытуемых, их отношения к процедуре тестирования и т.п.). Качественный тест не может быть создан без тщательного изучения этого важного аспекта измерения. Использование ненадёжных тестов, допуск большого количества ошибок в таком ответственном деле, каким является тестирование людей, могут стать причинами педагогических и административных ошибок, последствия которых трудно исправить.

В психологии термин “надёжность” применяется в двух значениях. Во-первых, тест называется надёжным, если он является внутренне согласованным. Во-вторых, тест называется надёжным, если он даёт одни и те же результаты для каждого испытуемого при повторном тестировании. Такая надёжность называется ретестовой.

Для тестов учебных достижений особую важность приобретает ретестовая надёжность, поскольку специфика заданий тестов учебных достижений делает внутреннюю согласованность достаточно прозрачной.

Н. Гронлунд отмечает: “Тесты по оценке результатов должны быть надёжными, и в связи с этим их обработка должна осуществляться очень тщательно. Если балл, полученный учеником в результате теста по оценке результатов, будет соответствовать той оценке, которую они **получили бы** при повторном прохождении того же теста или идентичного с ним по форме, то данная оценка считается высоко надёжной. Все тестовые результаты содержат некоторый процент ошибок (в связи с различием факторов, таких, как условия тестирования или студенческие ответы), но процент ошибок может быть уменьшен путём увеличения количества и усовершенствования качества вопросов, задаваемых в тесте. Чем длиннее тест, тем более надёжными и адекватными будут результаты”.

Надёжность определяется как коэффициент корреляции. Для этого нам необходимо получить два ряда оценок, в которых будут присутствовать результаты оцениваемого инструмента. Результаты должны быть получены в разных условиях. Теоретически может быть всего три варианта получения таких рядов: либо разнесение результатов по времени, либо разделение теста на две части и проведение этих частей на одинаковой выборке учащихся, либо разделение группы учащихся на эквивалентные подгруппы и тестирование их одним инструментом.

На практике используются три основных метода оценки надёжности тестов:

1. Повторное тестирование (ретестирование).
2. Расщепление теста (тестирование параллельной формой теста).
3. Расщепление группы.

Метод повторного тестирования (ретестирование) является основным при определении надёжности психологических тестов, но его применение к тестам достижений ограничено. Этот метод предусматривает повторное тестирование через некоторый промежуток времени. Однако за это время дети успевают подрасти, узнать что-то новое, иногда забыть известное. Таким образом, высокая динамика изменений объекта измерения ограничивает применение данного метода для тестов школьных достижений.

Поэтому при подготовке тестов школьных достижений для использования остаются два способа — разделение теста на части и тестирование эквивалентных групп. Технологию проведения разделения и описание методов расчёта надежности можно найти в литературе, список которой приведён в конце книги.

Источниками неудовлетворительной надёжности тестов могут быть:

— запоминаемость содержания задач и способов их решения;
— интересность и оригинальность задач;
— небольшое количество задач;
— небольшое время между первым и вторым проведением теста;
— причины, связанные с испытуемыми: усталость, скука, невнимательность, жара или холод, самочувствие, различная мотивация и т.д.

Повышение надёжности возможно двумя путями — ужесточением инструкций и повышением качества подготовки экспериментаторов.

Валидность

Однако одной надёжности для обоснования качества теста недостаточно. Ещё одной важнейшей характеристикой теста является его валидность. Валидность особенно важна для тестов, ориентированных на критерий, поскольку определение надёжности для них затруднено. По мнению Гронлунда, "в связи с тем, что традиционные оценки надёжности теста основаны на разнообразии баллов, возникают особые проблемы при разработке надёжного теста, не требующего такого разнообразия баллов, как это бывает в случае с тестами, ориентированными на критерий. В этом случае появляется более сильная зависимость от соответствия тестовых вопросов конкретным учебным задачам, что достигается путём использования достаточного числа вопросов для каждой изучаемой задачи и разработкой письменных вопросов, которые вызывают ожидаемый ответ".

Валидность и надёжность — связанные понятия. В литературе мы находим различные примеры, иллюстрирующие их связь. Вот один из них. Допустим, имеются два стрелка: А и В. Стрелок А выбивает 90 очков из 100, а стрелок В — только 70. Соответственно, надёжность стрелка В — только 0,7. Однако стрелок А всегда стреляет по чужим мишеням, поэтому на соревнованиях его результаты не засчитываются. Стрелок В всегда правильно выбирает мишени. Поэтому валидность стрелка А нулевая, а стрелка В — 0,7, то есть равна надёжности. Если стрелок А станет правильно выбирать мишени, его валидность тоже будет равна его надёжности. Если же он будет иногда путать мишени, то часть результатов не будет зачтена и валидность стрелка А будет ниже надёжности. В этом примере аналогом надёжности является меткость стрелка, а аналогом валидности — точность стрельбы по строго определённой "своей" мишени. В истории тестологии известны случаи, когда тест с низкой валидностью для измерения одних свойств (тех, для которых он создавался) оказывался валидным по отношению к другим. Ненадёжный тест не может быть валидным, и, наоборот, валидный тест всегда надёжен.

Понятие "валидность" очень часто вызывает путаницу не только среди педагогов, но и среди психологов. Причины этой путаницы носят исторический и лингвистический характер.

Валидность определяет, насколько тест отражает то, что он должен оценивать.

В США значение валидности в профессиональном тестировании обычно определяется набором стандартов, подготовленным совместным комитетом из трёх представителей основных организаций, профессионально занимающихся тестированием (Американская ассоциация образовательных исследований, Американская психологическая ассоциация и Национальный совет по измерениям в образовании), и зафиксированным в документе, который называется "Стандарты для образовательного и психологического тестирования".

Согласно этому документу существуют три подхода к валидности. Они представлены в таблице 7.

Этот документ поясняет ещё несколько особенностей определения валидности:

1. Валидность получается из экспертных оценок (не измеряется).
2. Валидность выражается степенью (высокая, средняя, низкая).
3. Валидность специфична для каждого конкретного использования.
4. Существует много способов определения валидности.

В современной тестологии выделяются следующие основные виды валидности:

1. Валидность по содержанию (содержательная).

2. Конструктная (концептуальная) валидность.

3. Валидность по критерию (критериальная или эмпирическая валидность).

1. Валидность
по содержанию
(содержательная)

Содержательная валидность устанавливается экспертами для деятельности, близкой или совпадающей с реальной.

Определение содержательной валидности — основное для тестов достижений и тестов профессиональной успешности, когда должен быть точно определён материал, применяемый для тестирования, и когда существует достаточная ясность смысла измеряемого параметра.

Очевидно, что содержательная валидность будет полезна только тогда, когда могут быть определены специальные навыки и особенности поведения. Это довольно легко можно сделать на элементарном уровне — при тестировании арифметических навыков (правил выполнения четырёх арифметических операций, правил вычислений с 0 и т.п.), знаний в области искусства (правил нотной записи, принципов архитектуры и др.), а также знаний базовых элементов для большинства научных дисциплин, в которых накоплен багаж фактических данных. Содержательная валидность определяется на основе экспертных методов.

П. Клайн предлагает следующую процедуру определения содержательной валидности для тестов достижений:

1. Укажите точно категорию лиц, для которой предназначен тест.

2. Определите навыки, подлежащие тестированию (возможно, вам потребуется их проанализировать). Составьте список.

3. Передайте этот список экспертам в данной области (учителям и т.п.) для проверки — нет ли упущений.

4. Преобразуйте этот список в перечень заданий, используя, когда это возможно, равное количество заданий на каждый навык.

5. Представьте эти задания экспертам для проверки.

6. Подвергните задания обычным процедурам конструирования тестов. В результате должен быть получен содержательно валидный тест.

2. Конструктная
(концептуальная)
валидность

Этот вид валидности определяется в тех случаях, когда представление об измеряемом феномене (конструкте) существует только в сознании исследователя. Разработчик теста может лишь строить гипотезу о существовании данного конструкта, его формах и характере проявления. Устанавливается концептуальная валидность путём доказательства правильности теоритических концепций, положенных в основу теста. Это особенно необходимо в тех случаях, когда результаты тестовых измерений используются не просто для предсказания поведения, а как основа для выводов о том, в какой степени испытуемые обладают некоторой характеристикой.

В.М. Мельников и Л.Т. Ямпольский предлагают проводить проверку концептуальной валидности в три основных этапа:

1. Определение некоторой теоретической концепции, которая предположительно объясняет выполнение валидизируемого теста.

2. Из теоретической концепции выводятся одна или несколько гипотез, связанных с тестом.

3. Выдвинутые гипотезы подвергаются эмпирической проверке.

Если эмпирические данные подтверждают гипотезу, то тем самым подтверждается концепция, положенная в основу теста, и способность теста служить инструментом изме-

рения данного конструкта. Ошибки при проведении валидности могут возникнуть как следствие неправильной теоретической концепции, положенной в основу теста, или отсутствия соответствия между тестом и теоретической концепцией, или ошибочного выдвижения гипотез.

Непосредственно для тестов учебных достижений этот вид валидности не используется, однако овладение этим методом может быть чрезвычайно полезно для системы мониторинга, поскольку он даёт возможность обоснования истинности, реальности существования понятий и явлений. Было бы чрезвычайно полезно провести определение конструктной валидности для тех уровней овладения учебным материалом, которые мы рассматривали у различных авторов.

3. Валидность по критерию
(критериальная или
эмпирическая валидность)

Суть её заключается в определении способности теста служить индикатором или предсказателем строго определённой психической особенности, формы поведения человека и др.

Валидизация теста по критерию состоит в сравнении баллов, полученных испытуемыми за решение теста, с данными по критерию и вычислении коэффициента корреляции тестового результата с внешним критерием. Например, школьный тест умственного развития (ШТУР) валидизировался на основе критерия школьной успеваемости — оценок детей по предметам. В качестве критерия может выступать любой показатель, независимо и бесспорно измеряющий ту же психологическую характеристику, что и валидизируемый тест.

Для тестов учебных достижений наибольшее распространение нашли такие способы определения внешнего критерия, как метод коллективной оценки, метод средневзвешенной оценки, метод ранжирования и метод парного сравнения.

В литературе можно встретить немало других видов валидности, которые получены классификацией по другим основаниям: очевидная (с точки зрения испытуемого), конкурентная (определяется по корреляции с результатами использования инструмента, определяющего тот же показатель, что и создаваемый инструмент), внешняя, внутренняя, прогностическая и т.д.

Для всех тестов учебных достижений должна быть определена содержательная валидность, а для тестов, используемых для аттестации учащихся, — содержательная и критериальная валидность.

Нормирование

В начале статьи было показано, что одно из преимуществ тестов по сравнению с другими видами измерений заключается в том, что они имеют основания для сравнения. Для тестов, ориентированных на критерий, это полученный на основе экспертных оценок критерий значимости, превышение которого учеником означает, что он успешно справился, готов, прошёл и т.д., в зависимости от целей тестирования. Для нормативно-ориентированных тестов основанием для сравнения служат статистические нормы. Возможно сравнение показателя некоторого испытуемого с показателями в генеральной совокупности или других релевантных группах, что в конечном счёте даёт возможность адекватной интерпретации полученного показателя. Из сказанного видно, что нормализация тестов наиболее важна в тех случаях, когда осуществляется явное или неявное сравнение показателей испытуемых, как, например, при профориентации или отборе в целях обучения, построении систем мониторинга в образовании.

Тестовые нормы представляют собой установленные на базе репрезентативной выборки эмпирические усреднённые количественные данные о результатах выполнения теста, полученные в стандартных условиях. Какими же бывают нормы? По широте охвата можно выделить:

— универсальные нормы — устанавливаются для широкого контингента людей и лишь в малой степени зависят от действия каких-либо признаков;

— национальные нормы — применяются для представителей конкретной народности или страны в целом и учитывают особенности культуры, норм и традиций обследуемых;

— региональные нормы;

— локальные нормы.

В образовании в настоящее время мы можем вести речь о региональных и локальных нормах — нормах для Калуги, нормах для Тульской области, нормах для Санкт-Петербурга и т.д. Возможно, что в будущем появятся организации, которые возьмут на себя задачу нормирования инструмента и на национальном уровне. Однако задача эта чрезвычайно дорогостоящая, что и является основным ограничением при выполнении таких работ.

При разработке тестовых норм необходимо учитывать следующее:

1. Нормы устанавливаются при разработке нового теста, адаптации или редактировании существующего, если он используется на выборке, отличающейся от стандартизированной по каким-либо критериям.

2. Введение нового типа нормировочного балла при разработке теста должно быть обосновано.

3. Стандартизированная выборка при разработке норм должна быть хорошо сбалансирована по составу и численности.

4. Все отклонения от процедуры нормирования тестовых результатов должны оговариваться в прилагаемых руководствах.

Дальнейшее рассмотрение будет носить преимущественно качественное описание. Поэтому, если читателю необходимы сведения о количественных характеристиках или описание процедуры нормирования, рекомендуем обратиться к списку литературы, который будет опубликован в следующем номере журнала.

Мы же попытаемся ответить на два вопроса.

Вопрос 1. Каким образом можно корректно сравнивать результаты тестирования, полученные в результате проведения разных тестов по разным предметам?

Вопрос 2. Каким образом на основе оценок тестирования можно выставлять оценки в привычной для нас пятибалльной системе, таким образом, чтобы эти оценки были обоснованы, чтобы ответ на вопрос, почему за 23 балла одного теста мы ставим оценку 4, а за 18 баллов другого — оценку 5, был однозначным.

Обычно показатели некоторого индивидуума сравниваются с показателями релевантной нормативной группы посредством такого преобразования, которое выявляет статус этого индивидуума относительно данной группы.

Итак, решая задачи сравнения в самом широком понимании этого слова, необходимо провести преобразование исходных данных в более удобный вид, который позволит нам ответить на первый из поставленных вопросов. Учитывая наше обещание минимально использовать математический аппарат и вычисления, попытаемся показать формы представления данных наглядно. Однако без некоторых сведений из классической теории ошибок нам не обойтись.

Для дальнейших рассуждений необходимо ввести три понятия: среднее, дисперсия (среднеквадратичное отклонение) и нормальное распределение.

Представим их в вербальном виде и поясним на примере.

Пусть в результате проведённого тестирования получен некий ряд “сырых” показателей (см. табл. 8).

$X_{ср.}$ — среднее значение показателя. Для его получения необходимо: сложить данные всех измерений и поделить их на количество измерений:

$$\frac{(13+15+16+12+19+14+18+15+15+12)}{10}=\frac{149}{10}=14.9$$

σ— *отклонение*, показывающее, насколько далеко данное значение показателя отстоит от среднего значения ряда показателей. Для его расчёта необходимо:

1. Рассчитать среднее значение показателя.

2. Вычесть из среднего значения показателя значение данного показателя.

Для показателя 1 отклонение будет равно: 14,9 -13=1,9.

Для показателя 2: 14,9 -15=

-0,1. (см. табл. 9).

σ^2— *среднеквадратичное отклонение,* мера разброса показателей относительно среднего значения. Если отклонения являются характеристикой одного показателя, то σ2 является характеристикой ряда показателей, оно показывает, насколько далеко значения показателя в данном ряду отстоят от среднего.

В крайнем случае, когда все значения нашего ряда равны 14,9, значение σ^2 будет минимальным. Если девять значений показателей будут равны 0, а одно равно 149 — то можно ожидать максимальное значение σ^2. Среднее значение и в том, и в другом случае будет 14,9.

Для расчёта среднеквадратичного отклонения необходимо:

1. Рассчитать среднее значение группы показателей.

2. Вычесть из среднего значения показателя значение каждого показателя.

3. Возвести каждое из полученных отклонений в квадрат:

$1,9^2=3,61$; $(-0,1)^2=0,01$;

$(-1,1)^2=1,21$; $2,9^2=8,41$;

$(-4,1)^2=16,81$; $0,9^2=0,81$;

$(-3,1)^2=9,61$; $(-0,1)^2=0,01$;

$(-0,1)^2=0,01$; $2,9^2=8,41$.

4. Сложить полученные значения:

3,61+0,01+1,21+8,41+16,81+0,81+9,61+0,01+0,01+8,41=48,9.

5. Поделить полученное число на количество показателей:

48,9 = 4,89.

6. Извлечь из полученной цифры квадратный корень:

$$\sqrt{4,89}=2.19\approx2.2$$

Путём этих несложных вычислений мы получаем характеристику ряда значений показателей с точки зрения их разброса. Конечно, вычисления по формулам даже с помощью калькулятора уже никто не проводит — для этого существуют компьютерные программы. В данном случае вычисления приведены с единственной целью — прояснить смысл понятия среднеквадратичного отклонения для ряда значений показателей.

Нормальное распределение — это распределение частот, которое подчиняется закону нормального распределения. Существование этого распределения обосновано эмпирически и математически.

По закону нормального распределения в большинстве случаев распределяются как чисто случайные величины, так и результаты выполнения тестов и их заданий. При этом необходимо сделать важное замечание: количество значений показателя должно быть достаточно большим, не менее 30. Это утверждение придётся принять на веру или обратиться к литературе по математической статистике.

График функции нормального распределения (см.рис. 1) симметричный, с асимптотически приближающимися к нулю ветвями. Вообще может быть, бесконечное множество графиков нормального распределения, но все они имеют одинаковый вид и отличаются друг от друга по двум значениям — среднему (оно характеризует положение графика относительно оси OX), и среднеквадратическому отклонению (характеризует “широту” или “крутизну” графика).

Рiс.Рис. 1. График функции нормального распределения

Попытаемся ответить на два наших вопроса на конкретном примере.

У нас в распоряжении есть два комплекта тестовых материалов: один по физике, второй по русскому языку. На основе данных апробации этих тестов на репрезантативной выборке учащихся были получены характеристики теста по физике — среднее количество заданий, с которым справлялись учащиеся $X_{ср}^{ф}=42$, а среднеквадратичное отклонение $^{2}_{ф}=8$; для теста по русскому языку — $X_{ср}^{р}=26$, $\sigma^2_{р}=5$.

Использовав эти тесты, мы провели оценку учеников нашего класса и взяли результаты двух учеников. Их результаты, выраженные в количестве правильно выполненных заданий (баллах), представлены в таблице 10.

Если оценивать эти результаты, исходя из количества выполненных заданий, то может показаться, что первый ученик справился с тестированием по двум предметам одинаково успешно, а вторая ученица блестяще выполнила тест по физике и провалилась по русскому. Однако такие оценки, как правило, не соответствуют действительности.

Для сравнения результатов тестирования “сырые” оценки обычно переводятся в стандартные. Перевод заключается в выполнении двух операций — центрирования и нормирования.

Центрирование. На рисунке 2 изображены два графика распределения показателей, которые получены на одной шкале. Здесь мы можем видеть один из недостатков первичных значений — они не совпадают по шкале. Для того чтобы сделать показатели сравнимыми, нам необходимо либо сдвинуть один на две единицы вправо, либо второй — на две единицы влево. На практике поступают несколько иначе: сдвигают оба графика в некоторую фиксированную точку, обычно это среднее или нулевое значение выбранной шкалы измерения. Для приведения данной кривой к нулевой необходимо вычесть из значений показателей $X_{ср}$. Действительно, если из значений показателей первой кривой вычесть её среднее значение 2, а из значений показателей второй — 4, то обе кривые окажутся симметричными относительно нулевой точки.

Нормирование. Его суть состоит в переходе к другому масштабу. Попытаемся проиллюстрировать эту операцию графическим примером. На рисунке 3 представлены два графика, оба они уже приведены к единой шкале — их среднее значения совмещены с нулевой отметкой. Однако форма этих графиков различна, что не даёт возможности провести сравнения признаков для этих двух распределений. Для приведения их к одинаковому виду нам следует уравнять их среднеквадратическое отклонение. Это приведёт к тому, что либо “сожмётся” один из графиков, либо расширится другой. Как и в случае центрирования, используют способ приведения графиков к стандартизированному виду с $\sigma^2=1$.

На практике это означает деление показателей на величину среднеквадратичного отклонения. В этом случае величина этого отклонения станет равна единице и результаты исследований можно будет сравнивать.

Таким образом, мы пришли к понятию стандартного Z— показателя, который характеризуется средним значением 0, средним квадратичным отклонением 1.

В нашем примере: для теста по физике — Хсрф=42, а среднеквадратичное отклонение $\sigma^2_{ф}=8$, для теста по русскому языку — $X_{ср}^{р}=26$, а $\sigma^2_{р}=5$. (см.табл.11, 12)

Таким образом, осуществив несложные преобразования и получив оценки в одних стандартных баллах, мы получили возможность некоторого сравнения результатов. Представим полученные результаты графически (рис.4).

Уже сейчас становится ясно, что оценки Мамина по физике и русскому очень разные, что казавшаяся провальной оценка Папиной по русскому языку лучше оценки Мамина по физике. Эти оценки начинают приобретать некоторый педагогический смысл.

Принципиальное значение для дальнейших рассуждений имеет определение площади под кривой нормального распределения.

Среднее значение находится под максимумом пика кривой. Среднеквадратичное отклонение (в случае стандартной Z-оценки) — расстояние от среднего значения до точки 1. Оно определено таким образом, что площадь графика, ограниченная Хср кривой и орди-

натной точки 1, составляет 34,13% ограничения, а площадь, ограниченная расстояниями А с двух сторон, равна 68,26%, а вся площадь составляет 100%. С точки зрения результатов выполнения тестов эти проценты представляют собой распределение учащихся, то есть если оценка ученика составляет единицу, то это означает, что лучше него справились 15,87% учащихся, а все остальные справились хуже.

Что мы можем сказать про ученика, который в единицах стандартных Z-оценок получил оценку “0”? С точки зрения нормального распределения, половина учащихся справляется лучше, чем он, а другая половина хуже. Что означает получение учеником в стандартных единицах оценки “1”? Это означает то, что хуже него справляются с работой 50+34,13=84,13% учащихся, а все остальные лучше; а оценка “2” будет означать, что лучше справляются чуть больше 2% школьников, а все остальные хуже. Аналогичные рассуждения мы можем провести и для других оценок.

Вернёмся к нашему примеру. Выразим оценки в количестве учащихся, которые справляются с той же работой лучше (см. табл. 13).

Для окончательного корректного перевода оценок наших учеников в привычные нам школьные необходимо знать статистическое распределение нашей выборки по школьным оценкам. В общем виде оно зависит от предмета, но не так сильно, как кажется. Если предположить, что распределение по оценкам соответствует приведённому в таблице, то нам необходимо посчитать коммулятивный (накопленный) процент, то есть количество учащихся, которые успевают лучше, чем учащиеся, имеющие данную оценку (см. табл. 14).

Таким образом, для школьных оценок можно получить следующие интервалы (см. табл. 15).

Нам осталось получить соответствующие выделенным интервалам стандартные Z-оценки. Это можно сделать, используя таблицу 16.

В таблице представлены значения для одной ветви положительных значений. Найдём значение для точки 8 процентов. Для этого из 50 (напомним, в таблице данные по одной ветви и, следовательно, только 50 процентов) вычтем 8, получим 42, или, если брать в долях, то 0,42. Найдём в таблице ближайшее значение, оно составит 0,4207 (значение выделено жирным шрифтом). По горизонтали найдём значение 1,4, по вертикали — значение сотых — 0,01. Сложив эти значения, получим 1,41 (см. табл. 16).

Аналогично найдём и другие значения для выделенных нами интервалов:(см. табл. 17).

В этом случае мы можем достаточно корректно выставить школьные оценки. Мамин по русскому получит оценку 5, потому что он выполнил работу так же, как выполняют её 8 процентов лучших учеников. Папина по физике получит оценку 4, потому что её результат выполнения не попал в 8 процентов лучших, но он попал в 53 процента четвёрочников и отличников. Мамин по физике и Папина по русскому получат оценку 3, потому что их результаты попадают в соответствующий интервал. Конечно, разрыв между двумя последними оценками достаточно велик, но такова наша система оценивания. Мы привели здесь примеры, выражая оценки в процентах, для того чтобы понятен смысл перевода. Гораздо удобнее пользоваться стандартными оценками, поэтому в таблице мы приводим все оценки (см. табл. 18).

Теперь становится достаточно ясно, что наши предварительные оценки результатов выполнения теста были ошибочны.

Завершая рассуждения, сделаем несколько важных замечаний и выводов.

Конечно, схему выставления оценок мы рассмотрели в большей мере качественно, без привлечения достаточного математического аппарата. Тем не менее совершенно очевидно, что нормирование даёт нам качественный способ корректного сравнения оценок, полученных в результате применения различных тестов, и выставления оценок в школьных баллах.

Хоть вычисление оценок проводит компьютерная программа, пользователь, особенно руководитель, должен иметь представление о тех преобразованиях, которые она делает, об

их корректности, должен доверять полученным в результате обработки оценкам.

В практической работе, в том случае, когда нет крайней необходимости, лучше пользоваться стандартными оценками, поскольку школьные оценки очень огрубляют полученные результаты, снижают дискриминативность инструмента.

Мы рассмотрели только одну стандартную оценку — Z, на самом деле стандартных оценок несколько, все они имеют свои плюсы и минусы. С точки зрения П. Клайна, для тестов с распределением (если не нормальным, то, по крайней мере, симметричным) Т-показатели со средним значением X=50 и со стандартным отклонением σ^2=10 является лучшей значимой оценкой. Получение этих оценок носит линейный характер и не должно представлять труда, в крайнем случае можно обратиться к литературе. Наглядное представление о переводе между шкалами можно получить из графика (см. рис. 5).

Наконец, последнее и, может быть, самое главное — при рассмотрении задачи нормирования мы нигде не использовали количество заданий в тесте. Для оценок учащихся и оценок теста это совершенно не важно. Показатель того, что учащийся из 70 заданий теста справился с 40, никак не характеризует ученика, потому что тест может содержать простые и сложные задания в разной пропорции. Точно так же не является характеристикой ученика время, которое ученик может затратить на выполнение заданий, поскольку задания могут быть очень сложными, но не требующими большого времени на выполнение. Может быть больше заданий, которые не представляют большой трудности, но на их выполнение уходит значительное время. (Попробуйте ответить на вопрос о том, как звали вторую жену Георга IV Английского, и заметьте, сколько времени вы потратили на поиски ответа). Количество заданий и время выполнения тестирования являются характеристиками теста как измерительного средства и не связаны с оцениванием результатов ученика.

Методическое оснащение

Методическое оснащение должно решать одну из основных задач объективности получаемых при тестировании результатов — обеспечивать одинаковость условий для всех испытуемых.

Методическое оснащение включает в себя две части — сведения, которые необходимо знать пользователю теста об инструменте, и указания, содержащие правила предъявления теста испытуемым.

Сведения для пользователей оформляются в виде спецификации. Она обязательна для тестов, предназначенных для внешнего использования. В ней излагается:

— классификационная характеристика теста (назначение и психолого-педагогическое содержание);

— ограничения и показания для применения;

— состав теста;

— описание существующих форм и модификаций;

— ссылка на апробацию теста;

— ключи;

— правила обработки данных;

— устройство шкал;

— данные о надёжности и валидности;

— правила интерпретации результатов.

Требования к процедуре проведения должны быть зафиксированы в инструкциях для исследователя (ведущего).

Требования к формальной стороне процедуры проведения могут быть следующие:

— обеспечение инструментарием в необходимых количествах в случае, когда используются простые материалы: карандаши, ластики, ручки, фломастеры (необходимо иметь их полуторакратный запас);

— наличие столов и стульев в количестве, необходимом для проведения исследования, в соответствии с инструкцией по проведению;

— размещение столов и стульев таким образом, чтобы к каждому испытуемому было удобно подойти;

— обеспечение удобного места за столом для каждого испытуемого путём подбора оснащённого мебелью помещения необходимых размеров;

— оборудование места с максимальным обзором для экспериментатора и, если это необходимо, для наблюдателя.

Особое место среди факторов, влияющих на индивидуальную и групповую работоспособность испытуемых, занимает время проведения теста и характер деятельности учащихся до тестирования. Наиболее благоприятно время с 9 до 12 или с 16 до 18 часов. Поскольку речь идёт о тестах учебных достижений, то наиболее приемлемо время второго или третьего урока первой смены. Авторы тестов могут потребовать не проводить тестирование после занятий физической культурой и спортом.

При организации проведения тестов важно учитывать ситуативные отвлекающие факторы. К ним относятся: шумы (с улицы, из других частей здания, радио- и телетрансляции и т.п.), звонки, стук, звук шагов, гудение неисправных ламп дневного света, запах (пищи, краски и пр.), мигание света, неопрятность столов, помещения и т.д.

Непосредственная подготовка к проведению теста заключается в проверке состояния помещения, его оснащения, пригодности для размещения испытуемых, а также устранении или уменьшении ситуативных отвлекающих факторов и проверке наличия, состояния и размещения тестовых установок и материалов.

Наиболее рациональным способом формализации процедуры проведения является написание сценария проведения. Особенно желателен сценарий для тестов, предназначенных для итоговой аттестации учащихся.

Сценарий проведения исследования в общем виде должен включать в себя следующие необходимые сведения, которые могут сообщаться испытуемым (в зависимости от условий тестирования что-то может быть сокращено или добавлено):

1. Объяснить, зачем нужен тест, какие результаты ожидаются.

2. Объяснить, почему испытуемые должны приложить максимум усилий для его выполнения, акцентировать внимание испытуемых на возможности проверки своих сил или подчеркнуть соревновательный мотив. Отметить, что слишком сильная мотивировка, равно как и слишком слабая, в одинаковой степени негативно сказывается на результативности выполнения задания.

3. Медленно, громко, чётко, без запинок, естественным голосом прочесть инструкцию к тесту с примерами, если они имеются. В данном случае возможен вариант, когда испытуемые самостоятельно следят по своим вариантам текста за инструкцией. При таком порядке возможно воспроизведение инструкции по памяти.

4. Дать возможность испытуемым потренироваться, решив самостоятельно одну или более из задач-образцов, если таковые имеются; проверить, правильно ли понята инструкция.

5. Сообщить о временном ресурсе, о правилах исправления допущенных ошибок, о том, чего не рекомендуется делать при решении задач, к кому обращаться в случае возникновения вопросов.

6. Вместе с испытуемыми или самому записать, если требуется, паспортные и биографические данные в регистрационных бланках. Проследить за правильностью их заполнения.

7. Ответить на имеющиеся вопросы.

8. Дать команду начать решение задач теста. Время начала записать самому или попросить сделать это испытуемых на регистрационном бланке.

9. Во время решения задач или ответов на вопросы следить:

— за временем решения, если это необходимо;

— за наличием отточенных карандашей и других материалов;

— за правильностью заполнения паспортной части регистрационных бланков (если

замечена ошибка, своевременно её устранить);

— за тем, чтобы испытуемые не писали на тестовых брошюрах, если иное не предусмотрено, не портили тестовых установок и приборов;

— за тем, чтобы соседи не общались между собой, не шептались, не мешали друг другу, не подглядывали друг у друга;

— за состоянием испытуемых;

— за тем, чтобы испытуемые своевременно получали ответы на вопросы, связанные с процедурой проведения (ответы не должны служить подсказкой для решения или нарушать указания инструкции, возможные варианты ответов должны быть предусмотрены).

10. После сигнала к окончанию решения задач теста при групповом проведении дать команду сложить брошюры и бланки для ответов в исходное положение или самому собрать их (если участвует не более 30 человек). Если участников тестирования больше 30, то рекомендуется попросить всех оставаться на своих местах, чтобы облегчить сбор материала. Затем попросить передать в начало или конец колонки (ряда) тестовые материалы в следующем порядке: бланки для ответов, брошюры, черновики. После этого пересчитать количество бланков и брошюр, проверить, чтобы их количество совпадало с числом испытуемых.

11. По окончании тестирования просмотреть все брошюры и стереть пометки на них. Если это невозможно — брошюры следует уничтожить.

Кроме этого, сценарий должен предусматривать процедуру приветствия и благодарности за выполненную работу, действия экспериментатора с опоздавшими учащимися, реакцию на просьбы учеников временно покинуть место проведения тестирования, ответы на наиболее часто встречающиеся вопросы и некоторые другие процедурные вопросы (по усмотрению авторов).

Для всех тестов учебных достижений, предназначенных для внешнего использования, обязательна фиксация требований к ведущему тестирование.

Проводить тестирование может только специально подготовленный человек. Он должен удовлетворять определённым профессиональным и личностным требованиям:

1. Должен быть в зафиксированном статусе в отношении к ученикам: это должен быть педагог, преподающий предмет, по которому проводится испытание, завуч, работающий или не работающий в данном классе, педагог иной школы и т.п. Более важным представляется то, чтобы при любом тестировании этот статус был бы одинаков.

2. Должен уметь контролировать себя, быть эмоционально уравновешенным, общительным, тактичным.

3. Должен понимать задачи эксперимента, быть компетентным в проведении теста, а если ему предстоит обработка результатов — то и в оценивании результатов.

Значительные возможности по стандартизации процедуры проведения тестирования даёт институт наблюдателей.

Наблюдатель — лицо, фиксирующее процедуру проведения и соответствие действий ведущего (исследователя) сценарию тестирования. Наблюдателю запрещается вмешиваться в процесс тестирования. Присутствие наблюдателя, несомненно, удорожает проведение исследования, однако достигаемый при этом значительный выигрыш в качестве компенсирует все дополнительные затраты: ведь в случае некачественного проведения все усилия могут оказаться напрасными. Наблюдатели могут присутствовать не на всех процедурах тестирования, но сама возможность их присутствия в значительной степени дисциплинирует исследователей.

Присутствие наблюдателей позволяет сравнить качество проведения тестирования у разных ведущих, а также в тех группах, где присутствовал наблюдатель, и там, где он не присутствовал. Наш опыт использования института наблюдателей говорит о том, что процедурные ошибки могут добавлять до 50% разброса в результативность выполнения тестов. Особенно важно присутствие наблюдателя на этапе апробации инструментария, поскольку он даёт возможность собрать материал для дальнейшей работы над инструмен-

том.

Деятельность наблюдателя заключается в заполнении анкеты наблюдателя, в которой, кроме фиксации времени и оценки правильности выполнения ведущим пунктов сценария, должны присутствовать вопросы по оценке поведения учащихся, корректности поведения ведущего, неординарных случаях. Институт наблюдателей предполагает наличие инструкции по использованию анкеты наблюдателя. В ней должны содержаться сведения о том, при каких нарушениях процедуры, зафиксированных наблюдателем, результаты тестирования могут быть аннулированы.

Основываясь на собственной практике, можно сказать, что сам факт присутствия наблюдателя настолько дисциплинирует ведущего, что за 4 года работы не было случая, когда пришлось бы воспользоваться этой инструкцией.

Артефакты и факторы, влияющие на результаты тестирования

В этом разделе мы рассмотрим влияние факторов, смещающих оценки. Учитывая, что для мониторинга важны результаты, обобщённые, по крайней мере, на уровне класса, мы не будем обсуждать особенности получения индивидуальных оценок. Таким оценкам мы уделили достаточно места в разделе, посвящённом нормированию тестов. Наше обсуждение будет касаться в большей мере современного состояния использования тестов и результатов тестирования. Скорее всего, здесь мы наметим некоторые проблемы, решение которых потребует усилий немалой части разработчиков систем измерений в образовании.

Примерный список факторов, смешивающих оценки, может выглядеть следующим образом.

Качество инструментария остаётся пока чрезвычайно низким, хотя обнадёживает значительный прогресс, достигнутый за последние несколько лет. Несмотря на то что с теоретической точки зрения проблемы измерения результатов учебного процесса достаточно ясны, приходится постоянно сталкиваться с фактами использования инструмента не по назначению (психологических тестов для оценки эффективности работы школы), попытками использовать одни оценки вместо других (усреднённые школьные оценки как показатель эффективности учебного процесса), использование инструмента, построенного на основах, не выдерживающих элементарной критики (уровневые контрольные работы) и т.д.

К основным проблемам можно отнести низкое ресурсное и кадровое обеспечение (для подготовки полноценного теста необходимы скоординированные усилия специалистов, по крайней мере, 12 профессий), что видно из таблицы 19.

Достаточно остро ощущается также недостаток материальных ресурсов (создание полноценного теста для итогового тестирования по одному предмету одного класса составляет по международным оценкам около 10 тыс. долларов, в наших условиях эта цифра, конечно, может быть снижена, однако её снижение до “голого энтузиазма” невозможно) и недостаток временных ресурсов (разработка полноценного теста занимает два-три года, а результат управленцы требуют уже сейчас).

Профессионализм и подготовленность людей. Тестирование минимизирует влияние субъективного фактора, связанного с личностью ведущего. При качественном методическом оснащении и соблюдении процедуры этот фактор можно не принимать во внимание.

Статистическая регрессия. Является достаточно специфичным явлением для крайних групп, для тестирования, несомненно, имеет самое непосредственное значение. Её учёт необходим в случае, когда по результатам тестирования проходит какое-то разделение. В качестве метода снижения её влияния может быть предложено увеличение надёжности инструментария, повторное тестирование, использование нетестовых форм получения оценок.

Цикличность. В том случае, если используется нормативно ориентированный тест, который проходил апробацию и нормирование в те же сроки, когда и используется, годичная цикличность рандомизируется. Однако вопрос о времени и возможных сроках использования тестов остаётся. Достаточно очевидно, что использовать тест через полгода и

сравнивать полученные результаты некорректно. Представляется корректным проведение тестирования с разницей в 1–2 дня. На сегодняшний день нет работ, которые доказательно могли бы определить календарный период использования тестов. При проведении тестирования необходимо учитывать годичную цикличность, которая даже в рамках небольшого календарного периода может оказать существенное влияние. Например, перед началом и сразу после каникул.

Конечно, практически не учитываемыми пока оказываются циклы кроме годичных и четвертных.

Значимость индикатора и его смещение. Этот фактор для системы образования вероятен как никакой другой. Именно из-за опасности такого смещения следует разнести проведение тестирования с контрольными и информационными целями. Это смещение в принципе непредсказуемо, и возможность принятия адекватных мер в рамках тестирования, несомненно, существует (варьирование временем, введение института наблюдателей, ужесточение процедуры и пр.). Однако это смещение индикатора порождает негативные процессы в рамках использования тестов с целью аттестации. Использование неадекватного учебному курсу тестового инструмента порождает изменения в преподавании. Например, если в состав итоговых тестов по биологии не включать схемы, то их изучение начнет сокращаться. Такое влияние тестирования на учебный процесс давно замечено специалистами, именно оно стало причиной отказа от тестов для аттестации учащихся в Германии.

Нарушение в информационных потоках. Своеобразным нарушением информационных потоков можно считать тенденцию вольного обращения с учебным материалом, когда часть базового учебного материала в образовательном учреждении или у конкретного учителя заменяется другим. Необходимо отметить, что опасность получения искажённых результатов, в виду того что данный материал не изучался или изучался не в должном объёме, в условиях нашей системы образования существует. Вероятно, такая тенденция будет наблюдаться до принятия и введения в действие полноценных образовательных стандартов, а это — срок не одного десятилетия.

Различная мотивация участников в естественных условиях. Учёт этого фактора несложен технически — сравнение результатов, полученных при разной мотивации, некорректно. Учесть же его влияние на этапе коррекции результатов навряд ли возможно. Как вариант можно предложить искусственное усиление мотивации: если какие-то данные были получены в рамках аттестационных процедур, то в процессе тестирования следует сообщить учащимся о том, что их результаты будут использованы для выставления им оценок. Однако такой способ не всегда можно использовать.

Изменение людей в процессе измерения. Для образования с его высокой динамичностью никому не приходит в голову сравнивать между собой результаты тестирования разных классов. Тем не менее во многих случаях возможно получение смещённых оценок. Например, при попытке сравнивать результаты тестирования шестиклассников в конце учебного года и семиклассников в начале.

Эффект повторного измерения. Повторное измерение тем же инструментом применительно к области образования представляется крайне сомнительным. С большей степенью уверенности можно говорить о том, что эффект повторного измерения и эффект развития делают такие измерения некорректными.

Изменения группы под влиянием отношений окружающих, вызванных экспериментальным воздействием. Вероятно, такое изменение может дать эффект при тестировании. Однако его влияние вряд ли можно учесть в рамках учебного процесса. Для выявления и возможного учёта в рамках этого фактора смещения оценок необходим учёт оценок социальной эффективности деятельности образовательного учреждения.

Групповая фальсификация результатов. Несомненно, вероятность этого фактора существует. Фальсификация в сторону увеличения оценок при корректной процедуре значительно затруднена, а в сторону занижения — вполне вероятна. В случае использования

нормированного инструмента это смещение можно выявить. Аномально высокие или низкие оценки должны насторожить исследователя и заставить его проверить гипотезу данного смещения.

Отбор испытуемых — весьма вероятная причина смещения. Однако в отличие от экспериментальной социальной и психологической работы, где существует проблема отбора в экспериментальные группы, связанная с добровольным участием и трудностями рандомизации, для нашего случая это проблема учёта отбора в классы и образовательные учреждения. При массовом характере отбора в те или иные классы в течение уже почти десятилетия мы не обладаем технологией учёта фактора отбора и оценки его влияния на результаты тестирования. Это смещение может иметь и характер скрытого отбора. Возможно, частично эту проблему удаёется решить на этапе нормирования, когда для одного и того же инструмента могут быть получены нормы для разного вида и уровня отбора классов. По причине влияния фактора отбора разница в результатах тестирования может достигать 2–3 раз. Конечно, в условиях любого отбора учёт этого фактора крайне необходим.

Изменение группы в процессе проведения эксперимента. Это достаточно легко учитываемый фактор, однако для таких небольших групп, как класс, выбытие или отсутствие 3–4 человек может существенно изменить результаты. При увеличении количества участников тестирования его влияние рандомизируется.

Естественное развитие. Для тестирования в рамках мониторинга этот фактор можно объединить с фактором развития в процессе эксперимента.

Социально-территориальные особенности групп. Чрезвычайно важный и опасный фактор. Как мы уже отмечали, социальная стратификация, которая сейчас идёт в школах, реализуется не по территориальному признаку престижности городской застройки, а по отдельным образовательным учреждениям, поэтому в наших условиях западный опыт учёта этого фактора не помогает. Дейл Манн вообще считает, что “результаты тестирования следует располагать в зависимости от социального статуса учащегося”.

Для нас это фактор не только внутригородского различия, но и различия между населёнными пунктами городской и сельской местности. Мы достаточно хорошо знаем, что в городских школах результаты учащихся выше, мы неплохо знаем, почему они выше, но мы совершенно не можем учитывать эту разницу в полученных результатах тестирования.

Мы можем предполагать, что проблема учёта этого фактора будет решена при создании разных норм. Ещё одно предположение заключается в том, что влияние фактора может распространяться на другие особенности, такие, как половые особенности, национальная специфика и т.д.

Различная внутренняя жизнь групп, разные события для разных групп и события, не связанные с воздействием, которые могут повлиять на результат. Эти три фактора, несомненно, могут оказать какое-то влияние на результаты тестирования, однако можно предположить, что их влияние не так велико, как в случае изучения социальных и внутригрупповых процессов.

Различная скорость протекания внутригрупповых процессов. Это весьма важное обстоятельство, особенно когда речь идёт об оценке эффективности различных технологий обучения. Промежуточные измерения могут давать разный результат в силу того, что процессы протекают неравномерно. Это может быть связано как с особенностями самой группы, так и с особенностями технологии. В какой-то мере промежуточное исследование в рамках мониторинга может оказаться информативным, но в специальных случаях лучше дожидаться окончательного результата.

Условия, вызывающие реакцию на эксперимент. Этот фактор отражает угрозу, которая может возникнуть, когда результаты эксперимента внедряют в практику и не получают тех результатов, что были достигнуты в экспериментальных условиях. Конечно, тестирование не может оказать существенного влияния на этот процесс, за исключением, может быть, тех случаев, когда речь идёт о внедрении самого тестирования.

Интерференция воздействий. Вполне реальный эффект, который может оказать влия-

ние на результаты тестирования, особенно в наших условиях. Например, один из классов мог пройти предварительное тестирование, обучившись навыкам тестирования, а другой — нет. В условиях, когда навыки тестирования у детей в массовом масштабе не сформированы, такая ситуация вполне может оказать влияние на результаты тестирования.

Синергизм и компенсаторность. Найти примеры влияния этих факторов на результаты тестирования нам не удалось, однако это совершенно не значит, что их не может быть.