Технология стандартизации дидактических тестов

Михайлычев Евгений Аркадьевич — заведующий кафедрой общей педагогики Таганрогского государственного педагогического института, научный руководитель Центра педагогических измерений Ростовского областного института повышения квалификации и переподготовки работников образования, профессор, доктор педагогических наук

Надёжность тестов и приёмы её определения

Надёжность теста — это согласованность показателей, полученных у тех же самых испытуемых при повторном тестировании, тем самым тестом или эквивалентной её формой. Надёжность может проверяться относительно временных изменений, выбора конкретных заданий, тестовой выборки поведения (каких-либо индикаторов), относительно роли индивидуальности экспериментатора или специалиста при обработке или же относительно других аспектов тестирования. Вычисление **ошибки измерений** — вероятных пределов колебаний измеряемой величины основано на понятии *надёжности*. Дисперсия ошибок может увеличиваться при любых изменениях условий проведения теста (если только они не имеют отношения к цели тестирования). Поэтому даже в оптимальных условиях ни один тест не является абсолютно надёжным.

Существуют различные типы надёжности и, соответственно, различные подходы к их вычислению. Один из наиболее распространённых способов проверки надёжности — это **ретестовая** надёжность, измеряемая при **повторном** проведении того же самого теста на том же контингенте, в тех же условиях.

Дисперсия ошибки (корреляция между данными двух сеансов тестирования) будет отражать случайные колебания в выполнении теста, вызываемые неконтролируемыми факторами (например, изменения погоды, неожиданные отвлекающие моменты и шумы, изменения в состоянии испытуемых, эмоционального напряжения из-за значимых для групп обследуемых событий и т.п.). Чем выше **ретестовая надёжность**, тем менее чувствительны результаты к обычным изменениям состояния испытуемого в обстановке тестирования. Следует указывать, в каком интервале времени она измерялась и произошли ли за этот период с испытуемыми какие-либо значимые события, способные повлиять на результаты тестирования. Возможен целый ряд коэффициентов ретестовой надёжности, полученных с разными интервалами между сеансами тестирования. Считается, что интервал между двумя последовательными применениями теста не должен превышать шести месяцев. В дидактических тестах на ретестовую надёжность влияет как процесс тренинга, так и обучаемость обследуемых. Напомним, что тренинг тестируемых — обычная и вполне легальная практика подготовки учащихся к ответственным испытаниям (вступительные или выпускные экзамены в вузе, а в школах — подготовка к срезовым контрольным работам или аттестации, к инспекторским проверкам). Реально его не избежать, но важно знать, имел ли он место и все ли обследуемые через него прошли.

Надёжность взаимозаменяемых форм позволяет избежать трудности, связанные с измерением ретестовой надёжности. Это корреляция между двумя считающимися равноценными формами одного и того же теста (их называют ещё *параллельными, взаимозаменяемыми, эквивалентными, сопоставимыми, дублированными, подобными*). Такой коэффициент надёжности одновременно измеряет временную стабильность теста и согласованность ответов по двум формам теста. Только если оба варианта теста следуют один за другим, мы в «чистом» виде получим надёжность форм теста относительно друг друга (без значительного разрыва во времени). Как и в первом случае, следует указать интервал времени проведения сеансов тестирования.

Параллельные формы теста могут быть независимо построенными тестами, отвечающими одним и тем же требованиям: содержать одинаковое число заданий, представленных в

одной форме и с однотипным содержанием, с одинаковым диапазоном и уровнем трудности заданий. На сопоставимость проверяются также инструкции, время работы испытуемого с тестовыми заданиями, формат бланков.

Метод расщепления используется, когда существует единственная форма теста. Тест разбивается на две сопоставимые части, и испытуемый получает два результата. Эта мера согласованности выборок содержания заданий теста позволяет получить **коэффициент внутренней согласованности**. При её применении затруднения методистов-разработчиков связаны с различиями в природе и уровне трудности заданий. Если задания теста расположены в порядке возрастания трудности, достаточно эквивалентные «половинки» теста можно получить, разбивая задания на чётные и нечётные. Сложности здесь встречаются, если тест содержит блочные, взаимосвязанные задания (несколько вопросов касается какого-либо фрагмента текста или диаграммы, а аналогичные задания не повторяются). Каждая такая группа заданий должна быть отнесена к какой-либо половине целиком и уравновешена заданиями аналогичного уровня сложности. Следует учитывать, что коэффициенты корреляции здесь будут относиться только к половине теста.

П.Ж. Рюлон разработал формулу определения надёжности методом расщепления:

$$r_{1I} = 1 - \frac{\sigma_d^2}{\sigma_x^2},$$

где σ^2_d — дисперсия разностей между результатами каждого испытуемого по обеим половинам теста, а σ^2_x — дисперсия суммарных результатов.

Метод Кьюдера-Ричардсона также основан на однократном применении теста. Он пригоден для случаев, когда выполнение заданий осуществляется только как «правильное — неправильное» (по принципу «всё или ничто»). Чем однороднее задания теста, тем выше согласованность результатов. Но такие тесты сравнительно редки — обычно дидактический тест содержит задания разного типа.

Для разнородных по характеру заданий тестов применяется формула:

$$r_{1I} = 1 - \frac{n(\sigma_t^2 - \sum_{pq})}{(n-1)\sigma_t^2},$$

где r_{1I} — коэффициент надёжности всего теста, n — число заданий, s^2_t — стандартное отклонение суммарных показателей теста, p и q — доля испытуемых, соответственно справившихся с каждым заданием, которые перемножаются по каждому заданию, а потом суммируются.

Коэффициент Кьюдера —Ричардсона обычно бывает меньше полученного методом расщепления. Для тестов, в которых испытуемый набирает баллы за ответы разной степени полноты и правильности (например, 4 — за «обычно», 3 — «иногда», 2 — «редко», 1 — «никогда» или по аналогичной шкале), имеется обобщённая формула коэффициента альфа:

$$r_{1I} = 1 - \frac{(\sigma_t^2 - \sum \sigma_t^2)}{(n-1)\sigma_t^2}.$$

Здесь находится дисперсия индивидуальных результатов по каждому заданию с последующим суммированием этих дисперсий по всем заданиям.

Надёжность теста на скорость

Для многих дидактических тестов значимым фактором является скорость выполнения заданий (особенно если задания моделируют какие-либо профессиональные ситуации, где скорость принятия решений регламентирована нормативами).

Перед разработчиком встаёт проблема согласования заданий на скорость. Без такой согласованности коэффициенты надёжности, основанные на количестве допущенных ошибок, окажутся явно завышенными на определённую величину. И она будет тем значимее, чем больше успешность выполнения заданий зависит от скорости работы испытуемого.

Для проверки надёжности тестов на скорость более эффективен *метод ретеста и метод*

взаимозаменяемых форм.

Применяются также приёмы разделения полного времени выполнения теста на четыре части с регистрацией результатов отдельно для каждой четверти. Для этого можно попросить обследуемых по сигналу отметить определённым знаком выполняемое в данный момент задание теста. Фиксировать затраты времени может помощник организатора тестирования.

Считается, что в качестве приблизительной меры значимости скоростного компонента в комплексе тестов может быть процент испытуемых, не успевших закончить тест в отведённое время. Как показывают данные методологических исследований, возможна ситуация, когда никто не уложился в заданное время, хотя фактор скорости здесь ни при чём.

Дидактическая тестология должна воспользоваться всем богатством методологического аппарата современной психодиагностики, особенно педагогической. Но педагогу-практику, апробирующему свой или чужой тест и перепроверяющему нормативы на своих классах, важнее всего владеть *ретестовой надёжностью и надёжностью эквивалентных форм*.

Эмпирический анализ полученных результатов дидактического теста

Предварительным этапом стандартизации (и обязательным этапом анализа эффективности любого теста, в том числе его первоначального варианта) является **эмпирический анализ теста**.

Для этого необходимо прежде всего получить сведения о степени трудности и сложности каждого задания, его селективности и о выверке **дистракторов** — отвлекающих вариантов ответа — подсказок в закрытом тесте с выбором вариантов.

Степень трудности заданий определяется процентом учащихся, получивших верный результат при выполнении каждого задания теста. Если задание правильно решили 70% испытуемых, степень его сложности (Р) равна 70. Чем легче задание, тем больше Р в числовом выражении. Поскольку степень трудности задания зависит и от его месторасположения в структуре всего теста (батареи), рекомендуется в проверочном испытании использовать несколько вариантов теста с различной последовательностью заданий*.

---

* Ингенкамп К. Педагогическая диагностика. М.: Педагогика, 1991. С. 131.

Следует различать понятия *трудность* и *сложность* задания. Трудность более субъективна и во многом определяется качеством предыдущей учёбы учащихся, особенностями его развития, психофизическим состоянием на сеансе тестирования.

Сложность задания закладывается в диагностических целях и отражается в формулировке разработчиком теста задания и инструкции к его выполнению, в выборе формы предъявления задания. Отчасти (как и на трудность) на сложность задания влияет и его месторасположение в структуре теста. Сложность задания задаётся разработчиками теста, а будет ли оно трудным для какой-то подгруппы учащихся и в какой степени — это определяется индивидуальными и групповыми особенностями диагностируемого контингента.

То, что будет трудным в обычной сельской школе (может быть, из-за нерутинности заданий и непривычности тестирования вообще), может быть несложным даже у слабых по данному предмету учащихся обычной городской школы. Особенно если её учителя часто применяют самодельные тесты для текущего контроля, а то и учат самих учащихся их разрабатывать, как В.М. Распопов*, практиковавший такой дидактический приём в ПТУ (что делает и ряд ростовских учителей).

---

* Распопов В.М. Программирование и организация самостоятельной работы учащихся. М.: Высшая школа, 1986.

Для определения степени сложности заданий мало пригоден какой-либо иной подход, кроме предварительного или последующего (за тестированием) экспертного оценивания сложности заданий теста группой высококвалифицированных педагогов-методистов первой и

высшей категорий, участниками творческих групп. Рекомендуемая некоторыми авторами методика учёта количества операций, выполняемых учащимися при решении конкретных тестовых заданий, на наш взгляд, пригодна только для несложных гомогенных тестов и примитивна, если не принимаются в расчёт качественные особенности заданий. Кроме того, методика слишком громоздка и трудоёмка, если учитывать все существенные характеристики заданий (число которых свыше десятка).

Экспертиза в нашей практике работы с творческими группами разработчиков измерителей по ВГОС проводилась следующим образом: 8–12 учителей первой и высшей категорий вместе с методистами ИПК и ПРО определяли первоначально самое простое задание в проекте теста. Здесь быстро удаётся достигнуть согласия экспертов. Чаще это задания на знание фактов, дат, на распознавание и элементарное сопоставление. Такому заданию присваивается коэффициент 1,0. Остальные задания оцениваются экспертами. Целесообразно это делать в режиме домашней самостоятельной работы для того, чтобы избежать давления авторитетов или стремящихся к лидерству участников группы. Затем методистом или руководителем группы данные усредняются и выводятся усреднённые оценки.

Такая методика позволяет использовать практически все средние величины и меры рассеяния (*моду, медиану, среднеквадратическое отклонение, вариационный размах*, а при желании и *корреляционный анализ* для определения разброса оценок).

В достаточно сложных случаях, при очень высоком разбросе оценок, целесообразно провести групповую мини-дискуссию для уточнения аргументов, позиций экспертов и внести по её итогам коррективы. Коэффициенты сложности заданий умножаются на «сырые» баллы, полученные учащимися при решении заданий теста (а при более глубоком анализе и на потерянные испытуемыми за ошибки баллы). Это позволяет даже в традиционном нормативном и относительно гомогенном тесте повысить качество измерения диагностируемых знаний и умений.

Для более точного научного анализа возможно введение дополнительных коэффициентов компетентности экспертов (хотя это целесообразно лишь в том случае, когда тест, по данным эмпирического анализа, разработчикам удался).

Напомним, что в нормативном тестировании принято считать, что овладение учащимися *ориентировочной основой действий* в какой-либо конкретной области достигается в том случае, когда *коэффициент усвоения* знаний или умений выше 0,7. **Коэффициент усвоения** показывает, сколько учащихся правильно выполнили операции из числа необходимых для решения теста.

В зарубежных педагогических концепциях успешным считается обучение, при котором учащийся на 95% вопросов даёт правильные ответы,* а в особо сложных и ответственных видах деятельности в профессиональной педагогике только обучение, дающее 100% усвоения (работа операторов на ТЭЦ, на АЭС, пилотов авиалайнеров).

---

* Беспалько В.П. Слагаемые педагогической технологии. М.: Педагогика, 1989.

Разумеется, определение необходимого минимального коэффициента усвоения знаний и умений зависит и от их структурной сложности, и от «удельного веса» значимости в общем комплексе диагностируемых знаний и умений. Для создания хорошего теста обычно стремятся к распределению степени сложности заданий от P=20 до P=80 со средней величиной P=50.

Коэффициент селективности заданий определяет взаимосвязь этого задания в пробной (первоначальной) форме тестов (корреляцию от -1 до +1). Если учащиеся, не выполнившие данное задание, не могут решить все остальные, коэффициент селективности приближается *к* +1. Если хорошие ученики не справляются с заданием так же часто, как и плохие, то *коэффициент селективности* задания равен нулю.

Если слабоуспевающие ученики решают задания лучше хорошо успевающих (например, при неточных формулировках, побуждающих «сильных» на поиск ошибочного пути), то коэффициент селективности может иметь отрицательный знак. Желательно, чтобы коэффи-

циент селективности был выше 0,30.

Одной из задач эмпирического анализа тестов является выявление по ответам учащихся **типичных ошибок** (затем их перечень и характеристики указать в «Руководстве к тесту»).

Ю. Гутцке и У. Волрабом разработана классификация ошибок, учитывающая особенности построения немецких диагностических программ и ориентированная на использование обучающих машин:

• случайные ошибки (которые невозможно отнести к другому виду);

• нераспознанные правила или ошибки переноса, когда правило, относящееся к предыдущим задачам, применяется в последующих задачах, на которые оно не распространяется;

• правильное понимание правила, но неполное его выполнение либо невыполнение в трудном случае;

• не логические, а «систематические» ошибки.

Подобная типологизация ошибок вполне приемлема для эмпирического анализа дидактических тестов по любым предметам. Разумеется, при условии их конкретизации, привязки к проверяемой области знания и к диагностическим целям каждого самостоятельного тестового задания. При эмпирическом анализе дидактических тестов одной из ключевых задач является также выявление того, насколько эффективно сработаны в «закрытых» заданиях и вопросах тестов различные варианты ответов — **дистракторы** — варианты ответов, отвлекающие от эталона.

В первоначальную версию теста, которая только начинает проходить апробацию, обычно включается больше, чем необходимо, дистракторов в вопросы либо же ставятся параллельные (эквивалентные) вопросы, в том числе в параллельной форме теста.

При анализе качества апробируемой методики тестологи придерживаются правила, выработанного эмпирически в первой половине XX века. В тестах с набором вариантов ответа ни один дистрактор не должен быть столь **невероятным**, чтобы его выбрали менее 5% опрошенных. Если такое произошло, дистрактор надо заменять или перестраивать данное тестовое задание и проверять его снова.

Имеются существенные различия в **обработке** и **анализе** качества нормативных и критериальных тестов. В первом случае подсчитываются баллы и с помощью таблицы норм превращаются в нормативные и стандартные величины. Во втором — при критериальном тестировании — необходимо выявить: а) когда можно утверждать, что достигнута учебная цель, отражённая в ряде заданий (субтесте); б) какой процент заданий решён каждым учащимся; в) к какой группе (категории) по успеваемости относится данный учащийся в зависимости от достижения им учебной цели, приближённости к этой цели*.

---

* Ингенкамп К. Педагогическая диагностика. М.: Педагогика, 1991. С. 137.

Разработка и перепроверка нормативов теста

Большинство применяемых в массовой зарубежной диагностической практике дидактических тестов в настоящее время являются нормативными. Недостаточная разработанность технологии стандартизации критериальных тестов, отсутствие у российских учителей и вузовских учёных опыта их создания приводят к ориентации разработчиков на морально-устаревшую, но проверенную практикой технологию создания всё тех же нормативных тестов. Их стандартизация обязательна, связана с выработкой (на основе эмпирического анализа выборки апробации) **нормативов**, ориентируясь на которые можно сказать, преуспел ли проверяемый учащийся в достижении учебной цели. А если преуспел, то насколько, какое место он занимает среди других учащихся класса, школы и прочих своих сверстников, прошедших тестирование по данной методике.

Нормативы позволяют оперативно ставить диагноз и применять коррекционные меры, управленческие решения (отсев, перевод в следующий класс и т.п.). Так, американские нормативы скорости чтения про себя: в начальном школьном возрасте — 80–150 слов в минуту, в

среднем школьном возрасте — 175–204 слов в минуту, старший школьный возраст — 214–250 слов, студенческий возраст — 250–280 слов в минуту.

Существует несколько распространённых подходов к выработке нормативов теста. Выбор разработчиком каких-либо из них не является случайным, а определяется типологическими характеристиками теста, его диагностическими целями и возможностями, имеющейся в распоряжении техникой (счётами или ПЭВМ типа «Пентиум»).

Современные тесты обязательно содержат данные **внутригрупповых норм**. Эти нормы дают возможность сравнить результаты школьника по тесту (например, по географии) с показателями по тому же тесту других учащихся одного с ними возраста из параллельного класса в той же или соседней школе. Все внутригрупповые показатели имеют единый и чётко определённый количественный смысл. С ними можно проводить большинство различных статистических операций, анализировать более или менее подробно — в зависимости от того, что требуется диагносту, на какой уровень глубины диагноза и анализа результатов применения методики он готов подняться. Одно дело — оперативная симптоматическая диагностика для выяснения, всё ли в порядке по интересующему педагога разделу. Другое — диагностика причин существенных отклонений в темпах обучения, в усвоении ключевых понятий, в овладении расчётными или иными навыками, замеченная в какой-либо группе классов.

Современные компьютерные программы обеспечивают возможность самых разнообразных подсчётов — от средних величин со среднеквадратичными отклонениями до представления данных в компьютерных вариантах в таблицах и графических формах. Так, йошкар-олинская программа «Резонанс» даёт 7 наглядных форм представления данных. Но, к сожалению, она ориентирована на примитивную ручную начальную обработку тестовых заданий и недостаточно совершенна, что ограничивает потенциал эмпирического анализа тестов с её помощью.

Внутригрупповые нормы (а их часто называют ещё соотносительными) исторически являлись первым видом нормативов и в психодиагностических, и в дидактических тестах. Их выработка и перепроверка (в адаптируемых переводных или исследовательских тестах) продолжается почти столетие.

Перенос известного теста на новый, ранее не изучавшийся с его помощью контингент влечёт за собой обычно выработку специфических для этого контингента внутригрупповых норм. Они должны войти в очередное переиздание теста, дополнив ранее существовавшие (для других контингентов) нормативы. Если учесть большие достижения современной психометрии и математической статистики, можно сказать, что этот подход пока не исчерпал себя, просто он может быть дополнен другим.

Соотносительная норма может быть трёх типов — при сопоставлении с результатами других учащихся (как социальная норма), при сравнении с прежними результатами того же ученика (индивидуальная), при сравнении с поставленными учебными целями (предметная).*

---

* Ингенкамп К. Педагогическая диагностика. М.: Педагогика, 1991. С. 45.

Для того чтобы осознанно разбираться в нормативах дидактических тестов, педагогу-разработчику и педагогу-пользователю необходимо знать азы математической статистики, применяемой в тестологии. Для этого достаточно объёма знаний по математике за основную школу — того минимума, которым обладает практически каждый учитель, даже если он давно не обращался к математическим расчётам.

Профессионально составленные дидактические тесты по инструктивно-методическому аппарату во многом схожи с такими же профессиональными психодиагностическими тестами. Их нормативы основываются на традиционной для математической статистики системе расчётов (и, соответственно, элементарных понятиях и общепринятых показателях).

Рассмотрим самые необходимые из этих понятий и показателей. Знакомство с ними позволяет педагогу-диагносту, разрабатывающему или адаптирующему дидактический тест, выработать основные нормативы теста.

Почти столетняя практика психологической и педагогической тестологии показывает,

что для репрезентативной **апробации** теста, при которой результаты могут считаться статистически значимыми и достоверными, обычно требуется на начальной стадии не менее 400 опрошенных учащихся. Если представить полученные результаты в алфавитном или любом другом списке, он отпугнёт малоопытного разработчика своей внушительностью. Для того чтобы сравнить данные, полученные учащимися одного класса между собой, или сравнивать данные разных классов (а их может быть от 10–12 до 20)— такой список неудобен.

В статистике принято в этом случае составлять таблицу частотного распределения. В ней все показатели, полученные всеми обследуемыми учащимися, группируются по заранее определённым тестологами **интервальным значениям**. Величину интервала определяет обычно сам разработчик. Допустим, в тесте на знание биологической терминологии (в объёме базовой школы) для девятиклассников было 60 заданий. Таблица показывает, что у нас получилось при интервале в пять заданий (см. табл. 1).

Таблица 1. Частотное распределение результатов 1000 учащихся средних школ (9-й класс) по тесту знания биологической терминологии

Интервал (классы)	**Частота**
56–60	100
51–55	30
46–50	40 110 170 190 160 130
41–45	90
36–40	60
31–35	10
26–30	
21–35	
16–20	
11–15	
6–10	
0–5	
Всего	**1000**

Принято два наиболее распространённых способа построения графика — **гистограмма** и **полигон частот**.

Для построения гистограммы надо вычертить над каждым интервалом столбец, высота которого будет соответствовать числу учащихся, решивших в тесте определённое количество упражнений. Полигон частот укажет количество испытуемых, решивших число заданий в пределах каждого интервала, точками над серединой интервала, которые будут соединены прямолинейными отрезками.

Можно ли обойтись без этих хлопот и получить тем не менее достаточно объективное представление о результатах обследования только на основе обычных, простых таблиц? Можно, но... Как отмечают специалисты по статистике в педагогике Дж. Гласс и Дж. Стенли, «часто график или таблица говорят больше, чем мы хотим или должны знать».*

---

* Гласс Дж., Стенли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976.

Эта же информация может быть представлена графически в виде кривой распределения.

По горизонтальной оси отложены интервалы тестового показателя, по вертикали — число случаев (учащихся из выборки апробации теста), приходящихся на данный интервал.

Часто используются 2–3 наиболее существенных показателя из числа **мер центральной тенденции**. Педагоги применяют их в своих отчётах по текущей работе (хотя далеко не полностью). И чаще всего не зная их научных названий — подобно герою Мольера, под старость лет узнавшего, что он всю жизнь говорил прозой.

Например, **выборочное среднее** (оно же *среднеарифметическое*), обозначаемое обычно как М. На глаз легко определяется **мода** (*модальное значение*) — наиболее часто встречающийся результат.

Третья мера центральной тенденции — **медиана**. Это значение, которое делит (по возрастающей или убывающей) последовательность результатов пополам. В нашем примере

интервал медианы приходится на интервал 26–30. В выборке из более сильных учащихся он бы был выше, а у более слабых — ниже (большее или меньшее число учащихся решило бы то или иное количество заданий).

Среди **мер разброса данных**, которые характеризуют величину (и степень) отклонения результата каждого ученика от средней и медианы, наиболее известны и необходимы начинающему тестологу четыре параметра.

При анализе контрольных работ учителя используют **размах распределения** (*вариационный размах*) — разность между самым высоким и самым низким результатом. Допустим, что попавшие в верхний интервал (56–60) учащиеся решили по 56 заданий, а «слабаки» в самом нижнем — по 5 заданий. Тогда размах распределения такой: 56 - 5 = 51. Это даёт информацию о диапазоне колебаний результатов в данной выборке. Но, как считают специалисты, размах распределения — неточная и недостаточная мера (уйди все они в кино во время тестирования — существенно изменился бы и размах распределения, сместившись на расположенные ближе к медиане интервалы).

Более точный метод измерения — учёт **разности** между индивидуальным результатом каждого обследованного и средним значением по всей выборке. Сумма отклонений всегда будет нулевой.

В статистике усредняют абсолютные значения отклонений, отбрасывают знаки (плюс или минус) и получают очень показательную меру — **среднее отклонение**, обозначаемое $|x|$. Этот показатель хорошо помогает описать распределение, но он малопригоден для математического анализа результатов тестирования. Здесь негативно сказывается то, что при подсчётах отбрасываются знаки (плюс-минус) и к тому же этот показатель не обладает рядом свойств, которые делали бы его удобным для математического анализа.

Две другие меры разброса — **стандартное отклонение** (сигма малая) и **дисперсия** (средний квадрат отклонения) будут рассмотрены в связи с выработкой нормативов теста — стандартных показателей.

Среднее отклонение во многом определяет **ошибку измерений** в тесте. Специалист по педагогической диагностике К. Ингенкамп считает, что для нормативных тестов ошибка в измерении может быть связана с количеством заданий в тесте и составлять плюс-минус 2 балла, если заданий менее 24, плюс-минус 3 балла при 24–47 заданиях, плюс-минус 4 балла при 48–89 заданиях.*

---

* Ингенкамп К. Педагогическая диагностика. М.: Педагогика, 1991. С. 39.

В тестологии широко применяются при анализе результатов тестирования и при стандартизации тестов **стандартное отклонение распределения** — мера разброса, обозначаемая буквой (сигма). Она позволяет сравнивать разброс результатов в разных группах обследуемых (в разных школах, классах, в группах, выделенных по уровню успеваемости, в половозрастных группах и т.д.).

Так как при её вычислении устраняются отрицательные знаки, ею удобнее пользоваться при выполнении многих операций анализа, чем средним отклонением.

$$\sigma = \frac{\sum |x|}{N},$$

где $|x|$ — отклонение индивидуального результата от среднего значения по группе, а N — число случаев.

Одним из самых распространённых в тестологии показателей является **дисперсия** (среднеквадратичное отклонение). Дисперсия хорошо показывает особенности разброса данных — насколько «плотно» результаты в классе «А» отличаются от классов «Б», «В» и «Г», т.е. насколько однородна основная масса обследуемых в решении того или иного теста, тестового задания, в выборе каких-то вариантов ответов (правильных или ошибочных).

Специалисты считают, что лучшей выборочной характеристикой дисперсии при небольших выборках (несколько десятков обследуемых) или несмещенной оценкой дисперсии является подсчёт её по формуле

$$\alpha^2 = \frac{\sum x^2}{N-1}.$$

Внутригрупповые нормы обычно в тестологии представляются в виде **процентилей** или **стандартных показателей**.

Процентили определяются как процентная доля обследуемых из **выборки стандартизации**, первичный результат которых оказывается ниже данного первичного показателя (числа решённых задач из общего их количества). Если 30% пятиклассников решают 20 заданий теста из 40, то первичному показателю 20 будет соответствовать тридцатипроцентный процентиль. Обычно он обозначается P30. Процентиль показывает, к какой подгруппе внутри выборки стандартизации относится каждый обследуемый, в зависимости от того, сколько заданий (из их общего количества в тесте) он решил. Это, по сути дела, ранговые градации с интервалом в единицу в шкале от 0 до 100%, когда отсчёт идёт сверху (со 100%), в то время как в традиционной ранговой шкале лучшим является тот, кто занимает 1-е место. Медиана — это P50, то есть 50-я процентиль. Все, что выше 50, — это показатели выше среднего, а все, что ниже, — сравнительно низкие.

Так как при измерениях нередко используют **квартили** (четвёртые части) — 25%, 50%, 75%, то 25 и 75-процентным квартилям соответствуют первые и третьи квартили (Q1 и Q3). Они выделяют из всего распределения данных нижнюю и верхнюю части. Вместе с медианой они помогают определить, какое положение среди всех обследуемых займёт тот или иной ученик, решивший определённое количество заданий.

Процентные показатели выполнения заданий отличаются от процентилей. Это вовсе не одно и то же. Процентиль указывает **долю** от общего числа членов группы тех учащихся, которые решили определённое количество заданий, т.е. не насколько хорошо Вася решил тест, а какой процент учеников из числа прошедших обследование по этому тесту оказался в таком же, как и он, положении. Процентный показатель отмечает совсем иное: сколько процентов заданий от общей суммы решил Вася.

Процентили позволяют проводить ряд сравнительных операций, наглядно иллюстрирующих результаты тестирования на уровне учебного класса школы. Это хорошо описано в статистике, в прикладной психологии и социологии.

Стандартные показатели широко используются в современных дидактических и психодиагностических тестах. Они достаточно универсальны. Смысл их заключается в том, что они диагностируют **отклонение результатов** конкретного учащегося от **средней нормы** в единицах, пропорциональных стандартному отклонению распределения. Существует два пути получения стандартных показателей.

Первый — **линейное преобразование**. Его преимущество заключается в том, что сохраняются соотношения между первичными показателями. Реально самые распространённые шкалы норм имеют мелкие интервалы (сколько заданий решил в тесте каждый ученик). Стандартные показатели вычисляются в этом случае **вычитанием из каждого первичного показателя** одной и той же величины. Результат делится на другую постоянную величину. Относительная величина разницы между стандартными показателями Маши и Даши при линейном преобразовании точно соответствует относительной величине различия первичных показателей этих учениц. В итоге, и с первичными, и со стандартными показателями можно проводить одни и те же вычисления при анализе результатов тестирования — никакого искажения конечных результатов не будет.

Линейно преобразованные показатели обычно называются **z-показателем:**

$$Z = \frac{(X - M)}{\sigma}.$$

Например, Маша решила 50 заданий, Даша — 42 — значит, $X_1 = 50$, а $X_2 = 42$. Среднеарифметическое решение заданий в 6-м «А» классе 45 ($M = 45$), σ^2 (сигма) = 5. Вычисляем значения стандартных показателей Маши и Даши (Z_M и Z_D).

$$Z_M = \frac{50-45}{5} = +1,$$

$$Z_D = \frac{42-45}{5} = -0{,}6.$$

Получившийся операциональный показатель демонстрирует, как и первичные данные, что результаты Маши выше среднего, а у Даши — ниже. Информация в данном случае почти та же. Однако первичные показатели трудно, а иногда бессмысленно сравнивать между собой. Особенно если в Ростове учащиеся работали при проверке знаний по биологии по одному тесту, а в Краснодаре и Волгограде по близким его модификациям. А вот стандартные показатели вполне сопоставимы — они не привязаны к конкретному числу заданий в каждом из тестов, а поэтому — универсальны.

При расчёте нормативов теста обычно используются **нормализованные стандартные показатели**, которые вычисляются с помощью широко известных статистических таблиц. В этих таблицах приводится процент случаев различных отклонений в единицах от среднего значения нормальной кривой.

Нормализованные стандартные показатели имеют ту же форму, что и линейно преобразованные стандартные показатели. Они соответствуют распределению, преобразованному так, что оно принимает вид нормальной кривой. Результат 0 у Маши будет соответствовать самой середине нормальной кривой (т.е. превосходит 50% результатов в её классе). Маша, получившая $Z_M = +1{,}0$, будет превосходить уже 84% учеников, а лентяйка Таня, у которой нормализованный показатель окажется — 1,0, будет иметь лучший результат только по сравнению с 16% ещё более ленивых или малоспособных учеников. Даша будет несколько выше её ($Z_D = -0{,}6$), но до Маши ей далековато — надо учить и учить то, что задавали педагоги.

Показатели тестов достижений в обучении часто интерпретируются с помощью понятия **эквивалентный класс** — ученик достиг уровня 7-го класса по орфографии, уровня 8-го класса по технике чтения, но остался на уровне 5-го класса по арифметике.

Это понятие столь же наглядно, как и умственный возраст в тестах интеллекта. Сколько бы о нём ни спорили, а всё равно используют, так как оно эффективно демонстрирует индивидуальные особенности достижений каждого отдельного учащегося в интеллектуальном развитии и позволяет сопоставлять учащихся между собой.

Нормы классов выводятся путём подсчёта среднего первичного результата при апробации теста на учениках данного класса (в условиях обеспечения относительно репрезентативной выборки). То, какой процент (среднее количество) учащихся решает задания теста, и будет определять норму для данного класса.

Несмотря на популярность среди исследователей, классные нормы имеют ряд недостатков. Во-первых, они применимы только к общеобразовательным предметам стабильных курсов, которые ведутся на всех уровнях образования, охватываемых данным тестом, как минимум несколько лет подряд. Они не годятся для старших классов, особенно для профилированных школ, лицеев, гимназий, в которых многие предметы изучаются полугодие, год, максимум — два. В результате данные не с чем сравнивать, особенно — если дело касается факультативов или альтернативных, нововведённых курсов, а также курсов, отражающих региональные стандарты образования («История донского казачества», «Культура Дона» и т.п.).

Во-вторых, изменения в содержании образования в целом (госстандарты и новые учебные планы) и в содержании, темпах изучения предметов по годам обучения приводят к тому, что становятся диспропорциональными (по сравнению с привычными и стабильными) часы, отводимые разными школами (в соответствии с правами Закона «Об образовании») на изучение различных предметов. Не всегда удаётся выявить определённую закономерность в изменении такой расчасовки предметов даже в одной школе — слишком много в оргработе управленцев школ по составлению учебных планов и сетки часов ситуативного, научно мало мотивированного.

Классные нормы — это не нормативы выполнения тестов. Они только показывают точку отсчёта от среднего уровня для данного класса. Отдельный учащийся может быть выше или ниже этой классной нормы, но сие не означает, что его надо срочно переводить в предыдущий или последующий класс. Просто он или отстаёт от своих одноклассников, или опережает их в успехах и развитии. Возможно, временно. Родители найдут хорошего репетитора — и будет резкий рывок вперёд. Если таких детей в классе несколько, резко подскочат вверх и классные нормы.

И при разработке дидактического теста, и при выработке его нормативов, и при анализе собранного по классу, школе материала интерес для диагноста представляют ещё одни специфические нормативы.

Промежуточные эквивалентные классы определяются путём интерполяции. Их можно получить и непосредственно тестируя детей несколько раз в учебном году. Поскольку учебный год длится 10 месяцев, на каждый месяц приходится по 0,1 года. Подсчёт ведётся следующим образом. Например, число 4,0 означает среднее по классу выполнение теста по родному языку (математике и пр.) в начале 5-го класса (года обучения) и данные эти были получены при сентябрьском тестировании. Тогда данные на уровне 4,6 будут соответствовать февральскому тестированию по той же самой методике, полученному спустя почти шесть месяцев. А если в соседнем классе на этот уровень ребята вышли в том же сентябре? Значит, они опережают своих сверстников на полгода.

Можно рассчитать и индивидуальное и групповое продвижение учащихся в каждом из классов, где определены данные нормативы. Только надо постоянно помнить, что это **условный** показатель, и не пытаться на его основе принимать скоропалительные решения (от выговоров учащимся за лень до критики учителей на педсоветах).

Одним из вариантов выработки нормативов является употребление **шкалы порядка**, основанной на психологическом подходе к процессу развития ребёнка и формированию у него психических качеств, умений, навыков. Есть определённая поэтапность в достижении какого-либо уровня развития (психической или психофизической функции, учебных умений, интеллектуальных операций, в усвоении знаний). Многие методики классика современной психологии Жана Пиаже построены на этом принципе, используются в психодиагностике. Но аналогичный подход может быть применён и в дидактическом тестировании. Особенно в тех предметных областях, где совместными исследованиями дидактов, методистов и психологов выявлены логически следующие один за другим этапы усвоения материала (понятия, идеи, закономерности алгоритма действий). Соотнесение этих этапов с возрастом, с объёмом и содержанием пройденной **стабильной** программы учебного курса (цикла, предметной области) и с определёнными **государственными стандартами** целями обучения (диагностично сформулированными применительно к каждому году обучения) даст обоснованные возрастные и поклассные нормативы для рубежного и итогового контроля.

Практическое решение задачи выработки нормативов при таком подходе будет зависеть только от официального принятия в законодательном порядке госстандартов, в которых цели обучения и развития должны быть обозначены предельно диагностично, с максимально полным, структурированным описанием ожидаемого **минимального** (а лучше и **оптимального**) результата обучения, а также от активности и профессионализма разработчиков тестов, что тоже является сегодня проблемой.

Для тестирования широко используется в выработке и представлении нормативов «Шкала Т-величин», для которой установлены средняя величина 50 и стандартное отклонение 10. В статистических справочниках имеются таблицы с Т-величинами, которыми можно пользоваться, если вычислены позиции учащихся при ранжировании или кумулятивная частотность (сколько случаев приходится на каждую отметку шкалы — например, сколько учащихся решило пятое задание). На практике встречаются такие шкалы Т-величин, которые простираются от Т-величины 20 до Т-величины 80. «Края» Т-величин обычно малоинформативны и используются лишь при процедурах выверки теста в специальных методологических экспериментах.

Реально самые распространённые шкалы норм имеют мелкие деления — процентная ранговая шкала от 1 до 100, шкала Т-величин — от 20 до 80. Различные значения шкалы достаточно точно отображают даже незначительные выражения диагностируемых признаков. Неточность измерений отражается в вычислении стандартной измерительной ошибки, которая показывает, в каких пределах можно доверять полученным результатам, какие возможны при данной методике колебания значений.

Для облегчения интерпретации данных разрабатывается нередко **лента норм**, или **лента Т-величин**, учитывающая в своих интервалах (от ...до...) величину стандартной измерительной ошибки. Это последовательная шкала, на которой фиксируется конкретный интервал, соотносимый с процентной ранговой шкалой и средней Т-величиной. Воспользуемся в качестве примера фрагментом такого соотношения для теста школьной успеваемости второклассников по книге К. Ингенкампа. В этом примере базовая шкала, дающая «сырые» баллы, имеет градацию до 124. Такие шкалы встречаются в тестах интеллекта и обучаемости и в дидактических тестах итогового и входного контроля (см. табл. 2).

Таблица 2. Пример перевода показателей шкалы в процентное ранжирование, средние Т-величины и ленту Т-величин

1 — показатели шкалы
2 — процентное ранжирование
3 — учащиеся средняя Т-велечина
4 — лента Т-велечин

1	**2**	**3**	**4**	
	58 – 66	53		54
75 – 77	62 – 73	54	2	56
78 – 80	66 – 79	56		58
81 – 83	73 – 82	58	3	59
84 – 86	79 – 84	59		60
87 – 89			4	
			6	
			8	

Если учащийся второго класса получил по базовой шкале 76 баллов, то при процентном ранжировании он оказывается в интервале 58–66 пунктов. Это значит, что его результаты при минимальном показателе (это было бы 75) столь же хороши, как и результаты 58 всех второклассников, или лучше их, а при максимальном показателе (77) — столь же хороши, как и результаты 66 сверстников, или лучше. Этот интервал может быть связан со средним значением шкалы Т-величин 53. Эта средняя Т-величина даётся для того, чтобы диагност мог просчитать средние величины с помощью ленты Т-величин, в которую заложена измерительная ошибка в 4,5 единицы.

Определение или перепроверка нормативов теста — довольно кропотливая работа. И хотя почти все её операции можно выполнить вручную или с самым простым калькулятором, лучше всего пользоваться современными ПЭВМ.

Педагог, проводящий апробацию дидактического теста, должен руководствоваться общим правилом — **не терять информацию**. Наиболее целесообразно вводить в компьютер **все** данные с бланка ответов учащихся, так как при определении (или перепроверке) надёжности теста, при его критериальной валидизации в целом и по отдельным субтестам (блокам заданий) «мелочи» могут играть существенную роль (включая ошибочные ответы испытуемых, не понявших суть задания и отмечавших несколько вариантов ответа там, где надо было выбрать один). При этом особое внимание следует обратить на ввод демографических характеристик и, если есть возможность, характеристик общей и предметной успеваемости учащихся.

Критериальная и прогностическая валидизация тестов

Критериальная валидизация теста проводится только после его применения, когда уже получены первые данные, проведён эмпирический анализ и проверена надёжность теста. Здесь во всей красе продемонстрирует «свои прелести» недостаточная репрезентативность выборки апробации и, может быть, невысокая надёжность методики. **Критериальная** валидность требует сравнения результатов данного теста с внешними критериями (например, оценками учащегося по предмету, процентом брака при выполнении производственных заданий) или же с результатами других аналогичных тестов. В последнем случае говорится о **конкурентной** валидности критериев (устанавливается путём определения корреляций между ранее применявшимися и новыми тестами). Критериальная валидность показывает возможность прогноза поведения, успешности обучения обследуемого в настоящем и будущем. Критериальная валидность дидактического теста во многом определяется особенностями его конструктной валидности: насколько адекватно в конструкте-модели теста отражена диагностируемая реальность — структура и существенные характеристики предметной или межпредметной области изучаемых знаний и умений, структура адекватных этим знаниям видов учебной и профессиональной деятельности. Конструктную и критериальную валидность теста часто в психодиагностике объединяют общим названием **эмпирической** валидности, которая для учителя и школьного психолога имеет важнейшее значение.

А.Г. Шмелев отмечает, что «прагматические тенденции западной текстологии привязали эмпирическую валидность теста к внешним для психологии социально-прагматическим критериям», в качестве которых выступают показатели, имеющие ценность для определённой практической области. Например, «успеваемость», которую надо повысить, или процент ошибок в решении профессиональных задач, которые надо понизить. Коррелируя тесты с этими показателями, разработчик одновременно «решает сразу две задачи: задачу измерения валидности и задачу измерения практической эффективности своей психодиагностической программы». Если получен значимый коэффициент корреляции, то можно считать, что решены с позитивным результатом сразу две задачи. Но если корреляции не обнаружено, то остаётся неясным: либо невалидна сама процедура, либо неверна гипотеза о наличии причинно-следственной связи между свойствами (знаниями и умениями) и измеряемыми социально значимым показателем — способностью решать профессиональные задачи определённого типа, для которых предполагалась необходимость этих знаний и умений, например, графических или математических.*

---

* Куписевич Ч. Основы общей дидактики. М.: Высшая школа, 1986.

В тех случаях (а в педагогике они обычны), когда на основании результатов тестирования принимаются коррекционные меры (дообучение, консультирование), то повышение показателей (достоверное по сравнению с контрольной группой в методическом эксперименте) показывает одновременно и эффективность принятых мер воздействия. Отрицательный результат вызывает дополнительную неопределённость — за счёт чего его можно отнести (низкая валидность диагностики или неадекватная методика обучения).

Сложности критериальной валидизации дидактических тестов учебных достижений заключаются в том, что не всегда удаётся найти внешние критерии, адекватно отражающие ту или иную предметную область. Здесь химикам и физикам намного легче, чем гуманитариям. Теоретические знания, диагностируемые тестом по химии или физике, можно проверить прямо в школе на лабораторно-практических занятиях, на уроках труда или производственного обучения. Хотя, конечно, электротехнические знания и умения будут легче подвергаться перепроверке на практике, чем понимание теории волн и т.д. Более того, существуют массовые виды профессиональной деятельности, в которых реализуются физико-химические (или биологические, языковые) знания и умения.

Намного сложнее гуманитариям — историки вряд ли смогут дождаться, пока кто-то из их учеников откроет Трою или проникнет в гробницу фараона. Знания обществоведения также имеют в основном мировоззренческий характер и напрямую не коррелируют ни с показате-

лями общественной активности учащихся (а также взрослых), ни с их успехами в продвижении по социальной лестнице, ни с экономическим благосостоянием. Это вряд ли кто-то серьёзно изучал — не исключено, что какие-то опосредованные связи здесь всё-таки могут существовать.

Преподаватели литературы, этики, культурологии также вынуждены или выбирать какие-либо некоррелирующие «очевидно» внешние критерии, или довольствоваться их заменителями — критериями учебной успеваемости.

Прогностическая валидность имеет как бы два вектора: **ретроспективная** валидизация, когда обследование по тесту уже проведено, и **проспективная** (называется иногда **перспективная**), когда обследование только ещё предстоит провести.

В первом случае обычно привлекают к проверке обследуемых, давших самые низкие и самые высокие результаты.

В случае перспективной валидизации выборку обследуемых надо составить с запасом, с учётом вероятности крайних (экстремальных) групп в будущем. При этом надо учесть данные ранее проводившихся аналогичных обследований и (или) мнения экспертов.

Ретроспективная валидизация позволяет доказать валидность самого измерения, а перспективная — как валидность измерения, так и наличие предполагаемой причинной связи.

Величина коэффициента прогностической валидности бывает различной, но весьма редко свыше 0,70. Так, **прогностическая валидность** в виде предсказания успеваемости через год (по математике) в диагностических программах Ю.Гутцке и У. Волраба составляет r = 0,63, а через два года r= 0,64, что можно считать очень высоким результатом.

Есть немалые различия в прогностической валидности дидактических тестов разного типа. Как считают специалисты, тесты обучаемости претендуют на лучший прогноз успешности будущего поведения и учения по сравнению с традиционными тестами интеллекта. Поэтому доминировавшая в большинстве ранее проводившихся исследований тестов обучаемости конкурентная валидизация теста с помощью школьных оценок и суждений учителей остаётся спорной и неудовлетворительной, даже если исходить из того, что эти внешние критерии отражают в определённой мере результаты прошлых процессов учения. Принимая во внимание основную цель теста обучаемости, желательнее было бы иметь критерии изменчивости, однако получить их чрезвычайно трудно.

Целый комплекс специфических методологических проблем связан с современным подходом к тестам обучаемости как не только диагностическому, но и дидактическому, формирующему средству. Различие результатов заключительного и предварительного тестов даёт ценную информацию об эффективности обучающих воздействий по корректировке учителями слабых позиций учащегося, выявленных в первом сеансе тестирования.

Критериальная и прогностическая валидность на практике просчитывается не только по тесту в целом, но и по его субтестам. Было замечено, что отдельные субтесты во многих известных методиках лучше работают на прогноз у одной подгруппы обследуемых (например, девочки из высокообразованных семей), чем у других (мальчики из сельских школ). В одних подгруппах коэффициенты критериальной валидности могут быть намного выше средних (средними для большинства американских тестов являются сравнительно невысокие 0,40–0,60).

В практике экспериментирования, как отмечалось выше, критериальная, прогностическая и другие виды валидности (за исключением содержательной) часто объединяют термином **эмпирическая валидность**. Она просчитывается при анализе тестов в очень широком диапазоне — от батареи тестов до уровня отдельных высказываний, входящих в состав того или иного теста.

Эмпирическая валидность высказываний, входящих в тест (а в тестах достижений — соответственно тестовых заданий и их отдельных вариантов ответов), может быть случайной или слишком опосредованной другими факторами. Концептуальная валидность может оказаться размытой из-за связи высказываний не только с интересующей исследователя концепцией, но и с другими концепциями. Чем больше будет собрано в одном тесте наиболее

валидных по одному критерию высказываний, тем выше валидность теста в целом.

Данные о надёжности и валидности теста позволяют с определённой степенью уверенности использовать его как диагностический инструмент. Методическая работа по совершенствованию теста, по повышению его надёжности и валидности почти неисчерпаема — достижение максимального показателя является, скорее всего, идеалом, к которому надо стремиться. Чем ближе разработчик к нему приближается, тем выше становится его методическая и диагностическая квалификация и тем точнее и эффективнее тест.

Педагогическая валидность дидактического теста

Г. Витцлак ставит задачу разработки таких методов, которые позволяют делать однозначные педагогические выводы по результатам диагностики, то есть имеют высокую **педагогическую** валидность.* Он вводит этот новый термин и ставит проблему, но не даёт ключи к её решению. Он прозрачно намекает на то, что такие ключи есть. А искать их надо там, где традиционная психодиагностика умывает руки и, уходя, оставляет, как подкинутого ребёнка, свои продукты — диагнозы педагогу. Педагог умный, ему надо, он разберётся.

---

* Вицлак Г. Принципы разработки и применения психодиагностических методов в школьной практике. М.: Прогресс, 1986. С. 141.

Говорить о педагогической валидности следует в первую очередь в том случае, когда создаются диагностико-коррекционные программы (синтетические методики по терминологии известного российского психодиагноста З.И. Калмыковой). Аналогичная идея неоднократно высказывалась и в прикладной психодиагностике. Большинству практических психологов представляется самоочевидным то, что психокоррекция после психодиагностического обследования должна в определённой мере как бы программироваться типовым диагнозом.

Система таких мер, высококоррелирующих с общими для всех подобных случаев типовыми диагнозами на уровне идеи витает в воздухе, постоянно обсуждается — но не делается почти ничего. Сама по себе идея прекрасна, хотя может вызывать и некоторые возражения. Они связаны с проблемами индивидуализации коррекционных мероприятий с учётом специфики личностных, психологических особенностей обследуемого (и обучаемого в дидактическом тестировании на всех уровнях и этапах контроля знаний).

В какой степени подобный подход, связанный с разработкой типовых диагнозов, целесообразен в дидактическом тестировании?

В дидактической диагностике он реально уже почти четверть века существует в работах Ю.К. Бабанского и учеников его школы оптимизации процесса обучения. В начале 70-х годов при разработке программы изучения реальных учебных возможностей учащихся с помощью педконсилиума предлагались **типичные педагогические коррекционные** мероприятия. Они рекомендовались при низких или средних «западающих» компонентах тех или иных показателей реальных учебно-воспитательных возможностей учащихся. Но измерение в педконсилиуме проводилось не дидактическими тестами, а экспертным оцениванием (рейтингом), сопровождаемым целенаправленной групповой дискуссией педагогов-участников. Явным плюсом этого подхода было сокращение затрат времени учителей (особенно молодых, с невысокой диагностической подготовкой) на выработку коррекционных мер, которые в типичных случаях могли дублироваться у учащихся одной учебной группы или одного тестируемого потока. Разумеется, в тех случаях, когда эмпирически доказана их эффективность.

В методологическом плане такой подход стимулирует решение мало разработанных в педагогике вопросов **типологизации** диагноза. Это требует экскурса в опыт технологии медицинского диагноза, где аналогичные проблемы сотни лет решаются практически. Однако даже медики отмечали, что одной типологизации и стандартизации коррекционных мер (у них — лечебных) на основе типовых диагнозов не хватает для излечения больного, особенно в сложных случаях. Необходимы ещё **мастерство** и **интуиция** диагноста.

При компьютерной диагностике мастерство и интуиция до сих пор не используются. Можно предсказать, что пока не созданы ЭВМ со всеми сущностными характеристиками самообучающегося интеллекта человека, проблема «остатков» будет существовать и опытного педагога-диагноста можно будет подменить самой лучшей программой для ПЭВМ только до определённого уровня. Пока достаточно примитивного. Учитывая, что в диагностической деятельности педагога и школьного психолога, а также управленцев школы много рутинных элементов — от инструктирования до заготовки бланков — и это было бы подспорьем. Могло бы освободиться время диагноста на дополнение, углубление, индивидуализацию диагноза дидактического теста (особенно — компьютерного). Но лишь при условии **надёжности** и **валидности** диагноза.

Выработка и соотнесение типовых мер педагогической валидизации дидактического теста с другими данными стандартизации и нормативами может идти по линии освоения опыта прикладной социологии. Она разрабатывала с 60-х гг. метод таксономии и проблематику распознавания образов (в связи с поисками путей оптимизации выборочных исследований, оценкой репрезентативности данных, формулировкой эмпирических выводов).

Скорее всего, педагогическая валидизация будет осуществляться в монопредметных тестах рубежного и текущего контроля. Сложнее определять педагогическую валидность в тестах для отбора более образованных контингентов обучающихся — абитуриентов, аспирантов, кандидатов в магистратуру. Особенно — в батареях тестов аттестации профессиональных знаний и общекультурного уровня, охватывающих циклы учебных предметов и межпредметные связи. То же — и в прогностически ориентированных тестах обучаемости (химии, физики и т.п.), где будет достаточно силён психодиагностический аспект.

Несколько менее сложным, но весьма трудоёмким будет этот процесс для диагностики завершённости определённых ступеней образования — начальной, основной, полной средней школы, базовой общенаучной подготовки в колледже или вузе.

Менее трудоёмкими будут, на наш взгляд, межпредметные тесты и тесты, диагностирующие завершённость обучения по какому-либо циклу учебных дисциплин, по определённой образовательной области федеральных или региональных стандартов образования. И далее — батареи тестов диагностики результатов обучения по годам учёбы (рубежный контроль при решении вопроса о переводе учащегося в следующий класс, о достигнутом за год результате работы государственного или альтернативного образовательного учреждения). Меньшими по трудозатратам будут необходимые, но пока что не первоочередные тесты более мелких ступеней рубежного контроля —полугодового, четвертного. Здесь как раз и пригодится упоминавшаяся нами выше технология выработки тестовых норм **промежуточных эквивалентных классов**, позволяющая определить, на каком уровне достижений в классе находится тот или иной учащийся на момент тестирования.

Проще всего будет с разработкой проблем педагогической валидности по тем учебным предметам, которые изучаются сравнительно недолго (полгода — год), не подвержены идеологизации. Пожалуй, к ним можно будет отнести такие предметы, как ботаника, зоология.

Тематические тесты рубежного и текущего контроля знаний и умений по крупным разделам каждого учебного предмета будут завершать снизу иерархию сложности задач разработки педагогической валидности дидактических тестов.

Включение в дидактические диагностико-коррекционные программы психодиагностических тестов на выявление особенностей обучаемости как психического свойства личности — чрезвычайно перспективное дело в дидактическом тестировании. Здесь только начинаются пробные разработки, в том числе и в связи с поисками в области построения комплексных компьютерных обучающе-контролирующе-коррекционных программ с дружественным интерфейсом. Чем более развёрнутыми будут исследования в области педагогической валидизации, тем более оснащённым будет педагог. Так что при всех ограничениях это весьма перспективное направление методических экспериментов, создания баз данных в дидактическом тестировании.

Процедуры и методы педагогической валидизации

Поскольку сама проблема педагогической валидизации пока только обозначена и реальный опыт её решения более чем ограничен, можем лишь предполагать возможные практические подходы к её решению. На наш взгляд, необходимо следующее:

1. Применение экспертного метода с участием опытных педагогов-методистов и практических психологов в выработке по тому или иному **стандартизированному дидактическому тесту** типовых коррекционных мер.

2. Сочетания исследования педагогической валидности известных и вновь создаваемых дидактических тестов с изучением корреляций по этим же тестам с различными внешними критериями, то есть с проверкой их перспективной и ретроспективной **прогностической валидности**, установлением текущей критериальной валидности различных разрабатываемых вариантов коррекционных мер с типовыми диагнозами по тесту.

3. Таксономический подход и распознавание образов для установления соотношения на уровне корреляций между многопараметровыми коррекционными мерами и типовыми диагнозами. Чем больше будет комплекс диагностических показателей, заложенный в дидактическом тесте, тем сложнее должно быть технико-математическое (компьютерное) обеспечение. Могут существенно возрасти требования и к объёму памяти ПЭВМ, к их программному обеспечению, к уровню квалификации и программиста, и пользователя-диагноста. Это отразится на темпах и стоимости выполнения работ в этой области дидактической тестологии.

4. Установление по каждому типу тестов целесообразного уровня типологизации диагнозов (по степени полноты и глубины, развёрнутости).

5. Определение целесообразности педагогической валидизации каждого конкретного дидактического теста мотивировалось в первую очередь минимально необходимой степенью прогнозирования объёма, прочности и глубины усвоения обучаемым знаний и умений, диагностируемых данным тестом. Возможно, что в тестах контроля особо сложных профессиональных знаний (операторы АЭС, лётчики и т.п.) этот подход потребует многолетней экспериментальной работы.

6. Строгая операционализация понятий и диагноза, коррекционных мер, создание терминологических словарей с операциональным толкованием педагогической терминологии.

7. Введение педагогической валидности в систему требований к стандартизации дидактического теста, это повысит его коррекционную ценность и усложнит инструктивный сопроводительный материал, но и приведёт к удорожанию процесса создания хорошего теста и росту его рыночной стоимости, потребует субсидий и грантов на разработку проблем дидактической тестологии.

Педагогическая валидность наименее исследована в настоящее время и представляет собой скорее **гипотетический вид** валидности, а не технологически разработанную процедуру.

Сам факт её проверки связан с рецептурной стороной педагогического диагноза. Сколько бы сторонники интуитивного подхода в обучении и в диагностике ни утверждали, что педагогика — это искусство, в обучении всегда есть определённая алгоритмичность и рецептурность, как бы её ни проклинали фанатики идеи глобально-творческой педагогической деятельности, какими бы неласковыми словами эту алгоритмичность ни обзывали, у педагога есть своя рутина, в том числе в формулировке типовых диагнозов и адекватных им коррекционных мер. Лучше от неё избавляться.

Разумеется, педагогическая валидность какого-либо теста не должна претендовать на исчерпывающий педагогический диагноз и подменять самостоятельную аналитическую работу учителя. Но в какой-то (и автор считает — немалой) своей части этот педагогический диагноз поддаётся достаточно жёсткой и последовательной алгоритмизации: «При таких-то данных надо, как минимум, делать то-то». Далее раскрываются возможности творчества, интуиции, поиска.

Проблемы технологии педагогической валидизации могут решаться только **опытно-экспертным путём**. Педагогическая валидность требует создания банков диагностиче-

ских данных, установления корреляции результатов диагноза по определённым показателям тестов и результатов и строго выполняемых по этим диагнозам корреляционных мер.

При накоплении достаточного количества данных по предметному или межпредметному тесту можно буде, в перспективе прогнозировать не только успешность будущей деятельности испытуемого, но и возможную степень эффективности **типового комплекса** коррекционых мероприятий. При получении данных о факторной зависимости этих мероприятий от характера диагноза и их взаимосвязи, удельном весе каждого из них и различных их сочетаний в достижении конечного коррекционного эффекта (при данном диагнозе по конкретным тестовым показателям) откроется возможность вводить поправочные коэффициенты к показателям критериального прогноза и повышать точность прогноза в нескольких вариантах:

а) оптимистическом (с учётом активной саморегуляции, самообразования, интенсивного и оптимального дообучения учащегося);

б) пессимистическом (возможность усиления негативного влияния отрицательных факторов, роли пробелов в знаниях и т.п. при сложившемся положении дел) и

в) реалистически-усредненном (с учётом возможности определённой самокорректировки и дообучения при обычной для педагогов и учащихся активности).

Методология и технология педагогической валидности ждёт своего решения и организации системы специальных методологических исследований. В перспективе, по мере накопления информации, можно будет ставить вопрос о включении данных педагогической валидности в проспекты и рекламные материалы по диагностическим тестам как необходимого и обязательного для всех разработчиков требования **стандарта теста**.