УДК 159.938+159.923
ББК Ю937

ТЕХНОЛОГИЯ РАЗРАБОТКИ ТЕСТОВ: ЧАСТЬ II

Н.А. Батурин, Н.Н. Мельникова

Статья является второй частью статьи, опубликованной в предыдущем номере журнала. В ней продолжается обсуждение пошаговой технологии разработки тестов и содержится описание еще двух этапов: III подготовительного и IV исследовательского. Эти два этапа являются центральными в разработке любого теста, независимо от того, создаётся он только для исследовательских целей или для психодиагностических. В статье рассматриваются вопросы, касающиеся стратегий отбора эффективных пунктов, а также вопросы проверки основных психометрических показателей тестов.

Ключевые слова: разработка теста, тестовые пункты, компоновка теста, согласованность и дискриминативность шкал, инструкции, надежность и валидность.

Введение

В первой части статьи [3] была представлена общая схема процесса разработки тестов, приведена последовательность этапов и шагов разработки и указаны основные задачи, решаемые на каждом из них. В первой части статьи рассматривались два начальных этапа: I организационный и II содержательный, которые выступают фундаментом создания любой методики, задавая стратегическую линию разработки и обеспечивая валидность методики на содержательном уровне. Вторая часть статьи посвящена этапам, которые концентрируют в себе большинство эмпирических процедур, привлекаемых к разработке теста. Именно в рамках этих этапов происходит непосредственная разработка и эмпирическая проверка самого тестового инструментария.

Этап III. Подготовительный

На подготовительном этапе осуществляется непосредственная подготовка всего материала, из которого будет состоять тест. Этот этап занимает существенное место в процессе разработки теста. Однако часто его важность недооценивается и деятельность по подготовке тестовых пунктов превращается в спонтанное «сочинительство» и мало обдуманное набрасывание вариантов, которые автор интуитивно считает «подходящими по содержанию», после чего случайный набор таких пунктов запускается для проверки валидности и надёжности. Такая недооценка важности подготовительного этапа и неграмотное выполнение соответствующих ему задач способны привести к неудаче всего предприятия.

Этап III «Подготовительный» состоит из 2 последовательных шагов: «Разработка пунктов» и «Сборка версий теста для апробации» (шаги 4 и 5 в общем процессе).

Шаг 4. Разработка пунктов

Шаг 4 «Разработка пунктов» включает три последовательных подзадачи: непосредственное формулирование пунктов(1), их профессиональная редакция (2) и создание банка пунктов (3).

Деятельность разработчика на подготовительном этапе опирается на материалы предыдущего этапа – «Содержательного». Напомним, что на предыдущем этапе работы были выбраны и обоснованы все основные формальные характеристики теста (в частности), формат пунктов, их необходимое количество, соотношение содержательных элементов в шкалах и т.д.). Эти характеристики, зафиксированные в «Спецификации» (см. шаг 3, описанный в I части статьи), становятся планом действий или «техническим заданием» для разработчиков пунктов.

1. Формулирование пунктов – первая задача подготовительного этапа. Она предполагает создание достаточного количества тестовых пунктов установленного формата (как правило, с запасом) для каждого обозначенного в спецификации аспекта содержания.

В деятельности ведущих зарубежных корпораций, занимающихся разработкой тес-

тов, давно утвердилась практика специального *отбора и подготовки разработчиков пунктов*. Считается, что хорошими разработчиками тестовых пунктов не рождаются, а ими становятся. Каждый тестовый формат предъявляет разработчику свои специфические требования, которые надо хорошо знать. Например, существенно различается специфика разработки пунктов, представляющих собой задания, предлагаемые для решения в тестах достижений, формулирование утверждений для опросников, создание пунктов для проективных тестов и т.д. За рубежом существуют достаточно объёмные методические разработки, посвящённые правилам и нюансам формулирования тестовых пунктов разных форматов. Особенно широко представлена и активно обсуждается специфика разработки заданий с множественным выбором, которые широко используются в тестах достижений и когнитивных способностей [16, 17]. В таких руководствах рассматриваются вопросы подбора дистракторов, их специфики для разных форматов заданий, последовательности предъявления дистракторов, расположения ключевого ответа и многие другие нюансы, о которых отечественные разработчики в большинстве случаев даже не задумываются.

При этом, хотя и имеется научная основа, касающаяся принципов написания пунктов для большинства общеупотребимых форматов, существует мнение, что разработка тестовых пунктов является больше искусством, чем технологией. Считается, что даже знание принципов формулирования не является гарантией того, что разработчик способен создать эффективные тестовые пункты [7, 16]. Во-первых, важно, чтобы разработчик был, если не специалистом, то хорошо осведомлённым в соответствующей содержательной области. Поэтому практикуется специализация разработчиков (например, для тестов достижений, интеллекта, личностных и т.д.). Во-вторых, немаловажное значение имеет опыт. Здесь важны практические навыки, которым нужно учиться. Опытный разработчик приобретает своеобразное «чутьё», способность прогнозировать, как та или иная формулировка отразится на характеристиках тестового пункта: например, на распределении ответов испытуемых; он способен улавливать тонкие нюансы, касающиеся социальной желательности формулировки или её смещения в иную содержательную область. В практике обучения новичков снабжают подробными инструкциями, они проходят стажировку, а опытные разработчики пунктов постоянно взаимодействуют друг с другом [12, 16].

2. Редактирование тестовых пунктов. Когда сформулировано достаточное количество пунктов, весь массив передаётся в руки профессиональных редакторов. В зарубежных источниках подчёркивается, что профессиональное редактирование тестовых пунктов – отдельная задача, которая должна выполняться другими людьми [9]. Редактор не только исправляет ошибки и опечатки, но и, что особенно важно, *корректирует отдельные пункты* с целью сгладить ненужные акценты, например, культурного или гендерного характера, а также социально желательные формулировки. Предполагается, что редактор должен иметь значительный опыт самостоятельной разработки тестовых пунктов. Здесь особое значение имеет то самое «чутьё», которое он приобретает в такой работе и которое можно обозначить как «спрессованный опыт». Редактор также отслеживает, чтобы в массиве поступивших на редакцию пунктов были в достаточной мере представлены все обозначенные в спецификации содержательные области. При необходимости для проверки соответствия пунктов рекомендуемому содержанию на этом этапе к проекту могут привлекаться независимые эксперты.

Работа, касающаяся формулирования и редакции пунктов, связана с особой ответственностью. Ведь именно тот (и только тот) материал, который заложен на этом этапе, и будет в дальнейшем подвергаться различным процедурам обработки и проверки. Исходный материал низкого качества обеспечивает такой же некачественный итоговый продукт.

3. Создание банка пунктов. Прошедшие редакцию пункты поступают в общую базу (или банк). В последнее время с расширением индустрии создания тестов за рубежом особое значение приобретают так называемые «банки тестовых пунктов» (pools) [15]. В такие банки входят прошедшие редакцию пункты, которые могут впоследствии использоваться для компоновки различных тестов. Ценность таких банков повышается, если для каждого пункта приводятся также и результаты его эмпирической апробации и другие данные, касающиеся его использования в составе различных тестов.

Примечательно, что многие такие банки размещены на открытых сайтах и ими могут воспользоваться любые разработчики, даже

не имеющие психологического образования (что само по себе порождает очень серьёзные проблемы для профессиональной психологической тестологии).

Создание таких банков на русском языке – сложная задача, решение которой, возможно, станет серьёзным подспорьем в развитии профессиональной психодиагностики в нашей стране. Однако, на наш взгляд, эти банки пунктов должны быть доступны только профессионалам.

Шаг 5. Сборка версий теста для апробации

Компоновка пробных версий теста из пунктов, помещённых в базу (банк), – основная задача следующего, 5-го шага разработки.

Специфика *пробных версий* состоит в том, что они являются предварительной заготовкой, которая имеет весьма узкое назначение: используется лишь для того, чтобы на следующем этапе после эмпирической проверки отобрать лучшие по статистическим характеристикам пункты. *Пробные версии* следует отличать от *рабочих версий* с утверждённым составом и структурой, процедурой и инструкцией, готовых для проверки надёжности, валидности и стандартизации.

Компоновка пробных версий теста может осуществляться автоматически, с помощью специальных программ, или же вручную. Однако следует помнить, что здесь метод случайного набора из базы нужного количества пунктов не является правильным решением. Общие требования к пробным версиям достаточно просты, но, тем не менее, требуют отдельного внимания:

1. Пробные версии должны включать избыточное количество пунктов, поскольку в ходе апробации предполагается отсев ряда пунктов, несоответствующих по статистическим характеристикам. Как правило, в пробные версии закладывается количество пунктов, которое не менее чем в 3 раза превышает запланированное для готовой формы теста.

2. В пробных версиях должна быть обеспечена репрезентативность выборки содержания: т.е. в них в достаточном объёме должны быть представлены все области содержания, описанные в спецификации.

3. Компоновка пробных версий должна быть хорошо сбалансирована. Например, если планируется создание опросника с прямыми и обратными утверждениями, то желательно включение в пробную версию примерно равного количества тех и других; или если разрабатывается тест достижений, использующий задания с множественным выбором, важно учитывать баланс расположения правильных ответов и т.п.

Поскольку для апробации запускается достаточно большое количество пунктов, пробные версии могут быть представлены в виде нескольких вариантов или блоков (не путать с параллельными формами теста).

Итогом 5-го шага разработки и всего подготовительного этапа является **утверждение версий теста для апробации пунктов.**

Этап IV. Исследовательский

Исследовательский этап занимает центральное место в процессе разработки теста. Он достаточно объёмен по содержанию и, как правило, требует значительного времени. На этом этапе исследуются эмпирические характеристики отдельных тестовых пунктов, шкал, определяются психометрические характеристики теста в целом. Исследовательский этап является критическим во всей разработке: здесь подвергаются проверке на практике все идеи, положения и материалы, которые были разработаны на предыдущих этапах.

Исследовательский этап состоит из 3 шагов: «Апробация пунктов и конструирование тестовых шкал» (шаг 6), «Уточнение процедуры тестирования» (шаг 7), «Изучение и проверка валидности и надёжности» (шаг 8). Практическим результатом исследовательского этапа является окончательная версия теста, готовая к стандартизации.

Шаг 6. Апробация пунктов и конструирование тестовых шкал

Основная цель 6-го шага разработки – получить шкалы (одну или несколько), обладающие двумя важными характеристиками: *внутренней согласованностью* и *дискриминативностью*. Любая шкала, предлагаемая тестом, во-первых, должна измерять только одно свойство, во-вторых – должна быть способна дифференцировать испытуемых по уровню изучаемого свойства. Если показатели внутренней согласованности и дискриминативности неудовлетворительны, то тест не сможет дать никаких интерпретируемых результатов и дальнейшая работа с ним, например проверка валидности, становится бессмысленной.

Величины показателей внутренней согласованности и дискриминативности во многом зависят от исходных характеристик отдельных пунктов. Поэтому вопрос отбора эффек-

тивных пунктов – один из самых важных в технологии разработки тестов [4, 5, 12]. И большая часть работы, осуществляемой в рамках 6-го шага, связана именно с *отбором эффективных пунктов*.

Информацию, полезную для отбора пунктов, предоставляет их эмпирическая апробация, в результате которой каждый из пунктов получает статистические характеристики, говорящие о его пригодности или непригодности для дальнейшей работы. Здесь могут использоваться разные методы и технологии, выбор которых зависит от целей тестирования, типа теста и формата пунктов. Ниже в качестве примеров мы рассмотрим 4 модели отбора пунктов, которые наиболее часто используются в современной практике разработки тестов: (1) отбор на основе классического анализа, (2) с помощью факторного анализа, (3) по критериальному принципу и (4) на основе «Item response theory». Эти модели используют разные статистические методы и опираются на разные правила принятия решений об отсеве или сохранении пунктов. По сути, конкретное содержание работы в рамках 6-го шага определяется выбранной моделью отбора пунктов.

Однако, несмотря на вариации в методах и технологиях, *общая последовательность действий для 6-го шага разработки* достаточно стабильна:

(1) сначала проводится предварительная апробация и отсев наиболее неудачных пунктов;

(2) затем оставшиеся пункты подвергаются целенаправленному отбору в соответствии с выбранной моделью;

(3) из прошедших отбор пунктов конструируются шкалы и собираются целостные формы теста;

(4) эти формы проходят эмпирическую проверку с целью получения итоговых показателей внутренней согласованности и дискриминативности;

(5) если шкалы теста выдержали такую проверку, то утверждается состав и структура теста.

Предварительная апробация пунктов проводится независимо от того, какой модели в дальнейшем будет следовать разработчик. Для предварительной апробации используются пробные версии теста, подготовленные на предыдущем этапе (см. шаг 5). Эти пробные версии предлагаются пилотажным выборкам, которые по качественному составу должны соответствовать планируемому контингенту тестируемых.

Цель предварительной апробации – уже *на ранней стадии отбраковать особо неудачные пункты*. Большинство разработчиков знакомы с общими принципами такого отбора и обычно следуют им. Эти принципы базируются на анализе статистик, характеризующих распределение ответов испытуемых на каждый пункт (таких, как меры центральной тенденции, характеристики разброса, асимметрия и т.д.). Традиционно исключаются пункты с малым разбросом (те, на которые все испытуемые отвечают почти одинаково) и с существенными асимметриями распределения. Пункты с малым разбросом будут заведомо снижать общую дискриминативность шкалы, в лучшем случае оставаясь «пустым балластом»; пункты с асимметричным распределением непригодны для большинства статистических процедур, которые подключаются на последующих стадиях отбора (например, для факторного анализа).

Иногда на стадии предварительной апробации могут потребоваться дополнительные статистические процедуры. Например, для формирования теста достижений может оказаться важным учёт сложности предлагаемых для выполнения заданий. Так, часто для тестов, использующих задания с открытым ответом, отбираются такие, которые при проверке показали от 40 до 60 % правильных ответов. Такая процедура позволяет отобрать для будущей шкалы задания примерно одинаковой сложности, которые могли бы при подсчёте баллов условно представлять равные единицы измерения свойства. Однако, несмотря на то, что эта процедура считается весьма полезной (особенно для тестов достижений и способностей), она подходит далеко не для всех случаев. Например, модель IRT изначально предполагает, что в экспериментальном массиве будут присутствовать задания разной сложности: это обязательное условие формирования эффективных адаптивных тестов.

Таким образом, на стадии предварительной апробации, прежде всего, проводится анализ первичных статистик по каждому пункту, а необходимость использования дополнительных статистических процедур, как правило, диктуется особенностями будущего теста и специально оговаривается в каждом конкретном случае.

Целенаправленный отбор пунктов в соответствии с выбранной моделью. После

предварительной апробации и первичного отсева оставшиеся пункты подвергаются целенаправленному отбору по технологии, соответствующей выбранной модели. Цель такой работы – отобрать пункты, которые могли бы составить шкалы (одну или несколько), обладающие внутренней согласованностью и дискриминативностью.

Разработчику важно знать, что разные модели отбора пунктов с неодинаковой успешностью обеспечивают вышеуказанные характеристики. В частности, модели, основанные на классическом и факторном анализе, «настроены» на обеспечение внутренней согласованности, а критериальный метод и, в особенности, модель IRT работают, прежде всего, на дискриминативность теста. Остановимся на этом подробнее, для примера рассмотрев особенности 4 вышеуказанных моделей отбора пунктов.

1) Классический анализ пунктов наиболее эффективно используется в тех случаях, когда планируется тест, состоящий из одной гомогенной шкалы. Анализ строится на оценке корреляций каждого пункта с общим баллом по тесту. При формировании шкал обычно изымаются те пункты, которые показывают низкие корреляции с общим баллом.

2) Отбор с помощью факторного анализа рекомендуется, если изначально планируется создать тест, состоящий из нескольких шкал. Факторный анализ полезен и в тех случаях, когда структура теста до конца не ясна, но допускается возможность существования нескольких относительно независимых параметров. При отборе заданий используется эксплораторный факторный анализ, который позволяет получить ответы на следующие вопросы [5]:

– Сколько отдельных шкал можно выделить в составе теста?

– Какие пункты принадлежат каким шкалам?

– Какие пункты должны быть удалены из теста?

Будет большой ошибкой формировать шкалы теста умозрительно, без такой эмпирической проверки. При использовании факторного анализа отбрасываются пункты, которым не удалось, как следует, нагрузить ни один из полученных факторов.

Поскольку принцип отбора и для классической, и для факторной моделей построен на оценках тесноты связи между отдельными пунктами, то *обе эти модели автоматически формируют шкалы с высокой внутренней согласованностью*. Напомним, что внутренняя согласованность шкалы определяется совместной изменчивостью компонентов, и именно на оценке этого свойства строятся различные математические формулы для вычисления надёжности по внутренней согласованности. Например, коэффициент α-Кронбаха чисто математически зависит как от количества пунктов в шкале, так и от средней величины корреляций между пунктами. При использовании классического анализа коэффициент α даже пересчитывают каждый раз заново, когда изымается очередное задание, добиваясь необходимой величины. Современные статистические программы (например, SPSS) предлагают удобную опцию («α, если пункт будет удалён»), которая позволяет заранее увидеть, повысится или нет общая согласованность, если отбросить конкретный пункт.

Однако такая простота выполнения таит серьёзные опасности. Часто разработчики стремятся «механически» повысить эту величину, отбирая для шкалы только вопросы с высокой взаимной корреляцией. Однако, как правило, пункты, дающие очень высокую корреляцию (более 0,7), чаще всего представляют собой простое перефразирование одного и того же утверждения или вопроса. В итоге в шкале остаются только пункты, представляющие одну какую-либо чрезвычайно узкую область содержания. Естественно, величина α-Кронбаха растёт, но повышается ли при этом качество шкалы? Как уже говорилось ранее, такая ситуация приводит к сужению содержания и накоплению систематической ошибки измерения, которая, в свою очередь, приводит к «смещению», «сдвигу» содержания [3].

Таким образом, основная опасность, подстерегающая разработчика, который использует для отбора заданий классическую модель или факторный анализ, состоит в соблазне *лёгкого повышения внутренней согласованности шкалы путём механического отбрасывания пунктов в ущерб репрезентативности содержания*. Поэтому, исключая отдельные пункты, необходимо постоянно сверяться со спецификацией, чтобы сохранить в итоговой шкале заданную пропорцию содержательных элементов. Во многих случаях вместо отсева приходится прибегать к переформулированию отдельных пунктов, чтобы избежать выхолащивания содержания.

Если говорить о *дискриминативности*

шкал, полученных посредством факторного или классического анализа, то следует отметить, что для них этот показатель в большей степени обеспечивается ещё на стадии предварительного отбора пунктов. Поскольку дискриминативность шкалы определяется разнообразием итоговых оценок (например, с помощью формул δ Фергюсона), то целесообразно ещё на ранней стадии исключить пункты, на которые большинство испытуемых дают одинаковые ответы. Если все пункты, из которых состоит шкала, имеют хороший разброс, то дискриминативность самой шкалы во многом будет определяться её внутренней согласованностью: в этом случае происходит «накопление» баллов у лучших испытуемых, что соответственно обеспечивает разброс показателей.

В противоположность моделям отбора, основанным на классическом и факторном анализе, критериальный принцип отбора и модель IRT «настроены», прежде всего, на достижение дискриминативности.

3) Отбор по критериальному принципу чаще всего применяется для конструирования тестов, предназначенных для прогноза и отбора. Также он удобен при разработке диагностических процедур, состоящих из объёмных, комплексных проб (например, кейсовых методов). В соответствии с этой моделью основанием для отсева или сохранения конкретного задания или пункта выступает его корреляция с внешним критерием.

Спецификой *модели отбора по критериальному принципу* является то, что она часто *продуцирует шкалы, имеющие очень низкую внутреннюю согласованность*. Бывает так, что пункты, из которых состоят такие шкалы, часто измеряют совершенно разные характеристики, хотя каждая из них в отдельности может быть важна для критериального признака. Результаты, полученные с помощью подобных шкал, очень трудно интерпретировать.

Как уже говорилось, сам механизм отбора по критериальному принципу в большей степени рассчитан на достижение дискриминативности. С этой целью для каждого пункта нередко вычисляется коэффициент дискриминации (не путать с дискриминативностью всей шкалы). Коэффициент дискриминации (или различительной силы) пункта отражает то, насколько данный пункт способен различать «лучших» и «худших» относительно критерия испытуемых [1, 4]. Для формирования шкалы отбираются пункты с высокими показателями дискриминации. Однако это не всегда решает поставленную задачу. Пункты, несогласованные между собой (даже если каждый из них в отдельности показал высокую дискриминативность), при объединении могут сложиться в шкалу с непредсказуемыми свойствами. Например, вследствие усреднения оценок по разнородным показателям, суммарный балл по шкале может не показать корреляции с критерием, а полученная шкала – достаточной дискриминативности. (На практике, внутренне несогласованные шкалы могут порождать как высокую дискриминативность, так и нулевую).

Отдельной серьёзной проблемой данного подхода является подбор репрезентативных групп с высокими и низкими показателями по критерию. Поэтому хотя модель отбора пунктов по критериальному принципу часто является единственно возможной для создания диагностических методов, состоящих из сложных объёмных проб (например, таких, как проба действием или диагностический эксперимент), этот метод не рекомендуется для создания традиционных психометрических шкал [4, 5].

4) Модель конструирования шкал на основе «Item response theory» используется чаще всего при разработке тестов достижений и способностей и особо продуктивна для создания программ компьютерного адаптивного тестирования [11, 14, 18]. В настоящее время модель IRT завоёвывает всё большую популярность в практике современного тестирования, распространяясь и на другие виды тестов, например, личностные [13, 14]. Однако возможность такого распространения на сегодняшний момент всё ещё является спорной.

Анализ пунктов на основе IRT реализуется с помощью специальных компьютерных программ. В результате такого анализа каждое задание может быть представлено в виде «характеристической кривой» (ICC), которая задаётся тремя параметрами: «a» – показатель дискриминации, «b» – уровень трудности, «c» – вероятность угадывания. Эти параметры служат основанием для отбора заданий. Считается, что оптимальная шкала должна включать много заданий разной сложности, но с высокими показателями дискриминации. (Три перечисленных параметра используются в трёхфакторной логистической модели 3ФЛ; более простые модели 1ФЛ и 2ФЛ ограничиваются одним или двумя параметрами).

Подчеркнём, что сам механизм построения процедуры тестирования с помощью IRT

настроен, прежде всего, на достижение высокой дискриминативности. И это наиболее ярко проявляется в реализации *компьютерного адаптивного тестирования*. Имея в запасе большой набор заданий разной сложности, адаптивная программа, учитывая успех/неуспех выполнения последовательных заданий, предъявляет каждому испытуемому больше заданий, тонко градуированных как раз в зоне его актуальной способности, тем самым обеспечивая особо точную оценку для каждого испытуемого. Тесты, созданные на основе IRT, чувствительны к минимальным различиям между испытуемыми.

Однако модель IRT так же, как и критериальная, слаба в плане обеспечения внутренней согласованности. *Необходимым условием применения модели IRT является изначальная согласованность пунктов*, которые берутся в анализ. Если модель анализа заданий на основе IRT применить к набору заданий, измеряющих несколько независимых свойств, то оценки параметров заданий, на основе которых и осуществляется их отбор (это особенно касается показателей дискриминации), будут некорректными. Поэтому на практике прежде чем начать трудоёмкий анализ на основе IRT, рекомендуется предварительно провести факторный, чтобы удостовериться, что выявляется только один фактор [5, 18]. Соответственно, для этой модели требуются дополнительные процедуры предварительного определения согласованности набора пунктов, поступающих в анализ.

Приведённые примеры не охватывают весь спектр нюансов и тонкостей, которые необходимо учитывать на стадии отбора пунктов. Однако позволяют продемонстрировать важность понимания разработчиком того, как его конкретные действия, в частности выбор модели отбора пунктов и применение тех или иных статистических процедур, могут непосредственно повлиять на достижение искомых показателей внутренней согласованности и дискриминативности.

Конструирование шкал и сборка теста. Из пунктов, прошедших отбор, конструируются тестовые шкалы и собираются целостные формы теста. Обычно (за исключением модели IRT) компоновка осуществляется вручную.

Компоновка тестовых форм – очень ответственная задача, поскольку как состав, так и последовательность расположения пунктов могут серьёзно повлиять на психометрические характеристики шкал и теста в целом. Здесь много тонкостей, которые необходимо учитывать. Например, если в тестовой форме нарушен баланс прямых и обратных вопросов в пользу прямых, то склонность испытуемых чаще давать ответы «да», чем «нет», в итоге приведёт к тому, что «прямые» вопросы покажут при проверке более высокую корреляцию с общим баллом, чем обратные. Или другой пример: если нескольких пунктов, касающихся одного и того же аспекта содержания располагаются подряд друг за другом, то между ними за счёт эффекта контекста увеличивается корреляция. Кроме искусственного завышения коэффициента внутренней согласованности, это способно сформировать более «плотный» фактор, чем он есть на самом деле, или даже выделить несуществующий фактор. «Статистические артефакты», возникшие из-за необдуманной компоновки, легко могут ввести в заблуждение даже самого автора теста.

Что особенно важно на данной стадии работы – это *проверить для каждой шкалы сохранность пропорции содержательных элементов, обозначенной в спецификации* (см. I часть статьи). Как уже говорилось, очень часто отсев неудачных по статистическим характеристикам пунктов приводит к значительному искажению этой пропорции, что нарушает репрезентативность выборки содержания и способно существенно снизить валидность теста. Хорошо, если на этой стадии разработчик имеет в своём распоряжении «запасные» пункты (напомним, что в пробные формы неслучайно закладывается избыточное количество пунктов). Однако во многих случаях для восстановления содержательного равновесия приходится прибегать к переформулированию отвергнутых пунктов или даже созданию новых. Естественно, новые пункты также должны пройти апробацию (а это значит повторение всех процедур 6-го шага сначала).

В результате такой работы формируются почти готовые тестовые формы, где пункты располагаются в том порядке, в каком их планируется предъявлять испытуемым и в дальнейшем. На самом деле – конструирование шкал – критический этап при разработке теста. Неудачная компоновка может свести на нет все достижения предыдущих шагов; грамотно проделанная – во многом обеспечивает высокие показатели надёжности и валидности. Неслучайно, в отличие от отдельных пунктов, их подборка в шкалы уже является объектом авторских прав.

Проверка внутренней согласованности и дискриминативности шкал. Далее полученные формы теста предъявляются новой выборке испытуемых с целью получить итоговые значения показателей внутренней согласованности и дискриминативности, которые вычисляются для каждой шкалы теста.

Следует подчеркнуть, что названные показатели *одинаково важны для качества теста*. Именно в сочетании они дают необходимый эффект: если один из них не выдерживает проверки, то высокое значение второго, само по себе уже не имеет смысла – такая шкала непригодна для пользования. (Если шкала не обладает внутренней согласованностью, то полученные с её помощью данные невозможно интерпретировать; если шкала не способна дифференцировать испытуемых, – то интерпретировать просто будет нечего).

Отдельную проблему составляет вопрос о том, какой величины должны достичь показатели внутренней согласованности и дискриминативности, чтобы быть признанными удовлетворительными. Надо сказать, что требования, принятые на сегодня в психологическом сообществе, можно считать достаточно лояльными. Так, например, EFPA (см. форму рецензии) предлагает считать «отвечающим требованиям» коэффициент внутренней согласованности не ниже 0,7; «хорошим» – от 0,8 до 0,89 и «отличным» – 0,9 и более. Если рассмотреть в свете этих требований такой популярный показатель, как α-Кронбаха, то достаточно 6 пунктов со средней корреляцией между ними, равной 0,3, чтобы получить величину α = 0,72; при количестве пунктов, равном 10, α возрастает до 0,81. Для «отличного» результата (α = 0,91) необходимы 10 пунктов со средней корреляцией 0,5 или 15 пунктов – с корреляцией 0,4. Как видим, требования вполне осуществимые.

На самом деле, *требования к конкретному тесту могут варьировать* в зависимости от его особенностей. Приведём несколько примеров. Первый: шкалы осведомлённости, предполагающие охват качественно различных областей содержания, как правило, будут иметь несколько меньшую согласованность, чем шкалы, «сконцентрированные» на измерении однородных по содержанию навыков (например, арифметических). Другой пример: при создании тестов, предназначенных для отбора, разработчики иногда стремятся к бимодальному распределению итоговых показателей, чётко дифференцирующему претендентов на две группы. Однако для теста с таким распределением дискриминативность, вычисленная с помощью δ Фергюсона, будет всего лишь около 0,55. И ещё один пример: для адаптивных тестов, использующих модель IRT, вовсе не вычисляются традиционные коэффициенты внутренней согласованности и дисриминативности (например, такие, как α и δ), поскольку конкретные величины этих коэффициентов имеют смысл только по отношению к фиксированному набору пунктов. Специфика же адаптивного тестирования на основе IRT в том, что каждый раз каждому испытуемому предъявляются разные наборы пунктов, состав которых определяется ходом тестирования. При этом считается нормой, что на разных участках выраженности измеряемого свойства тест может обладать разной дискриминативностью. Например, конкретный тест может быть более информативным в зоне высокой трудности и менее информативным – в зоне низкой [11, 14].

Поэтому хотелось бы предостеречь разработчиков от механической, бездумной эксплуатации любой, даже самой популярной формулы. Автоматически полученная величина без понимания того, из чего она сложилась, может ввести в заблуждение и исследователя, и пользователя.

В целом, при проверке внутренней согласованности и дискриминативности шкал могут быть использованы разные статистические процедуры, а требования к величинам этих показателей обосновываются в каждом конкретном случае. Очень важно, чтобы в технических отчётах о проделанной на этой стадии работе были приведены подробные данные, позволяющие судить о том, из чего сложилась конкретная величина.

Если в результате проведённой проверки показатели внутренней согласованности и дискриминативности признаны удовлетворительными, то **состав и структура теста утверждаются** для дальнейшей работы.

Шаг 7. Уточнение процедуры тестирования

Следующий шаг разработки направлен на решение задач, связанных, прежде всего, *с вопросами администрирования теста*. Среди них: уточнение последовательности действий в ходе тестирования, определение времени тестирования, разработка и апробирование

инструкций, уточнение алгоритмов обработки данных. В итоге процедура должна быть максимально формализована, а тест приведён в рабочую форму, готовую к широкомасштабным психометрическим исследованиям.

Процедура тестирования представляет собой такую же важную составляющую методики, как и тестовый материал, поэтому должна быть тщательно продумана, апробирована, описана и стандартизована. Опытные диагносты знают, что даже незначительное изменение в процедуре или инструкции может существенно отразиться на результатах, снизив валидность теста. К сожалению, в большинстве случаев этой стороне разработки методик уделяется слишком мало внимания. Важно, чтобы все аспекты процедуры были не только тщательно прописаны, но и *эмпирически проверены*.

Следует заметить, что для разных тестов процедура тестирования может сильно варьировать по сложности. Сложность её определяется теми действиями, которые должен производить психолог-диагност в процессе тестирования. В одних случаях процесс тестирования может состоять лишь в зачитывании инструкции, раздаче и сборе бланков и подсчёте баллов с помощью ключа; в других – требовать организации сложной диагностической ситуации, включать регистрирование поведения методом наблюдения, контент-анализ при обработке данных и т.д. Поэтому, в зависимости от особенностей методики, этот шаг для разработчика также может существенно различаться и по наполнению, и по сложности. Для простых тестов работа в рамках 7-го шага заключается лишь в проверке и уточнении инструкции и алгоритмов обработки данных; для сложных – может потребовать серьёзных дополнительных исследований.

Не имея возможности в данной статье останавливаться на всех подробностях и вариантах действий, которые могут потребоваться на этой стадии разработчику, сконцентрируем внимание лишь на ключевых позициях, актуальных практически для всех случаев.

Одна из таких позиций – **инструкция** испытуемому. Очень часто текст инструкции составляется автором лишь на основе здравого смысла и просто «присоединяется» к тесту. Однако такой подход может быть очень рискованным. Инструкция выполняет в процессе тестирования ряд важных функций, которые должны быть реализованы в достаточной мере. Основная функция инструкции – разъяснительная (что и как следует делать): текст инструкции должен предоставлять полную информацию и не вызывать разночтений. Идеально, если инструкция «работает» без дополнительных пояснений со стороны диагноста. Для проверки этой функции инструкции проводят посттестовые интервью с испытуемыми, в ходе которых выясняются вопросы понятности инструкции и изучаются реакции на отдельные фразы и слова. Если тест предполагает *тренировочные задания*, то они также апробируются в рамках уточнения инструкции.

Вторая функция, которую часто несёт инструкция, – установочная: инструкция настраивает испытуемого на определённый образ действий. Хрестоматийным является пример с изменением времени реакции из-за установочных акцентов в инструкции: время реакции увеличивается, если в инструкции даётся сенсорная установка («как можно быстрее увидеть сигнал»), и уменьшается, если моторная («как можно быстрее нажать на клавишу»). В некоторых методиках инструкция выполняет функцию моделирования экспериментальной ситуации, выступая центральным звеном, определяющим качество результатов. Известно, что на результаты влияет и то, как испытуемые понимают цель тестирования, которая обычно также оговаривается в инструкции.

То, как «сработает» конкретный текст инструкции, часто невозможно предсказать заранее, поэтому важно подвергнуть инструкцию эмпирической проверке. Для этого планируются специально организованные эксперименты, где варьируются параметры инструкции. Проверенные содержание и форма подачи инструкции фиксируются для пользователя и не должны подвергаться изменениям.

Параллельно с инструкцией проверяется и уточняется **формат регистрации данных** (например, форма бланков, интерфейс компьютерной программы, способы регистрации данных для качественных методов и т.д.).

Здесь же уточняются **алгоритмы обработки данных**: прописываются ключи, утверждаются схемы контент-анализа, фиксируются формулы для получения производных показателей (коэффициентов, индексов) и т.д. Если тест использует сложные методы регистрации данных (например, наблюдение) или качественные методы обработки (например, контент-анализ), то обязательно вычисляется *надёжность оценщика* (через согласован-

ность показаний нескольких оценщиков). При необходимости процедура обработки корректируется и совершенствуется до тех пор, пока согласованность оценок экспертов не достигнет нужной величины.

В рамках уточнения процедуры фиксируется примерное **время**, необходимое для тестирования. Особого внимания требуют тесты, где вводятся временные ограничения или проводится регистрация времени. Этот приём используется чаще всего в тестах достижений и способностей. Здесь время (или, точнее, скорость выполнения) становится дополнительным параметром, отражающим уровень достижения или способности. Как правило, такие тесты предъявляют особые требования к составу заданий: желательно, чтобы они были одной сложности или же располагались в порядке возрастания сложности. В этом случае временные ограничения добавляют полезный фактор, который способен увеличить дискриминативность теста. Именно с этой позиции и выбирается длительность временного интервала: должен быть выбран такой временной интервал, когда тест обеспечивает максимальную дискриминативность.

Ещё раз отметим, что содержание и объём работы, которые предполагает 7-й шаг «Уточнение процедуры тестирования», зависят от характера и сложности разрабатываемой методики. Для некоторых тестов важно формализовать условия проведения, для некоторых – прописать перечень дополнительных уточняющих вопросов, для некоторых – автоматизировать обработку результатов и т.д.

После всех проведённых проверок **утверждается рабочая версия теста**, в состав которой входят: подробно описанная процедура, зафиксированная инструкция, тестовый материал, формы для регистрации данных и алгоритмы обработки. Только после этого тест готов к дальнейшим испытаниям.

Шаг 8. Изучение и проверка надёжности и валидности

Восьмой шаг является завершением исследовательского этапа и «кульминацией» всего процесса разработки теста. Здесь проверяется успешность всех действий, предпринятых ранее для обеспечения эффективности теста. Основные задачи 8-го шага связаны с изучением базовых психометрических характеристик теста: надёжности и валидности. Методы и технологии, применяемые для этих целей, достаточно широко представлены в современной литературе. Более сложными и, часто, спорными являются вопросы, касающиеся выбора и обоснования необходимых процедур. Поэтому имеет смысл остановиться, прежде всего, на обсуждении именно этих вопросов.

Проверка надёжности теста. Первый вопрос, который требует прояснения, вызван путаницей, проистекающей из существования *нескольких видов надёжности*. Традиционно выделяют: надёжность по внутренней согласованности (к ней же относится надёжность эквивалентных половин теста), надёжность взаимозаменяемых форм, надёжность оценщика и, наконец, ретестовую надёжность [1, 4, 14]. Распространено заблуждение, что при психометрической проверке теста достаточно вычислить какой-либо один показатель надёжности. Однако важно понимать, что перечисленные виды надёжности *не заменяют друг друга*. Они имеют разную природу и отличаются друг от друга источниками дисперсии ошибок. Поэтому если того требуют особенности теста, он должен сопровождаться несколькими коэффициентами надёжности.

Показательно, что надёжность разных видов проверяется на разных шагах процесса, что определяется самой логикой разработки теста. (Отметим сразу, что в рамках 8-го шага речь идёт лишь о проверке ретестовой надёжности).

Надёжность по внутренней согласованности является закономерным итогом работы, проводимой на 6-м шаге (отбор пунктов и конструирование шкал), и проверяется в рамках этого же шага разработки. Полученный коэффициент отражает согласованность выборки содержания и, несмотря на то, что требования к его величине могут варьировать, является необходимым показателем для любого теста. Для методик, предполагающих наличие параллельных форм, их эквивалентность проверяется в рамках этого же 6-го шага, поскольку также связана с согласованностью выборки содержания в 2 или более формах теста.

Надёжность оценщика актуальна лишь для тестов, которые используют слабо формализованные качественные методы обработки, и связана с тем, удалось ли найти такие способы кодирования эмпирических показателей, которые бы однозначно трактовались разными людьми, обрабатывающими тест. Этот вид надёжности проверяется на 7-м шаге в рамках уточнения алгоритмов обработки.

Благодаря такой последовательности действий, к 8-му шагу в руках разработчиков имеется форма теста, где уже проконтролированы такие источники ненадёжности, как несогласованность содержания и возможные ошибки измерения, вносимые субъективным фактором при обработке. Теперь можно приступать к измерению надёжности, зависящей от динамических, временных факторов, т.е. к ретестовой.

Несмотря на техническую простоту вычисления коэффициента ретестовой надёжности, не так просто грамотно выстроить сам режим исследования. Дело в том, что этот вид надёжности имеет сложную природу. На получаемый показатель независимо друг от друга влияют два фактора: (1) стабильность измерительного инструмента (теста) и (2) стабильность самого измеряемого явления. Чтобы оценить качество теста, необходимо минимизировать влияние второго фактора. Именно поэтому проверку ретестовой надёжности часто рекомендуют проводить с небольшим интервалом между замерами, что особенно актуально для тестов, измеряющих свойства, изменчивые по своей природе. Однако проведение ретеста в сжатые сроки провоцирует искажения, связанные с повторным использованием одного и того же тестового материала. Избежать таких эффектов часто возможно только лишь посредством применения 2 параллельных форм теста, эквивалентность которых необходимо проверить ранее, что само по себе привносит дополнительные затраты и сложности в процесс разработки теста.

Поэтому при организации процедуры ретеста очень *важно правильно выбрать и обосновать длительность временного промежутка*, в течение которого ожидается сохранение показателей, полученных при первом тестировании. Если природа измеряемого свойства относительно стабильна, то может быть выбран достаточно большой интервал между замерами, что избавит от необходимости разработки параллельных форм только для нужд ретестового исследования. На самом деле, существуют психологические свойства, которые достаточно стабильны во времени. Например, можно ожидать, что хороший тест, измеряющий экстраверсию, должен давать высокую согласованность результатов и для двух замеров, проведённых с интервалом в полгода и более. При этом мы знаем, что для некоторых тестов, измеряющих особо изменчивые характеристики (например, многие психические состояния), определение надёжности посредством ретеста будет вообще неадекватным.

На самом деле, для практики было бы весьма полезно иметь конкретную информацию о своеобразном «сроке годности» результатов теста. Например, сколько времени руководитель может ориентироваться на результаты проведённого в сентябре исследования мотивации сотрудников, замеренного при помощи теста Х? Серия последовательных замеров могла бы предоставить необходимую информацию. Интервал времени, достаточный для того, чтобы результаты очередного замера потеряли свою согласованность с исходными показателями, можно считать своеобразным «сроком годности» результатов теста. Такая информация могла бы быть полезной как для определения периодичности тестирования при планировании массовых диагностических обследований, так и для целей индивидуальной диагностики.

Проверка надёжности может доказать, что набор тестовых пунктов, объединённых в шкалу, стабильно измеряет некоторое конкретное свойство. Тест с низкой надёжностью не может быть валидным. Однако проверка надёжности не способна пролить какой-либо свет на сущность измеряемого свойства. Для подтверждения того, что тест действительно измеряет запланированное содержание, необходимо соотнесение его результатов с внешней по отношению к тесту реальностью. На этих принципах и построены процедуры его валидизации (это верно как для критериальной, так и для конструктной валидности).

Проверка критериальной валидности

На первый взгляд, процедура проверки критериальной валидности очень проста: подбирается внешний критерий, который бы отражал содержание, связанное с тем, что измеряет тест, и с этим критерием соотносятся полученные результаты. Однако такая лёгкость выполнения обманчива и таит в себе несколько проблем весьма непростых для практического решения. Три наиболее существенных из них: *(1) проблема выбора критерия, (2) проблема качества критерия, (3) проблема определения достаточной величины взаимосвязи между тестом и критерием*.

Проблема выбора критерия возникает из-за того, что, в действительности, критериев для соотнесения может быть несколько, они качественно разнообразны и требуют неоди-

наковых методических решений. В качестве таких критериев могут использоваться данные аналогичного теста, результаты наблюдения за поведением в конкретных жизненных ситуациях, показатели успешности какой-либо деятельности, оценки экспертов и т.д. (В зависимости от характера применяемого критерия выделяются разные виды критериальной валидности: конкурентная, прогностическая и др.).

Естественный вопрос, который возникает: какой критерий будет более адекватен для проверки валидности данного теста? Ответ на этот вопрос следует искать, обратившись к цели тестирования: как, для чего и на каких популяциях планируется использовать тест. Валидность одного и того же теста в зависимости от цели его применения может устанавливаться разными способами [1, 5]. На самом деле, в рамках применения критериального подхода будет вообще неверным говорить о некоторой общей «абстрактной» валидности. Важно отметить, для каких именно целей валиден тест. Например, один и тот же тест может быть валиден для отбора работающих программистов, но невалиден для прогноза успешности обучения студентов на факультете информатики. В итоге именно критерий определяет «*область валидности*» теста. И эта область должна соответствовать исходной цели тестирования, обозначенной ещё на I этапе разработки, при «планировании проекта». К примеру, если цель теста – прогноз успеваемости в вузе, то адекватным критерием будут оценки успеваемости, полученные через некоторое время после тестирования. Если цель теста – клинический диагноз, то в качестве критерия могут быть использованы, например, данные анамнеза или наблюдения за поведением, позволяющие установить тот же диагноз другим способом.

Проблема качества критерия связана с тем, что в большинстве случаев сложно найти адекватные и надёжные методы замера критерия. Требования к качеству критерия в отношении его надёжности и валидности ничуть не ниже, чем требования к самой методике. Если нет возможности достоверно измерить критериальный признак, то вся процедура соотнесения с ним результатов теста теряет смысл. Неслучайно стандарты рецензирования методик EFPA требуют предоставления подробной информации о качестве измерительных инструментов, используемых как критерии или маркеры. Нередко попытка найти удовлетворительный критерий ставит разработчика перед необходимостью создания метода измерения критерия (что, по сути, равно разработке ещё одного нового теста).

Особые сложности возникают, когда в качестве критериев используются характеристики какой-либо деятельности. Дело в том, что подобные критерии являются сложными по своей структуре. Например, работа руководителя предполагает умение распределять обязанности, стратегическое планирование, владение конструктивными моделями взаимодействия с подчинёнными и т.д. При этом не совсем ясна относительная важность этих разнообразных функций и в большинстве случаев отсутствуют надёжные и валидные средства их замера.

Отдельную проблему составляет оценка полученной величины корреляции между тестом и критерием [6]. Какого числового значения должен достичь коэффициент корреляции, чтобы прогноз на основе критерия можно было бы считать достоверным? Согласно стандартам EFPA (см. форму рецензии), «отвечающим требованиям», считается коэффициент r=0,2. В каких случаях можно считать эту величину достаточной? Естественно, ожидания будут различаться для тестов, целью которых является текущая оценка достижений, отсроченный прогноз или жёсткий отбор. При этом важно учитывать характер критерия, его сложность и надёжность метода измерения, которые также способны повлиять на итоговую величину корреляции. В каждом отдельном случае теоретически ожидаемая величина должна обосновываться, исходя из цели тестирования и особенностей критерия.

Распространено мнение, что один из самых простых и «верных» способов валидизации по критерию – это соотнесение данных нового теста с уже существующим, аналогичным по содержанию (конкурентная валидность). Однако этот способ, так же как и все другие, имеет весьма серьёзные ограничения, связанные с тремя вышеобозначенными проблемами.

Здесь также существуют сложности, возникающие при подборе теста, который планируется использовать в качестве критерия. Как правило, выбор, ориентированный только на название, оказывается почти равнозначен случайному. Требуется детальный анализ содержания, позволяющий удостовериться, что исходная трактовка теоретического концепта и выборка содержания у двух тестов анало-

гичны. На практике в большинстве случаев бывает очень трудно найти тест, который бы полностью совпадал по содержанию с вновь создаваемым. Как правило, приходится довольствоваться тестами, лишь близкими по содержанию, а не аналогичными.

Несмотря на использование готовых тестов, проблема качества критерия при проверке конкурентной валидности также остаётся актуальной. Используемый как критерий тест должен быть высокого психометрического качества. А это, учитывая состояние отечественной психодиагностики, само по себе – редкость.

Если и коснуться вопроса об ожидаемой величине коэффициента корреляции с аналогичными тестами, то в стандартах EFPA обозначен нижний допустимый порог: это r = 0,55 (рекомендуется использовать выборку размером не менее 100 человек); «отличными» считаются результаты проверки, если r>0,75. При этом здравый смысл подсказывает, что если получен слишком высокий коэффициент корреляции, то новый тест практически дублирует уже имеющийся. В этом случае нелишним будет требование обосновать необходимость создания нового теста. Новый тест будет полезен, если он более прост в применении, способен более тонко и детально представлять измеряемое содержание или же существует актуальная потребность в создании дублирующих друг друга средств измерения какого-либо свойства.

При использовании близких по содержанию, но не аналогичных тестов получаемые величины корреляций обычно удерживаются в средних пределах. При этом разница в содержании приводит к тому, что их сложно интерпретировать. В таких случаях полезно выяснить, какие аспекты содержания нового теста отвечают за полученную корреляцию. Для этого можно провести дополнительный детальный анализ взаимосвязей (по пунктам). При существенных различиях в содержании корректнее было бы вообще говорить не о проверке конкурентной валидности, а об изучении конструктной, что предполагает совсем другие критерии качества.

Также, оценивая полученную величину коэффициента корреляции, следует помнить, что максимальная возможная корреляция между двумя тестами ограничена величиной их надёжности. Поэтому, выбирая некоторый тест в качестве критерия, имеет смысл обратить внимание не только на соответствие его содержания, но также и на показатели надёжности.

Таким образом, использование для валидизации теста критериального подхода при нынешнем состоянии отечественной психодиагностики привносит больше проблем, чем решений. По всей видимости этот подход оправдан лишь тогда, когда соотнесение с критерием выступает основным источником валидности теста и диктуется самой целью тестирования (например, для тестов отбора). В остальных случаях решение о выборе критериального подхода для валидизации теста в значительной степени *определяется наличием отвечающего требованиям метода измерения критерия*.

Изучение конструктной валидности теста. Конструктная валидность, или точнее валидизация теста посредством идентификации конструкта считается наиболее важным аспектом валидности. По мнению многих авторов, именно конструктная валидность определяет обоснованность смысловой интерпретации результатов теста и поэтому наиболее точно соответствует самому пониманию того, что есть валидность [1, 5, 10, 14].

Валидизация конструкта требует постепенного накопления информации из разных источников. Поэтому обычно говорят не о проверке, а именно об изучении конструктной валидности: широком и разностороннем изучении того, что представляет собой эмпирический конструкт, полученный с помощью разработанного теста. Для изучения конструктной валидности необходимо проведение ряда исследований, *направленных на проверку конкретных и хорошо продуманных гипотез*.

Печально, что на деле изучение конструктной валидности часто выливается в хаотическое коррелирование результатов теста с тем, что попалось под руку. Такая стратегия порождает отчёты, содержащие большое количество данных, значительная часть которых не несёт полезной информации для понимания сущности конструкта.

Грамотно спланированное исследование базируется на теоретических основах, заложенных и прописанных на 2-м содержательном этапе разработки теста. Именно из понимания природы измеряемого явления должны выводиться предположения об особенностях его функционирования. Эти предположения и формулируются в качестве гипотез для проверки на исследовательском этапе проекта.

Традиционно с изучением конструктной валидности связывают корреляционные исследования, позволяющие определить место изучаемого конструкта среди других психологических переменных. В рамках конструктной валидности различают так называемые конвергентную и дискриминантную валидность, первая из которых связана с гипотезами о наличии корреляции конструкта с содержательно близкими переменными, вторая – с отсутствием таковой с содержательно далёкими. Однако корреляционные исследования далеко не исчерпывают список методов, которые могут быть использованы для изучения конструктной валидности. Методы, привлекаемые к исследованию, определяются характером выдвигаемых гипотез, а они могут быть чрезвычайно разнообразны.

Например, могут быть сформулированы гипотезы о возрастных или гендерных различиях в измеряемых тестом характеристиках, что предполагает планы исследований, построенные на сравнении групп. Возможны гипотезы о динамике изменений во времени, проверка которых требует лонгитюдных исследований. В рамках изучения конструктной валидности могут проверяться гипотезы и о структуре конструкта, что потребует факторного анализа.

Особое место при изучении конструктной валидности занимают экспериментальные проекты, где в качестве независимых переменных выступают факторы, предположительно способные воздействовать на измеряемое тестом свойство. В специально моделируемых ситуациях регистрируются изменения, возникающие в показателях теста после экспериментального воздействия. Например, если тест измеряет эмоциональное напряжение, свидетельством его валидности может служить изменение в результатах, появившееся у испытуемых после пребывания в экспериментально созданной эмоционально напряжённой ситуации. Основная цель таких экспериментов – определить, будут ли оценки теста варьировать в соответствии с теоретическими ожиданиями.

При оценке качества теста в стандартах EFPA есть параметр, учитывающий количество исследований, проведённых для изучения конструктной валидности, и разнообразие применяемых методов (см. форму рецензии EFPA). Однако следует помнить, что основным критерием здесь выступает всё же не количество методов само по себе, а обоснованность применения того или иного метода, что определяется грамотной формулировкой гипотез. Также важно понимать, что математические величины, ожидаемые в ходе исследования конструктной валидности теста (например, величины коэффициентов корреляции) не могут быть однозначно заданы и опять же определяются конкретными гипотезами.

Для уточнения конструкта могут быть использованы и сведения, собранные в процессе разработки теста [1]. Например, данные о факторной структуре, полученной при отборе пунктов, о корреляциях между шкалами теста, о результатах соотнесения с критерием. Дополнительной детализации при описании конструкта могут способствовать данные о надёжности теста на разных временных интервалах и для разных условий и групп, а также анализ норм, полученных на разных группах. По большому счёту при идентификации конструкта могут пригодиться любые данные, проливающие свет на природу рассматриваемого свойства или на условия, от которых зависит его развитие и проявление [1, 5, 10, 14]. Однако важно, чтобы используемые данные были осмыслены в свете теоретических построений, описывающих природу конструкта.

Бывают случаи, когда изучение конструктной валидности вносит поправки и уточнения в само понимание изучаемого явления, обогащая его концептуальную содержательную область. Это особенно актуально для сложных, малоизученных явлений, имеющих неявную структуру и содержание. Здесь возникает ситуация, когда требуется совмещение разработки диагностических методов с исследовательскими проектами. История создания общеизвестных многофакторных личностных опросников (начиная с Р. Кэттелла) – тому прекрасная иллюстрация. Дело в том, что без эмпирических исследований (требующих наличия метода), мы не можем чётко определить содержание, а без ясного понимания содержания – сконструировать хороший метод. По большому счёту это – циклический процесс. Иногда возникает несколько возвратных повторяющихся циклов, включающих все шаги со II по IV этап включительно, когда каждый раз создаются новые, всё более совершенные версии теста, базирующиеся на всё более точной и ясной концептуальной основе.

Изучение конструктной валидности также продолжается и после публикации готового

теста. Опыт его практического применения и проведённые с его помощью исследования постепенно добавляют новую полезную информацию, в свете которой всё более выкристаллизовываются возможности и ограничения теста и сфера его применения.

Шестой «Исследовательский» этап процесса разработки теста завершается **утверждением окончательной версии теста**, которая после тщательной **корректорской проверки** может быть допущена к стандартизации.

Литература

1. Анастази, А. Психологическое тестирование / А. Анастази, С. Урбина. – СПб.: Питер, 2001. – 668 с.

2. Батурин, Н.А. Современная психодиагностика России / Н.А. Батурин // Вестник ЮУрГУ. Серия «Психология». – 2008. – Вып. 2. – С. 4–9.

3. Батурин, Н.А. Технология разработки тестов: часть I / Н.А. Батурин, Н.Н. Мельникова // Вестник ЮУрГУ. Серия «Психология». – 2009. – Вып. 6. – С. 4–14.

4. Клайн, П. Справочное руководство по конструированию тестов: введение в психометрическое проектирование / П. Клайн; под. ред. Л.Ф. Бурлачука. – Киев: Изд-во ПАН Лтд, 1994. – 688 с.

5. Купер, К. Индивидуальные различия / К. Купер; под ред. И.В. Равич-Щербо. – М.: Аспект Пресс, 2000. – 527 с.

6. Шмелёв, А.Г. Психодиагностика личностных черт / А.Г. Шмелёв. – СПб.: Речь, 2002. – 480 с.

7. Abedu, J. Language issues in item development / J. Abedu // Handbook of test development / ed. by S.M. Dowing, T.M. Haladyna. – Lawrence Associates, 2006. – P. 377–398.

8. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. – Washington, DC: American Educational Research Association, 1999. – 101 p.

9. Baranowski, R.A. Item editing and editorial review / R.A. Baranowski // Handbook of test development / ed. by S.M. Dowing, T.M. Haladyna, 2006. – P. 349–358.

10. Cronbach, L.J. Construct validation after thirty years / L.J. Cronbach // Intelligence: measurement, theory, and public policy / ed. by R.E. Linn. – Urbana: University of Illinois Press, 1989. – P. 147–171.

11. Davey, T. Designing computerized adaptive tests / T. Davey, M.J. Pitoniak // Handbook of test development / ed. by S.M. Dowing, T.M. Haladyna, 2006. – P. 543–574.

12. Dowing, S.M. Twelve steps for effective test development / S.M. Dowing // Handbook of test development / ed. by S.M. Dowing, T.M. Haladyna. – 2006, ed. by Lawrence Associates, P. 3–25.

13. Embretson, S.E. Item response theory for psychologists / S.E. Embretson, S.R. Reise. – Mahwah, NJ: ed. Lawrence Erlbaum Associates, 2000.

14. Furr, R.M. Psychometrics: an introduction / R.M. Furr, V.R. Bacharach. – Sade Publications, Inc., 2008. – 349 p.

15. Goldberg, L.R. The international personality item pool and the future of public-domain personality measures / L.R. Goldberg, J.A. Johnson et al. // Available online 25 October, 2005.

16. Haladyna, T.M. Developing and validating multiple-choice test items / T.M. Haladyna. – Hillsdale, NJ: Lawrence Erlbaum associates, 2004.

17. Haladyna, T.M. A taxonomy of multiple-choice items-writing rules / T.M. Haladyna, S.M. Dowing // Applied Measurement in education. – 1989. – № 1. –P. 37–50.

18. Hambleton, R.K. Item response theory: principles and application / R.K. Hambleton, H. Swaminathan. – Boston: Kluwer-Nijhoff, 1985.

Поступила в редакцию 14 сентября 2009 г.

Батурин Николай Алексеевич. Доктор психологических наук, профессор, декан факультета психологии Южно-Уральского государственного университета: nikbat@list.ru.

Nikolay A. Baturin. PsyD, professor, the dean of the Faculty of psychology, head of chair «Psychological diagnostics and Counselling» of South Ural State University: nikbat@list.ru.

Мельникова Наталья Николаевна. Кандидат психологических наук, доцент кафедры социальной психологии ЮУрГУ: MNN17@yandex.ru.

Natalia N. Melnikova. Candidate of Psychological sciences, docent of department of social psychology of South Ural State University: MNN17@yandex.ru.