УДК 159.922
ББК Ю937

ТЕХНОЛОГИЯ РАЗРАБОТКИ ТЕСТОВ: ЧАСТЬ III

Н.А. Батурин, Н.Н. Мельникова

Статья является продолжением описания технологии разработки тестов, содержащейся в двух предыдущих номерах журнала. В ней в рамках дальнейшего обсуждения пошаговой технологии разработки тестов описывается пятый этап, который называется в общей схеме «стандартизационным». В статье рассматриваются разные методы стандартизации тестов и особенности различных видов норм. В частности, обсуждаются вопросы, касающиеся формирования репрезентативных выборок стандартизации и принципы стратификации при формировании групповых норм. Приводятся положения, обеспечивающие эффективность работы экспертов при установлении предметно-ориентированных норм. Обсуждаются проблемы, связанные с выработкой критериально-ориентированных норм.

Ключевые слова: разработка теста, стандартизация теста, нормы, репрезентативные выборки стандартизации, стратификация, «секущий балл», вероятностные таблицы ожиданий.

Введение

В первой части цикла статей была представлена общая схема процесса разработки тестов, приведена последовательность этапов и шагов разработки и указаны основные задачи, решаемые на каждом из них [2]. Там же были описаны два начальных этапа: этап I – организационный и этап II – содержательный, которые являются фундаментом создания любой методики, задавая стратегическую линию разработки и обеспечивая валидность методики на содержательном уровне. Во второй части цикла статей обсуждались этапы, в рамках которых происходит непосредственная разработка и эмпирическая проверка самого тестового инструментария: этап III – «подготовительный» и этап IV – «исследовательский», которые были посвящены вопросам формулирования и отбора эффективных тестовых пунктов, там же обсуждались проблемы проверки основных психометрических показателей тестов: надёжности и валидности [3]. Данная статья продолжает поэтапное описание универсальной технологической процедуры разработки тестов. Она посвящена вопросам стандартизации разработанного, апробированного и проверенного теста.

Этап V. Стандартизационный

Шаг 9. Стандартизация методики

Этап, на котором происходит стандартизация методики, – один из наиболее существенных для формирования пользовательских характеристик разработанного теста. Конечная цель стандартизации – это всегда получение норм, которые позволяют перевести полученные в процессе тестирования «сырые» данные в стандартные, чтобы затем корректно интерпретировать итоговые тестовые показатели конкретных людей. Считается, что нормы нужны, прежде всего, для практического применения теста в диагностических целях. При применении того же теста для исследовательских целей корректнее использовать «сырые» показатели, не подвергнутые никаким преобразованиям, и нормы здесь практически не добавляют полезной информации [1, 5, 14]. Поэтому методики, специально разрабатываемые для использования исключительно в научных целях, могут обходиться без стандартизации. Тесты же, которые планируется применять главным образом как диагностический инструментарий, без надёжных стандартов практически непригодны. Исключение до недавнего времени составляли проективные, клинические и другие виды методик, относящиеся к категории экспертных, которые ис-

пользовались без проведения процедуры стандартизации. Однако в последнее время разработчики и издатели тестов ориентированы не только на проверку надежности и валидности всех видов тестов, но также на проведение их стандартизации.

Как уже отмечалось, к процедуре стандартизации допускаются лишь готовые тесты, с тщательно выверенной и утверждённой процедурой проведения, полностью прошедшие проверку надёжности и валидности (см. этап IV – исследовательский). Если в тесте существуют серьёзные недоработки, то никакая стандартизация, даже на очень представительных выборках, не способна исправить ситуацию: тест будет непригоден для пользования, а временные и материальные средства, потраченные на стандартизацию, пропадут впустую.

В отечественной практике разработки тестов процедуры стандартизации обычно не связывают с особыми теоретическими сложностями. Традиционный подход – «чем больше выборка стандартизации и широта охвата контингента, тем лучше» – ставит акценты на трудностях, обусловленных нехваткой материальных и временных ресурсов. Однако такой подход является весьма односторонним и приводит к недооценке серьёзности и ответственности данной задачи и необходимости осмысления и тщательного планирования всех действий и методов, используемых на этапе стандартизации теста.

На самом деле, объём выборки стандартизации далеко не первый по важности критерий качества полученных норм, а при грамотной организации процесса стандартизации можно обойтись и меньшими ресурсами, получив в то же время более корректные результаты.

Чтобы грамотно организовать процесс стандартизации, важно в первую очередь понимать, *для какой практической цели создаются тестовые нормы* [11, 17]. Существует много разных методов стандартизации тестов и несколько видов норм, которые предназначены для разных практических целей. Выше уже отмечалось, что нормы, в конечном счете, нужны для интерпретации результатов теста. Именно они определяют правила принятия практических решений. Поэтому так важно, чтобы разработанные в процессе стандартизации нормы были действительно полезны для практики и помогали принять правильное решение.

По сути, нормы – это *основания для оценки «сырых» показателей*, полученных конкретным лицом при выполнении теста. Важно иметь ясность относительно того, что именно необходимо оценить. Например, при использовании теста учебных достижений разработчика может интересовать: а) положение конкретного студента относительно других студентов того же года обучения; б) уровень усвоения им знаний по данному предмету или в) вероятность того, что этот студент будет успешно справляться с некоторыми видами профессиональной деятельности.

Приведённый пример иллюстрирует ***3 вида норм***, наиболее явно отличающихся друг от друга и предназначенных для разных практических целей.

1. *Групповые нормы* (иногда их называют выборочными, статистическими, относительными) – отображают выполнение теста в выборке стандартизации. При использовании таких норм «сырые» показатели конкретного испытуемого соотносят с эмпирически полученным распределением оценок в выборке стандартизации, что позволяет узнать, какое место он занимает в этом распределении. Подобные нормы называют относительными, поскольку они выявляют статус испытуемого относительно некоторой группы, взятой в качестве нормативной. Для удобства соотнесения при получении таких норм используют различные преобразования исходных показателей (процентили, Z-показатели, Т-баллы, станайны и др.)

Традиционно групповые нормы используются для большинства тестов (особенно тестов интеллекта и личности). Часто в рамках одного теста такие нормы рассчитывают отдельно для разных категорий обследуемых (например, для групп, различающихся по возрасту, полу, социальному положению и т. д.). Нормы данного вида не только позволяют сравнивать отдельных людей между собой, но также дают возможность соотносить результаты разных тестов, выполненных одним человеком.

2. *Предметно-ориентированные* (domain-referenced) *нормы* – призваны отражать уровень мастерства (mastery), которое показал тестируемый (количество знаний, качество освоения навыков и т. д.). Учитывая специфику этого вида норм, их иногда называют ориентированными на содержание и абсолютными. Для установления таких норм определяется «стандарт исполнения», который выража-

ется либо в проценте выполненных заданий, либо через «секущий» балл (cut score), который подразделяет протестированных людей на группы по принципу «прошёл/не прошёл». (При использовании нескольких секущих оценок могут выделяться несколько категорий, например: «начальный уровень усвоения знаний», «базовый» и «продвинутый»). В современной практике существует более десятка методов получения предметно-ориентированных норм, к установлению которых часто привлекаются эксперты [11]. Предметно-ориентированные нормы актуальны, прежде всего, для тестов достижений и широко используются в системе образования, а также для таких целей, как лицензирование и сертификация.

3. *Критериальные* (criterion-referenced) *нормы* – отражают вероятность того, что испытуемые, получившие некоторый балл по тесту, достигнут критериального показателя. Критериальные нормы получают посредством эмпирического соотнесения тестовых показателей с оценками по критерию и обычно представляют в виде так называемых «таблиц ожидания». Критериальные нормы могут быть рассчитаны как для тестов достижений, так и для тестов, измеряющих способности и личностные черты. При этом основная сфера применения таких норм связана с практическими задачами отбора и прогноза.

Несмотря на различия в способах получения и представления, указанные виды норм не противоречат, а наоборот, дополняют друг друга. Хорошо, если один и тот же тест сопровождается разными видами норм: это делает его применение более универсальным, а интерпретацию более «тонкой». Напомним, что выбор вида норм определяется практической целью и осуществляется в самом начале разработки, ещё на этапе планирования проекта.

Независимо от того, нормы какого вида планируется получить, стандартизация любой методики предполагает следующую *последовательность действий*: (1) формирование выборки стандартизации, (2) эмпирические процедуры установления стандартов, (3) фиксация способов перехода к нормам.

Однако для получения норм разных видов существуют различия в способах выполнения этих трёх задач: используются разные методы, по-разному распределяются акценты. Поэтому необходимо отдельно рассмотреть специфику процесса стандартизации для каждого из вышеназванных видов норм, более подробно останавливаясь на моментах, заслуживающих особого внимания разработчика.

Стандартизация для получения групповых норм

Как уже отмечалось раньше, групповые нормы (выборочные, статистические) воспринимаются как традиционные и используются в большинстве известных тестов. Напомним, что такие нормы указывают относительное положение обследуемого на фоне нормативной выборки и позволяют оценить полученный им результат в сравнении с результатами других людей. Специфика этих норм в том, что они всегда *зависят от характеристик конкретной совокупности людей, на которых они рассчитывались*. Поэтому задача формирования выборки стандартизации для этого вида норм особенно актуальна. *Здесь качество выборки – определяющий аспект стандартизации*. Если выборка сформирована некорректно, нормирование теста не только будет бесполезным, но даже может оказаться вредным, вводя пользователя в заблуждение.

Формирование выборки стандартизации

Две важные переменные, определяющие качество выборки, – это её *объём и репрезентативность*. При этом репрезентативность выборки считается более важным показателем, чем её объём, хотя и к объему выборки предъявляются достаточно строгие требования [1, 5, 14]. Достаточно большое количество людей в выборке необходимо для уменьшения значения стандартной погрешности. Качественные же характеристики выборки определяют основы для интерпретации: на фоне какой группы будут рассматриваться получаемые при использовании методики результаты.

Если говорить о количественной стороне выборки стандартизации, то согласно Стандартам Европейской федерации психологических ассоциаций (EFPA) для любых тестов выборки меньше 150 человек считаются «не соответствующими требованиям» (см. форму рецензии). Эта величина определяется по большей части из чисто статистических соображений. При этом верхняя граница чётко не оговаривается: хорошими могут считаться выборки и в количестве 300 человек, и достигающие нескольких тысяч испытуемых. Вопрос о том, какую величину выборки можно считать «достаточной» для стандартизации

конкретного теста, не решается изолированно от вопроса о её качественном составе. А решение этого вопроса, в свою очередь, зависит от практического назначения теста.

Формирование выборки следует начинать с точного описания популяции, на которую планируется распространить полученные нормы («целевой популяции»), что следует сделать еще на начальных этапах разработки теста. Есть тесты, которые изначально предназначены для диагностики представителей очень общих популяций, таких, как «взрослые работающие люди» или «студенты вузов». Другие тесты узко специализированы и ориентированы на конкретные, чётко обозначенные группы, например, «офисные работники компаний, занимающиеся продажей недвижимости». Естественно, ориентация теста на более широкие и неопределённые популяции требует более объёмных и разнородных по составу выборок для стандартизации. Нормы для ограниченных популяций могут быть получены на меньших по объему выборках.

Однако и для тестов, ориентированных на общие популяции, объём выборки можно значительно сократить без потери качества результата, если хорошо поработать над её репрезентативностью. В настоящее время грандиозные выборки, собранные методом случайного отбора, которые были так популярны несколько десятилетий назад, сменяются меньшими по объёму, но более продуманными стратифицированными выборками.

Стратификация (от «страта» – «слой») – процесс выделения в общей популяции некоторых категорий (слоёв), характеризующихся качественно различными признаками (например, возраст, пол, социальное положение). *Стратифицированная выборка* – это выборка, составленная с учётом представленности разных категорий (страт) в популяции. Стратифицированная выборка считается более репрезентативной, чем случайно отобранная, поскольку отражает пропорции разных социально-демографических групп в общей популяции. Например, если в популяции «мужчин от 20 до 30 лет» 50 % имеют среднее образование, 20 % являются студентами вузов и 30 % имеют законченное высшее образование, то и в выборке испытуемые должны быть представлены в той же самой пропорции.

Однако составление стратифицированной выборки непростая задача. Основная сложность состоит в *правильном выделении признаков для формирования страт*. Общую популяцию можно разделить на группы сотней различных способов. Традиционно для выделения страт учитываются такие параметры, как пол, возраст, иногда профессиональный статус, уровень дохода и другие социально-демографические показатели. Возникает вопрос: всегда ли это оправдано, и какие именно из многих доступных показателей следует выбрать в качестве стратифицирующих признаков при стандартизации конкретного теста? Нелишним будет напомнить, что использование только 2-х дихотомических признаков (например, пол (м/ж) и образование (среднее/высшее)) уже предполагают 4 группы, если добавить к этому 3-й признак (например, семейное положение) – будет уже 8 групп, и т. д. Поэтому из практических соображений при формировании выборки имеет смысл ограничиться небольшим количеством наиболее важных признаков. Как их определить?

Существует правило, согласно которому в основе формирования страт должны лежать *признаки, которые наиболее связаны с измеряемой тестом переменной*. Например, если стоит задача стандартизировать тест, направленный на диагностику лояльности к организации, то следует поработать над тем, чтобы обозначить переменные, способные провоцировать различия в уровне лояльности. Так, если есть веские теоретические (или эмпирические) основания утверждать, что лояльность к организации должна различаться у мужчин и женщин, то признак пола целесообразно использовать как стратифицирующий при формировании выборки стандартизации. Но может быть, здесь более полезными окажутся такие признаки, как характер организации (коммерческая/бюджетная), стаж работы или должностной статус?

Поэтому ещё раз отметим, что шаблонная стратификация, в основу которой положены стандартные характеристики (такие, как пол, возраст) далеко не всегда приводит к желаемым результатам, поскольку изучаемая тестом переменная может быть нечувствительна к этим признакам, и выделение соответствующих групп практически ничего не даст. Цель стратификации – учесть возможные вариации диагностируемой тестом переменной в популяции и отразить это в выборке посредством пропорциональной представленности соответствующих групп. Именно это обеспечивает репрезентативность выборки стандартизации. Таким образом, при формировании выборки стандартизации для кон-

кретного теста признаки, которые будут положены в основу стратификации, выбираются в зависимости от его содержания и практического назначения. Соответственно для теста, измеряющего мыслительные функции, это будут одни признаки, для теста, измеряющего экстраверсию, – другие, честность – третьи.

Если удалось найти реальные основы для стратификации, то две выборки, извлечённые из одной генеральной совокупности, не должны давать заметно отличающихся норм. Это будет говорить о том, что они действительно репрезентативны для популяции. Если же какой-то важный признак для стратификации остался неучтённым, то две выборки могут продуцировать различные нормы, и это может быть спровоцировано как раз вариацией в них неучтённого признака. В тех случаях, когда сложно выделить обоснованные признаки для стратификации, задачу получения устойчивых норм можно решить только за счёт увеличения объёма выборки стандартизации. И здесь случайный метод отбора из популяции большой по объёму выборки может быть оправданным.

Для тестов, ориентированных на общие популяции, часто проводится *дифференцированная стандартизация*, т. е. тест обеспечивается целым набором норм, полученных для разных групп. Так, многие интеллектуальные тесты нормируются отдельно для разных возрастных групп, личностные – часто сопровождаются различными нормами для мужчин и женщин. Также нормы могут быть представлены отдельно для групп, отличающихся уровнем дохода, образования или по национальному признаку.

Дифференцированная стандартизация повышает пользовательский потенциал методики и считается одним из важных показателей качества теста (см. рецензию EFPA). Такие нормы могут использоваться отдельно или наряду с общегрупповыми показателями.

Если к стандартизации была привлечена большая стратифицированная выборка, то выделение из неё подгрупп и подсчёт норм для каждой из них не представляют особой сложности. Однако надо помнить, что в каждой такой подгруппе должно быть достаточное количество человек (не менее 150 согласно стандартам EFPA), иначе полученные нормы будут непригодны из-за высокого значения стандартной погрешности.

В связи с этим возникает вопрос практического плана: до какой степени детализации следует «дробить» общую выборку на подгруппы? В каких случаях более выгодно (и корректно) будет остановиться на объединённых нормах (например, совместно для мужчин и женщин или для разных возрастных групп)? Ответ на этот вопрос достаточно прост. После того, как проведено массовое обследование и получены данные для всей выборочной совокупности, необходимо проверить, существуют ли значимые различия в показателях по тесту между интересующими разработчика группами. Для этого удобно использовать многомерный дисперсионный анализ (программы, выполняющие такой анализ, представлены почти в каждом современном статистическом пакете, например SPSS) [6]. Дисперсионный анализ поможет выделить группы, нуждающиеся в раздельной стандартизации. Если же по полученным для конкретного теста данным различия (например, между мужчинами и женщинами) отсутствуют, то и нет необходимости в дифференцированной стандартизации этих групп. Здесь следует принять решение в пользу объединённой выборки стандартизации, которая от этого выиграет по объёму.

Ещё один фактор повышения практической ценности теста и одновременно уменьшения затрат на стандартизацию – это *изначальное ограничение совокупности обследуемых, на которую планируется распространить использование теста после его разработки* [1, 5, 14]. Во многих случаях более полезными бывают специфические нормы для групп, непосредственно соответствующих характеру и области практического применения теста. Например, узко профильный тест может быть стандартизирован специально для таких совокупностей, как «первокурсники технических вузов» или «рабочие промышленных предприятий». На самом деле, большинство тестов по своему содержанию и назначению не претендуют на универсальность и не требуют норм, полученных на широкомасштабных выборках. Допускается даже, что конкретные организации могут накапливать свои нормы для внутренних целей (например, это могут быть нормы при тестировании претендентов на определённые должности внутри организации или нормы в приёмной комиссии конкретного колледжа).

Другая сторона этого вопроса состоит в том, что в связи с организационными и материальными сложностями получения универсальных норм для больших популяций многие

авторы идут по пути переопределения генеральной совокупности на этапе стандартизации, умышленно сужая её и определяя в терминах групп, которые удалось привлечь к стандартизации теста. Учитывая всю условность такого подхода, его всё же следует признать более корректным (и, по крайней мере, честным по отношению к пользователю), чем распространение норм на широкие популяции при отсутствии соответствующей выборки стандартизации.

В любом случае границы нормативной выборки должны быть чётко определены и приведены вместе с нормами в Руководстве к тесту, чтобы эта информация была доступна для пользователя. Если возникнет потребность в расширении сферы применения теста, он может быть повторно стандартизирован другой группой исследователей на других выборках. Такое развитие событий на этапе эксплуатации теста вполне естественно и отвечает современным тенденциям перехода процесса разработки тестов от индивидуальных инициатив к широкомасштабным общественным проектам. Например, для координации таких проектов в США еще в 1993 году по инициативе APA и Министерств обороны, образования и труда был создан Государственный совет по тестированию и оценке.

Процедуры установления групповых норм

Способы получения групповых норм всегда эмпирические. Когда сформирована выборка стандартизации, проводится массовое обследование с помощью разработанной методики. Переход к нормированным данным основан на преобразовании сырых показателей в стандартную шкалу, ориентированную на эмпирически полученное распределение показателей в выборке стандартизации. Существует несколько способов преобразования сырых оценок в стандартные (среди них процентили, Z-показатели, Т-баллы, станайны и др.). Подробные описания статистических процедур, применяемых при разных способах преобразования, широко представлены в современной литературе и доступны любому интересующемуся лицу. Единственный вопрос, которого следует коснуться в данной статье, – это вопрос о том, какой из многочисленных вариантов следует выбрать при стандартизации конкретного теста.

Чаще всего метод преобразования выбирают из соображений удобства пользования или же «по привычке» (характерно, что многие авторы обычно используют для стандартизации своих тестов один-два «любимых» метода). Следует отметить, что если эмпирическое распределение, полученное на выборке стандартизации близко к нормальному, то все способы дают практически идентичные результаты, и полученные нормы легко, без искажений могут быть переведены из одной шкалы в другую. Однако, учитывая возможность получения «неидеальных» распределений, полезно знать некоторые особенности каждого способа преобразования, чтобы подобрать наилучший для конкретного теста. Рассмотрим наиболее популярные стандартные показатели, акцентировав внимание на условиях их применения и возможных ограничениях.

Процентили – самый простой и понятный для пользователя способ преобразования, который отражает значение индивидуального показателя через процент случаев «лежащих» (расположенных) в выборке стандартизации ниже данного балла. Несмотря на наглядность и простоту получения, процентили имеют ряд существенных недостатков, которые ограничивают возможности их использования [1, 5, 14]. Во-первых, процентили представляют собой значения порядковой шкалы, поэтому не пригодны для многих процедур статистического анализа. Во-вторых, распределение процентилей по своей природе равномерное (прямоугольное). Поэтому при наложении его на данные, близкие по своему распределению к форме нормальной кривой (а большинство эмпирически полученных распределений именно таковы), результаты могут быть сильно искажены. В частности, в середине распределения небольшие отклонения от среднего будут увеличиваться процентилями, а значительные (на краях кривой) – неадекватно сжиматься. По сути, процентили отражают относительное положение каждого индивидуума в выборке и не способны показать величину различия между тестовыми оценками.

Более совершенные способы преобразования выражают тестовый балл в терминах «расстояния от среднего», а это расстояние, как правило, измеряется в единицах стандартного отклонения. Такие способы преобразования «выравнивают» все различия в исходных единицах измерения и позволяют легко соотносить друг с другом не только результаты отдельных испытуемых, но и показатели, полученные по разным тестам. Поэтому такие показатели по праву называют «*стандартными*».

Для их получения существуют «линейные» и «нелинейные» способы преобразования. Специфика стандартных показателей, полученных способом «*линейного*» преобразования, состоит в том, что они в точности воспроизводят все особенности распределения «сырых» баллов. «*Нелинейный*» способ преобразования призван «подогнать» не совсем нормальное эмпирическое распределение к заданному типу (обычно – к типу нормальной кривой).

Классическим образцом линейного преобразования является *Z-показатель.* Используя Z-баллы, мы получаем информацию о том, выше или ниже среднего находится показатель конкретного испытуемого и насколько (в единицах стандартного отклонения). Однако, несмотря на «статистическую прозрачность», Z-баллы весьма неудобны для пользователя. Во-первых, Z-шкала имеет малое количество целых позиций (от –3 до 3), что делает необходимым введение дробных значений. Во-вторых, данная шкала использует отрицательные значения, что не приветствуется при сообщении результатов клиенту или заказчику. Поэтому на практике чаще используются вторичные линейные преобразования Z-показателя (например, Т-баллы), которые также сохраняют все особенности исходного эмпирического распределения, но избавлены от 2 вышеназванных пользовательских недостатков. Например, Т-баллы более удобны и понятны для пользователя, поскольку имеют среднее отклонение, равное 50, а стандартное – 10. Это позволяет увеличить разброс шкалы и избежать дробных и отрицательных величин.

Есть одно условие, определяющее правомерность использования стандартных показателей, полученных посредством линейного преобразования. Важно, чтобы эмпирическое распределение в выборке стандартизации действительно соответствовало форме нормальной кривой. Если распределение существенно отклоняется от параметров «нормального», то полученные нормы не позволяют провести адекватную интерпретацию результатов.

Небольшие отклонения эмпирического распределения от «нормы» можно откорректировать посредством «нелинейного преобразования». Для этого «сырые» баллы, например, сначала преобразуют в процентильные эквиваленты, а затем переводят в стандартные, используя таблицы значений функции плотности нормального распределения. Такой способ преобразования, как и другие подобные ему, называется «нормализацией», а полученные единицы – «*нормализованными*». Примером нормализованных показателей могут быть *станайны* и *стены*, которые представлены в удобной для пользователя форме (например, 9-балльная шкала станайнов или 10-балльная шкала стенов).

Нормализация как способ формирования тестовых норм может показаться очень привлекательным методом, поскольку позволяет «сгладить» недостатки в разработке теста, выправив несовершенное эмпирическое распределение. Однако следует избегать чрезмерной эксплуатации этого полезного свойства.

В целом нормализацию рекомендуется проводить, если:

а) существуют лишь незначительные отклонения от формы нормального распределения;

б) есть веские теоретические основания предполагать наличие нормального распределения по изучаемому признаку в целевой популяции;

в) есть основания считать, что отклонение эмпирического распределения от нормального произошло в силу определённых недостатков в формулировке пунктов теста, а не из-за особенностей выборки, или других факторов, влияющих на исследуемое свойство [1, 5, 14].

На самом деле, если применить нормализацию к сильно смещённым распределениям, то недостатки будут видны сразу. Например, из-за дефектов распределения компьютерные программы нормализации «выдают» стандартные баллы с пропуском или повтором одних и тех же значений. В таких случаях приходится прибегать к выравниванию шкал «вручную», экстраполируя некоторые шкальные значения. При этом необходимо подчеркнуть, что такая экстраполяция допустима, только если у разработчика есть уверенность в том, что указанные проблемы нормализации являются следствием незначительного несовершенства выборки стандартизации (например, недостаточность ее величины).

В целом стандартизация с целью получения групповых норм (независимо от того, какой способ преобразования использовался) всё же ориентирована на эмпирическое распределение, приближенное к нормальному. Если в выборке стандартизации получено очень далёкое от нормального распределение, то установление рассматриваемого вида норм

будет затруднено. Конечно, абсолютно нормальное распределение («стандартная нормальная кривая») является в реальности, скорее, теоретическим идеалом. Эмпирические распределения всегда несколько отличаются от идеальной формы: бывают несколько асимметричные или уплощённые и т. п. Однако речь идёт о том, что они все-таки могут быть отнесены к этому классу распределений. Естественно, нередки случаи, когда разработчик изначально стремился получить распределение другой формы (исходя из целей тестирования или учитывая природу конструкта). В этих случаях модель нормального распределения и основанные на ней методы стандартизации не смогут обеспечить адекватно интерпретируемых результатов и следует искать другие способы стандартизации.

Но если есть веские основания предполагать, что распределение изучаемого свойства в генеральной совокупности всё же должно быть нормальным, то причину отклонений в форме эмпирического распределения следует искать либо в ошибках при отборе пунктов, либо в дефектах выборки стандартизации.

Нормы, полученные для разных групп или для выборочной совокупности, в целом фиксируются в виде ***таблиц перевода сырых оценок в стандартные***. В Руководстве для пользователя обязательно должны указываться количественные и качественные характеристики выборок, на которых были получены нормы.

Стандартизация с целью получения предметно-ориентированных (domain-referenced) норм

Рассмотренные выше групповые нормы позволяют интерпретировать тестовый балл как относительно высокий или низкий, однако они не отражают абсолютную величину психологической переменной. Например, если конкретный студент при выполнении экзаменационного теста показал результат на одно стандартное отклонение выше среднего, это говорит о том, что он знает предмет лучше, чем средний студент, но ничего не сообщает о том, «сколько» знания у него имеется в абсолютных единицах.

При этом такая информация является весьма важной, например, для вынесения адекватной оценки степени усвоения знаний учащимися или студентами после прохождения некоторого курса. Особенно актуально это становится, когда стоит задача сертификации или лицензирования специалистов. Можно, конечно, при выдаче разрешения на профессиональную деятельность ориентироваться на относительные показатели выполнения квалификационного теста в выборке прошедших обучение… Однако представьте себе последствия такого подхода, например, при выдаче лицензии на использование нового метода в хирургической практике или же при принятии решений о допуске к полётам летчиков-испытателей? Очевидно, что в сферах, где требуется адекватная оценка качества усвоения знаний или овладения некоторыми умениями и навыками, не очень важно, как конкретный претендент выглядит на фоне общей выборки.

В сфере образования и для целей сертификации и лицензирования чаще используются *предметно-ориентированные нормы, где в качестве интерпретационной системы используется не некоторая совокупность людей, а чётко определённая содержательная область*. При этом интерпретация результатов теста производится не посредством сравнения с показателями, полученными другими людьми, а с точки зрения его смыслового содержания, т. е. определяется, что тестируемый знает или что он может делать.

Исторически предметно-ориентированные тесты намного старше ориентированных на групповые нормы. Ещё в начале прошлого века делались попытки детально описать уровни выполнения теста с точки зрения его содержания. Для этого результаты конкретного испытуемого оценивались на основе сопоставления с набором стандартных образцов. Тогда в педагогике было популярно понятие «mastery» в смысле овладения чем-либо или усвоения определённых учебных единиц [13]. Однако позже в связи с распространением статистических методов такой способ нормирования вытеснили групповые нормы. В последние годы происходит все большее распространение тестирования в области образования и разработка многочисленных тестов достижений в других областях практики, вследствие чего интерес к предметно-ориентированному нормированию тестов снова активизировался. В настоящее время создаются разнообразные методы установления таких норм, например, Г. Сайзек в своей обзорной статье описывает 10 таких методов [11].

Существует мнение, что практико-ориентированные нормы применимы лишь для тестов достижений, где чётко и полно

описано содержание требуемых знаний. Лучше всего они подходят для проверки конкретных навыков (например, таких, как базовые арифметические навыки), вся совокупность которых является чрезвычайно фиксированной и определённой и представляет небольшой объем содержания, который полностью покрывается тестом. Для таких тестов 85 % правильных ответов будет говорить о 85% освоении заданных навыков [5]. При этом тест позволит выделить «области» ошибок (например, покажет, какие конкретно арифметические операции недостаточно освоены). Однако, если тест представляет собой лишь выборку из всей совокупности знаний (а таких тестов большинство), то здесь ещё надо доказать, что выполнение теста на 85 % соответствует 85%-ному усвоению знаний. Кроме того, подобный способ оценки сложно использовать к более высоким уровням знаний (например, в процентном эквиваленте практически невозможно оценить уровень понимания).

Однако сложности методической реализации не уменьшают практический запрос на абсолютную форму оценки. Во многих областях практики из соображений безопасности требуется установление критических точек, отражающих минимальный адекватный уровень в исполнении деятельности. В современном понимании предметно-ориентированные нормы не столько предоставляют информацию об абсолютной величине знаний, сколько отвечают на вопрос о том, «*сколько необходимо получить баллов, чтобы уровень знаний считался достаточным*» [11, 12].

Таким образом, при стандартизации, направленной на получение предметно-ориентированных норм, устанавливается некоторый стандарт выполнения теста (performance standard), который соотносится с задаваемым уровнем компетенций [11, 17, 20]. Этот стандарт выполнения фиксируется в виде некоторого «секущего» балла (cut score), иногда он же называется проходным баллом (passing score). По сути, «секущий» балл призван провести различия между двумя или более уровнями выполнения. Соответственно нормирование теста предполагает выделение некоторых категорий тестируемых, отличающихся уровнями выполнения (или достижения). Для одних тестов могут быть использованы лишь две категории, такие, как «прошёл /не прошёл», «аттестован /не аттестован»; для других требуется более градуированная классификация, например: «начальный уровень, базовый, продвинутый» и др.

В современной практике существует более десятка методов установления предметно-ориентированных норм [10, 11]. Практически все они требуют привлечения экспертов.

Методы установления предметно-ориентированных норм

По сути, сам процесс установления предметно-ориентированных норм представляет собой тщательно спланированную и организованную процедуру, направленную на получение квалифицированного экспертного мнения. Разные методы реализуют эту задачу разными способами.

Большинство методов установления предметно-ориентированных норм основаны на концепции так называемого *«рискованного экзаменуемого»* (borderline student) – имеющего одинаковую вероятность (50/50) сдать экзамен или провалить его. (Иногда используются и другие версии этой концепции: например, человек, обладающий минимальным необходимым уровнем способностей или компетенций – «minimally acceptable person», «minimally competent examinee» и т. д.).

Имея в виду такого рода гипотетических испытуемых, эксперты выносят суждения о том, сможет ли подобный испытуемый правильно ответить на каждый конкретный пункт теста, либо прогнозируют вероятность правильного ответа и т. д. Экспертов также могут просить подразделять конкретных экзаменующихся на группы относительно пограничной линии или сортировать выполненные ими работы (например, тексты или творческие задания). При этом мнения экспертов могут комбинироваться с реальными статистическими данными, например, о сложности пунктов или о распределении показателей в выделенных группах испытуемых.

Формат данной статьи не предполагает подробного описания конкретных технологий реализации разнообразных методов. Для ознакомления с ними можно обратиться к соответствующим источникам [9–10, 11, 13, 15, 18–20]. Здесь же приведём несколько примеров для иллюстрации основных подходов, используемых при установлении предметно-ориентированных норм.

Наиболее известный метод принадлежит У. Ангоффу и носит его имя [9]. В настоящее время существует много модификаций и усовершенствований этого метода, однако изна-

чальная оригинальная версия, несмотря на её некоторую «наивность», наиболее подходит, чтобы проиллюстрировать базовый принцип, характеризующий всю группу родственных методов. Специально отобранным экспертам предлагается просмотреть тест по пунктам и, имея в виду гипотетического «рискованного экзаменующегося», решить, мог бы такой человек ответить правильно на каждый из пунктов теста. Пункты, доступные, по мнению экспертов, для такого испытуемого, получают «1» балл, а усреднённая по группе экспертов сумма баллов для всего теста определяет «секущий» балл. Как правило, для надёжности такая процедура проводится два раза.

Вариации метода Ангоффа состоят в модификации задач для экспертов (например, они могут оценивать вероятность правильного ответа) и включают дополнительные процедуры организации групповой деятельности (дискуссии для стабилизации представления о «пограничном испытуемом», обмен мнениями между сериями оценок, обратную связь о полученных нормативах и т. д.) [15, 16, 18].

Иногда для облегчения когнитивной задачи экспертам предоставляются статистические данные, характеризующие пункты теста. Например, при использовании «Метода маркировки» (Bookmark method) эксперты снабжаются специальными буклетами (ordered item booklet – OIB), в которых пункты расположены по одному на странице в порядке возрастания их трудности. (Показатели сложности пунктов могут быть получены ранее, на этапе отбора пунктов при помощи IRT-технологии или традиционным способом). Экспертам предлагается поместить маркер на той странице буклета, где, по их мнению, вероятность гипотетического пограничного испытуемого ответить на вопрос правильно опускается ниже заданной величины (например, станет ниже 75 % или 67 %) [19].

Более известные у нас методы, такие, как «Метод контрастных групп» и «Метод пограничной группы» акцентируют внимание экспертов не на оценке пунктов теста, а на классификации экзаменуемых. Например, первый из названных методов предлагает экспертам (знающим экзаменуемых), выделить из них две группы: «masters» и «nonmasters». «Секущий балл» здесь получают способом, более ориентированным на традиционные статистические методы: он выбирается на пересечении эмпирических распределений тестовых оценок двух полученных групп так, чтобы наилучшим образом дифференцировать эти группы [11].

В настоящее время за рубежом активно создаются новые методы установления предметно-ориентированных норм и совершенствуются уже известные. В целом новые модификации методов направлены на то, чтобы обеспечить (1) воспроизводимость результатов, (2) достоверность источников информации, (3) соответствие полученных норм целям тестирования [20]. Эти эффекты достигаются несколькими способами. Во-первых, за счёт введения процедур, повышающих надёжность экспертных суждений (предварительная тренировка экспертов, несколько раундов оценивания и т. д.). Во-вторых, это использование ресурсов групповых процессов в работе экспертов (групповые обсуждения, обмен мнениями, обратная связь в процессе принятия решений и т. д.). В-третьих, за счёт комбинации экспертных процедур с анализом реальных эмпирических данных (статистических характеристик пунктов, распределений тестовых оценок и т. д.).

Тем не менее все методы установления предметно-ориентированных норм не обходятся без критики. На самом деле, когда получен конкретный результат в виде «секущего» балла, практически невозможно оценить его достоверность. Качество получаемых норм этого вида целиком зависит от качества работы экспертов. Поэтому в стандартах AERA, APA & NCME, регламентирующих деятельность по установлению предметно-ориентированных норм, особое внимание уделяется *качеству организации самого процесса получения норм с помощью экспертов* [8].

Организация процесса получения предметно-ориентированных норм

В зарубежной практике существуют хорошо обоснованные и проработанные положения, касающиеся организации процесса получения предметно-ориентированных норм. Эти положения направлены на повышение эффективности работы экспертной группы и в итоге обеспечивают качество получаемых норм. Независимо от применяемого метода процесс установления предметно-ориентированных норм предполагает чётко определённую *последовательность стадий* [11]. К сожалению, такое пристальное внимание непосредственно к экспертной процедуре нетипично для отечественной практики разработки тестов, и это, на наш взгляд, является серь-

ёзным упущением. Поэтому остановимся на описании этих стадий чуть подробнее.

1. *Выбор метода стандартизации.* Выбор метода установления стандартов определяется двумя факторами: (1) целью, для которой создаются стандарты, (2) спецификой формата пунктов теста. Так, выделяют 4 типичных цели, для которых могут быть использованы предметно-ориентированные нормы: (а) сопровождение процесса обучения, (б) выпускной экзамен, (в) подготовка отчётности о качестве образовательного процесса, (г) сертификация и лицензирование.

Разное назначение теста предполагает формирование разных классификационных групп и соответственно разные требования к «секущему» баллу. Также обозначенные цели различаются по оценке затрат, связанных с возможными неверными решениями (некоторые из них допускают значительный риск, другие – должны максимально исключить его). Формат пунктов также следует учитывать при выборе метода, поскольку существуют методы, рассчитанные на дихотомические формы, на задания с множественным выбором или же предназначенные для тестов использующих задания с открытыми ответами и пробы деятельности.

2. *Планирование деятельности по установлению стандартов.* На этой стадии составляется подробный план экспертной процедуры и подготавливаются все необходимые материалы и технические средства, которые будут востребованы в процессе работы экспертов. Например, производится разработка и настройка программного обеспечения, подготовка печатных материалов (буклетов, бланков, оценочных листов). Также при необходимости формируются группы экзаменующихся, которые могут быть привлечены к тестированию.

3. *Подготовка описаний для разных категорий исполнения.* Принято, что при установлении предметно-ориентированных норм заранее обозначаются категории, которые планируется выделить посредством секущего балла (performance level labels – PLLs). Для этих категорий составляются подробные текстовые описания (performance level descriptions – PLDs). Эти описания помогают экспертам ориентироваться в поставленной задаче, а впоследствии могут быть использованы также и для интерпретации результатов.

4. *Отбор экспертов.* Для предметно-ориентированных норм задача формирования выборки стандартизации заменяется задачей отбора экспертов (participants). В зависимости от специфики метода для работы отбираются эксперты (как правило, не менее 10 человек), которые хорошо знакомы с соответствующей содержательной областью. В отдельных случаях требуется также, чтобы эксперты были знакомы с людьми, привлекаемыми к эмпирическому исследованию в качестве испытуемых (examinee).

5. *Подготовка экспертов.* Группы экспертов, привлекаемые к установлению стандартов, проходят специальную подготовку. Подготовка включает подробное ознакомление с предстоящей процедурой, тренировку навыков оценки по заданной схеме и т. д. Если метод предполагает групповой процесс принятия решений, то в программу подготовки экспертов включают специальный социально-психологический тренинг, направленный на формирование навыков работы в группе.

6. *Сопровождение процедуры установления стандартов.* Сопровождение работы экспертов – очень ответственная задача. Специально подготовленные модераторы управляют процессом работы экспертов. Они отслеживают последовательность действий, собирают и обрабатывают полученные оценки, предоставляют обратную связь участникам и т. д. Во многих случаях такие специалисты должны владеть и социально-психологическими навыками работы с группой, поскольку (как того требуют отдельные методы) им приходится выступать в роли фасилитатора, управляя дискуссией и процессом принятия решений в группе.

7. *Оценка процесса установления стандартов.* В большинстве случаев качество готовых нормативов можно оценить, только оценив качество самого процесса их получения. Поэтому этой заключительной стадии процесса уделяется особое внимание. Оценке подвергаются все стадии процесса, для чего используется специально подготовленная документация. Для оценки корректности процесса используются и другие источники информации, такие, как видеозаписи групповых дискуссий и обратная связь от участников процесса (для чего им раздаются специальные анкеты). По результатам комплексной оценки процесса производится утверждение или отклонение полученных нормативов.

Повышенное внимание и активный социальный интерес к этой форме стандартизации

тестов вызваны тем, что на основе предметно-ориентированных норм принимаются важные решения, серьёзно влияющие на судьбу конкретных людей. Поэтому возможные ошибки, которые проходят незаметными при чисто описательных интерпретациях, здесь должны быть сведены к минимуму. В то же время сама условность метода экспертной оценки заставляет максимально жёстко регламентировать и контролировать его.

Полученные ***предметно-ориентированные нормы для конкретного теста фиксируются*** в виде констатации «секущего» балла (баллов), обозначения категорий обследуемых, выделяемых на основе нормирования (performance level labels – PLLs) и их подробного описания (performance level descriptions – PLDs). Также в Руководстве для пользователя указывается точное целевое назначение полученных нормативов.

Стандартизация с целью получения критериальных (criterion-referenced) норм

Два вида норм, описанные выше, служат полезными ориентирами для интерпретации: групповые нормы позволяют сравнивать друг с другом показатели отдельных испытуемых и данные по разным тестам, предметно-ориентированные нормы дают возможность соотносить уровень индивидуальных достижений с требуемым стандартом исполнения. Тем не менее всё ещё остаются неясными психологические следствия того или иного балла. Например, что значит, иметь уровень выше среднего по тестам доброжелательности или честности: как это будет отражаться на реальном поведении? Или какова будет успешность профессиональной деятельности для людей, преодолевших необходимый барьер при оценке качества усвоенных ими знаний? Для полноты интерпретации не хватает *соотнесённости тестовых баллов с некоторыми внешними по отношению к тесту явлениями.*

Этот пробел восполняют критериальные нормы, которые связывают тестовые показатели с внешними параметрами, выступающими в качестве критерия. Такие нормы отражают вероятность того, что испытуемые, получившие некоторый балл по тесту, достигнут критериального показателя. Критериальные нормы широко применяют для целей прогноза и отбора. При этом они могут быть рассчитаны как для тестов достижений, так и для тестов, измеряющих способности и личностные черты.

Методы установления критериально-ориентированных норм

Наиболее распространенный способ представления критериальных норм – это таблицы ожидания [1, 4, 5]. В *таблицах ожидания* приводятся вероятности различных критериальных исходов для лиц, получивших тот или иной тестовый балл. Значения этих вероятностей определяются экспериментально. Для испытуемых, составляющих выборку стандартизации, получают показатели по тесту (который в этом случае называется «предиктором») и по критерию. Затем с помощью специальных статистических процедур соотносят эти показатели.

Статистические процедуры, которые применяются для построения таблиц ожидания, не представляют большой сложности. Самый простой способ использует двумерное распределение, связывающее значения тестовых показателей с оценками по критерию. Для удобства тестовые показатели и оценки по критерию предварительно группируют в интервалы или же представляют в порядковой или номинальной шкале (например: ниже среднего /средний /выше среднего). В таблице ожиданий столбцы обычно соответствуют интервалам тестовых баллов, а строки – нескольким уровням критерия. В каждой ячейке таблицы фиксируется число случаев (в процентах), соответствующих каждому конкретному сочетанию тестовых результатов и оценок по критерию. Например, из такой таблицы можно получить сведения о том, сколько процентов учащихся, получивших при тестировании логического мышления результаты «ниже среднего», «средние» и «выше среднего», попадают в группу с низкой, средней и высокой успеваемостью по математике. (По существу, таблицы предсказания в некотором роде репрезентируют корреляционное поле, отражающее взаимосвязь тестовых и критериальных оценок).

Более детальные таблицы могут быть рассчитаны с учётом каждого тестового балла. При этом для составления таблиц ожидания рекомендуется использовать тестовые показатели, уже стандартизированные другим способом (например, переведённые в стены или Т-баллы). Иногда для прогноза критериальных показателей используются уравнения регрессии. Однако в таблицах ожиданий, основанных на уравнениях регрессии, прогнозируемые показатели следует сопровождать значениями стандартных по-

грешностей (так как в этом случае они могут быть немалыми).

Критериальные нормы, как правило, разрабатывают для тестов, уже подтвердивших (на IV-м «Исследовательском» этапе) свою валидность по отношению к выбранному критерию. Однако для того, чтобы были получены полезные для практики нормы, необходим достаточно *высокий коэффициент корреляции между тестом и критерием*. В то время как для целей проверки критериальной валидности (согласно стандартам EFPA) «отвечает требованиям» уже коэффициент взаимосвязи с критерием r=0,2, такой показатель недостаточен для того, чтобы можно было получить «работоспособные» критериальные нормы. При таком уровне взаимосвязи таблицы ожидания не принесут существенной пользы для практики, поскольку будут «сильно размыты», особенно в середине.

Часто, исходя из практических целей, ориентированных на отбор и прогноз, таблицы ожиданий дополняются выделением критериального балла. *Критериальный балл* – это тестовый показатель, с этого балла прогнозируется успешность искомой деятельности не ниже заданной вероятности (например, успеваемость на уровне «не ниже среднего» с вероятностью 95 %). Это своего рода «секущий» балл, который маркирует заданную вероятность критериального поведения. Для конкретного теста критериальный балл, например, может быть равен 6 или 8 стенам или др.

Процедурно задача установления критериального балла заключается в том, чтобы «отсечь» с правой стороны таблицы ожиданий область тестовых результатов, которые гарантируют необходимый уровень выполнения деятельности с заданной вероятностью. Балл, через который проходит «секущая линия», и определяется как критериальный. (Конкретные технологии получения такого рода критериальных баллов подробно описаны в широко известных источниках: см., например, Анастази А., 2001, Готтсданкер Р., 1982).

Требования к ***формированию выборки стандартизации***, используемой при получении критериальных норм, по существу, те же, что и для групповых норм. Качество таблицы ожидания зависит от репрезентативности и объёма конкретной выборки. При неадекватном формировании выборки стандартизации эффективность прогноза будет сомнительна из-за больших выборочных погрешностей или из-за невозможности переноса полученных результатов на целевую популяцию [1, 4, 5].

Однако критериальные нормы редко рассчитывают для широких популяций. Поскольку тесты, использующие такие нормы, максимально практикоориентированы, то пользователя обычно интересует прогноз, затрагивающий вполне конкретные группы. Поэтому критериальные нормы обычно устанавливаются для локальных целевых групп, соответствующих назначению теста. Часто конкретные организации даже формируют свои нормы для внутренних целей отбора кандидатов. Однако иногда в качестве целевой может быть обозначена и достаточно широкая популяция (например, выпускники школ). В таких случаях при формировании выборок стандартизации необходимо ориентироваться на описанные выше принципы стратификации.

К одному тесту могут прилагаться несколько таблиц ожиданий, предназначенных для разных групп (например, для мужчин и женщин или для представителей разных профессий). Проблемы дифференцированной стандартизации подробно рассматривались при обсуждении групповых норм. Они же остаются актуальными и при подготовке норм критериальных.

Для одного теста также может быть получено *несколько разных критериальных норм,* связывающих показатели теста с разными критериями. Например, тест интеллекта может соотноситься и с успешностью обучения, и с успешностью некоторой профессиональной деятельности.

Серьёзная проблема, с которой сталкиваются разработчики, выбирающие для стандартизации этот вид норм, связана с *качеством критерия*. Во II части статьи вопрос качества критерия подробно рассматривался в разделе, касающемся проверки критериальной валидности теста [3]. Не повторяя детали обсуждения, напомним, что, учитывая современное состояние отечественной психодиагностики, в большинстве случаев сложно найти адекватные и надёжные методы замера критерия. Особые трудности возникают, если в качестве критериев используется сложная по своей структуре деятельность, где не совсем ясна относительная важность отдельных функций и отсутствуют проверенные средства их измерения.

Однако если эта проблема решена на этапе проверки критериальной валидности, то она уже не будет актуальной на этапе стан-

дартизации. На самом деле, для расчета критериальных норм можно воспользоваться уже имеющимися данными (при условии, что критериальная валидность теста проверялась на репрезентативной и достаточной по объёму выборке). Нередко при формировании норм существует необходимость расширить выборку. При этом для замера критерия имеет смысл оставить тот же метод, который использовался и при проверке критериальной валидности.

В настоящее время критериальные нормы широко применяются на практике, однако, это не избавляет их от критики. Таблицы ожидания иллюстрируют с особой ясностью проблему практической психологии – а именно – *сложность предсказания индивидуальных результатов* на основе статистического прогноза [1, 5, 7]. Например, если конкретному показателю теста соответствует вероятность 0,45 достижения критерия (например, успешной сдачи экзамена), то это действительно означает, что 45% испытуемых с таким показателем сдадут экзамен, но кто именно составит эти 45%?

Тем не менее, несмотря на неизбежные ошибки предсказания, касающиеся конкретных людей, критериальные нормы и принятые на их основе решения всё же следует признать полезными. Как известно, применение теста на практике оправдано, если ошибка оценки, полученная при использовании его результатов, меньше той, которая могла бы иметь место при принятии решений совсем без тестирования. Например, если при отборе кандидатов на вакантное место критериальные нормы позволяют уменьшить долю ошибочно принятых на работу специалистов, не обладающих необходимыми компетенциями, то такой тест может быть признан не только полезным, но и экономически выгодным.

Напомним, что для многих видов деятельности ошибочное принятие несоответствующих кандидатов (false-positive decision) представляется более нежелательным, с точки зрения социальной безопасности, чем ошибочное отклонение тех, кто мог бы справиться с работой на должном уровне (false-negative decision). Поэтому критериальный балл (о котором говорилось выше), отмечающий необходимый для желаемого поведения тестовый показатель, обычно устанавливают в соответствии с достаточно высокими значениями вероятности (как правило, не ниже 80 %).

Полученные ***критериальные нормы фиксируются*** в виде развёрнутых таблиц ожидания (которые могут быть рассчитаны для разных групп и для разных критериев). При необходимости каждая таблица сопровождается указанием на соответствующий критериальный балл. В Руководстве для пользователя обязательно приводятся количественные и качественные характеристики выборок, на которых были получены нормы. Также указывается точное целевое назначение критериальных норм и при необходимости описываются правила принятия решений на основе полученных нормативов.

В заключение стоит отметить, что, как уже говорилось выше, конкретный тест может сопровождаться разными видами норм. Например, тестовые показатели могут быть переведены в Т-баллы, а на основе Т-баллов построены таблицы ожиданий, связывающие тест с критериальным поведением. Также тест может быть одновременно снабжён и предметно-ориентированными нормами, отражающими уровень освоения некоторых навыков, и критериально-ориентированными, показывающими вероятность успешности в разных видах деятельности для людей с разными тестовыми показателями. Напомним, что выбор вида норм, прежде всего, определяется практическим назначением теста. Однако использование одновременно разных видов норм и их соотнесение предоставляет больше возможностей для интерпретации, что повышает практическую ценность теста.

Литература

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

2. Батурин, Н.А. Технология разработки тестов: часть I / Н.А. Батурин, Н.Н. Мельникова // Вестник ЮУрГУ. Серия «Психология». – 2009. – Вып. 6. – № 30(163). – С. 4–14.

3. Батурин, Н.А. Технология разработки тестов: часть II / Н.А. Батурин, Н.Н. Мельникова // Вестник ЮУрГУ. Серия «Психология». – 2009. – Вып. 7. – № 42(175). – С. 11–25.

4. Готтсданкер, Р. Основы психологического эксперимента / Р. Готтсданкер. – М.: МГУ, 1982. – 464 с.

5. 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.

6. Наследов, А.Д. SPSS: компьютерный анализ данных в психологии и социальных науках / А.Д. Наследов. – СПб.: Питер, 2007. – 416 с.

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

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. Angoff, W.H. Scales, norms, equivalent scores / W.H. Angoff // Educational measurement / ed. by R.L. Thorndike. – Washington, DC: American Council on Education, 1971. – P. 508–600.

10. Cizek, G.J. Setting performance standards: concepts, methods, and perspectives / G.J. Cizek. – Mahwah, NJ: Lawrence Erlbaum Associates, 2001.

11. Cizek, G.J. Standard Setting / G.J. Cizek // Handbook of test development / ed. by Steven M. Downing, Thomas M. Haladyna. – Mahwah, NJ: Lawrence Associates, 2006. – P. 225–258.

12. Downing, S.M. Twelve steps for effective test development / S.M. Downing // Handbook of test development / ed. by Steven M. Downing, Thomas M. Haladyna. – Mahwah, NJ: Lawrence Associates, 2006. – P. 3–25.

13. Ebel, R.L. Essentials of educational measurement / R.L. Ebel. – Englewood Cliffs, NJ: Prentice Hall, 1979.

14. Furr, R.M. Psychometrics: an introducetion / R. Michael Furr, Verne R. Bacharach. – Sade Publications, 2008. – 349 p.

15. Hambleton, R.M. Using an extended Angoff procedure to set standards on complex performance assessments / R.M. Hambleton, B.S. Plake // Applied Measurement in Education. – 1995. – №8. – P. 41–56.

16. Impara, J.C. Standard Setting: an alternative approach / J.C. Impara, D.S. Plake // Journal of Educational Measurement. – 1998. – №35. – P. 353–366.

17. Kane, M. Validating the performance standards associated with passing scores / M. Kane // Review of Educational Research. – 1994. – №64 (3). – P. 425–461.

18. Kingston, N.M. Setting performance standards using the body of work method / N.M. Kingston, S.R. Kahl, K. Sweeney, L. Bay // Setting performance standards: concepts, methods, and perspectives / ed. by G.J. Cizek. – Mahwah, NJ: Lawrence Erlbaum Associates, 2001. – P. 219–248.

19. Mitzel, H.C. The bookmark procedure: Psychological perspectives / H.C. Mitzel, D.M. Lewis, R.J. Patz, D.R. Green // Setting performance standards: concepts, methods, and perspectives / ed. by G.J. Cizek. – Mahwah, NJ: Lawrence Erlbaum Associates, 2001. – P. 283–312.

20. Zieky, M.D. So much has changed: how the setting of cut scores has evolved since the 1980s. / M.D. Zieky // Setting performance standards: concepts, methods, and perspectives / ed. by G.J. Cizek. – Mahwah, NJ: Lawrence Erlbaum Associates, 2001. – P. 159–174.

Поступила в редакцию 24 января 2010 г.

Батурин Николай Алексеевич. Доктор психологических наук, профессор, декан факультета психологии, заведующий кафедрой психодиагностики и консультирования 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», 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.