

Науменко А.С., Орел Е.А. А судьи кто? Индивидуальные особенности разработчиков и характеристики тестовых заданий

2010 - N 4(12)

English version: Naumenko A.S., Orel E.A. Individual attributes of item writers and item characteristics

Южно-Уральский государственный университет, Челябинск, Россия

Государственный университет – Высшая школа экономики, Москва, Россия

Представлен обзор зарубежных исследований, анализирующих влияние индивидуальных особенностей разработчика тестов на конструируемые им тестовые задания. Факт существования такого влияния кажется очевидным, однако работ, посвященных этому вопросу, относительно немного. Наличие такого влияния во многих случаях несет угрозу валидности разрабатываемого инструмента. Рассматривается отражение личности автора тестовых заданий в тестах знаний, личностных опросниках и инструментах для оценки профессиональных навыков. Сформулированы рекомендации по минимизации и/или компенсации влияния индивидуальности разработчика на его творческую продукцию.

Ключевые слова: конструирование тестов, разработка тестов, разработчик теста, автор тестовых заданий, тестовые задания, систематические искажения, личностный опросник, тест знаний, оценка профессиональных навыков

Оглавление

1 Тесты знаний

1.1 Минимизация влияния автора заданий

1.1.1 Навыки успешного выполнения тестов

1.1.2 Алгоритмизация процесса составления тестовых заданий

1.2 Эксперты в предметной области как авторы и рецензенты тестовых заданий

1.2.1 Составление тестовых заданий экспертами в предметной области

1.2.2 Оценка тестовых заданий экспертами в предметной области

Заключение по разделу 1

2 Тесты профессиональной компетентности

2.1 Операционализация конструктов и разработка заданий

2.2 Отбор и анализ пунктов

2.3 Подсчет баллов и интерпретация результатов

Заключение по разделу 2

3 Личностные опросники

4 Общие рекомендации по минимизации искажений

4.1 Детальные спецификации для разработчиков

4.2 Команды разработчиков

4.3 Качественная психометрическая подготовка авторов и экспертов

4.4 Обратная связь авторам и экспертам о качестве заданий

4.5 Супервизия авторов заданий

Любой тест – личностный опросник, тест интеллекта или достижений – это продукт авторского труда разработчика. Всякий разработчик тестов, как любой другой живой человек, обладает определенными индивидуальными особенностями (личностными чертами, спецификой мышления, опытом), которые неизменно сказываются на его творческой продукции, в частности и на тестовых заданиях, которые он составляет. Это влияние может выражаться на формальном или на более концептуальном уровне. Например, согласно расхожему мнению, тесты интеллекта измеряют *то, что их автор считает* интеллектом. Ссылаясь на диссертационное исследование И.С. Кострикиной, А.Н. Поддьяков отмечает, что математический склад ума разработчиков первых тестов интеллекта повлиял на их содержание: в то время, как вербальная логика представлена в тестах заданиями средней сложности, логико-математические задания являются трудными или очень трудными [Поддьяков, 2003]. Если разработчик считает главным в

интеллекте способности к математике и логике, то его тест окажется более содержательно богатым именно в логико-математической части. Это концептуальный вопрос. Что касается формы, то, к примеру, в тестовых заданиях со множественным выбором один автор может постоянно использовать более длинные, развернутые формулировки для ключевых ответов и более короткие – для дистракторов. Другой разработчик может всегда располагать верный ответ в середине списка альтернатив (например, вторым или третьим). Однако каким конкретно образом индивидуальные особенности разработчика влияют на характеристики создаваемых им тестовых заданий? Насколько существенно это влияние? Должны ли индивидуальные особенности предварительно измеряться и учитываться при отборе экспертов или формировании команд разработчиков?

Проведенные до настоящего времени исследования не дают исчерпывающего ответа на эти вопросы. Р.Торндайк и Н.Хаген называли конструирование хорошего теста искусством, напоминающим одновременно сочинение сонета и изготовление торта, когда, с одной стороны, автор относительно свободен в выборе содержания, а с другой – действует по определенному алгоритму [Thorndike, Hagen, 1991]. Множество современных руководств и учебников стремятся дать разработчикам квалификационных тестов эти алгоритмы. Они содержат рекомендации по выбору наиболее «удачного» содержания и формы тестовых заданий и призваны в определенном смысле ограничить творчество и проявление индивидуальности авторов, которые могут быть распознаны «опытными» испытуемыми и использованы для угадывания правильных ответов. Большинство подобных руководств имеют сугубо практическую направленность, и хотя сам факт их существования является красноречивым признанием индивидуальности разработчиков, они не направлены на детальное изучение или рассмотрение связи особенностей авторов с характеристиками заданий.

Исследования, которые посвящены изучению именно этой связи, достаточно редки, однако они есть, и в данной статье будет представлен их обзор. В следующих разделах будут отдельно рассмотрены работы, в которых изучалось влияние индивидуальных особенностей разработчиков на создаваемые ими 1) тесты знаний, 2) инструменты для оценки профессиональной компетентности и 3) личностные опросники. Читателя может удивить отсутствие в этом списке тестов интеллекта, как отдельного класса диагностических инструментов. Однако, по нашему мнению, эту группу методик невозможно рассматривать все общего контекста проблемы интеллекта, что несколько выходит за рамки данного обзора.

Прежде чем приступить к обзору, стоит сказать несколько слов об актуальности подобных работ. Помимо того что влияние особенностей разработчиков тестовых заданий на их характеристики – это слабо исследованная тема, в которой вопросов гораздо больше, чем ответов, на наш взгляд, более важен другой аспект. Забегая вперед, отметим, что главный вывод подавляющего большинства проанализированных нами статей заключается в том, что особенности разработчиков оказывают существенное влияние на конструируемые ими задания, а следовательно, на результаты тех, кто потом будет выполнять эти задания в качестве респондентов. Этот важный источник искажений до сих пор не был систематически описан ни в классических учебниках, ни в современных обзорах по психометрике. Обращая внимание на эту проблему, мы бы хотели еще раз подчеркнуть важность подготовки профессиональных тестологов, которые владели бы технологией создания тестов, в меньшей степени подверженных такого рода искажениям.

Сегодня уже сложно представить учебный курс, в котором не использовалась бы тестовая форма контроля знаний, или крупную организацию, в которой совсем не использовались бы опросники и тесты. Но кто является автором подобных разработок? Чаще всего это сами педагоги, HR-специалисты – те, кто потом принимает решения на основе полученных результатов. Проанализированные нами исследования показывают, что, выступая в роли автора заданий, они сами неосознанно вносят искажения в получаемые тестовые баллы. Таким образом, преимущества тестирования как объективного метода несколько снижаются.

Насколько нам известно, в российской психологической литературе А.Н. Поддьяков первым поднял вопрос об индивидуальных ценностных ориентациях, личностных предпочтениях и стилях разработчиков тестов [Поддьяков, 2003, 2004, 2007]. Своим обзором мы бы хотели продолжить обсуждение, начатое А.Н. Поддьяковым, и обратить внимание профессионального сообщества на данную проблему. Мы надеемся, что в будущем эмпирические исследования уточнят характер и предложат способы минимизации влияния индивидуально-личностных особенностей авторов тестовых заданий.

1 Тесты знаний

1.1 Минимизация влияния автора заданий

Достаточно давно ученых посетила мысль о том, что автор тестовых заданий – «тоже человек», а значит, будет неизбежно проецировать свои личностные черты, особенности мышления, опыт и видение мира на создаваемые им тестовые задания. В ряде достаточно ранних исследований была сделана попытка: 1) отследить связь между гипотетической идиосинкразией разработчика и

успешностью выполнения тестов испытуемыми и 2) предложить теоретически обоснованный способ уменьшить это «авторское» влияние. В следующих подразделах мы кратко остановимся на этих двух направлениях исследований.

1.1.1 Навыки успешного выполнения тестов

Достаточно обширный пласт зарубежных исследований в 1950–1980-е годы был посвящен «тестовой мудрости», или навыку успешного выполнения тестов. Впервые данный конструкт был предложен в работе Р.Торндайка [Thorndike, 1951]. В работе Д.Миллмана и коллег [Millman et al., 1965] тестовая мудрость определяется как способность испытуемого использовать особенности и формат тестового задания и/или ситуации тестирования для получения более высокого балла по тесту. Навык успешного выполнения тестов логически независим от знаний испытуемого в тестируемой области. То есть при одинаковых знаниях предметной области испытуемый, обладающий данным навыком, получит более высокий балл, чем испытуемый, им не владеющий. В большинстве англоязычных работ навык успешного выполнения тестов обозначают термином test-wiseness [Gibb, 1964; Millman et al., 1965], но также и (test) insight [Thorndike, 1951], testmanship [Huff, 1961], test sophistication [Anastasi, 1976; Erickson, 1972] и test wisdom [Preston, 1964]. Пристальное внимание к «тестовой мудрости» обусловлено прежде всего тем, что она может являться дополнительным источником дисперсии тестовых баллов и потенциальным фактором, снижающим валидность теста.

В классической работе Д.Миллмана, К.Бишоп и Р.Эбея [Millman et al., 1965] выделяются две большие группы принципов «тестовой мудрости»: 1) элементы, независимые от разработчика и/или цели тестирования, и 2) элементы, связанные с разработчиком или целью тестирования. В число элементов, связанных с разработчиком, входят:

* распознавание намерений и особенностей автора (угадывание ответа, который имел в виду автор, понимание уровня сложности, предполагаемого автором, и т.д.);

* использование «подсказок», которые неосознанно дает автор (ключевой ответ короче / длиннее дистракторов; ключевой ответ расписан более тщательно, чем дистракторы; ключевой ответ располагается в определенном месте среди дистракторов, среди других аналогичных (или, наоборот, противоположных) утверждений, использована определенная стереотипная фразеология и т.д.).

Как мы видим, исследователи тестовой мудрости описали определенные особенности тестовых заданий, в которых может выразиться индивидуальность автора. Тем не менее они, как правило, не ставили своей целью изучение связей конкретных элементов заданий с конкретными характеристиками их автора. Исследования тестовой мудрости были в основном сосредоточены на анализе ее компонентов и коррелятов, а также создании методов ее измерения [Nilsson, Wedman, 1974; Diamond, Evans, 1972; Millman, 1966];

- оценке определенных тестов (стандартизованных или, наоборот, разработанных непрофессионалами) на наличие «подсказок» для испытуемых, отличающихся тестовой мудростью [Brozo et al., 1984; Metfessel, Sax, 1958; Mehrens, Lehmann, 1973];
- изучении возможности повышения тестовых баллов по стандартизованным тестам после целенаправленного обучения навыкам успешного выполнения тестов [Oakland, 1972; Gaines, Jongasma, 1974; Gross, 1976; Callenbach, 1973; Omvig, 1971; Crehen et al., 1974; McPhail, 1984; Sarnacki, 1979];
- разработке и апробации практических программ обучения навыкам успешного прохождения тестов [Gifford, Fluitt, 1980; Lange, 1978; McPhail, 1984; Parrish, 1982; Roznowski, Bassett, 1992].

1.1.2 Алгоритмизация процесса составления тестовых заданий

Д.Бормут [Bormuth, 1970] был одним из первых, кто предложил заменить конструирование тестовых заданий, основанное на интуиции и субъективном опыте, наукой разработки тестовых заданий. Эту идею подхватили и другие исследователи. Например, Д.Миллман утверждал, что необходимо и возможно структурировать правила порождения тестовых заданий таким образом, чтобы любые два автора с их помощью создавали по сути одинаковые задания [Millman, 1974]. Начиная с этого момента различные исследователи начали работу над теоретически обоснованными методами алгоритмизации и автоматизации процесса создания тестовых заданий. Уже в 1980 году Г.Ройд и Т.Халадина подводят «промежуточные» итоги этих усилий: они упорядочивают методы создания тестовых заданий на пятиуровневом континууме, простирающемся от неформального субъективного способа до алгоритмизированных компьютерных техник [Roid, Haladyna, 1980].

Исследования, однако, указывают на то, что даже использование более формализованных методов создания тестовых заданий не гарантирует свободу от индивидуальных особенностей автора. Так, Г.Ройд и Т.Халадина [Roid, Haladyna, 1978] показали, что два опытных разработчика формулировали задания существенно разной сложности, хотя и получили одинаковые инструкции относительно целей теста и правил создания тестовых заданий. В другой работе [Roid, 1980] сравнивались характеристики тестовых заданий, разработанных с использованием различных методов, а также различными авторами (трое были профессиональными разработчиками тестов, трое – учителями начальной школы). Выяснилось, что при использовании субъективного неформального метода двое из трех профессиональных авторов составляли слишком сложные задания по сравнению с учителями, у которых получались задания средней сложности. При использовании алгоритмизованных техник профессиональные разработчики, напротив, составляли слишком простые задания.

В этом исследовании также измерялась «чувствительность» заданий к обучению, то есть сравнивались результаты тестирования учеников до и после прохождения обучения. По мнению авторов, чувствительность к обучению является важным показателем качества задания, поскольку указывает на способность задания выявлять реальные различия в уровне знаний, появляющиеся в результате обучения. Задания, составленные профессионалами, оказались менее чувствительны к обучению, чем задания, составленные учителями. Таким образом, в данном эксперименте учителям удалось составить несколько более качественные задания, чем профессиональным авторам.

1.2 Эксперты в предметной области как авторы и рецензенты тестовых заданий

Авторы тестовых заданий могут быть разделены на две большие группы: профессиональные разработчики тестов и случайные (например, школьные учителя или другие специалисты, которым приходится время от времени составлять тестовые задания) [Osterlind, 1998]. В данном разделе мы бы хотели остановиться на исследованиях, касающихся «случайных» разработчиков, многие из которых не имеют специальной тестологической подготовки, однако хорошо владеют предметной областью, знания в которой тестируются. Насколько хорошо эксперты в предметной области справляются с задачей составления и/или оценки тестовых заданий?

1.2.1 Составление тестовых заданий экспертами в предметной области

Использование нестандартизованных тестов весьма распространено в сфере образования. Контроль знаний является важной частью работы школьных учителей и университетских

преподавателей, и многие из них самостоятельно составляют задания для оценки усвоения учебного материала. По приблизительным прикидкам, в типичном американском классе учителя используют 54 «самодельных» теста в год [Marso, Pigge, 1988]. В целом ряде стран тесты, самостоятельно разработанные преподавателями, фактически являются единственным инструментом оценки учебных достижений: они применяются для разбиения учеников на группы («норма», одаренные, со специальными потребностями), определения программы обучения (обычной или углубленной), предоставления обратной связи родителям об успехах их детей, корректировки содержания обучения и т.д. При выставлении отметок учителя в большей степени склонны полагаться на собственные тесты, чем на инструменты, разработанные другими, стандартизованные тесты (!) или прочие источники информации [Boothroyd et al., 1992; Fennessey, 1982; Stiggins, Bridgeford, 1985; Williams, 1991].

Таким образом, важность использования учителями качественных инструментов сложно переоценить. Ряд исследований в 1980–1990-х годах был посвящен анализу качества учебных тестов и компетенций преподавателей в качестве авторов тестовых заданий [Boothroyd et al., 1992; Brozo et al., 1984; Carter, 1984; Jozefowicz et al., 2002; Marso, Pigge, 1989, 1991, 1992; Oescher, Kirby, 1990; Pigge, Marso, 1988; Valentin, Godfrey, 1996 и др.]. В целом ученые сходятся во мнении, что качество тестовых заданий, составленных учителями, далеко от идеала. Это связано прежде всего с тем, что при наличии достаточных знаний в предметной области и способности обозначить цель тестирования учителям не хватает знаний в области измерения и конструирования тестов [Wise et al., 1991].

В исследовании Р.Марсо и Ф.Пиджи учителя, завучи школ и супервизоры учебных программ оценивали навыки создания тестов у учителей [Marso, Pigge, 1989]. Выяснилось, что наиболее высоко свои навыки оценивают сами преподаватели, чуть ниже – завучи и наиболее низко – супервизоры. В других исследованиях самооценка учителей в качестве авторов тестовых заданий также оказалась высока [Oescher, Kirby, 1990; Wise et al., 1991]. Интересно, однако, что между реальным качеством заданий и самооценкой учителя связи обнаружено не было [Boothroyd et al., 1992], либо она и вовсе была отрицательной [Marso, Pigge, 1989]. Важно также, что профессиональный опыт учителей (в исследовании Р.Марсо и Ф.Пиджи они были разделены на три группы по опыту работы – от 1 до 3 лет, от 4 до 6 лет, от 7 до 10 лет) не влиял на качество заданий, которые они составляли. То есть новички, проработавшие в школе всего один год, так же успешны (или неуспешны) в качестве авторов тестовых заданий, как и более опытные учителя с 10-летним стажем. Данный факт можно интерпретировать в свете высказанного ранее предположения: дело не в профессионализме, педагогическом опыте учителя или его знаниях предметной области, а в навыках конструирования тестовых заданий.

В относительно недавнем исследовании А.Джафарпура [Jafarpur, 2003] изучалось влияние автора заданий на результаты выполнения испытуемыми теста на понимание иностранного текста. В качестве разработчиков выступили шесть преподавателей английского языка как иностранного: каждый из них составлял вопросы на понимание двух из шести текстов, использованных в эксперименте. Никаких предварительных инструкций и рекомендаций авторам не давали. Выяснилось, что все шесть авторов составили задания разной сложности. Причем авторам, «тяготеющим» к более простым заданиям, удавалось составлять простые вопросы и к изначально сложным для понимания текстам, так, что они становились более пригодными для тестирования плохоуспевающей подгруппы учеников. И наоборот, авторы, склонные к более сложным заданиям, могли составить настолько сложные задания к простому тексту, что он становился пригодным для тестирования более способных студентов.

Кроме того, задания, подготовленные разными авторами, оказались направлены на диагностику различных (суб)навыков: например, одного из авторов больше интересовало умение испытуемого понимать подразумеваемое содержание, а другого – извлекать информацию, явно содержащуюся в тексте. Таким образом, различное содержание тестовых заданий стало следствием различной операционализации конструкта «понимание текста» у разных разработчиков.

Исследование одного из авторов данной статьи [Naumenko, 2009] было проведено в форме конкурса авторов тестовых заданий в области оценки персонала (первая серия) и юридического обеспечения управления персоналом (вторая серия). Конкурсантов просили составлять тестовые задания средней трудности – такие, с которыми справилось бы от 40 до 70% специалистов, для которых предназначался тест. В отличие от зарубежных исследований, описанных в данном обзоре, в нашей работе были измерены как определенные особенности авторов заданий (уровень осведомленности в предметной области), так и характеристики самих заданий (трудность). Для всей выборки авторов (33 человека) была обнаружена значимая корреляция между уровнем экспертного знания и трудностью составленных заданий. То есть наиболее эрудированные участники составляли слишком сложные задания, а задания средней (оптимальной) трудности представили участники со средним уровнем знаний в предметной области. С одной стороны, такая корреляция выглядит логичной. С другой стороны, она позволяет сделать достаточно неожиданный вывод: для составления заданий средней сложности (а, как правило, именно такие пункты преобладают в тестах знаний) есть смысл привлекать профессионалов, показывающих средние (но не наиболее высокие) результаты в интересующей теме.

1.2.2 Оценка тестовых заданий экспертами в предметной области

В некоторых странах (например, в США) для того, чтобы тест знаний был признан соответствующим Стандартам психологического и педагогического тестирования, необходимо провести эмпирическое исследование для определения проходного балла (балла отсека). Существуют различные методы определения проходного балла, один из наиболее широко используемых – метод Ангофа [Angoff, 1971]. В нем группа экспертов в предметной области «на глазок» оценивает трудность тестовых заданий и дает приблизительную оценку доли минимально компетентных испытуемых, которые справятся с заданием. Оценки для каждого пункта усредняются по всем экспертам, а затем суммируются для получения сырого проходного балла.

Несмотря на популярность данного метода, в ряде исследований было показано, что на самом деле экспертной комиссии достаточно сложно адекватно оценить трудность тестовых заданий, то есть прогноз экспертов относительно процента испытуемых, которые не справятся с тестом, оказывается далек от эмпирических данных, получаемых впоследствии. Например, в работе Д.Импара и Б.Плэйк [Impara, Plake, 1998] 26 учителей должны были сделать прикидку того, какой процент их студентов не справится с тестом по естественным наукам. Учителя отдельно оценивали успешность их собственной группы и успешность гипотетической «граничной» (минимально компетентной) группы. Оба прогноза оказались далеки от идеала, однако своих собственных студентов (по всей видимости, в силу персонального знакомства) педагогам удалось оценить точнее. В других исследованиях [Bejar, 1983; Shepard, 1994] были получены в целом сходные результаты.

В работе голландских ученых Верховена и коллег [Verhoeven et al., 2002] сравнивалась точность оценок трудности медицинского теста, предложенных недавними выпускниками и авторами тестовых заданий. Выяснилось, что для получения надежного проходного балла по всему тесту необходимо более 39 авторов тестовых заданий или не менее 10 выпускников. Иными словами, более приемлемым оказался стандарт, предложенный выпускниками. В другом исследовании [Koens et al., 2005] недавних выпускников и опытных практикующих врачей просили оценить задания медицинского теста, составленного преподавателями научных и клинических дисциплин, на предмет отражения ими ключевых знаний, необходимых медику. Из клинических вопросов 82,4%, а из базовых научных вопросов только 42,4% были оценены как отражающие необходимые знания. Между усредненными оценками опытных докторов и выпускников наблюдалась удивительная степень согласия (коэффициент корреляции $r = 0,975$). То есть разработчики теста,

стремившиеся в своих заданиях отразить ключевые знания, оказались более далеки в своих суждениях от реально практикующих специалистов, чем выпускники.

Заключение по разделу 1

Подводя итог обсуждению, представленному в данном разделе, можно сказать, что специалисты в предметной области далеко не всегда могут адекватно отразить собственные индивидуальные особенности (например, уровень эрудиции), влияющие на их творческую продукцию, и предоставить задания с определенными запрашиваемыми характеристиками (например, задания определенной сложности). Также экспертам сложно оценить свою успешность в качестве авторов тестовых заданий. Наиболее вероятными причинами исследователи называют, во-первых, недостаточность знаний и навыков в области конструирования тестов [Carter, 1983; Stiggins, Bridgeford, 1985; Wise et al., 1991], а во-вторых, отсутствие обратной связи: например, в школьной системе ни сами учителя, ни другой школьный персонал не занимаются оценкой психометрического или иного качества тестовых заданий [Gullickson, Ellwein, 1985; Oescher, Kirby, 1990]. Другие ученые, однако, утверждают, что ни многократная и детальная обратная связь авторам тестовых заданий о том, как испытуемые справляются с тестом, ни тренинги в области конструирования тестов не помогают разработчикам составлять задания «нужной» трудности [Verhoeven et al., 1999]. Авторам настоящего обзора, однако, неизвестны систематические исследования по данной теме, поэтому последнее утверждение, с нашей точки зрения, требует более детальной проверки.

2 Тесты профессиональной компетентности

Еще одна важная сфера применения тестирования – оценка персонала. В ней широко используются как личностные опросники, так и тесты интеллекта и достижений. И часто, чтобы оптимизировать и удешевить процесс оценки, специалисты в этой области сами разрабатывают методики оценки профессиональной квалификации. В этом разделе мы попробуем разобраться, какие искажения они вносят в создаваемые ими пункты.

В работе К.Бенкс и Л.Роберсон «Специалисты по оценке персонала как разработчики тестов» (Performance appraisers as test developers) [Banks, Roberson, 1985] проведена аналогия между разработкой критериев и методов оценки профессиональной деятельности и конструированием стандартизированной тестовой методики. Такое сравнение авторы оправдывают вот чем. Разг. Часто процедуры оценки персонала предусматривают лишь общие направления работы оценщика, а конкретные критерии эффективности сотрудников зависят от специфики работы конкретной организации и поэтому разрабатываются уже под конкретную задачу оценки или отбора. Из-за того что в разных организациях работа специалиста на одной и той же позиции может иметь свои

особенности, кадровые консультанты и прочие внешние специалисты в сфере оценки персонала не могут точно прописать, что же оценщик должен сделать, чтобы оценить конкретного специалиста на конкретной должности в конкретной компании. Все критерии эффективности должны прописываться теми, кто проводит оценку. От правильно выбранных критериев в конечном итоге зависит и результат оценочных процедур.

Бенкс и Робертсон утверждают, что хорошими оценщиками могут быть разработчики тестов: именно они обладают нужным навыком выбора валидных индикаторов профессиональной успешности. Разработчик теста пытается измерить гипотетический конструкт, собирая примеры конкретного поведения, в которых он воплощается. Его основная задача заключается в том, чтобы детальным образом определить измеряемое явление и затем создать банк заданий, который бы его описывал целиком и полностью. Успех всей измерительной процедуры зависит от того, сумеет ли разработчик описать конструкт, составить адекватные ему задания и разработать схему подсчета результатов и их интерпретации.

Аналогии между действиями разработчика теста и специалиста в области оценки персонала очевидны. Специалисты в области оценки персонала, по словам Бенкс и Робертсон, могут так же точно определить сферу профессиональной деятельности, чтобы потом подобрать к ней конкретные поведенческие индикаторы эффективности ее выполнения. Точно так же, как разработчик тестов формирует банк вопросов, специалист в сфере оценки формирует пул индикаторов «хорошего» выполнения деятельности до того, как начнет собирать информацию о конкретном специалисте. Он определяет, что именно оцениваемый должен будет сделать или сказать, чтобы получить определенный балл за выполнение своей профессиональной деятельности. То есть специалист в области оценки персонала операционализирует конструкт профессиональной деятельности в виде описания определенного поведения, личностных черт, установок, результатов работы и разнообразных комбинаций всего перечисленного.

Таким образом, делают вывод Бенкс и Робертсон, процесс оценки персонала достаточно близок к разработке тестов, по крайней мере на первом этапе. И в той, и в другой деятельности операционализация измеряемых конструктов играет важную роль. Разумеется, и в оценке персонала, и в разработке тестов существуют различия в операционализации одних и тех же конструктов разными разработчиками, что как раз и служит источником искажений, вносимых автором в процесс измерения.

Попробуем теперь сравнить процедуры создания тестов и разработки системы оценки персонала более подробно. Традиционно процесс конструирования стандартизированного теста принято разделять на следующие этапы: операционализация конструкта и разработка заданий, отбор и анализ пунктов, создание системы подсчета баллов и интерпретации результатов. Бенкс и Роберсон обозначают возможные искажения, которые автор заданий может вносить на каждом этапе:

* одно задание определяет содержание всех остальных – это ситуация, когда одно (удачное, на взгляд автора) задание будет определять содержание всех последующих и, таким образом, измеряемый конструкт не будет охвачен полностью;

* чувствительность заданий к эмоциональному состоянию автора – возможность влияния эмоционального состояния автора на содержание и отбор заданий, а также на подсчет и интерпретацию результатов;

* чувствительность к субъективности в подсчете баллов – степень прозрачности и однозначности системы подсчета баллов в методике;

* чувствительность к индивидуальным особенностям автора – степень, в которой отбор заданий и интерпретация результатов зависят от устойчивых индивидуальных особенностей автора.

Обратимся более подробно к каждому из перечисленных выше этапов разработки.

2.1 Операционализация конструктов и разработка заданий

Выше мы уже писали о том, что операционализация измеряемого конструкта является важной составляющей работы специалиста в области оценки персонала, а разработку заданий можно сравнить с описанием поведенческих индикаторов успешной профессиональной деятельности.

Исследования показывают, что специалисты в сфере управления персоналом используют разные критерии для оценки одной и той же деятельности на разных организационных уровнях. Так, в исследовании Бормана [Borman, 1974] было продемонстрировано, что вместо того чтобы использовать общие критерии эффективности для одной и той же деятельности на разных должностных позициях, оценщики используют для каждого уровня свой набор критериев. Исследование Бормана было проведено на материале оценки секретарей и преподавателей секретарского дела. Автор предположил, что, несмотря на то что в их деятельности много общего,

оценщики будут концентрироваться только на том уровне, который представляют сами, и предлагать критерии оценки, применимые только для него. Чтобы проверить эту гипотезу, Борман собрал две группы оценщиков с той и с другой позиции и предложил им разработать критерии эффективности только для тех сфер (секретарь и преподаватель секретарского дела), в которых, по их мнению, сотрудники их уровня могут оценить работу секретаря.

Критерии, предложенные теми, кто оценивал работу секретаря, практически не пересекались с критериями, предложенными оценщиками преподавателей (несмотря на инструкцию оценивать только ту деятельность, которая является общей для обеих должностей). Для сравнения, первая, «секретарская», группа в качестве параметров оценки выдвинула «Знание работы», «Организованность», «Сотрудничество с коллегами» и «Ответственность». Вторая, «преподавательская», группа предложила другие критерии: «Рассудительность», «Техническая компетентность» и «Добросовестность». Более того, когда испытуемым предложили оценить специалистов с обеих позиций по критериям, предложенным ими, и затем по критериям, предложенным второй группой, оказалось, что согласованность оценок выше в том случае, когда оценщики пользуются теми критериями, которые разработали сами. Таким образом, положение оценщика в организационной иерархии оказывается источником искажений при операционализации оцениваемого конструкта. Данные, подтверждающие этот вывод, были получены также в исследовании Лэнди, Фарра, Саала и Фрайтага [Landy et al., 1976] на материале оценки офицеров полиции.

Также, по словам Бенкс и Робертсон, распространена ситуация, когда оценщики не учитывают большое количество критических инцидентов (поведение в ситуациях, высоко релевантных выполняемой деятельности), когда адаптируют уже готовые оценочные шкалы под нужды своей организации. Это происходит из-за того, что разные эксперты по-разному категоризируют образцы поведения и по-разному оценивают их важность для выполняемой деятельности.

О том, как влияет прошлый опыт на используемые оценщиком критерии эффективности, известно немного, однако существуют теории [Hogarth, 1980], в которых утверждается, что они связаны. Эти предположения подтверждает эмпирическое исследование Бенкс [Banks, 1982], в котором доказано, что оценщики, различающиеся по своему профессиональному опыту, используют разные критерии (а также разное количество критериев), оценивая одну и ту же профессиональную деятельность.

Таким образом, исследования свидетельствуют о том, что причины несогласованности оценок одной и той же деятельности зачастую кроются в том, что каждый оценщик по-своему операционализирует оцениваемые конструкты и по-своему оценивает вес каждого критерия в эффективности оцениваемой профессиональной деятельности.

2.2 Отбор и анализ пунктов

В правильно разработанном измерительном инструменте окончательный набор заданий (поведенческих критериев оценки) должен быть апробирован на большой выборке, релевантной той популяции, на которой методика будет использоваться в дальнейшем. Для создания психометрических тестов это необходимый этап. Специалисты в области оценки персонала, так же как и разработчики тестов, проводят отбор используемых ими критериев, но не подтверждают свой выбор эмпирическими данными. Предполагается, что отобранные ими критерии являются согласованными и скоррелированными между собой [Abelson, 1976; Hogarth, 1980], однако в большинстве случаев они не проходят предварительного тестирования на репрезентативных гетерогенных выборках. Оценщики основываются на собственных знаниях и профессиональном опыте [Banks, Roberson, 1985], поэтому валидность и надежность выбранных критериев не подтверждается эмпирически.

Необходимо отметить, что в целом эмпирические данные о валидности и надежности критериев оценки оказываются не столь важны для работы специалиста по оценке персонала. Исследования показывают, что они часто формируют свое мнение на основе таких не слишком релевантных профессиональной деятельности свойств, как раса [Landy, Farr, 1976], внешняя привлекательность оцениваемого [Cann et al., 1981] и даже сходство с самим оценщиком [Leonard, 1976].

Так, в исследовании Канна, Зигфрида и Пирс [Cann et al., 1981] было показано, как внешняя привлекательность и пол кандидатов влияют на решение о принятии их на вакантную должность. Исследователи выяснили, что мужчины более высоко оценивают профессиональную квалификацию привлекательных женщин. В работах Леонарда [Leonard, 1976] и Маникс и Нил [Mannix, Neal, 2005] рассматривалась социально-психологическая концепция привлекательности по сходству (*similarity – attraction paradigm*) и оказалось, что люди, обладающие ценностями и нормами поведения, близкими к нашим собственным, оцениваются как более привлекательные и обладающие более высокой квалификацией. Этот вывод можно также применить и к специалистам по оценке персонала.

Валидность критериев оценки также подвергается угрозе со стороны других особенностей социальной перцепции. Например, в исследовании Кэррол и Шнайдера [Carrol, Schneider, 1982] было доказано, что судьи воспринимают негативную информацию об обвиняемом как более важную, чем ту, которая свидетельствует в его пользу. Аналогии с процессом оценки очевидны.

Из проанализированных нами исследований можно сделать вывод, что эмпирическая проверка валидности и надежности выбранных критериев оценки эффективности профессиональной деятельности, хотя и не считается обязательной, сильно влияет на качество и объективность процедуры оценки.

2.3 Подсчет баллов и интерпретация результатов

Как известно, требование к любой стандартизованной тестовой методике – унификация процедур проведения и подсчета баллов. Однако когда дело касается оценки эффективности, все далеко не так однозначно. Так, эмпирические данные, полученные в исследовании Шмитта и Хилла [Schmitt, Hill, 1977], свидетельствуют о том, в зависимости от состава групп, в которых происходит оценка, пол и раса влияют на принятие оценщиком решения об уровне квалификации оцениваемого. Авторы показали, что если черные женщины входили в группу, преимущественно состоявшую из белых мужчин, их профессиональную квалификацию оценивали ниже, чем когда они представляли группы, разнородные по полу и цвету кожи.

Другие исследования показывают, что в ситуации, когда у оценщиков есть возможность варьировать критерии оценки и последующую интерпретацию результатов (то есть когда процедура прописана нестрого), они часто изменяют процедуру оценки и принятия решения в зависимости от особенностей оцениваемого, не относящихся к его профессиональной деятельности. Так, в работе Снайдера и Свонна [Snyder, Swann, 1978] выяснилось, что интервьюеры, выстраивающие беседу согласно своему внутреннему представлению об оцениваемых, задают им разные вопросы и таким образом ставят в разные условия. Чаще всего процедура оценки и задаваемые вопросы не бывают прописаны. А значит, отсутствует возможность проанализировать и унифицировать ее.

Заключение по разделу 2

Процессы создания и проведения оценки персонала вполне можно сопоставить с разработкой стандартизованных тестовых методик. И, если привести их в соответствие с критериями,

применяемыми к психометрическим инструментам, эти системы оценки только выиграют в качестве. Проанализированные нами исследования прямо указывают на то, что на каждом этапе, начиная от операционализации исследуемых понятий и заканчивая проведением оценки, те, кто эту оценку проводят, являются источниками искажений, которые сказываются на ее результатах. Поскольку решения, принимаемые в процессе оценки профессиональной квалификации, оказываются важными в судьбе оцениваемого, эти искажения следует обязательно принимать во внимание.

Вывод, который можно сделать из проанализированного нами материала, однозначен: специалисты в области оценки персонала далеко не всегда могут быть хорошими разработчиками психометрических тестов. Однако овладение навыками конструирования тестов может существенно улучшить качество их собственной профессиональной деятельности.

3 Личностные опросники

В данном разделе мы рассмотрим несколько исследований, посвященных анализу различных подходов к конструированию личностных опросников. В двух работах 1970-х годов ученые пытались выяснить, могут ли относительно наивные авторы создавать валидные тестовые шкалы. Первое подобное исследование принадлежит С.Эштону и Л.Голдбергу [Ashton, Goldberg, 1973]: они сравнивали набор шкал, подготовленный 15 непрофессионалами и 15 студентами-психологами, со шкалами Калифорнийского личностного опросника (California Psychological Inventory, CPI) и Формы по изучению личности Д.Джексона (Personality Research Form, PRF). В качестве оценки внешней валидности трех исследованных шкал («Коммуникабельность», «Достижение» и «Доминантность») использовались оценки людей, хорошо знакомых с испытуемыми. Валидность шкал, разработанных непрофессионалами, в среднем оказалась несколько ниже валидности шкал CPI, однако наиболее надежные шкалы непсихологов были сравнимы по валидности со шкалами CPI. Валидность шкал, разработанных студентами, оказалась в среднем выше валидности шкал CPI, а наиболее надежные студенческие шкалы и вовсе были сравнимы по валидности с более тщательно и скрупулезно разработанными шкалами PRF.

Д.Джексон [Jackson, 1975] отчасти повторил и одновременно расширил исследование Эштона и Голдберга. Джексон использовал другой набор личностных конструктов («Социальная включенность», «Самооценка», «Толерантность») – менее очевидных для непрофессионалов, чем для профессиональных авторов тестовых заданий. В его эксперименте участвовали 22 студента-психолога (в основном третьекурсники). В качестве меры внешней валидности были использованы результаты самооценки и результаты оценки испытуемых их соседями по комнате.

Валидность студенческих шкал оказалась, как и в работе Эштона и Голдберга, гораздо выше, чем валидность шкал CPI, и была почти сравнима с валидностью шкал Личностного опросника Джексона [Jackson Personality Inventory, JPI]. Однако показатели социальной желательности оказались у студенческих шкал несколько ниже, чем у шкал JPI. Шкалы CPI показали более низкую валидность по сравнению со студенческими шкалами и по критериям самооценки, и по оценкам соседей.

Результаты двух описанных работ говорят о том, что непрофессиональные авторы (в частности, студенты-психологи) могут выступать в качестве разработчиков личностных опросников и создавать инструменты с валидностью, сравнимой с валидностью стандартизованных опросников.

В недавнем исследовании К.Шарпли и Д.Роджерс [Sharpley, Rogers, 2006] сравнивались задания, составленные непрофессионалами, профессиональными авторами тестовых заданий, и пункты стандартизованной Шкалы самооценки тревожности Цунга. Выяснилось, что не-психологам удалось составить задания, по уровню валидности сопоставимые с пунктами, сформулированными профессиональными психологами.

4 Минимизация влияния индивидуальных особенностей разработчика тестов

В заключение нашего обзора мы бы хотели рассмотреть некоторые рекомендации, направленные на минимизацию искажений, связанных с особенностями разработчика тестов. Эти рекомендации также основаны на зарубежном опыте, и их можно разделить на несколько содержательных пунктов.

4.1 Детальные спецификации для разработчиков

Распространенные сейчас руководства для разработчиков тестов включают весьма конкретные и часто очевидные рекомендации по формулированию пунктов – в большинстве своем они элементарны, основаны на здравом смысле и устоявшихся традициях (например, «Избегайте двусмысленных пунктов» или «Формулируйте вопрос четко»). Как иронически отмечает Нитко [Nitko, 1984]: «Пожилые авторы тестовых заданий передают новичкам списки, которые они и их предшественники добыли с помощью искусства, эмпирического исследования и практического опыта». Действительно, мета-анализ, проведенный Б.Фреем и коллегами [Frey et al., 2005], показал, что далеко не все рекомендации из учебников и руководств подтверждены эмпирически и, следовательно, валидны.

Как мы обсуждали, наличие у разработчиков таких весьма общих руководств не гарантирует отсутствие влияния индивидуальных особенностей авторов на содержание и психометрические характеристики заданий. Необходима унификация процедуры оценки и требований к авторам заданий. То есть перед тем, как авторы приступят к разработке, они должны получить детальные спецификации, в которых прописано, для каких целей будет применяться методика, какие навыки, черты или свойства будет измерять, на какие группы респондентов рассчитана, и т.д. Все это – стандартные требования к психометрическим тестам. Однако если взглянуть на них с позиций исследуемого предмета, они обретают несколько иной смысл: таким образом мы ограждаем результаты респондентов от тех искажений, которые не зависят от его личностных или мотивационных особенностей. В ситуациях, когда разработчиком тестов выступает школьный учитель, оценивающий своих учеников, или специалист по персоналу той компании, где работает испытуемый (то есть лица, в некоторой степени заинтересованные в определенных результатах), это представляется особенно важным.

4.2 Команды разработчиков

Еще один способ борьбы с подобного рода искажениями – создание команд разработчиков [Nichols, 1994]. В команде индивидуальные особенности каждого будут несколько сглаживаться. А чтобы не получилось ситуации, когда сходные особенности авторов, напротив, в команде усилятся, рекомендуется формировать их из разных специалистов: психологов, которые корректно операционализируют измеряемый конструкт; специалистов в предметной области, которые обеспечат грамотное содержание; специалистов по статистике, которые разработают адекватную процедуру обработки результатов, а также других экспертов, которые смогут критически оценить результаты работы. Роль лидера в такой команде – координировать действия этих специалистов, чтобы получившийся в результате измерительный инструмент отвечал всем психометрическим требованиям.

При этом отбор составителей заданий и экспертов в подобные группы должен учитывать описанные нами результаты эмпирических исследований. Например, как мы обсуждали, принцип «чем больше профессиональный опыт, тем лучше составляемые задания» в области психометрики, по всей видимости, не работает. Как показывает практика, для тестов знаний лучшие задания (то есть обладающие наибольшей дифференцирующей способностью) составляют специалисты со средним уровнем профессиональной подготовки, студенты-психологи могут выступить в качестве разработчиков личностных шкал, а недавние выпускники могут быть лучшими экспертами для оценки содержания и трудности заданий, чем их педагоги.

4.3 Качественная психометрическая подготовка авторов и экспертов

Здесь встает вопрос о профессиональной психометрической подготовке различных специалистов, занимающихся оценкой. Специализированные программы по подготовке тестологов в нашей стране только открываются. В основном все образование в этой области заканчивается обычным курсом психодиагностики и самостоятельным освоением того, о чем преподаватель успевает только упомянуть. Поэтому качество подготовки специалистов по тестированию как в психологии, так и в образовании пока оставляет желать лучшего. Однако это не мешает нам в очередной раз поднять вопрос о том, что специальная подготовка профессионалов в сфере тестирования в нашей стране необходима. Такие специалисты будут востребованы как в системе образования (и школьного, и вузовского, и постдипломного), так и в психологии, и в работе с персоналом.

4.4 Обратная связь авторам и экспертам о качестве заданий

Важная составляющая текущей работы разработчика тестов – это постоянная обратная связь о качестве составленных им заданий. Если автор теста не забывает о своем продукте сразу после того как рассчитает нормы, то постоянный контроль качества заданий как раз и обеспечивает такую обратную связь. Поэтому соблюдение полной технологии в разработке теста автоматически снижает вероятность искажений, вносимых самими авторами заданий.

4.5 Супервизия авторов заданий

Для того чтобы не проецировать на клиентов свои «неотработанные» эмоции и переживания, психологи-консультанты и психотерапевты проходят регулярную супервизию. Система подобной супервизии должна быть разработана и стать доступна авторам тестовых заданий. Она должна, во-первых, включать обратную связь о качестве заданий (см. предыдущую рекомендацию), а во-вторых, анализ «потенциально опасных» индивидуальных особенностей разработчика, который будет предшествовать разработке. Здесь необходимо отметить, что в российской психологии проблема супервизии разработчиков тестовых заданий уже поднималась в работе А. Н. Поддьякова [Поддьяков, 2004].

Литература

Поддьяков А. Н. Тест творчества – "синяя птица" психологии // Знание - сила. 2003. № 5. С. 101-104

Поддьяков А.Н. Психодиагностика интеллекта: выявление и подавление способностей, выявление и подавление способных // Психология. Журнал Высшей школы экономики. 2004. Том.1. № 4. С. 75-80. То же [Электронный ресурс] // [Сайт ГУ-ВШЭ] URL: http://new.hse.ru/sites/psychology_magazine/rus/issues/v1_n4/Poddyakov_1-04pp75-80.pdf (дата обращения: 20.08.2010).

Поддьяков А.Н. Тестирование интеллекта, конкуренция и рефлексия // Рефлексивные процессы и управление. 2007. N 2. С. 46-56.; То же [Электронный ресурс]. URL: http://www.intelros.ru/pdf/rpu/01_2008/6.pdf (дата обращения: 20.08.2010).

Abelson R.P. Script processing in attitude formation and decision making // J.C.Callol, J.W.Payne (Eds.). Cognition and social behavior. Hillsdale, Nj: Erlbaum, 1976. P. 33–45.

Anastasi A. Psychological testing. N.Y.: Prentice Hall, 1968. 665 p.

Angoff W.H. Scales, norms, and equivalent scores // Thorndike R.L. (Ed.) Educational measurement. 2nd ed. Washington, DC: American Council on Education. 1971. P. 508–600.

Ashton S.G., Goldberg, L.R. In response to Jackson's challenge: The comparative validity of personality scales constructed by the external (empirical) strategy and scales developed intuitively by experts, novices, and laymen // Journal of Research in Personality. 1973. Vol. 7, N 1. P. 1–20.

Banks C.G. Cue selection and evaluation elicited during the rating process // Department of Management Working Paper N 82–83, University of Texas, Austin, TX, 1982. 37 p.

Banks C., Roberson L. Performance Appraisers as Test Developers // *Academy of Management Review*. 1985. Vol. 10(1). P. 128–142.

Bejar I. Subject matter experts' assessment of item statistics // *Applied Psychological Measurement*. 1983. Vol. 7. P. 303–310.

Boothroyd R.A., McMorris R.F., Pruzek R.M. What do teachers know about measurement and how did they find out? (Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA, 1992.) // ERIC Document Reproduction Service. No. 351 309. 24 p.

Borman W. The Rating of Individuals in Organizations: An Alternate Approach // *Organizational Behavior & Human Performance*. 1984. Vol. 12, N 1. P. 105–124.

Bormuth J.R. On the theory of achievement test items. Chicago: University of Chicago Press. 1970. 163 p.

Brozo W.G., Schmelzer R.V., Spires H.A. A study of test-wiseness clues in college and university teacher-made tests // *Journal of Learning Skills*. 1984. Vol. 3. P. 56–68.

Callenbach C. The effects of instruction and practice in content-independent test-taking techniques upon the standardized reading test scores of selected second grade students // *Journal of Educational Measurement*. 1973. Vol. 10. P. 25–30.

Carter K. Do teachers understand principles for writing tests? // *Journal of Teacher Education*. 1984. Vol. 35. P. 57–60.

Carter K. Tackling the testing issue: Test-wiseness for teachers and students (Paper presented at the annual meeting of the American Educational Research Association, Montreal, 1983.) // ERIC Document Reproduction Service No. 482 315. 19 p.

Crehan K.D., Koehler R.A., Slakter M.J. Longitudinal studies of test-wiseness // *Journal of Educational Measurement*. 1974. Vol. 11. P. 209–212.

Diamond J.J., Evans W.J. An investigation of the cognitive correlates of test-wiseness // *Journal of Educational Measurement*. 1972. Vol. 9. P. 145–150.

Erickson M.E. Test sophistication: An important consideration // *Journal of Reading*. 1972. Vol. 16. P. 140–144.

Fennessey D. Primary teachers' assessment practices: some implications for teacher training (Paper presented at the Annual Meeting of the South Pacific Association for Teacher Education, Frankston, Victoria, Australia, 1982.) // ERIC Document Reproduction Service No. 229 346. 21 p.

Gaines W.G., Jongsma E.A. The effect of training in test-taking skills on the achievement scores of fifth grade pupils // Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Illinois, April 1974. 7 p.

Gibb B.G. Test-wiseness as secondary cue response: Ph.D. Thesis / School of Education, Stanford University. 1964. 99 p.

Gifford C.S., Fluitt J.L. How to make your students test wise // *American School Board Journal*. 1980. Vol. 29. P. 29–40.

Gross L.J. The Effects of Test-Wiseness on Standardized Test Performance // *Scandinavian Journal of Educational Research*. Vol. 21, Issue 1. 1977. P. 97–111.

Gullickson A.R., Ellwein M.C. Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice // *Educational Measurement: Issues and Practice*. 1985. Vol. 4. P. 15–18.

Hogarth R.M. Judgment and choice. N.Y.: Wiley, 1987. 324 p.

Huff D. Score: The strategy of taking tests. N.Y.: Appleton-Century-Crofts, 1961. 119 p.

Impara J.C., Plake B.S. Teacher s' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method // Journal of Education Measurement. 1998. Vol. 35, N 1. P. 69–81.

Jackson D.N. The Relative validity of scales prepared by naive item writers and those based on empirical methods of personality scale construction // Educational and Psychological Measurement 1975. Vol. 35. P. 361–370.

Jafarpur A. Is the test constructor a facet? // Language Testing. 2003. Vol. 20, N. 1. P. 57–87.

Jozefowicz R.F., Koeppen B.M., Case S., Galbraith R., Swanson D., Glew R. The quality of in-house medical school examination // Academic medicine. 2002. Vol. 77, Issue 2. P. 156–161.

Koens F., Rademakers J.J.D.J.M., Ten C. Validation of core medical knowledge by postgraduates and specialists // Medical Education. 2002. Vol. 39. P. 911–917.

Landy F.J., Farr J.L. Performance rating // Psychological Bulletin. 1980. Vol. 87. P. 72–107.

Landy F.J., Farr J.L. Police performance appraisal // JSAS Catalog of Selected Documents in Psychology. 1976. Vol. 6. P. 83–97.

Landy F., Farr J., Saal F., Freytag W. Behaviorally Anchored Scales for Rating the Performance of Police Officers // Journal of Applied Psychology. 1970. Vol. 61, N 6. P. 750–758.

Lange R. Flipping the coin: Test anxiety to test-wiseness // *Journal of Reading*. 1978. Vol. 22. P. 274–277.

Leonard D. Cognitive complexity and the similarity-attraction paradigm // *Journal of research in personality*. 1976. Vol. 10. P. 83–88.

Mannix E., Neale M. What Differences Make a Difference? // *Psychological Science in the Public Interest*. 2005. Vol. 6. N 2. P. 31–55.

Marso R.N., Pigge F.L. The status of classroom teachers' test construction proficiencies: assessment by teachers, principals, and supervisors validated by analyses of actual teacher-made tests (Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Francisco, 1989.) // ERIC Document Reproduction Service No. 306 283. 39 p.

Marso R.N., Pigge F.L. An analysis of teacher-made tests: item types, cognitive demands, and item construction errors // *Journal of Contemporary Educational Psychology*. 1991. Vol. 16. P. 279–286.

Marso R.N., Pigge F.L. A summary of published research: Classroom teachers' knowledge and skills related to the development and use of teacher-made tests (Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1992.) // ERIC Document Reproduction Service No. ED 346 148. 29 p.

McPhai I. Coaching, test-wiseness and test scores // *NAPW Journal*. 1984. Vol. 1, N 2. P. 19–26.

Mehrens W.A., Lehmann I.J. *Measurement and evaluation in education and psychology*. 4th ed. N.Y.: Wadsworth Publishing, 1991. 592 p.

Metfessel N.S., Sax G. Systematic biases in the keying of correct responses on certain standardized tests // *Educational and Psychological Measurement*. 1958. Vol. 18. P. 787–790.

Millman J. Criterion-referenced measurement // W.J.Popham (Ed.), *Evaluation in education: Current applications*. Berkeley, California: McCutchan Publishing Co., 1974. P. 311–397.

Millman J., Bishop C.H., Ebel R. An analysis of test-wiseness // *Educational and Psychological Measurement*. 1965. Vol. 25. P. 707–726.

Naumenko A.S. Selecting experts for developing multiple-choice tests // *The 11th European Congress of Psychology. A Rapidly Changing World – Challenges for Psychology*. Oslo, Norway, 7–10 July, 2009. Final Program. P. 133.

Nilsson I., Wedman I. On test-wiseness and some related constructs // *Educational Reports, UMEA*, 1974. N 7. P. 147–159.

Oakland T. The effects of test-wiseness materials on standardized test performance of preschool disadvantaged children // *Journal of School Psychology*. 1972. Vol. 10. P. 355–360.

Oescher J., Kirby P.C. Assessing teacher-made tests in secondary math and science classrooms (Paper presented at the Annual Meeting of the National Council on Measurement in Education. Boston, MA, 1990.) // *ERIC Document Reproduction Service No. 322 169*. 36 p.

Omvig C.P. Effects of guidance on the results of standardized achievement testing. *Measurement and Evaluation in Guidance*. 1971. Vol. 4. P. 47–52.

Osterlind S.J. *Constructing test items: multiple-choice, constructed-response, performance and other formats*. Boston, MA: Kluwer Academic Publishers, 1998. 352 p.

Parrish B.W. A test to test test-wiseness // *Journal of Reading*. 1982. Vol. 25. N 7. P. 672–75.

Pigge F.L., Marso R.N. Supervisors agenda: identifying and alleviating teachers' test construction errors // Paper presented at the Annual Conference of the Ohio Association for Supervision and Curriculum Development (Columbus, OH, November 3–4, 1988). 50 p.

Preston R. Ability of students to identify correct responses before reading // *Journal of Educational Research*. 1964. Vol. 58. P. 181–183.

Roid G. A comparison of item-writing methods for criterion-referenced tests // Paper presented at the joint Annual Meetings of the American Educational Research Association and the National Council on Measurement in Education (Boston, MA, April 7–11, 1980). 24 p.

Roid G., Haladyna T. The emergence of an item-writing technology // *Review of educational research*. 1980. Vol. 50, N 2. P. 293–314.

Roznowski M., Bassett J. Training test-wiseness and flawed item types // *Applied Measurement in Education*. 1982. Vol. 5, N 1. P. 35–48.

Sarnacki R.E. An examination of test-wiseness in the cognitive test domain // *Review of Educational Research*. 1979. Vol. 49, N 2. P. 252–279.

Schmitt N., Hill T. Sex and race composition of assessment center groups as a determinant of peer and assessor ratings // *Journal of Applied Psychology*. Vol. 62, N 3. P. 261–264.

Sharpley C.F., Rogers H.J. Naive versus sophisticated item-writers for the assessment of anxiety // *Journal of Clinical Psychology*. 2006. Vol. 41, Issue 1. P. 58–62.

Shepard L. Implications for standard setting of the NAE evaluation of NAEP achievement levels // Paper presented at the Joint Conference on Standard Setting for Large Scale Assessments, Washington, DC. 1994. 21 p.

Snyder M., Swann W.B. Hypothesis-testing processes in social interaction // *Journal of Personality and Social Psychology*. 1978. Vol. 36, N 11, P. 1202–1212.

Stiggins R.J., Bridgeford N.J. The ecology of classroom assessment // *Journal of Educational Measurement*. 1985. Vol. 22. P. 271–286.

Thorndike E.L. *Educational Measurement*. Washington, D.C.: American Council on Education, 1971. 768 p.

Thorndike R.L., Hagen E. *Measurement and evaluation in psychology and education*. 8th ed. N.Y.: Macmillan Publishing Company, 2009. 528 p.

Valentin J.D., Godfrey J.R. The reliability and validity of tests constructed by Seychellois teachers. A paper presented at the 1996 joint conference organised by Educational Research Association (Singapore) and Australian Association for Research in Education held in Singapore from November 25th to 29th, 1996. 25 p.

Verhoeven B.H., Verwijnen A.M.M., Muijtjens G.M., Scherpbier A.J.J.A., van der Vleuten C.P.M. Panel expertise for an Angoff standard setting procedure in progress testing: item writers compared to recently graduated students // *Medical Education*. 2002. Vol. 36, N 9. P. 860–867.

Verhoeven D.H., van der Steeg A.F.W., Scherpbier A.J.J.A., Muijtjens A.M.M., Verwijnen G.M., van der Vleuten C.P.M. Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges // *Medical Education*. 1999. Vol. 33. P. 832–837.

Williams J.M. (1991). Writing quality teacher-made tests: a handbook for teachers // ERIC Document Reproduction Service No. 349 726. 1991. 48 p.

Wise S.L., Lukin L.E., Roos L.L. Teacher beliefs about training in testing and measurement // Journal of Teacher Education. 1991. Vol. 42. P. 37–42.

Поступила в редакцию 20 июня 2010 г. Дата публикации: 26 августа 2010 г.

Сведения об авторах

Науменко Анна Сергеевна. Кандидат психологических наук, доцент кафедры психологической диагностики и консультирования, факультет психологии, Южно-Уральский государственный университет, пр. Ленина, д. 76, 454080 Челябинск, Россия.

E-mail: a.s.naumenko@gmail.com Данный адрес e-mail защищен от спам-ботов, Вам необходимо включить Javascript для его просмотра.

Орел Екатерина Алексеевна. Кандидат психологических наук, старший преподаватель кафедры организационной психологии, факультет психологии, Государственный Университет – Высшая школа экономики, Волгоградский проспект, д. 46б, 109316 Москва, Россия.

E-mail: eorel@hse.ru Данный адрес e-mail защищен от спам-ботов, Вам необходимо включить Javascript для его просмотра.

Ссылка для цитирования

Науменко А.С., Орел Е.А. А судьи кто? Индивидуальные особенности разработчиков и характеристики тестовых заданий [Электронный ресурс] // Психологические исследования: электрон. науч. журн. 2010. N 4(12). URL: <http://psystudy.ru> (дата обращения: чч.мм.20гг).

[Номер госрегистрации статьи в Реестре электронных научных изданий ФГУП НТЦ "Информрегистр" ожидается.]