Организационная психология. 2016. Т. 6. № 4. С. 30–52.

ОРГАНИЗАЦИОННАЯ ПСИХОЛОГИЯ

www.orgpsyjournal.hse.ru

Эффективность центров оценки: историческая перспектива

ЖУКОВ Юрий Михайлович
Московский государственный университет имени М. В. Ломоносова, Москва, Россия

В статье прослеживается эволюция концепции эффективности в психологической практике с особым акцентом на проблемы валидности центров оценки. Ранние методы и процедуры центров оценки возникли и были развиты в военно-психологических и военно-психиатрических службах некоторых стран Европы и Америки перед Второй мировой войной и во время ведения боевых действий. Первые исследования эффективности центров оценки были начаты уже после окончания войны. Проводились измерения надежности и валидности. Центр оценки в его современном виде оформился в результате проведения исследований профессиональной карьеры в телекоммуникационной корпорации AT&T. Тогда же развернулись широкомасштабные исследования валидности центров оценки. Это было время господства тринитарной доктрины валидности (а именно, комплекса, состоящего из содержательной, критериальной и конструктной валидности). Как правило, при проведении центров оценки обеспечивались высокие показатели содержательной и критериальной валидности. Проблема была в том, что не удавалось получить веских доказательств наличия конструктной валидности. Исследования раз за разом обнаруживали, что поведенческие индикаторы разных компетенций в одних и тех же упражнениях более сильно коррелируют между собой в отдельных упражнениях (эффект упражнений), чем поведенческие показатели одних и тех же компетенций в разных упражнениях (эффект измеряемых параметров). Этот феномен получил название парадокса конструктной валидности центров оценки. В настоящее время тринитарную доктрину сменяет концепция унитарной валидности, что открывает возможности для переформатирования данной проблемы с целью создания более эффективных центров оценки. Далее проводится анализ намечающихся тенденций решения сходных методологических проблем в других областях психологической практики, таких как качественные исследования и исследования действием. Рассматриваются пути повышения валидности, справедливости, приемлемости и действенности центров оценки для подбора и развития персонала.

Ключевые слова: тринитарная доктрина; парадокс конструктной валидности; концепция унитарной валидности; справедливость; приемлемость; действенность.

Введение

Большинство специалистов сходятся во мнении, что технологии центров оценки (от англ. *assessment center*) зародились к концу первой мировой войны в немецких военно-психологических лабораториях. В годы второй мировой войны эти технологии получили свое развитие на британских островах силами армейских психиатров, а за океаном — сотрудниками станций оценки американских спецслужб. В Великобритании эти технологии нашли свое применение сразу после окончания военных действий при подборе персонала для гражданских служб,

Адрес: 125009, Москва, Моховая ул., д. 11, Стр. 9. **E-mail:** zhukovury@rambler.ru

а в США начали использоваться с 1950–х годов для решения кадровых вопросов в коммерческих организациях (MacKinnon, 1987; Munchus, McArthur, 1991; Барышникова, 2013). Что касается оценки эффективности, то масштабные исследования в этом направлении были развернуты несколько позже. В настоящее время при обсуждении проблем качества методов и методик психологической диагностики используются такие понятия, как «надёжность», «валидность», «справедливость», «применимость», «затратность». На стадии зарождения технологий центров оценки (ЦО) основной акцент делался на таких свойствах методического инструментария, как *надёжность* и *валидность*. При этом для обозначения различных аспектов и компонентов, как надёжности, так и валидности употреблялись в разные времена разнообразные наименования. И на самых ранних этапах становления ЦО–технологий, и в настоящее время для определения надёжности используются преимущественно меры согласованности оценок экспертов. Хотя в литературе можно обнаружить попытки оценить надёжность через гомогенность и попытки создания эквивалентных форм и примеры применения повторных измерений (Попов, Лурье, 2012). Что касается валидности, то на первых порах решались проблемы соответствия видов активности, проявляемой в испытательных упражнениях, тем задачам, которые характерны для ситуаций профессиональной деятельности. То есть тому, что стало именоваться *валидностью по содержанию*. Эта проблема хорошо осознавалась британскими и североамериканскими специалистами, в то время как германская модель строилась на теоретическом фундаменте характерологии, что позволяло игнорировать ситуационную специфичность той деятельности, для которой предназначался отбор. Практически то же самое можно сказать и в отношении задач прагматического характера, а именно, определения практической эффективности разработанных и применяемых в Великобритании и США подходов при проведении отбора кандидатов, то есть того, что охватывалось понятием *«критериальная валидность»*. Здесь специалисты центров оценки успешно следовали маршрутами, проложенными психометрической теорией и практикой использования психологического, образовательного и профессионального тестирования. Вплоть до середины 1950–х годов валидизация разрабатываемого диагностического инструментария ограничивалась установлением валидности по содержанию и критериальной валидности. Но все изменилось при введении в профессиональный дискурс, а затем и в выдвижении на передний план понятия *«конструктная валидность»*, то есть того, насколько тестовые баллы могут интерпретироваться как эмпирический показатель определенного теоретического конструкта (Furr & Bacharach, 2008). Довольно скоро конструктная валидность стала рассматриваться как основная в триаде: содержательная — критериальная — конструктная (Guion, 1980; Shepard, 1993). И с этого момента стали возникать проблемы в деле оценки валидности ЦО–технологий. Если до сей поры ориентация на теорию и методологию психологического тестирования и стандарты психометрики скорее облегчали, чем затрудняли проведение оценки качественности результатов ЦО, то повсеместное принятие тринитарной доктрины создало трудности, к преодолению которых практика ЦО оказалась не готова. Речь идет о так называемом *парадоксе конструктной валидности* ЦО, который заключается в том, что при обеспечении высокой содержательной и критериальной валидности психометрический инструмент должен обязательно обладать и высокой конструктной валидностью, но применительно к ЦО это не удалось показать в ходе специально проведенных исследований (Arthur, Woehr, Maldegen, 2000). Более того, произошедший на рубеже веков фактический отказ от тринитарной доктрины, то есть от признания относительной самостоятельности содержательной, критериальной и концептуальной валидностей в пользу унитарной модели, где декларируется принцип использования различных и разно-

родных источников получения сведений, говорящих в пользу единой и неделимой валидности, в ещё большей мере усложнил ситуацию. Была закрыта возможность локализовать и отдельно рассматривать неудобные вопросы, которые прежде относились исключительно к валидности конструктов центра оценки. А теперь обо всем по порядку.

О практике немецких военно-психологических центров

Если обратиться к процедурам, используемым в немецкой модели, то попыток полномасштабной оценки качества результатов деятельности военно-психологических центров отбора курсантов и аттестации офицеров по сути и не предпринималось. И тому, как минимум, есть две причины, одна из которых относится к разряду теоретико-идеологических, а вторая — носит преимущественно организационный характер.

Под теоретико-идеологической здесь понимается уже упоминавшаяся ориентация военных психологов в Германии на характерологию, в которой акцент делался на понимание целостного характера путем анализа некоторой совокупности разнообразных поведенческих и эмоциональных манифестаций характерологических особенностей, свойственных тому или иному индивиду. Качество выносимого специалистами вердикта определялось в первую очередь квалификацией и мастерством людей, проводящих оценочные процедуры, а не соблюдением установленного регламента и скрупулезным подсчетом баллов, набранных кандидатами в результате выполнения тех или иных упражнений и тестовых заданий.

Другой причиной отсутствия впечатляющих результатов при оценке качества работы германской военно-психологической службы является временно́й фактор. Если обратиться к источникам информации об эффективности деятельности британских прототипов центров оценки, то можно увидеть, что почти все они датируются годами, несколько отстоящими от времени окончания войны, когда появилась возможность провести такого рода работу. В частности, использовать те документы, к которым в период военных действий по вполне понятным причинам доступ был закрыт. У немецких специалистов для этого совсем не было ни времени, ни возможностей. Одни проходили процедуру денацификации, другие были заняты своим трудоустройством. У специалистов, относящихся к стану союзнической коалиции, хватало своих проблем, да вдобавок и мало кто из них владел немецким языком и знанием характерологии на таком уровне, чтобы суметь разобраться с теми материалами, которые уцелели после окончания боевых действий.

Эффективность деятельности военных отборочных комиссий Великобритании

Представляется, что наиболее ценный и внушительный по объему материал о работе военно-психологических служб в области определения качества ассессмента мы находим в Великобритании. Но прежде чем приступить к анализу результатов работы британских специалистов, необходимо сделать два важных замечания. Во-первых, надо сказать, что безотносительно к тому, чем закончились попытки, что называется, «по горячим следам» измерить качество методик для работы британских военных отборочных комиссий, они выполнили свою основную задачу — обеспечили бесперебойную подготовку командного состава для действующей армии в условиях, где традиционные процедуры отбора кандидатов в военные училища попросту не могли работать. Во-вторых, в то время, когда началась работа по определению эффективности технологий ассессмента, ещё не сложилась традиция

употребления тех или иных показателей надёжности, валидности, несмещённости данных и приемлемости тех или иных форм обратной связи при предъявлении результатов оценки. Неудивительно, что без должного обоснования применялись вперемежку как параметрические, так и непараметрические методы обработки данных, а использование процентов наблюдалось не только при решении задач дескриптивной, но и индуктивной статистики. Последнее во многом объясняется тем, что рассчитывать на понимание сущности принятых в психометрике показателей, таких как коэффициент линейной корреляции или регрессии, было бы нереалистичным не только в случае предъявления результатов широкой публике, но и при оформлении отчетов для руководителей военных и гражданских служб, которые являлись заказчиками исследований. Более того, значительная часть лиц, организующих и проводящих оценку, не имела опыта оперирования статистическими показателями. Имеются в виду прежде всего армейские психиатры, ориентированные на психоанализ. Неудивительно, что из текстов отчетов (если отдельно не проводить тщательный анализ приложений к основному тексту) не всегда ясно, какой метод анализа — корреляционный, факторный, регрессионный или дисперсионный — применялся в том или ином случае, когда говорилось о процентах «объяснения» вариативности целевой переменной на основании данных о вариации переменной-предиктора. Но несмотря на видимые с позиций сегодняшнего дня многочисленные недочеты и несуразицы, все эти незрелые и кажущиеся где-то наивными попытки дать оценку качества используемого инструментария заслуживают нашего внимания, так как именно они предопределили направления развития квалификационной работы и способствовали становлению современной системы исследований методического характера.

Первоначально оценка качества процедур ассессмента проводилась в Великобритании на базе центров подготовки и переподготовки офицерского состава всех видов вооруженных сил, и в качестве объектов оценки выступали не только кандидаты на поступления в военные училища, но и военнослужащие, имеющие опыт командования подразделениями в бою. Работа по валидизации в отборочной комиссии началась с исследования сопоставимости заключений военных психиатров, с оценками, сделанными преподавательским составом Центров подготовки, сформированным из опытных кадровых военных. Психиатры на основе результатов психологического тестирования и глубинного интервью давали краткую характеристику курсанту и выносили свое суждение о степени его пригодности для выполнения роли полевого командира. Независимо от них то же самое делали сотрудники Центра подготовки на основании своего опыта взаимодействия с теми же курсантами в процессе обучения. При сопоставлении двух рядов данных оценивалась степень их совпадения. В самой первой серии исследований были получены такие результаты: из 48 сопоставлений 26 (55%) были признаны в основном совпадающими, 12 (25%) частично сходными и в 10 случаях (20%) отмечались существенные различия. В следующих сериях исследования были получены существенно более высокие показатели совпадения оценок, вынесенных начальствующим составом центров обучения, и заключений военных психиатров — до 80% и даже до 90% (Murrey, 1990, p. 47–48). Прогресс в согласованности оказался связанным как с включением в процедуру испытаний когнитивных тестов, так и со специально проведенной работой по согласованию критериев оценки между преподавательским составом центров обучения и психиатрами. Полученные на данном этапе результаты вправе рассматривать в качестве свидетельств наличия не столько надёжности, сколько валидности по содержанию.

Для установления надёжности процедур оценки кандидатов была использована следующая схема: две независимые параллельно работающие комиссии одновременно наблюдали процесс и результаты прохождения тестирования одними и теми же кандида-

тами. Затем обе этих комиссии выносили свои вердикты, не согласовывая их между собой. В каждой комиссии были председатель, войсковой представитель в ранге офицера, психиатр и психолог. Сопоставлялись как независимые оценки всех участников, так и согласованные внутри каждой из комиссий интегральные оценки каждого кандидата. Средняя интеркорреляция всех оценок оказалась равной 0.60, между двумя председателями разных комиссий 0.65, между психиатрами тоже 0.65, между офицерами 0.86 и между двумя психологами 0.87. Коэффициент корреляции между интегральными, согласованными внутри каждой комиссии, оценками достигал величины 0.80 (Morris, 1949). Полученный результат был признан вполне удовлетворительным.

Несмотря на несомненную важность, которую представляет собой оценка надёжности, наибольшие усилия были направлены на установление критериальной валидности. Прямое сопоставление оценок кандидатов, сделанных военно-отборочной комиссией, с оценками успехов тех же индивидов, выставленных преподавателями военных училищ и тренинговых центров, приводило к величинам коэффициентов корреляций в диапазоне от 0.4 до 0.5 в большинстве исследований, проведенных в конце 1940–х годов. Хотя большинство валидизационных исследований проводилось на материале, собранном в различных тренинговых центрах, несколько изысканий подобного типа были осуществлены в действующей армии. В этом случае предметом сопоставления были оценки отборочной комиссии с оценками, сделанными командованием тех частей, в которые были направлены аттестованные офицеры. В одном из самых масштабных исследований такого рода, проведенном в 1945 г. на выборке из 500 офицеров действующей армии, была получена корреляция 0.35[1] (Vernon, Parry, 1949).

Оценка качества методического аппарата, используемого Управлением стратегических служб США

Информация об эффективности методик, применяемых для подбора агентов Управления стратегических служб, ограничена как тематически, так и по временному диапазону. Первое, по-видимому, связано с повышенной секретностью этой организации, а второе — с относительной кратковременностью её существования в первоначальном формате, так как в 1947 году Управление стратегических служб (УСС)[2] было преобразовано в Центральное разведывательное управление (ЦРУ). Тем не менее, сотрудники УСС успели провести масштабную работу по оценке качества как технологии в целом, так и её отдельных компонентов. Наиболее ценным источником информации об особенностях работы кадровых служб УСС является опубликованный в 1948 году официальный отчет о пяти годах деятельности этих подразделений. В более чем пятисотстраничном отчете, авторами которого являются все участники проекта и его консультанты, фигурируют числа, предназначение который, по-видимому, состояло в том, чтобы дать представление об общей эффективности работы станций оценки (OSS Assessment Staff, 1948). В частности, приводится такой показатель, как процент корректных селекций, величина которого устанавливается на уровне 63% (за 50% берется результат случайного отбора). Интересно, что в исследовании Д. Вигинса, проведенного через полтора десятка лет, фигурирует уже скорректированный на надёжность и

1 В самом отчете об исследовании фигурирует число 0.165, но после коррекции на смещённость выборки и поправкам, сделанным на основе оценок надёжности методов получилась результирующая величина 0.35.

2 Управление стратегических служб, УСС — от англ. Office of Strategic Services, OSS.

смещённость выборки показатель, численное значение которого после пересчета оказалось равным уже 77% (Wiggins, 1973, p. 536).

Спустя три года после запуска программ оценки были опубликованы некоторые данные о валидности ассессмента двух из четырех станций УСС. На станции S, где проводилась углубленная трехдневная программа оценка, коэффициенты критериальной валидности варьировали для разных групп, каждая из которых насчитывала несколько десятков человек, в диапазоне от 0.08 до 0.37 с медианным значением, равным 0.21. Станция W, в которой осуществлялась более короткая однодневная программа, неожиданно получила даже несколько лучшие результаты: коэффициенты валидности в разных группах, насчитывающих в каждой до сотни оцениваемых лиц, находились в интервале от 0.15 до 0.53 с медианой 0.26 (OSS Assessment Staff, 1948, p. 423).

Помимо определения эффективности общей оценки, проводилась работа по выявлению вкладов оценок по отдельным измеряемым параметрам или критериям. Это оказалось возможным в связи с тем, что в американской модели ЦО был изменен общий дизайн – произошел отказ от структурирования процесса оценки на основе функциональных задач и ролей (как это было задано в эталонной британской модели) в пользу системообразующих измеряемых параметров (*dimensions*). Такое новшество было введено в связи с тем, что для разработчиков был в то время во многом неизвестен состав тех задач, которые должны были решать агенты во время проведения тайных операций за рубежом. Перечень выделенных измеряемых параметров выглядел следующим образом: *мотивация, практический интеллект, эмоциональная стабильность, социабельность, лидерство, физические данные, внимательность, способности пропагандиста, скрытность.* Проведенный специальный анализ накопленных за несколько лет эмпирических данных показал, что из всего набора предварительно выделенных параметров самым лучшим предиктором успешности деятельности агентов оказался *практический интеллект*. Различия в интеллектуальности «дают объяснение» 10 процентам дисперсии показателей успешности. За интеллектуальностью сразу следует *лидерство* (3%), а вслед за лидерством идет *мотивация* (1%). Авторы считают, что эти переменные на самом деле вносят более весомый вклад в успешность деятельности оцениваемых лиц, но вместе с тем указывают, что для обоснования этого положения нет надежных свидетельств (OSS Assessment Staff, 1948).

Эффективность британской модели подбора в гражданскую службу

Функционирующая в настоящее время система отбора гражданских служащих в Великобритании была заложена в 1945 году и была первоначально скопирована с системы, разработанной для отборочных военных комиссий. Данные об её эффективности скудны, отрывисты и противоречивы. По-видимому, одним из первых, кто попытался установить степень продуктивности процедур, разработанных в армейской среде, для задач комплектования гражданской службы, был Ф. Вернон. В журнальной статье 1950 года Ф. Вернон приводит некоторые оценки качества методического аппарата, используемого при подборе служащих. Так, критериальная валидность оказалась плавающей в диапазоне между 0.44 и 0.49 (Vernon, 1950).

В опубликованной через полтора десятка лет статье, Э. Энсти, главный психолог отборочной комиссии для гражданских служб в 1960–х и 1970–х годах, прослеживает судьбу 301 гражданского служащего, принятого на работу 30 лет назад (это — те же люди, что и у Ф. Вернона). На основе проведенного анализа Э. Энсти утверждает, что процедуры отбора, использовавшиеся сразу после войны, показали высокую эффективность в целом. Из 301

принятого на работу лишь 21 не дослужились до ранга старшего служащего и только трое были уволены в связи с профессиональной непригодностью. Рассчитанный автором показатель прогностической валидности, где критерием выступали не оценки эффективности работы, как у Ф. Вернона, а степень продвижения по служебной лестнице, оказался равным 0.66 (Anstey, 1977). К сожалению, другие надежные источники информации об эффективности центра оценки для британских специальных служб труднодоступны.

Ещё меньше информации о валидности ЦО для *Стремительного потока* — системы подбора и карьерного продвижения выпускников вузов Великобритании в организациях и учреждениях гражданских служб. Достоверно известно, что исследование его валидности проводил И. Робертсон, но отчет, представленный им в 1999 году, так и не был опубликован. Тем не менее, ссылки на его работу можно встретить в обзорных статьях, использующих мета-анализ данных (использование результатов, взятых из неопубликованных диссертаций и научных отчетов часто практикуется при проведении мета-аналитических исследований). Так, мы можем узнать, что объем выборки в его исследовании — 105 человек, а коэффициент критериальной валидности оказался равным 0.23 (Robertson, 1999).

Дефицит информации о качественных характеристиках ситуационно-поведенческого обследования применительно к подбору госслужащих заставляет особенно внимательно проанализировать текст отчета, сделанного К. Флетчером, директором компании с ограниченной ответственностью «Оценка персонала»[3] по заказу Комитета гражданской службы Великобритании. Перед К. Флетчером была поставлена задача выделения наиболее приемлемых, надежных и валидных методов и методик, которые могли быть использованы при подборе высшего персонала для частного, государственного и некоммерческого секторов, фокусируя внимание на таких методах как интервьюирование, использование ассессмент центров, а также методиках психологического и психометрического тестирования. В качестве показателей эффективности рассматриваются такие, как валидность, справедливость, впечатления кандидатов и стоимость. Окончательный ранжированный список Флетчера выглядит следующим образом: *центры оценки; структурированные интервью; рабочие задания; личностные опросники; биографические сведения; рекомендации; когнитивные тесты; неструктурированные интервью* (Fletcher, 2005). Центр оценки в нем занимает первое место, несмотря на то, что его стоимость самая высокая, а по критериальной валидности этот метод, на основе тех данных, которыми располагал в то время К. Флетчер, не превосходил такие методы, как структурированное интервью и когнитивные тесты. Все дело в том, что для системы отбора в гражданскую службу Великобритании самой важной характеристикой является не валидность, а *справедливость*, то есть несмещаемость оценок в зависимости от пола, возраста и расовой принадлежности. И, кроме того, *релевантность* ситуационно-поведенческих упражнений реалиям служебной деятельности не вызывает у кандидатов сомнений, чего нельзя сказать, например, о тестах когнитивных способностей.

Об исследованиях эффективности центров оценки для коммерческих организаций

Первые систематические исследования эффективности технологий ассессмента для бизнеса были проведены при активном участии руководителя проекта Исследование управленческого прогресса телекоммуникационной компании AT&T и создателя современного

3 Personnel Assessment Ltd is a company set up by Professor Clive Fletcher, one of the UK's best-known occupational psychologists. URL: http://www.personnel-assessment.com/frameset.html

формата центра оценки, Д. Брея. Если в текстах, предназначенных для потенциальных заказчиков и широкой публики, их авторы оперировали абсолютными числами и процентами, то при публикациях в научных изданиях использовался стандартный аппарат математической статистики. В ряде статей, появившихся на свет в 1960–х годах, Д. Брей с соавторами приводят конкретные числовые показатели критериальной валидности технологий ЦО. В одном из этих исследований в качестве внешнего критерия был взят такой показатель, как прирост заработной платы. Рассчитанные для четырех выборок коэффициенты корреляции расположились в диапазоне от 0.41 до 0.52 (Bray, Grant, 1966). В другом исследовании общий рейтинг 78 человек, полученный в результате прохождения ассессмент-центра (*overall assessment rating, OAR*) коррелировал с производственными показателями со значением коэффициента корреляции, равном 0.51 (Bray, Campbell, 1968). В одном из первых и часто цитируемом мета-аналитическом исследовании, проведённом уже в 1980–х годах, его авторам удалось получить статистически значимые показатели критериальной валидности ассессмент-центра, со средним значением коэффициента корреляции, равным 0.37 (Gaugler, Rosenthal, Thornton, Bentson, 1987).

Но с годами эта картина начала мало-помалу терять свою очевидную убедительность. Ряд крупномасштабных мета-аналитических исследований, проведенных уже в XXI веке, со всей беспощадностью выявил тенденцию снижения показателей критериальной валидности центров оценки. Типичными выглядят уже такие значения коэффициентов корреляции как 0.26 и 0.28 (Thornton, Gibbons, 2009, p. 172). Усилия по выявлению причин наблюдаемого тренда не привели к однозначному результату. Но большинство специалистов приходят к выводу, что все дело в том, что в связи с высокой затратностью ЦО–технологий они применяются в основном на заключительных этапах многоступенчатой процедуры отбора среди множества претендентов на вакантные должности (Куприянов, 2011, с. 55). В качестве примера можно сослаться на упомянутую выше модель ассессмента для *Стремительного потока*, в котором ключевые технологии ЦО — ситуационно-поведенческие упражнения включались только на четвертой, заключительной стадии ассессмента, к началу которой претенденты с предположительно меньшей выраженностью требуемых компетенций были отсеяны на предыдущих этапах. В ситуации искусственного сжатия разброса измеряемых показателей стандартные статистические приемы расчета эффективности дают заниженную величину «истинной» взаимосвязи. Несмотря на весьма заметное снижение градуса восторженности при обсуждении проблем критериальной валидности ЦО, целесообразность использования этой технологии пока что не ставится под сомнение. Расчеты показывают, что применение ЦО оправдывает себя даже при тех не весьма высоких показателях критериальной валидности, которые имеют современные ассессмент-центры (Куприянов, 2011, с. 50–51).

Но если проблемы с критериальной валидностью на какое-то время можно считать хотя бы частично проясненными и, во всяком случае, не фатальными для современной практики использования центров оценки, то не так обстоит дело с тем, что называлось и продолжает по старинке называться валидностью конструктной. Если исходить из того, что в серии различных упражнений центров оценки одни и те же измеряемые параметры (одни и те же компетенции или диспозиции одного и того же индивида) будут оцениваться сходным образом, а разные компетентности или диспозиции одного и того же индивида в одном и том же упражнении будут получать не тождественные оценки, то анализ финальной матрицы результатов должен был бы показать высокие корреляционные связи между оценками одних и тех же компетенций или диспозиций в различных упражнениях. В свою очередь, оценки разных компетенций или диспозиций, проявленных в одном и том же упражнении, должны

обнаружить более слабые корреляции. Однако, похоже на то, что всё происходит с точностью до наоборот. Более чем тридцатилетняя история изучения этого вопроса приводит к неутешительному выводу: в большинстве проведённых исследований были выявлены высокие корреляции между оценками разных компетенций и диспозиций в одних и тех же упражнениях, а также низкие корреляции между оценками одних и тех же компетенций и диспозиций в разных упражнениях. Это приводит большинство специалистов к выводу об отсутствии конвергентной и дискриминантной валидности центров оценки, а, следовательно, и валидности конструктной (Thornton, Gibbons, 2009, p. 173). Данное обстоятельство ставит под сомнение претензии на строгость и научную обоснованность применяемого в ЦО методического аппарата.

Можно, конечно же, продолжать, как это уже и происходило в течение многих лет, не обращать особого внимания на этот омрачающий разум факт. В конце концов, для заказчиков центра оценки на первом месте стоит прагматика (что обеспечивается критериальной валидностью), для клиентов (оцениваемых лиц) важно воспринимать соответствие испытаний реалиями деловой жизни (а это проходит по линии содержательной валидности). А что касается причудливости взаимосвязи различных диспозиций, критериев, компетенций и прочих переменных между собой, то пусть об этом волнуются специалисты по дизайну центра оценки. И, похоже, что именно так всё и происходило вплоть до сегодняшнего дня, а регулярно вспыхивающие и тут же гаснущие дискуссии по поводу неопределённости этой самой конструктной валидности происходили для поддержания сложившегося в околопрофессиональных кругах мнения. Это мнение заключалось в том, что центр оценки — это не только наукоемкая, но и научно обоснованная технология, а специалисты, её реализующие, готовы во всеоружии разбираться с проблемами методологического свойства при соблюдении процедур и правил, прописанных во всякого рода релевантных стандартах и ответственных рекомендациях[4].

Некоторые шаги к прояснению того положения дел, которое сложилось в области оценки качества ЦО, можно осуществить обратившись к истории появления самого понятия «конструктная валидность». В середине 1950–х гг. у разработчиков первых стандартов психологического тестирования конструктная валидность играла роль «бедной родственницы», которая в случае необходимости могла быть использована для замены более строгих критериев. «Конструктная валидность обычно применяется, когда у лица, проводящего тестирование, отсутствуют более надежные способы оценки тех черт и свойств, которые интересуют исследователя, и в этом случае он использует косвенные измерения для проверки своей теории» (American Psychological Association, 1954, p. 14). Но прошло некоторое время, и вот уже А. Анастази определяет конструктную валидность как суперординантную категорию, включающую в себя и содержательную, и критериальную валидности (Anastasi, 1986). А ближе к концу прошлого века Л. Шепард в авторитетном руководстве по тестированию пишет: «сегодня мы утверждаем, что все тестовые интерпретации включают в себя конструктную валидизацию» (Shepard, 1993, p. 417). И если невзрачную Золушку можно было не замечать, то это никак нельзя делать по отношению к королеве бала.

А затем подошло время, когда тринитарная доктрина, доминирующая десятки лет в области теории тестирования и исправно игравшая роль щита или ширмы, позволяя затягивать решения назревающих, а то и перезревших проблем валидности ЦО–технологий,

4 Помимо следования собственным национальным и международным стандартам для центров оценки специалисты, работающие в данном направлении, считают необходимым соотноситься со стандартами, разработанными в сопредельных областях, например, The Principles for the Validation and Use of Personnel Selection Procedures и The Standards for Educational and Psychological Testing.

перестала выполнять эти функции. К настоящему времени проявилось несколько обстоятельств, делающих невозможным дальнейшее затягивание в решении проблем валидности центров оценки. Назовём наиболее существенные из них.

В наши дни уже представляется неуместным использование стремительно устаревающего понятийного аппарата для описания и структурирования проблематики валидности ЦО, то есть того аппарата, который стал достоянием истории прошлого века. Всё идёт к тому, что использование понятий «содержательная», «критериальная» и «конструктная валидность» довольно скоро будет считаться признаком дурного тона и говорить об обскурантизме тех лиц, с уст которых слетают соответствующие слова. Разумеется, речь идёт не о «высокой моде», и не об очередной смене риторических вех, а о необходимом усовершенствовании языковых средств. Усовершенствовании, которое бы отражало произошедшие изменения в сущностном понимании указанной проблематики. В Стандартах образовательного и психологического тестирования[5] образца 1999 года из текста исчезают наименования отдельных видов валидности, остается единое и неделимое понятие «валидность теста». Те реалии, что ранее выступали в качестве отдельных целостных составных частей этого единства, объявляются не более чем разнообразными источниками сведений, аргументов, свидетельств и доказательств валидности проведенного тестирования. В качестве таких источников назывались такие, как *содержание* теста, *процесс тестирования*, *внутренняя структура* теста, связь с *внешними* переменными, *последствия* тестирования (Standards for Educational and Psychological Testing, 1999).

Ещё одна примета назревающего кризиса — возобновилась критика положения о необходимости держаться *измеряемых параметров (dimensions)*, в качестве основного ключевого компонента внутренней структуры технологии, её несущей конструкции, её архитектоники, сердцевины самого дизайна ЦО, в конце концов, её «фирменной матрицы»: *измеряемые параметры × моделирующие упражнения*. Снова становятся вполне уместными рассуждения о роли личности и ситуации в истории дисциплин, изучающих детерминанты поведения человека (Росс, Нисбетт, 1999). Вновь делаются попытки радикального отказа от личностных диспозиций и компетенций в качестве основного концепта методологии ЦО в пользу ключевых задач и профессиональных ролей (Lance, 2008), то есть переменных преимущественно ситуационного характера. Тем более что именно ситуационный принцип структурирования деятельности ЦО был положен в основу дизайна отборочных комиссий британских вооруженных сил, и успешность этого подхода доказана историей. Но в той же истории можно обнаружить и серьезные резоны для отказа от данного подхода, точнее говоря, выявлению условия, при которых этот подход становится малопродуктивным. Как известно, оперирование понятиями «функциональная задача» и «роль» может существенно затруднить работу по определению детерминант успешности выполнения заданий в новых неисследованных ситуациях, что произошло в конце 1940–х годов на станции S Управления стратегических служб США при создании программы оценки готовности кандидатов к выполнению тайных операций за рубежом (OSS Assessment Staff, 1948). Разумеется, данный случай не исключает возможностей плодотворного использования функционально-ролевого подхода в другие времена и при иных обстоятельствах. Тем более что идут активные поиски путей синтеза технологий, ориентированных на компетенции и ориентированных на задачи (Melchers, Wirz, Kleinmann, 2012). Независимо от успехов и неудач в деле интеграции различных подходов, то есть от того, будут ли они развиваться, взаимодействуя друг с другом, или идти раздельными

5 Нравится нам это или вызывает активное несогласие, но именно этот стандарт задаёт если не идеологию и методологию психологической диагностики, то уж как минимум лексику дискурса по поводу валидности психодиагностики в целом и валидности центров оценки в частности.

путями, можно не сомневаться, что процедуры, используемые для валидизации инструментария ЦО, с необходимостью будет подвергаться ревизии и трансформациям.

Озабоченность по поводу складывающейся ситуации наблюдается и в стане «умеренных консерваторов» – сторонников опоры на ставшие уже привычными *измеряемые параметры* и ряд сопряженных с ними понятий, таких как «аспект», «диспозиция», «компетенция», «критерий», «образец» и др. Эта озабоченность проявляется в том, что всё чаще и громче звучат призывы к масштабному развёртыванию работы по совершенствованию технологии во всех аспектах и направлениях. А именно: лучшая тренировка наблюдателей и увеличение их численности, уменьшение количества и укрупнение измеряемых параметров, более вдумчивый подбор и тщательная проработка упражнений, использование новейших разработок в области статистики. И, что немаловажно, ужесточение стандартизации, включая постоянное совершенствование процедур, регламентов, инструкций, правил и алгоритмов (Hoffman et al., 2011; Ерофеев, 2013). Упование на стандартизацию как на панацею от существующих несовершенств присуще в первую очередь представителям академических кругов, хотя не обходит стороной и некоторых профессионалов, казалось бы, имеющих значительный опыт кадрового консультирования организаций разного типа. Типичным примером демонстрации такого рода позиции является публикация трёх авторов, представляющих три университета южной части США: Техаса, Теннесси и Оклахомы (Arthur, Day, Woehr, 2008). Основной посыл произведения, сотворённого этим трио — полное и последовательное внедрение и неукоснительное следование правилам, процедурам и стандартам, созданным в рамках психометрического тестирования, для безукоризненного решения задач, стоящих перед центрами оценки. Прежде всего, имеется в виду многоступенчатая разработка и предварительная валидизация всех инструментов, задействованных в технологических схемах центров оценки. Кроме того, подчёркивается необходимость тщательной и вдумчивой проработке формулировок при определении содержания конструктов (измеряемых параметров). Авторы сетуют на то, что в подавляющем большинстве известных им проведённых проектов используется набор конструктов по принципу *ad hoc* — только для данного случая — что затрудняет сопоставление результатов работы разных ЦО и даже различных проектов, выполненных в рамках одного центра. Так, перечень измеряемых параметров в исследованных ими 34 центрах оценки насчитывал 168 наименований, а в другом случае — в 48 центрах оценки таковых оказалось не намного меньше — 129 (Arthur, Day, Woehr, 2008, p. 107). Хуже того, даже если названия измеряемых параметров звучат идентично, это ещё не означает, что их содержания являются эквивалентными. Все это вместе взятое не позволяет полномасштабно использовать такое эффективное средство обобщения информации, получаемой при валидизации центров оценки, каким является мета-анализ данных.

Сетования на избыточность и неупорядоченность наборов компетенций и иных видов измеряемых параметров, предназначенных для оценки, осмыслены и понятны, но улучшить положение дел за счет ужесточения норм представляется нереалистичными. Поскольку и существующая практика, и её отражение в нормативных документах основываются на том, что при проведении ЦО абсолютно необходимым считается учет специфики ситуации, в которой находится организация заказчика оценки, а также характера задач, которые стоят перед ней. Это, в частности, выражается в том, что набор измеряемых параметров (в частности, компетенций) определяется той моделью профессиональной деятельности, которая существует в организации, а если её нет, то предписывается проведение специальных видов работ, для определения перечня оцениваемых параметров исходя из специфики организации (Российский стандарт..., 2014, с. 5–6). Даже те специалисты, которые ратуют за ужесточение норм проведения оценки и являются явными сторонниками более высокой

регламентации деятельности ЦО, отмечают необходимость учитывать специфику организации заказчика. Так, в самом начале статьи А. К. Ерофеева и Т. Ю. Базарова, посвящённой обсуждению способов построения моделей компетентностей, мы встречаем заявление о том, что «авторы убеждены в необходимости в каждом конкретном случае создавать именно ту модель компетенций/компетентностей, которая релевантна условиям конкретной организации» (Ерофеев, Базаров, 2014, с. 74). Работа, ориентированная на различающиеся наборы компетенций означает в частности постоянную готовность к тому, что для каждого нового ЦО нужно будет подбирать или даже изобретать состав, содержание и форму методических документов, включая стимульный материал, инструкции, бланки и т.д. и т.п. А это говорит о том, что дизайнерам ЦО никогда не достичь идеала, а именно: предложить методический пакет, не нуждающийся в дальнейшем усовершенствовании. То ли дело пятна Роршаха или прогрессивные матрицы Равена! Пятна со временем если и выцветают, то не расползаются, а матрицы, если и прогрессируют, то никому это не заметно.

Стремление к бóльшей стандартизации очевидно родилось не как пустопорожняя затея, но все мы знаем, куда иной раз ведут дороги, устланные благими намерениями. Многое говорит о том, что проявление осторожности в данном случае будет не лишним. И здесь самым удивительным является то, что даже профессионалы время от времени упускают из виду тот непреложный факт, что центр оценки не является ни тестом, ни тестовой батареей, хотя и может включать в себя некоторый набор тестов, но в только в качестве отдельных и вовсе не обязательных компонентов. Его основой изначально были поведенческие (в Германии), а в дальнейшем ситуационно-поведенческие (в Великобритании и США) упражнения-испытания, как правило, дополняемые в той или иной степени структурированным или фокусированным интервью. Ни эти упражнения, ни фрагменты интервью не являются аналогами пунктов личностного или мотивационного опросника или заданий теста способностей или достижений. Измерять качество ЦО общим психометрическим аршином необходимо с величайшей осторожностью. И если в деле оценки надёжности понимание специфики ЦО технологии встречается не столь уж редко (мало кому сейчас придет в голову измерять надёжность ЦО с помощью коэффициента «альфа Кронбаха»), то для определения валидности используется без явных опасений весь «джентльменский набор инструментов», наработанный современной психометрической мыслью. Утешает то, что отдельные просветления, тем не менее, наблюдаются. В частности, наступает понимание того, что матрица *измеряемые параметры × моделирующие упражнения* ЦО не является аналогом МТММ–матрицы[6] и подход Кемпбелла–Фиске к оценке конструктной валидности тестов на основе сопоставлений показателей конвергентной и дискриминантной валидности, неплохо зарекомендовавший себя при оценке качества образовательных и психологических тестов (Campbell, Fiske, 1959), весьма ограниченно пригоден для апробации центров оценки. В настоящее время для этих целей вновь стали использоваться *номологические сети* Кронбаха–Мила (Cronbach, Meehl, 1955), предложенные за несколько лет до появления МТММ–подхода, но временно преданные забвению в связи с большей сложностью. Справедливости ради надо отметить, что, несмотря на явную привлекательность подхода Кронбаха–Милла, связанную прежде всего с его гибкостью, каких-то впечатляющих успехов его реализация пока что не принесла. Большие надежды возлагаются на возможности разработанной Л. Кронбахом с сотрудниками расширенной статистической теории тестирования (G–теорией). На её основе оказалось возможным проводить эффективный сравнительный анализ влиятельности нескольких классов переменных в их сочетаниях между собой на *общий оценочный рейтинг*

6 МТММ – таблица с двумя входами, где строки занимают черты (*traits*), а в столбцах представлены методы (*methods*).

(OAR) (Cronbach, Rajaratnam, Gleser, 1963), а не рассматривать вариации, связанные с различными ситуациями проведения испытаний, всего лишь в качестве источников ошибок тестирования. Возможности данного подхода, проявившиеся при его использовании для оценки качества образовательных тестов (Протасова, Толстобров, Коржик, 2014), всё чаще находят свое применение и при установлении валидности ЦО–технологий (Cahoon, Bowler, Bowler, 2012; Jackson et al., 2016).

Общее обсуждение и выводы

Широкомасштабные исследования, имеющие отношение к проблеме валидности ЦО, уже сейчас дали ощутимые результаты в плане совершенствования технологий, углубления понимания тонкостей взаимосвязи ситуационных и личностных переменных, определяющих линию и стилевые характеристики поведения, и много ещё чего-то важного и нужного. Но, все эти приобретения выглядят как приобретения, бенефициариями которых предстают представители академического сообщества. Как-то в пылу полемики ушла на второй план весьма и весьма важная задача, ради которой специалисты ЦО ввязались в затеи несколько отвлеченного характера в виде вопроса о трансситуативности личностных диспозиций. А именно: как за счет повышения конструктной валидности добиться повышения валидности критериальной, или, в более современной трактовке, получить многообразные свидетельства всесторонней эффективности (унитарной валидности) центров оценки. За спорами о том, что лучше «объясняет» дисперсию общего оценочного рейтинга (OAR), забыли о том, что всё-таки более важным является вопрос о том, насколько этот OAR определяет производственные показатели и прогнозирует успешность карьеры. Обнадёживает то, что в последнее время возобновились исследования, в которых одновременно рассматривается динамика показателей, традиционно относящихся к конструктной и критериальной валидности (Merkulova et al., 2016). В более современной терминологии речь идёт об источниках сведений о валидности, получаемых из анализа информации о характеристиках *процесса* тестирования, *внутренней структуры* теста, связи с *внешними* переменными, и *последствия* тестирования.

Усилия многочисленных специалистов, озабоченных повышением качества психодиагностического инструментария и следующих по пути совершенствования различных сторон ЦО-технологий, можно только приветствовать. Важно лишь не поддаваться искушению ходить торными дорогами, точнее, дорогами, проторёнными в других направлениях психодиагностики, которые возникли раньше и успели накопить значительный опыт работы в этих направлениях. Опыт — палка с двумя концами, вдвойне опасными, если это опыт чужой. Восприятие технологии ЦО как одной из разновидностей психометрического тестирования глубоко укоренено в сознании специалистов, работающих в сфере личностной диагностики. Так, Ф. Ливенс и Дж. Конвей без тени сомнений утверждают, что «центры оценки изначально задумывались как метод измерения устойчивых свойств индивидов. И именно эти свойства, обозначаемые как измеряемые параметры, рассматривались в качестве составных частей центров оценки» (Lievens, Conway, 2001, p. 1202). Схожие заявления можно найти в текстах, созданных командой исследователей из Цюрихского университета: «Центры оценки обычно предназначены для измерения кросс-ситуационных стабильных свойств индивидов (в том числе аналитических или межличностных умений) которые рассматриваются в качестве измеряемых параметров в различных тестовых ситуациях или упражнениях» (Merkulova et al., 2012, p. 2). Знакомство с историей возникновения и становления центров оценки никак не позволяет согласиться с тем, что они «изначально задумывались как метод измерения

устойчивых свойств индивидов». И в момент зарождения, и в настоящее время, предназначение ЦО состояло и состоит вовсе не в том, чтобы достоверно, точно, валидно и без заметного ущерба для здоровья испытуемых и испытателей измерять интересующие нас (или кого-то иного) свойства определенных людей. Оно состояло и состоит в том, чтобы успешно справляться с социальными и жизненно важными проблемами, как то: снижение отсева из учебных заведений, прогнозирование академической успеваемости абитуриентов и успешности адаптации к рабочему месту выпускников вузов, оценка перспектив профессионального, социального и служебного роста, снижение аварийности и травматизма, предупреждения явлений профессионального выгорания или срыва адаптации и т.д. И технология ЦО складывалась для решения именно данных задач, не имея в качестве основной цели вспомоществование научным учреждениям в деле порождения теоретического знания. Иначе говоря, проведение центров оценки есть разновидность профессиональной консультативной, а не исследовательской практики, поэтому технология ЦО «настроена» на решение актуальных задач, а не на преумножение вечных истин, или, в стиле риторики Ф. Бэкона, она есть средство получения не столько чего-то светоносного, но преимущественно приобретения чего-то плодоносного.

Все вышесказанное имеет самое непосредственное отношение к проблеме валидности ЦО. Можно и дальше следовать по пути, проложенном «старшими братьями», в качестве которых в данном случае выступают специалисты по теории и практике психологического и образовательного тестирования, соглашаясь тем самым, что технологии центров оценки в своих сущностных чертах являются не более чем разновидностью психометрического тестирования. И в этом случае — ориентироваться на стандарты, созданные иными людьми для решения иных задач. Но есть возможность посмотреть на проблему валидности исходя из того, что ЦО представляет собой уникальную технологию, являющуюся конгломератом или сплавом разнообразных методов, важнейшими из которых являются три: моделирующие упражнения, интервью и когнитивные тесты. Первые два не только не являются тестами, но и не имеют перспективу ими стать. И иметь в виду, что понятие «валидность» относится не столько к инструментам, которые используются в процедурах ассессмента, сколько к самой этой технологии, называемой центром оценки. Вот из этого и стоит исходить, и посмотреть на ситуацию свежим взглядом, и вернуться к изначальному пониманию валидности как к некой измеряемой или иным образом квантифицируемой и квалифицируемой способности теста или процедуры делать то, для чего этот тест или процедура предназначались. Мало того, необходимо сообразовываться со всем тем, что традиционно не считается входящим в понятие «валидность», но играет важную роль в определении эффективности диагностических технологий, одной из которых является технология ЦО.

Британские военные психиатры создали систему оценки, о валидности которой в то время приходилось только гадать, и в то же самое время этой системе нельзя было отказать в *действенности* — её применение позволило решить неразрешимую традиционными средствами задачу. Использование её модифицированного варианта в гражданской службе Великобритании показало, что валидность иной раз является не только не единственным, но и не самым важным компонентом, определяющим её *востребованность* — здесь более существенной оказалась *справедливость*. Применение моделирующих заданий при отборе кандидатов в Управление стратегической службы США выявило необходимость учитывать такое свойство оценочных процедур, как их *приемлемость* для оцениваемых лиц. В центрах оценки для производственных и коммерческих организаций существеннейшую роль играла и продолжает играть такая характеристика оценочных процедур, как их воспринимаемая *релевантность*, то есть убежденность в том, что диагностируемое в центрах оценки

имеет отношение к тому, что необходимо для успешного осуществления профессиональной деятельности и способствует служебной карьере.

Если возвратиться к собственно проблематике валидности, то пора уже подумать над тем, а не стоит ли отказаться от безоглядного следования извивам и изгибам психометрической мысли, и в первую очередь там, где обсуждаются вопросы валидности. Необходимо ли принимать те новшества, которые возникли в последнее время в лоне психометрики, или следует искать свои пути? Здесь имеются в виду последние вариации на тему унитарной валидности, которая возникает уже не в качестве места слияния потоков содержательной, конструктной и критериальной валидностей, а выстраивается с опорой на информацию, получаемую из таких источников, как наблюдение за процессом тестирования, анализ внутренней структуры теста, прослеживание связей с внешними переменными и учёт последствий тестирования. При этом следует принимать во внимание то, что данная новация, появившаяся в стандартах тестирования в 1999 году, уже подвергается ревизии (Standards for Educational and Psychological Testing, 2014) и данный перечень источников и составных частей теории тестовой валидности нельзя рассматривать как окончательный и не подлежащий обжалованию. Возможно, стоит присмотреться к тому, что происходит на смежных территориях психологической практики, когда она сталкивается с проблемами установления эффективности своих инструментов и процедур.

В первую очередь это разработки, сделанные в последнее время в таком направлении практико-ориентированных исследований, каким выступает исследование действием (Жуков, 2015). Здесь можно найти сторонников умеренно-консервативного подхода, которые предлагают разделить исследования действием на канонические и все прочие. При этом под каноническими понимаются исследования, следующие принципам традиционно понимаемой научности и проводимые по определённым правилам. Именно к ним, с некоторыми оговорками, приложимы критерии, разработанные в лоне позитивистской методологии (Davison, Martinsons, Kock, 2004). А что касается неканонических исследований, то они требуют какого-то не совсем ясного, но, безусловно, иного подхода.

Существенно более радикальная позиция заключается в том, чтобы выстроить другую, принципиально отличную от позитивистской, систему критериев оценки эффективности исследований действием, учитывающих интересы всех участников того или иного исследовательского проекта, таких как заказчики, клиенты, сами исследователи и профессиональное сообщество, к которому принадлежат последние. Небезынтересными в этом отношении представляются предложения, которые можно обнаружить в работе Г. Андерсона и К. Герр, опубликованной на рубеже веков и посвящённой разработке новой парадигмы проведения валидизации для методов исследований действием. Статья содержит описание пяти видов валидности исследований действием, которые обозначаются как *результативная, процессуальная, демократическая, каталитическая и диалогическая* валидности. Валидность результата определяется степенью, в которой реализованный проект достигает поставленных целей. В процессуальной валидности отражается уместность избранного подхода к исследуемой проблеме. Понятием «демократическая валидность» описывается уровень вовлечённости в исследование всех заинтересованных лиц. Каталитическая валидность устанавливается степенью, в которой исследовательский процесс преобразует участников, углубляет их понимание проблем и побуждает предпринять активные действия. Диалогическая валидность позволяет участникам исследования действием вступать в критический и рефлексивный диалог со своими коллегами (Anderson, Herr, 1999, p. 16).

Другим источником, где имеет смысл черпать информацию для размышления над проблемами эффективности, является то направление исследований, которые именуются

качественными. Здесь также можно наблюдать довольно-таки пёструю картину. Встречаются как последовательные приверженцы позитивистской традиции, так и радикально настроенные ниспровергатели установленных канонов. К примеру, в статье Н. Пандита при оценке инструментария для качественных исследований используются весьма традиционная терминология, и в ней можно наряду с надёжностью встретить внешнюю, внутреннюю и конструктную валидность (Pandit, 1996). В то же время в публикации К. Хенвуд и Н. Пиджена если и встречается существительное «валидность», то в непривычном сочетании с прилагательным «респондентная», а при описании других свойств методик используются такие понятия, как «прозрачность», «убедительность», «близость к данным», «интегрированность», «рефлексивность» (Henwood, Pidgeon, 2003). Принципиально иную новацию можно обнаружить в работе О. Мельниковой и Д. Хорошилова. Эти авторы предлагают проводить валидизацию, ориентируясь на последовательно развёртываемые этапы качественного исследования, такие как планирование, сбор и анализ данных, их интерпретация и презентация результатов. Кроме того, отдельно рассматривается вопрос об *этической* валидности. Все это предваряется анализом основных подходов к проблемам валидности качественных исследований и заканчивается выводом, что речь должна идти не просто о введении специфичных для качественных исследований представлений о валидности, но о концептуальном переформатировании проблематики на основе разведения и критического переосмысливания способов позиционирования участников исследования (Мельникова, Хорошилов, 2014).

Научный канон прекрасно смотрится в сказке, в которой учёные удовлетворяют свое личное неуёмное любопытство за счет налогоплательщиков. В пространствах этой сказки можно помечтать об остроумно задуманных и блестяще проведённых исследованиях, соответствующих лучшим мировым стандартам, что обеспечивает их объективность и беспристрастность. А ответственность при этом исследователь несёт только перед Её Величеством Наукой и перед своей совестью. Реалии наукоемкой практики иные, и при планировании, реализации и даже при представлении результатов уже завершённых проектов необходимо действовать таким образом, чтобы сделанная совместными усилиями работа отвечала запросам всех её участников, будь то клиенты, заказчики, посредники, подрядчики и субподрядчики, исполнители и соисполнители, испытуемые и испытатели, то есть всех тех, кто находится в системе ответственной взаимозависимости. Соответственно, и оценка эффективности должна проводиться под углами зрения всех только что названных и ещё не поименованных групп людей, которых в последнее время принято называть стейкхолдерами. Взгляд на более чем полувековую историю усилий, предпринятых для оценки качества работы ЦО, позволяет обнаружить возможности для выхода за ставшими чрезмерно узкими рамки надёжности и валидности, сколь бы важными они не представлялись. И выразить надежду, что в обозримом будущем удастся выстроить такие системы для оценки эффективности ЦО, в которых найдут свое место среди иных и такие понятия, как *«действенность», «востребованность», «справедливость», «приемлемость», «релевантность»*.

Литература

Барышникова, Е. И. (2013). *Оценка персонала методом ассессмент-центра*. М.: Манн, Иванов и Фербер.

Ерофеев, А. К. (2013). Центр оценки. Особенности метода и принципы стандартизации программ оценивания. *Организационная психология, 3*(4), 18–42.

Ерофеев, А. К., Базаров, Т. Ю. (2014). Авторские технологии разработки моделей компетенций. *Организационная психология, 4*(4), 74–92.

Жуков, Ю. М. (2015). Научная теория и наукоемкая практика. *Организационная психология, 5*(4), 29–38.

Куприянов, Е. А. (2011). Стоит ли игра свеч: валидность Центров оценки. *Организационная психология, 1*(1), 50–58.

Мельникова, О. Т., Хорошилов, Д. А. (2014). Современные критериальные системы валидности качественных исследований в психологии. *Национальный психологический журнал, 2*, 36–48.

Попов, А. Ю., Лурье, Е. В. (2012). Те же люди, другое время: валидность и надёжность Центров оценки, динамика развития оцененных компетенций. *Организационная психология, 2*(4), 43–58.

Протасова, И. В., Толстобров, А. П., Коржик, И. А. (2014). Методика анализа и повышения качества тестов в системе электронного обучения Moodle. *Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологи, 3*, 61–72.

Российский стандарт центра оценки (2013). *Организационная психология, 3*(2), 8–32.

Росс, Л., Нисбетт, Р. (1999). *Человек и ситуация. Перспективы социальной психологии*. М.: Аспект Пресс.

American Psychological Association (1954). Technical recommendations for psychological tests and diagnostic techniques. *Washington DC US. Psychological Bulletin, 51*(2), 1–38.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1–15.

Anderson, G. L., Herr, K. (1999). The new paradigm wars: is there room for rigorous practitioner knowledge in schools and universities? *Educational researcher, 28*(5), 12–40.

Anstey, E. (1977). A 30-year follow-up of the CSSB procedure, with lessons for the future. *Journal of Occupational Psychology, 50*, 149–159.

Arthur, W., Day, E. A., Woehr, D. J. (2008). Mend it, don't end it: an alternate view of assessment center construct-related validity evidence. *Industrial and Organizational Psychology, 1*(1), 105–111.

Arthur, W., Woehr, D. J., Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: a conceptual and empirical reexamination of the assessment center construct-related validity paradox. *Journal of Management, 26*(4), 813–835.

Bray, D. W., & Campbell, R. J. (1968). Selection of salesmen by means of an assessment center. *Journal of Applied Psychology, 52,* 36–41.

Bray, D. W., & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs: General and Applied, 80*(17), 1–27.

Cahoon, M. V., Bowler, M. C., Bowler, J. L. (2012). A reevaluation of assessment center construct-related validity. *International Journal of Business and Management, 7*(9) 3–19.

Campbell, D. T., Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin, 56*(2), 81–106.

Cronbach, L. J., Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin, 52*(4), 281–302.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology, 16*(2), 137–163.

Davison, R., Martinsons, M. G., Kock, N. (2004). Principles of canonical action research. *Information systems journal, 14*(1), 65–86.

Fletcher, C. (2005). Final report on a research study relating to effective selection of staff for senior posts in the civil service. *Queen's Printer and Controller of HMSO*, 3–38.

Furr, R. M., Bacharach, V. R. (2008). *Psychometrics: an introduction*. Thousand Oaks, Sage.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511.

Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology, 11*(3), 385–398.

Henwood, K., Pidgeon, N. (2003). Grounded theory in psychological research. In P. M. Camic, J. E. Rhodes, & L. Yardley (eds.). *Qualitative research in psychology: expanding perspectives in methodology and design* (131–155). Washington, DC: American Psychological Association.

Hoffman, B. J., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011). Exercises and dimensions are the currency of assessment centers. *Personnel Psychology, 64*(2), 351–395.

Jackson, D. J. R., Michaelides, M., Dewberry, C., & Kim, Y. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Psychology, 101*(7), 976–994.

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology, 1*(1), 84–97.

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: a large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86,* 1202–1222.

MacKinnon, D. W. (1987). *How assessment centers were started in the United States: the OSS assessment program*. Pittsburgh: Development Dimensions International.

Melchers, K. G., Wirz, A., & Kleinmann, M. (2012). Dimensions and exercises: theoretical background of mixed-model assessment centers. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (eds.). *The psychology of assessment centers* (237–254). New York: Routledge.

Merkulova, N., Melchers, K. G., Kleinmann, M., Annen, H., & Szvircsev Tresch, T. (2013). *Not just a myth? Testing the generalizability and nomological network of a new conceptual model for assessment center ratings.* Poster presented at the 28th Annual Conference of the Society for Industrial and Organizational Psychology, Houston, USA.

Merkulova, N., Melchers, K. G., Kleinmann, M., Annen, H., & Szvircsev Tresch, T. (2016). A test of the generalizability of a recently suggested conceptual model for assessment center ratings. *Human Performance, 29*(3), 1–25.

Morris, B. S. (1949). Officer selection in the British army 1942–1945. *Occupational Psychology, 23,* 219–234.

Munchus, III G., McArthur, B. (1991). Revisiting the historical use of the assessment centre in management selection and development. *Journal of Management Development, 10*(1), 5–13.

Murray, H. (1990). The transformation of selection procedures: the war office selection boards. In E. Trist; H. Murray; B. Trist (eds.). *The social engagement of social science: A Tavistock anthology, Vol. I: The socio-psychological perspective* (46–67). Baltimore, MD: University of Pennsylvania Press

OSS Assessment Staff (1948). *Assessment of men: Selection of personnel for the Office of Strategic Services*. NY: Rinehart & Co.

Pandit, N. R. (1996). The creation of theory: a recent application of the grounded theory method. *The qualitative report, 2*(4), 1–15.

Robertson, I. (1999). *Predictive validity of the general fast stream selection process*. Unpublished validity report, school of management, UMIST.

Shepard, L. A. (1993). *Evaluating test validity. Review of research in education, 19*, 405–450.

Standards for Educational and Psychological Testing (1999). Joint Committee on Standards for Educational Evaluation.

Standards for Educational and Psychological Testing (2014). Washington, DC: American Educational Research Association.

Thornton, III G. C., Gibbons, A. M. (2009). Validity of assessment centers for personnel selection. *Human Resource Management Review, 19,* 169–187.

Vernon, P. E. (1950). The validation of Civil Service Selection Board procedures. *Occupational Psychology, 24*, 75–95.

Vernon, P. E., Parry, J. B. (I949). *Personnel selection in the British forces.* London: University of London Press.

Wiggins, J. S. (1973). *Personality and prediction: principle of personality assessment.* Reading, Mass.: Addison-Wesley Publishing Company.

ORGANIZATIONAL PSYCHOLOGY

Assessement center effectiveness: historical perspective

Yury ZHUKOV
Lomonosov Moscow State University, Moscow, Russia

Abstract. The article reviews the evolution of the concept of effectiveness in the psychological practice, with a particular emphasis on assessment centers (AC) validity. Early AC-methods and processes were originated and developed by military psychological and psychiatric services in Europe and America before and during the WW2. The first studies of AC effectiveness were initialized after the war. There were some attempt to measure AC reliability and validity. The assessment center method, in its modern form, came into existence as a result of the AT&T Management Progress Study. At that time large-scale investigations of AC-validity were successfully developed. The trinitarian view of validity (namely content, criterion-related, and construct) has dominated psychology for almost a second half of the XX century. ACs have traditionally demonstrated strong content and criterion-related validity. However, researchers have been puzzled with the lack of evidence concerning construct validity. Researchers have consistently revealed that different behavioral ratings within simulations are more strongly related to one another (exercise effects) than the same dimension rating across simulations (dimension effects). This phenomenon was named as the AC construct-related validity paradox. Now the trinitarian doctrine has been replaced by the unitarian validity concept and we have the opportunity to reformate the problem for creating more effective ACs. Further, we examine some emerging trends in resolving similar methodological problems by other branches of the psychological practice, such as qualitative investigations and action researches. Conclusions about the validity, fairness, acceptability and efficacy of ACs for personnel selection and development are offered.

Keywords: trinitarian doctrine; construct-related validity paradox; unitarian validity concept; fairness; acceptability; efficacy.

References

American Psychological Association (1954). Technical recommendations for psychological tests and diagnostic techniques. Washington DC US. *Psychological Bulletin, 51*(2), 1–38.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1–15.

Anderson, G. L., Herr, K. (1999). The new paradigm wars: is there room for rigorous practitioner knowledge in schools and universities? *Educational researcher, 28*(5), 12–40.

Anstey, E. (1977). A 30-year follow-up of the CSSB procedure, with lessons for the future. *Journal of Occupational Psychology, 50,* 149–159

Arthur, W., Day, E. A., Woehr, D. J. (2008). Mend it, don't end it: an alternate view of assessment center construct-related validity evidence. *Industrial and Organizational Psychology, 1*(1), 105–111.

Address: 11, build. 9, Mokhovaya Str., Moscow 125009, Russia **E-mail:** zhukovury@rambler.ru

Arthur, W., Woehr, D. J., Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: a conceptual and empirical reexamination of the assessment center construct-related validity paradox. *Journal of Management, 26*(4), 813–835.

Baryshnikova, E. I. (2013). *Otsenka personala metodom assessment-tsentra* [The Assessment center as a method for personnel assessment]. M.: Mann, Ivanov i Ferber.

Bray, D. W., & Campbell, R. J. (1968). Selection of salesmen by means of an assessment center. *Journal of Applied Psychology, 52*, 36–41.

Bray, D. W., & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs: General and Applied, 80*(17), 1-27.

Cahoon, M. V., Bowler, M. C., Bowler, J. L. (2012). A reevaluation of assessment center construct-related validity. *International Journal of Business and Management, 7*(9) 3–19.

Campbell, D. T., Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin, 56*(2), 81–106.

Cronbach, L. J., Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin, 52*(4), 281–302.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology, 16*(2), 137–163.

Davison, R., Martinsons, M. G., Kock, N. (2004). Principles of canonical action research. *Information systems journal, 14*(1), 65–86.

Erofeev, A. K. (2013). Tsentr otsenki. Osobennosti metoda i printsipy standartizatsii programm otsenivaniya [Assessment Center. Features of the method and principles of standardization of evaluation programs]. *Organizational Psychology, 3*(4), 18–42.

Erofeev, A. K., Bazarov, T. Yu. (2014). Avtorskie tekhnologii razrabotki modelei kompetentsii [Author's technology development competency models]. *Organizational Psychology, 4*(4), 74–92.

Fletcher, C. (2005). Final report on a research study relating to effective selection of staff for senior posts in the civil service. *Queen's Printer and Controller of HMSO*, 3–38.

Furr, R. M., Bacharach, V. R. (2008). *Psychometrics: an introduction*. Thousand Oaks, Sage.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493–511.

Guion, R. M. (1980). On Trinitarian doctrines of validity. Professional Psychology, 11(3), 385–398.

Henwood, K., Pidgeon, N. (2003). Grounded theory in psychological research. In P. M. Camic, J. E. Rhodes, & L. Yardley (eds.). *Qualitative research in psychology: expanding perspectives in methodology and design* (131–155). Washington, DC: American Psychological Association.

Hoffman, B. J., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011). Exercises and dimensions are the currency of assessment centers. *Personnel Psychology, 64*(2), 351–395.

Jackson, D. J. R., Michaelides, M., Dewberry, C., & Kim, Y. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Psychology, 101*(7), 976–994.

Kupriyanov, E. A. (2011). Stoit li igra svech: validnost' Tsentrov otsenki [Is it all worth it: Assessment centers validity]. *Organizational Psychology, 1*(1), 50–58.

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology, 1*(1), 84–97.

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: a large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86*, 1202–1222.

MacKinnon, D. W. (1987). H*ow assessment centers were started in the United States: the OSS assessment program*. Pittsburgh: Development Dimensions International.

Melchers, K. G., Wirz, A., & Kleinmann, M. (2012). Dimensions and exercises: theoretical background of mixed-model assessment centers. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (eds.). *The psychology of assessment centers* (237–254). New York: Routledge.

Mel'nikova, O. T., Khoroshilov, D. A. (2014). Sovremennye kriterial'nye sistemy validnosti kachestvennykh issledovanii v psikhologii [Modern system of criteria of validity of the qualitative research in psychology]. *Natsional'nyi psikhologicheskii zhurnal, 2*, 36–48.

Merkulova, N., Melchers, K. G., Kleinmann, M., Annen, H., & Szvircsev Tresch, T. (2013). *Not just a myth? Testing the generalizability and nomological network of a new conceptual model for assessment center ratings*. Poster presented at the 28th Annual Conference of the Society for Industrial and Organizational Psychology, Houston, USA.

Merkulova, N., Melchers, K. G., Kleinmann, M., Annen, H., & Szvircsev Tresch, T. (2016). A test of the generalizability of a recently suggested conceptual model for assessment center ratings. *Human Performance, 29*(3), 1–25.

Morris, B. S. (1949). Officer selection in the British army 1942–1945. *Occupational Psychology, 23*, 219–234.

Munchus, III G., McArthur, B. (1991). Revisiting the historical use of the assessment centre in management selection and development. *Journal of Management Development, 10*(1), 5–13.

Murray, H. (1990). The transformation of selection procedures: the war office selection boards. In E. Trist, H. Murray, B. Trist (eds.). *The social engagement of social science: A Tavistock anthology, Vol. I: The socio-psychological perspective* (46–67). Baltimore, MD: University of Pennsylvania Press

OSS Assessment Staff (1948). *Assessment of men: Selection of personnel for the Office of Strategic Services*. NY: Rinehart & Co.

Pandit, N. R. (1996). The creation of theory: a recent application of the grounded theory method. *The qualitative report, 2*(4), 1–15.

Popov, A. Yu., Lurie, E. V. (2012). Te zhe lyudi, drugoe vremya: validnost' i nadezhnost' Tsentrov otsenki, dinamika razvitiya otsenennykh kompetentsii [The same people, different time: the validity and reliability of assessment centers, the dynamics of the evaluated competences]. *Organizational Psychology, 2*(4), 43–58.

Protasova, I. V., Tolstobrov, A. P., Korzhik, I. A. (2014). Metodika analiza i povysheniya kachestva testov v sisteme elektronnogo obucheniya Moodle [Methods of analysis and improve test quality in e-learning system Moodle]. *Vestnik Voronezh. gos. un-ta. Ser. Sistemnyi analiz i informatsionnye tekhnologi, 3*, 6172.

Robertson, I. (1999). *Predictive validity of the general fast stream selection process.* Unpublished validity report, school of management, UMIST.

Ross, L., Nisbett, R. (1999). *Chelovek i situatsiya. Perspektivy sotsial'noi psikhologii* [Person and the Situation. Perspectives of Social Psychology]. M.: Aspekt Press.

Rossiiskii standart tsentra otsenki [Russian Standard for Assessment Centers] (2013). *Organizational Psychology, 3*(2), 8–32.

Shepard, L. A. (1993). *Evaluating test validity. Review of research in education, 19,* 405–450.

Standards for Educational and Psychological Testing (1999). Joint Committee on Standards for Educational Evaluation.

Standards for Educational and Psychological Testing (2014). Washington, DC: American Educational Research Association.

Thornton, III G. C., Gibbons, A. M. (2009). Validity of assessment centers for personnel selection. *Human Resource Management Review, 19,* 169–187.

Vernon, P. E. (1950). The validation of Civil Service Selection Board procedures. *Occupational Psychology, 24,* 75–95.

Vernon, P. E., Parry, J. B. (I949). *Personnel selection in the British forces.* London: University of London Press.

Wiggins, J. S. (1973). *Personality and prediction: principle of personality assessment*. Reading, Mass.: Addison-Wesley Publishing Company.

Zhukov, Yu. M. (2015). Nauchnaya teoriya i naukoemkaya praktika [A scientific theory and knowledge-based practice]. *Organizational Psychology, 5*(4), 29–38.