Проблемы оценки качества измерений

Игорь Дубина
Алтайский государственный университет
igor_dubina@yahoo.com

Опубликовано в ж. «Педагогические Измерения», №2, 2007 г.

В статье представлены подходы и методы оценки качества измерений, разработанные как в рамках классической теории измерений, так и на основе параметров, определяемых моделью Раша. Рассматриваются основные критерии качества измерений – точность, валидность и надежность. Обсуждаются вопросы, связанные с разграничением характеристик качества результатов измерений и качества измерительных инструментов.

Введение

Несмотря на то, что история целенаправленных исследований проблем качества измерений в целом и проблем оценки качества измерений в частности насчитывает уже более ста лет[1], эти проблемы остаются актуальными и сегодня, особенно с методологической точки зрения. В теории измерений разработано довольно много методов и подходов к оценке качества измерений. Использование таких подходов для обоснования полученных результатов в зарубежной исследовательской практике считается обязательным. Проблема качества измерений приобретает актуальность не только после их осуществления, но и на этапе проектирования, в том числе стадии разработки методов измерения. К сожалению, в нашей литературе этим методам и подходам уделяется явно недостаточное внимание. Это обстоятельство является одной из основных причин и некачественных эмпирических социально-психологических исследований, описание которых иногда встречается в литературе, и создания сомнительных тестов, которые порой используются в педагогической практике.

Качество измерений характеризуется их *обоснованностью* или *валидностью* (*validity*)[2] и *надежностью* (*reliability*). Качество полученных результатов зависит также от *точности* (*accuracy*) проводимых измерений. В некоторых работах эти понятия (особенно надежность) используются не только для характеристики качества *измерений*, но и для характеристики

1 Началом системных исследований в этой области считаются работы Г.Гельмгольца (1821-94), связанные с вычисление ошибок измерений в естественнонаучных исследованиях. Большинство «классических» методов оценки качества измерений было разработано в первой половине 20 в.

2 В русскоязычной литературе, связанной с рассматриваемой темой, используется калька с англоязычного термина *«валидность»*. Поскольку этот термин представляется уже достаточно известным, закрепившимся и привычным, он также будет использован в этой статье.

моделей и *методов измерений*. Подобное «расширенное толкование» понятий, характеризующих качество измерений, встречается преимущественно в «прикладных» статьях, где предлагается и описывается новый метод или инструмент для измерения тех или иных психологических или социальных феноменов. По качеству результатов измерений заключают о качестве примененного метода[3].

В ряде зарубежных теоретико-методологических и учебных работ, посвященных вопросам социальных измерений, также можно встретить определения надежности и валидности как меры качества инструмента измерений или даже его отдельных составляющих, например вопросов анкеты или заданий теста[4]. В этих работах термины «валидный» или «надежный» относятся к инструменту в том смысле, что этот инструмент обеспечивает валидные и надежные измерения.

В.С. Аванесов полагает методологически неправомерным использование характеристик качества измерений для характеристики качества инструментов измерения. Он считает, что правильнее обсуждать вопрос не надежности или валидности педагогических тестов, а надежности или валидности тестовой информации (результатов), поскольку результаты зависят не только от качества теста, но и от условий, в которых он применяется, от выборки испытуемых и т.д.[5] Автор данной статьи разделяет эту точку зрения. Действительно критерии валидности, надежности и точности должны относиться к результатам, а не к методам и инструментам измерения. Характеристики методов оцениваются через анализ результатов измерений, но прямой перенос не всегда корректен. Хорошим методом могут быть получены некачественные измерения и наоборот, некачественный метод в ряде случаев может продуцировать результаты, которые с формально-статистической точки зрения можно считать хорошими. Поэтому все характеристики и методы оценки качества, рассматриваемые в данной статье, в первую очередь надо относить к результатам измерений.

Оценка точности измерений

[3] Примерами такого подхода являются работы по измерению характеристик организационного климата и когнитивных стилей решения проблем: Kirton, M. (1976) Adaptors and Innovators: A Description and Measure, Journal of Applied Psychology, No. 61, pp. 622-629; Mathisen, G.E. and Einarsen, S. (2004) A Review of Instruments Assessing Creative and Innovative Environments Within Organizations, Creative Research Journal, Vol. 16, No. 1, pp. 119-140; Amabile, T. M., Burnside, R. M. and Gryskiewicz, S. S. (1999) User's manual for assessing the climate for creativity. Greensboro, NC: Center for Creative Leadership; Basadur, M., Graen, G. and Wakabayashi, M. (1990) Identifying individual differences in creative problem solving style, Journal of Creative Behavior, No. 24, pp. 111-131.

[4] Litwin, M.S. (1995) How to measure survey reliability and validity. SAGE Publications; Shuman, H. (2004) 'The random probe: A technique for evaluating the validity of closed questions', in Bulmer, M. (Ed.) Questionnaries. Vol. 3. SAGE Publications, pp. 389–396; Black, T.R. (1999) Doing Quantitative Research in the Social Sciences: An Integrated Approach to Research Design, Measurement and Statistics. SAGE Publications; Cooper, D.R. and Shindler, P.S. (1995) Business Research Methods. Irwin/McGraw-Hill.

[5] Аванесов В.С. Проблема качества педагогических измерений // Педагогические измерения, №2, 2004, с. 3-27.

Точность измерений – это величина, характеризующая качество *выборочных* измерений. Эта характеристика обычно оценивается по среднему значению и стандартной ошибке измерений, связанной с численностью выборки. Точность интервальной оценки параметра, измеряемого при выборочном исследовании, определяется двумя показателями:

а) интервалом, в котором ожидается обнаружить оцениваемый параметр;

б) вероятностью обнаружения этого параметра в данном интервале.

Эти два показателя объединяет понятие *доверительного интервала*. Процесс определения доверительного интервала основан на центральной предельной теореме – одной из основных теорем теории вероятностей и статистики. Согласно этой теореме, распределение средних значений выборок, извлекаемых из одной и той же совокупности, соответствует нормальному распределению. Более того, когда выборки становятся достаточно большими, то выборочные средние подчиняются нормальному закону, даже если исходная переменная не распределена по нормальному закону. Среднее значение всех выборочных средних равно среднему значению генеральной совокупности (*M*), стандартное отклонение выборочных средних арифметических (σ_x) определяется по формуле:

$$\sigma_x = \frac{\sigma}{\sqrt{n}},$$

где σ – стандартное отклонение по генеральной совокупности; *n* – объем выборки.

Величина σ_x называется *стандартной ошибкой среднего арифметического* (*standard error of the mean*)[6]. Вычисление стандартной ошибки среднего основывается на предположении нормальности измеряемой переменной величины. Если это предположение не выполнено, то оценка может оказаться неверной, особенно для малых выборок.

Естественным образом возникает вопрос о том, какой объем выборки может считаться «достаточно большим». Известно эмпирическое правило, согласно которому принимается, что если объем выборки (*n*) равен 100 или более, то применима центральная предельная теорема, и допущение о нормальности распределения всех возможных выборочных средних может быть принято. Показано, что при увеличении объема выборки до 100 и более, качество оценки стандартной ошибки среднего улучшается и без предположения нормальности выборки. Если же *n* меньше 100, то нужно иметь веские доказательства нормальности распределения генеральной совокупности. И только в этом случае можно полагать, что распределение, которому подчиняются выборочные статистики, является нормальным.

Поскольку в большинстве случаев значение стандартного отклонения по генеральной совокупности (σ) неизвестно, его заменяют выборочным стандартным отклонением (*s*), и

[6] Термин впервые ввёл Юл (Yule) в 1897 г.

стандартная ошибка среднего арифметического рассчитывается как $\sigma_x = \frac{s}{\sqrt{n}}$. Предполагается, что выборка формируется в результате случайного повторного отбора.

Отсюда следует, что стандартное отклонение по выборке определяет интервал попадания среднего значения по всей генеральной совокупности. Стандартная ошибка среднего зависит от стандартного отклонения по выборке и ее объема. Например, если стандартное отклонение по выборке уменьшается в *два* раза, то оцениваемое изменение измеряемого параметра по генеральной совокупности также уменьшается в *два* раза. При увеличении численности выборки в *четыре* раза, при том же самом значении стандартного отклонения по выборке, мы можем обеспечить увеличение точности лишь в *два* раза.

При бесповторном случайном отборе стандартное отклонение выборочных средних рассчитывается как $\sigma_x = \frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$. Очевидно, что для применения этой формулы должна быть известна численность генеральной совокупности *N*.

Для нормального распределения существует универсальное соотношение между относительной частотой встречаемости в генеральной совокупности значений *x,* средним значением (*M*) и стандартным отклонением (σ)[7]. Это соотношение удобно представить для *стандартного нормального распределения*[8] (или *z-распределения*) в виде $f(z) = \frac{1}{\sqrt{2\pi}}\ell^{-\frac{z^2}{2^2}}$. Любое нормальное распределение может быть сведено к *z*-распределению с помощью простого преобразования: $z = \frac{x-M}{\sigma}$. Последняя формула называется *стандартным z-преобразованием*, переводящим измерения в *стандартную z-шкалу*. В результате такого преобразования значения *z* выражаются в единицах стандартного отклонения от среднего.

Важным практическим следствием этого является возможность однозначного определения для любого *z* площади под кривой любого нормального распределения вне зависимости от величины среднего значения и стандартного отклонения. Так, например, для *z* = 1 около 68,26% всех значений признака располагаются в пределах одного стандартного отклонения по обе стороны от среднего значения при любом нормальном распределении. Это означает, что с вероятностью 0,6826 значение параметра, оцениваемого по элементу, случайно

---

[7] Закон нормального распределения: $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\ell^{-\frac{(x-M)^2}{2\sigma^2}}$.

[8] Стандартное нормальное распределение имеет среднее значение, равное 0, и стандартное отклонение, равное 1. Поэтому для обозначения стандартного нормального распределения также используется термин *единичное нормальное распределение.*

извлекаемому из генеральной совокупности, будет попадать в интервал $M \pm \sigma$. Для z=2 значение вероятности составит 0,9544, т.е. в 95,44% случаев значение параметра будет попадать в интервал $M \pm 2\sigma$. Для z=3 значение вероятности составит 0,9972, т.е. в 99,72% случаев значение параметра будет лежать в интервале $M \pm 3\sigma$. Другие значения z и соответствующие им значения вероятности можно взять из статистических таблиц, включаемых практически во все учебники по теории вероятностей и математической статистике.

На основе этого важного свойства нормального распределения можно оценить точность измерений при выборочных исследованиях. Так, если известно среднее арифметическое значение по выборке ($\overline{X}$) и выборочное стандартное отклонение (s), легко определить стандартную ошибку среднего σ_x. Используя соответствующую статистическую таблицу и задавая необходимое значения вероятности (требуемый *уровень статистической значимости*), можно определить значение z, которое соответствует заданному значению вероятности попадания среднего значения параметра по генеральной совокупности в интервал $\Delta = \overline{X} \pm z\sigma_x$. Величина Δ называется *доверительным интервалом* (*confidence interval*), а величина $\delta = \pm z * \sigma_x$ называется *предельной ошибкой среднего*. Доверительный интервал фактически характеризует *точность оценки* измеряемой величины. Таким образом, для оценки точности выборочных измерений достаточно определить среднее значение и стандартное отклонение по выборке, а также задать уровень значимости.

Очевидно, что с увеличением значения z возрастает вероятность попадания среднего в доверительный интервал Δ, но при этом диапазон оценки становится неопределеннее и размытее, что уменьшает точность оценки[9]. Поэтому не следует стремиться задавать очень большое значение вероятности. Вполне достаточным является 90% или 95% уровень значимости. Поскольку стандартное отклонение средних значительно меньше стандартного отклонения индивидуальных откликов, приемлемым считается даже 68% доверительный интервал[10].

В случае, когда выборка состоит из менее 100 элементов или когда нет достаточных оснований считать выборочное распределение нормальным, для определения доверительного интервала рекомендуется использовать другое теоретическое распределение – t-распределение Стьюдента. В этом случае процесс определения доверительного интервала аналогичен случаю больших выборок, но вместо значения z используется значение t-критерия Стьюдента (зависит от объема выборки и задаваемого уровня вероятности).

[9] Чем менее определенным является прогноз, тем с большей вероятностью он осуществится.

[10] Traub, R.E. (1994) Reliability for the social sciences: theory and applications. SAGE Publications. p. 42.

Валидность как характеристика измерительных инструментов

Обоснованность, или *валидность* (*validity*), – это эквивалентность результатов измерений характеристикам измеряемых объектов. Другими словами, это мера соответствия оценок, получаемых в процессе измерения, представлениям о сущности свойств исследуемых объектов и их роли в исследуемых процессах. Оценивая валидность измерений, мы отвечаем на вопрос: «Действительно ли мы измеряем то, что предполагаем измерять?»

В общем случае валидность – это интерпретационная и не формализуемая характеристика[11]. Оценка и аргументация валидности измерений носит преимущественно описательный характер, хотя в некоторых случаях используются и математико-статистические методы.

В литературе встречается достаточно большое количество (свыше 10) различных терминов, обозначающих типы валидности. При этом их четкая классификация отсутствует. Иногда одним и тем же термином в разных источниках обозначается разное содержание, а иногда, наоборот, различные термины наполняются одинаковым содержанием, характеризующим валидность. Поскольку русскоязычная терминология в этой области окончательно не сложилась, в рассматриваемых ниже типах валидности, мы всюду указываем исходный англоязычный термин.

Внешняя валидность (*face validity*) характеризует восприятие заданий теста или вопросов анкеты непрофессионалами в той области, в которой планируется проводить измерения. В то же время, это люди – объекты той генеральной совокупности, которую предполагается исследовать. Это понятие характеризует, как задания воспринимаются и понимаются респондентами (испытуемыми).

Содержательная валидность (*content validity*) показывает, насколько вопросы анкеты или задания теста соответствуют сути (содержанию) измеряемых показателей. Это, как и внешняя валидность, не формализуемая характеристика, но в отличие от предыдущей, она оценивается экспертами, т.е. специалистами в той области, в которой проводится измерение.

Проверка внешней и содержательной валидности – это первые и обязательные элементы при разработке любого измерительного инструмента. Обоснование содержательной валидности может предшествовать проверке внешней валидности.

Критериальная валидность (*criterion-related validity*) характеризует качество измерений с позиций двух эмпирических критериев, а именно:

[11] Parry, H.J. and Crossley, H.M. (2004) ‘Validity of responses to survey questions’, in Bulmer, M. (Ed.) Questionnaires. Vol. 3. SAGE Publications, pp. 351–372.

а) возможность предсказывать те или иные результаты на основе полученных измерений – *прогностическая валидность* (*predictive validity*);

б) соответствие полученных результатов неким «золотым стандартам», т.е. результатам измерений этого же свойства, полученным ранее (разного рода психометрические или социологические индексы, статистические распределении и т.п.); либо соответствие результатов измерений результатам, полученным уже испытанным и признанным инструментом, используемым параллельно проводимым измерениям – *согласованная валидность* (*concurrent validity*).

Эти два вида критериальной валидности могут оцениваться с помощью статистических показателей связи, например, коэффициента корреляции. Считается, что значение коэффициента корреляции, превышающее 0,7, свидетельствует в пользу валидности полученных измерений[12].

Концептуальная валидность (*construct validity*) обозначает соответствие результатов измерений тому концепту (свойству), для измерения которого проводилось исследование. Нам представляется не совсем адекватной калька с англоязычного термина *construct validity* («конструктная валидность»), которую используют некоторые авторы в русскоязычных изданиях, поэтому в этой статье, придерживаясь терминологии В.С. Аванесова[13], используется термин «концептуальная валидность».

Это наиболее важный и наиболее сложно оцениваемый аспект валидности. Другими словами, этот термин характеризует логическое соответствие измеряемых показателей изучаемому понятию (концепту): действительно ли с помощью выделяемых показателей мы можем характеризовать то свойство, которое мы изучаем. В литературе выделяется два вида концептуальной валидности:

а) *конвергентная валидность* (*convergent validity*) предполагает, что если с помощью различных методов измерений мы получаем близкие результаты, то эти измерения обоснованы (валидны);

б) *дивиргентная* или *дифференцирующая валидность* (*discriminant validity*) связана с возможностью выделения и отделения различных показателей изучаемого свойства (концепта). Для оценки этого вида валидности может быть использован факторный анализ, но основное внимание здесь должно уделяться содержательному анализу изучаемого свойства или феномена.

---

[12] Litwin, M.S. (1995) How to measure survey reliability and validity. SAGE Publications. p. 45.

[13] Аванесов В.С. Проблема качества педагогических измерений // Педагогические измерения, №2, 2004, с. 3-27.

Оценка надежности измерений и измерительных инструментов

Надежность (*reliability*) – это характеристика, отражающая устойчивость и согласованность получаемых результатов измерения. В повседневном общении мы очень часто используем слово «надежность» (надежный человек, надежный компьютер, надежный автомобиль и т.д.). Основные смыслы, которые при этом вкладываются в эту характеристику, – это стабильность, безотказность, повторяемость, предсказуемость, регулярность. Примерно такие же смыслы вкладываются и в понятие «надежность» как характеристики измерения.

Надежность характеризует, насколько измерения свободны от случайных ошибок. В отличие от оценки валидности, оценка надежности измерительного инструмента всегда осуществляется с помощью математических операций. Общий подход к оценке надежности заключается в оценке степени связанности результатов измерения с помощью либо *параллельных испытаний*, либо разнесения измерений во времени, либо соотнесения данных, полученных по разным фрагментам одного инструмента.

Для определения надежности используются три основных подхода, основанных на трех разных аспектах понимания надежности:

1. *Надежность-устойчивость* (*stability*) характеризует стабильность результатов во времени.

2. *Надежность-эквивалентность* (*equivalence*) характеризует идентичность результатов, полученных несколькими аналогичными инструментами.

3. *Надежность-согласованность* (*internal consistency*) характеризует внутреннюю согласованность результатов, полученных одним инструментом.

Рассмотрим эти подходы подробнее. Для оценки надежности в смысле устойчивости результатов во времени проводится повторное измерение тем же инструментом по той же выборке через определенный промежуток времени (*метод «тест-ретест»*). Результаты двух измерений, как правило, сравниваются путем определения коэффициента корреляции или другой меры связи, а также средних значений по двум испытаниям. В случае, если получается высокий коэффициент корреляции (близкий к единице) и средние значения по первому и второму тестированию близки, то это свидетельствует о надежности измерений в смысле их воспроизводимости и стабильности. Если по результатам первого и второго испытаний средние значения различаются достаточно сильно, но в целом те испытуемые, которые имели высокие баллы при первом тестировании, получили также высокие баллы во втором, то в этом случае коэффициент корреляции принимает достаточно высокие значения, что указывает на определенную надежность измерений. На практике статистически значимый коэффициент корреляции выше 0,8 считается свидетельством достаточной надежности измерений, хотя в

некоторых работах[14] в качестве приемлемого значения указывается 0,7. При этом также следует указывать уровень статистической значимости полученного результата. Плохая воспроизводимость результатов предыдущего тестирования приводит к низкой корреляции результатов, что свидетельствует о низкой надежности.

Достоинство этого метода заключается в сравнительной простоте его использования, ясности основных посылок, лежащих в определении надежности, и простоте расчетов. Сложности возникают при определении временного интервала между двумя испытаниями. Если ретест проводится слишком рано, испытуемые могут запомнить ответы, которые они давали при первом испытании. При слишком позднем проведении повторного испытания измеряемые характеристики могут измениться (например, знания, способности, опыт испытуемых, отношения респондентов и т.д.). Приемлемым считается интервал между тестированиями от 2-х недель до 2-х месяцев. Кроме того, сама по себе высокая корреляция не может однозначно свидетельствовать о воспроизводимости результатов, поэтому результаты повторного тестирования рекомендуется контролировать другими методами. Например, можно сравнивать ранги испытуемых, и если они в основном не изменились, то появляются дополнительные основания в пользу надежности измерений, но только в смысле их стабильности, так как возможен тренд, т.е. систематическое увеличение или уменьшение результатов от одного тестирования к другому. Возможно использование процедур проверки статистической гипотезы о равенстве средних значений и достоверности различий дисперсий по первому и повторному тестированиям.

Для оценки надежности-эквивалентности используется метод параллельного тестирования, или альтернативных тестов (*parallel forms*), проводимых либо одновременно, либо с небольшим интервалом. Данный метод оценки надежности применим только тогда, когда имеются параллельные (сходные, но не одинаковые) формы одного инструмента. Одной и той же группе испытуемых предлагается вначале одна форма, затем после некоторого перерыва (до одной-двух недель) – другая. Коэффициент корреляции, полученный по результатам двух тестов, называется *коэффициентом эквивалентности результатов измерения*. Если между предъявлением обеих форм имеется значительный временной интервал (свыше двух недель), то полученный коэффициент называется *коэффициентом эквивалентности и стабильности результатов измерений*.

Статистически значимый коэффициент корреляции выше 0,8 считается свидетельством достаточной надежности тестируемого инструмента. Однако вычисление коэффициента корреляции может оказаться недостаточным в случае больших различий в средних значениях и

[14] Например, Litwin, M.S. (1995) How to measure survey reliability and validity. SAGE Publications. p.8.

дисперсиях по параллельным тестам. Еще одна сложность применения данного метода заключается в невозможности обеспечить *полную* эквивалентность двух разных тестов. Рекомендуется в качестве альтернативного теста использовать тот же самый инструмент с переформулированными или сходными по уровню сложности заданиями.

Наиболее часто надежность измерений оценивается по *согласованности* (гомогенности) полученных результатов. Такие измерительные инструменты как анкеты и тесты состоят из большого числа отдельных составляющих: вопросов, утверждений, заданий и т.п. Каждый из пунктов направлен на косвенное выяснение какой-то одной стороны, отдельного фрагмента общего целого, вследствие чего он является частичным индикатором измеряемого фактора (свойства). Предполагается, что когда мы принимаем во внимание всю совокупность индикаторов и определенным образом интегрируем косвенную информацию, которую несет каждый из индикаторов, наши выводы становятся более надежными и обоснованными. Но при этом надежное измерение должно быть внутренне непротиворечиво.

Применяются различные способы интегрирования информации из частных индикаторов (суммирование баллов, полученных по каждому заданию; использование модели Раша и др.). Однако прежде чем интегрировать данные по индикаторам, необходимо соблюдение условия, что эти индикаторы отражают одно и то же, имеют нечто общее. Если это не так, тогда операция получения комплексной оценки просто не имеет смысла. Надежность-согласованность как раз и показывает, в какой степени результаты измерений внутренне согласованы.

Для оценки надежности-согласованности разработано несколько методов. Рассмотрим базовые предпосылки, лежащие в их основе. Убедиться в том, что два задания измеряют нечто общее, можно путем определения коэффициента корреляции между ответами на эти задания. Достаточно высокое (>0,8) значение коэффициента корреляции (r) между двумя переменными может свидетельствовать о том, что имеется какой-то скрытый (латентный) фактор, общая причина, которая стоит за каждой из них. Именно на этом соображении может строиться проверка такого качества измерений, как согласованность. Но такой подход позволяет сравнивать отклики (ответы) попарно. Как можно на этой основе получить универсальный показатель для результатов измерения в целом?

Одним из первых методов, разработанных для решения этой задачи, был *метод раздельного коррелирования* (*split-half*). Он заключается в разбиении откликов по всем пунктам инструмента (ответов на задания теста) на две половины и расчете коэффициента корреляции по соответствующим двум наборам данных – суммарным баллам по каждому заданию. Суммирование баллов в двух сформированных группах дает два набора данных, корреляция между которыми и характеризует надежность-согласованность измерений. Если результаты измерений совершенно надежны, то следует ожидать, что обе части абсолютно коррелируют

(т.е. *r = 1,0*). Впрочем, такая «абсолютная надежность» является гипотетической и на практике встречается исключительно редко. Если результаты измерений не являются абсолютно согласованными, то коэффициент корреляции будет меньше единицы.

Преимущество метода раздельного коррелирования перед методом параллельного тестирования заключается в том, что он позволяет оценить надежность при однократном тестировании. Однако использование этого метода предполагает допущение об эквивалентности не только отдельных форм, но и заданий теста. Еще одна сложность использования метода раздельного коррелирования заключается в том, что два набора тестовых заданий можно получить разными способами, причем количество возможных вариантов разделения «астрономически» возрастает с количеством заданий. Если у нас, например, 20 заданий в тесте, можно первые 10 включить в одну группу, а остальные 10 – в другую; также можно в первую группу включать задания с нечетными номерами, а во вторую – с четными (это наиболее распространенная на практике процедура) и т.д. Тест, состоящий, например, из 20 заданий, может быть поделен на две половины для раздельного коррелирования $\frac{20!/2}{10!10!} = 92378$ разными способами. Понятно, что коэффициент корреляции будет зависеть от способа разбиения.

Наиболее правильным считается разбиение, производимое случайным образом, что позволяет избежать искусственных эффектов. Тем не менее, показатель надежности-согласованности, полученный таким методом, будет варьироваться всякий раз при формировании групп. Проблемами этого подхода является также неэквивалентность заданий (например, одни задания могут быть сложнее, чем другие, и наоборот), а также то обстоятельство, что испытуемый может вообще не выполнить какие-то задания или, дойдя до половины теста, ответить небрежно на оставшиеся вопросы (задания).

Еще одной проблемой применения этого метода является то, что коэффициент корреляции рассчитывается не по всем заданиям теста, а по половине. Для корректировки полученного значения используется *формула Спирмена-Брауна,* предложенная независимо друг от друга К. Спирменом (C. Spearmen) и У. Брауном (W. Brown) в 1910 г.:

$$r_{SB_2} = \frac{2r}{1+r},$$

где r_{SB2} – скорректированный показатель надежности-согласованности по методу *split-half*;
r – коэффициент корреляции между двумя наборами пунктов инструмента.

Данная формула обобщается на случай теста, состоящего не из 2, а из *k* эквивалентных частей, по каждой из которых известен коэффициент корреляции *r* (*обобщенная формула Спирмена-Брауна)*:

$$r_{SB} = \frac{kr}{1+(k-1)r}.$$

На следующем уровне обобщения эту формулу можно использовать для оценки согласованности измерений, предполагая, что k – это количество всех заданий (а не блоков заданий) теста, а r – усредненный коэффициент корреляции между всеми заданиями[15]. В этом случае снимается проблема многообразия способов формирования групп. Проиллюстрируем использование этой формулы на простом примере. Пусть мы имеем тест из 3 заданий, на которые получены ответы 4 испытуемых (оцениваемые по шкале от 0 до 2).

Испытуемые	**Задания**		
	1	**2**	**3**
1	0	1	1
2	1	2	2
3	2	1	2
4	0	1	1
	0	1	1

	1	*2*	*3*
Корреляционная матрица	1		
	0,25	1	
	0,918559	0,612372	1
Усредн. коэф. корреляции	0,593644		

Вычисляется корреляционная матрица (матрица коэффициентов корреляции). Затем определяется среднее арифметическое трех коэффициентов корреляции – усредненный коэффициент корреляции (для нашего случая 0,594). Далее по обобщенной формуле Спирмена-Брауна определяется индекс согласованности (0,814).

Такой подход предполагает равенство дисперсий в двух коррелируемых группах. Известен аналог коэффициента раздельного коррелирования Спирмена-Брауна, который не предполагает такого равенства дисперсий. Он вычисляется по *формуле Рулона* (*Rulon formula*)[16]:

$$r_R = 2\frac{\sigma_t^2 - \sigma_1^2 - \sigma_2^2}{\sigma_t^2},$$

где σ_t^2 – общая дисперсия по всем данным (первой и второй группам); σ_1^2 – дисперсия по первой группе; σ_2^2 – дисперсия по второй группе.

Другие подходы к определению внутренней согласованности основаны на вычислении коэффициентов *KR20 Кадера-Ричардсон*, *альфа Кронбаха* и *лямбда Гутмана*. Рассмотрим

[15] HR-Лаборатория Human Technologies – www.ht.ru.

[16] Traub, R.E. (1994) Reliability for the social sciences: theory and applications. SAGE Publications. pp. 80-85.

базовые аксиомы, на основе которых разработаны формулы для вычисления этих коэффициентов.

Каждое измерение включает в себя как истинное значение, так и частично неконтролируемую, случайную погрешность

$$O = T + E,$$

где O – наблюдаемое (измеряемое) значение (*observed score*); T – истинное значение (*true score*); E – случайная ошибка (*random error*).

Более полно, с учетом систематической ошибки (B), имеем: $O = T + E + B$.

Изменчивость измеряемого признака может быть связана с «естественной» изменчивостью самого признака (например, различие в подготовленности), но определенный вклад может внести то, как мы измеряем, т.е. изменчивость ошибки измерения. Запишем это как

$$\sigma_O^2 = \sigma_T^2 + \sigma_E^2$$

(систематическая ошибка не учитывается, так как считается, что ее изменчивость равна 0).

Тогда надежность измерения (ρ) может характеризоваться отношением изменчивости истинных значений к изменчивости наблюдаемых значений, т.е. $\rho = \frac{\sigma_T^2}{\sigma_O^2}$.

Разумеется, мы не можем знать истинные значения и их изменчивость (иначе нам бы не пришлось проводить никакие измерения), но мы можем исключить их из рассмотрения, представив как $\sigma_T^2 = \sigma_O^2 - \sigma_E^2$. Тогда получаем

$$\rho = \frac{\sigma_O^2 - \sigma_E^2}{\sigma_O^2} = 1 - \frac{\sigma_E^2}{\sigma_O^2}.$$

В знаменателе мы имеем не что иное, как дисперсию измеряемых значений. Определить значение в числителе сложнее, но понятно, что оно должно иметь смысл дисперсии ошибок наших измерений. Однако заметим сразу, что использование подобного подхода не предполагает разбиения пунктов инструмента на группы, поэтому снимается проблема зависимости результата от способа разбиения. Это весьма абстрактная идея о надежности измерений воплотилась в нескольких конкретных вариантах расчетных моделей.

Впервые конкретная реализация подобных рассуждений была предложена Кадером и Ричардсон (1937 г.) и получила название *формулы Кадера-Ричардсон-20* (*Kuder-Richardson 20*), или KR_{20}. Несколько необычное название формулы связано с тем, что авторы предложили несколько различных формул, обозначаемых разными индексами; двадцатая оказалась наиболее удачной. Эта формула была предложена для вычисления коэффициента

согласованности для дихотомической шкалы (т.е. для переменных, принимающих только два значения, например для ответов истинно/ложно):

$$KR_{20} = \frac{k}{k-1}(1 - \frac{\sum_{i=1}^{k} p_i q_i}{\sigma_t^2}),$$

где p_i – доля первого варианта ответа на i-й вопрос; $q_i = (1 - p_i)$ – доля второго варианта ответа на i-й вопрос; σ_t^2 – дисперсия сумм измеряемых значений (суммирование осуществляется по всем заданиям теста для каждого респондента); k – количество вопросов.

Например, получены ответы 5 испытуемых по трем заданиям теста с дихотомической шкалой. В нижеследующей таблице приведены ответы, а также все промежуточные результаты для вычисления коэффициента KR_{20} . Полученное значение (0,86) свидетельствует о хорошей согласованности измерений, продуцируемых данным тестом.

Испытуемые	Задания			Сумма	
	1	2	3		
1	0	1	1	2	
2	1	1	1	3	
3	0	0	0	0	
4	1	1	1	3	
5	1	1	0	2	
σ_t (ст. отклонение суммарных баллов)					1,22
σ_t^2 (дисперсия суммарных баллов)					1,5
p_i	0,6	0,8	0,6		
q_i	0,4	0,2	0,4		
$p_i q_i$	0,24	0,16	0,24	0,64	
KR_{20}					0,86

Для порядковых шкал с большим количеством позиций (например шкалы Лайкерта), а также для более мощных шкал (например интервальных) Л. Кронбах предложил другую формулу для определения согласованности измерений (1951 г.). Показатель согласованности, рассчитанный по этой формуле, получил название *коэффициент альфа Кронбаха* (*Cronbach's Coefficient Alpha*). Большинство современных статистических пакетов (SPSS, SAS, STATISTICA и др.) включают процессы вычисления коэффициента альфа Кронбаха. Несложно посчитать этот коэффициент и с помощью стандартных функций Excel. Формула выглядит следующим образом:

$$\alpha = \frac{k}{k-1}(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_t^2}).$$

где σ_i^2 – дисперсия ответов по каждому заданию; σ_t^2 – дисперсия суммарной шкалы (дисперсия суммы ответов каждого респондента на задания); k – количество пунктов.

Формула Кронбаха является расширенной аналогией формулы Кадера-Ричардсон и отражает следующую идею. Если есть несколько субъектов, отвечающих на вопросы анкеты, то можно вычислить дисперсию для каждого вопроса и суммарной шкалы. Дисперсия для суммарной шкалы будет меньше, чем сумма дисперсий каждого отдельного вопроса в том случае, когда вопросы измеряют (оценивают) *одну и ту же* изменчивость между субъектами, т.е. если они измеряют некоторое истинное значение. Если не измеряется реальное значение, а только случайная погрешность в ответах на вопросы (следовательно, ответы полностью не коррелированны между субъектами), то дисперсия суммы будет такой же, как сумма дисперсий отдельных пунктов. Поэтому коэффициент альфа будет равен нулю. Если все вопросы измеряют один и тот же объект (истинную метку), то коэффициент альфа равен 1.

Рассмотрим использование формулы Кронбаха на примере. Возьмем те же данные, которые мы использовали для иллюстрации применения обобщенной формулы Спирмена-Брауна. Определим дисперсии по ответам на каждый вопрос и их сумму. Получим значение 1,3. Дисперсия агрегированных оценок по каждому вопросу составит 2,7. Отсюда значение коэффициента альфа Кронбаха 0,778.

Если мы сравним полученное значение с коэффициентом согласованности, определенным по обобщенной формуле Спирмена-Брауна, то увидим, что коэффициент альфа меньше: 0,778 < 0,814. Это связано с тем, что обобщенная формула Спирмена-Брауна вычисляет коэффициент согласованности как если бы измерения были стандартизованы, т.е. приведены к одной шкале с нулевым средним значением и единичной дисперсией. Часто (но не всегда) стандартизация исходных данных приводит к возрастанию надежности измерений.

Коэффициент альфа Кронбаха принимает значения в диапазоне от 0 до 1. Приемлемыми считаются значения $\alpha>0,8$. Однако, заключая о надежности-согласованности измерений, следует принимать во внимание и объем выборки: чем меньше выборка, тем меньше может быть коэффициент альфа. Поэтому для небольших выборок (меньше 20 элементов) приемлемым может считаться значение $\alpha>0,7$[17]. Высокое значение коэффициента указывает на наличие общего основания у набора вопросов (заданий), но не говорит о том, что за ними стоит именно тот фактор, который предполагается измерять, поэтому предварительно необходимо обосновать валидность измерений.

Надежность-согласованность, определяемая по формуле Кронбаха, будет зависеть также от количества и качества заданий, входящих в тест. При исключении любого задания коэффициент альфа будет изменяться (уменьшаться или увеличиваться). При исключении заданий, которые не противоречат другим заданиям теста (в том смысле, что все они

[17] Black, T.R. (1999) Doing Quantitative Research in the Social Sciences: An Integrated Approach to Research Design, Measurement and Statistics. SAGE Publications. p. 280.

направлены на измерение общего фактора), коэффициент альфа Кронбаха уменьшается. И напротив, при исключении заданий, которые не согласуются с другими, значение коэффициента альфа будет увеличиваться. Теоретически, при оценке надежности измерений мы должны определить коэффициент альфа Кронбаха при условии, что одно из заданий исключается (и так для всех заданий). Это весьма трудоемкая задача, особенно если в измерительном инструменте много пунктов. Поэтому для решения этой задачи используют специальные статистические пакеты (SPSS, STATISTICA и др.).

Задания, при исключении которых коэффициент альфа увеличивается достаточно сильно, следует убрать из теста. К сожалению, однозначных критериев того, что значит «достаточно сильное увеличение», не существует. Однозначно нельзя сказать, что если при удалении задания коэффициент альфа увеличился на столько-то, то это задание должно быть исключено. Единственное общепринятое правило заключается в том, что если альфа (или другой показатель, характеризующий надежность-согласованность измерений) меньше 0,7, то измерение не может считаться надежным. Если при этом в тесте есть задания, при исключении которых коэффициент надежности увеличивается до 0,7 и выше, то такие задания необходимо удалить. Если, например, мы определили, что коэффициент альфа при исключении некоторого задания теста возрастает с 0,65 до 0,75, то это задание лучше исключить. Но если коэффициент альфа для исходного набора заданий составляет 0,8, а при исключении какого-то задания увеличивается до 0,9, нужно обратить особое внимание на это задание (например, на то, как оно сформулировано), но исключать его не обязательно, так как и с данным заданием надежность-согласованность измерений приемлема.

Коэффициент альфа Кронбаха можно рассматривать как оценку корреляции измерений данным инструментом с измерениями всеми другими инструментами, составленными из такого же числа индикаторов, которые случайным образом извлекли из множества всех возможных индикаторов измеряемого свойства. Его можно также интерпретировать как корреляцию между измерениями данным инструментом и «истинными» измерениями, полученными, если бы испытуемый выполнил *все* возможные задания, направленные на измерение изучаемого свойства. Коэффициент альфа может также применяться и для решения гораздо более широкого круга задач. Например, с его помощью можно измерять степень согласованности экспертов, оценивающих тот или иной объект, стабильность данных при многократных измерениях, качество различных шкал и т.д.

Еще один подход к оценке согласованности данных был предложен в 1945 г. Л. Гутманом, который составил формулы для вычисления шести коэффициентов, наиболее важными из них

являются первые три (L_1, L_2, L_3)[18]. Первый коэффициент определяет нижнюю границу надежности, второй коэффициент – «лучшую» из возможных оценок нижней границы надежности, а третий формально эквивалентен коэффициенту альфа Кронбаха. Доказано, что коэффициент L_2 всегда больше либо равен коэффициенту альфа Кронбаха. Мы не будем приводить здесь формулы для расчета коэффициента Гутмана в силу их достаточной громоздкости, что делает весьма трудоемким расчет этих коэффициентов без специальных статистических пакетов. По-видимому, громоздкость формул и является основной причиной того, что этот подход получил значительно меньшее распространение на практике, чем формула Кронбаха, хотя подход Гутмана был описан в литературе на 6 лет раньше. С помощью современных статистических программ коэффициенты Гутмана вычисляются так же просто, как и коэффициент альфа Кронбаха, но в силу «привычки» и того обстоятельства, что в литературе они описаны гораздо реже, коэффициенты Гутмана в исследовательской практике используются не так часто, как коэффициент альфа Кронбаха.

Отметим, что рассмотренные показатели (α, KR_{20}, L_1, L_2, L_3 и др.) не обязательно всегда неотрицательны. Возможны ситуации, когда каждый из этих коэффициентов будет иметь отрицательные значения (это произойдет в случае, если сумма ковариаций между ответами на задания теста отрицательна). В отличие, например, от коэффициента корреляции, отрицательные значения коэффициентов надежности-согласованности не несут никакой дополнительной информации, кроме той, что из-за слабой согласованности измерения не могут считаться надежными.

При использовании процедур проверки валидности и надежности измерений может возникнуть определенная сложность, связанная с тем, что инструменты, используемые для измерения тех или иных интересующих исследователя признаков, часто включают в себя несколько различных блоков (например, заданий теста), которые не только сформулированы по-разному, но и используют различные измерительные шкалы. Как в таком случае оценить валидность и надежность?

Рекомендуется измерительный инструмент делать гомогенным (однородным). Прежде всего необходимо, чтобы инструмент измерял некий единый концепт. Должны использоваться одинаковые шкалы для каждого пункта. Желательно, чтобы задания теста были одинаковы по форме. Если исследователь все же использует разнородный инструмент, то можно оценивать обоснованность и надежность измерений по блокам, определяемым смысловыми категориями (концептами), на изучение которых направлено исследование. Но для корректной оценки надежности измерений рекомендуется все же использовать однородный инструмент.

[18] Traub, R.E. (1994) Reliability for the social sciences: theory and applications. SAGE Publications. pp. 87-94.

Оценка качества измерений на основе модели Раша

Рассмотренные выше процедуры оценки надежности-согласованности измерений были разработаны в рамках классической теории измерений. Серьезным ее недостатком является то, что во многих случаях при использовании процедур оценки не принимается во внимание вид измерительной шкалы. В частности, для данных в порядковых шкалах используются те же процедуры, что и для интервальных шкал.

Поэтому другой (не альтернативный, но дополняющий) подход к оценке качества измерений построен на основе измерительной модели Раша. С помощью этой модели можно ответить на вопросы: «Насколько задания теста согласованы в плане измерения единого концепта?», «Измеряют ли они некий единый фактор или различные факторы?», «Насколько исходные данные подходят для измерения на основе используемой модели?». Модель Раша показывает, насколько каждое задание теста подходит (*fits*) для измерения той или иной характеристики предмета исследования. Надежность измерения можно оценивать как по заданиям (*item reliability index*), так и по испытуемым (*person reliability index*)[19]. Первый показатель характеризует повторяемость результатов для заданий: если эти же задания будут предложены другой группе испытуемых, будут ли получены аналогичные результаты? Второй показатель характеризует повторяемость результатов для испытуемых: если этой же группе испытуемых будут предложены другие задания, измеряющие тот же концепт, будут ли получены аналогичные результаты? На основе модели могут быть получены ошибки измерений по испытуемым, а также степень соответствия их ответов модели. Измерения, не соответствующие модели, не могут рассматриваться как надежные, и должны быть исключены из анализа.

Оценка этих показателей может быть осуществлена с помощью программы WINSTEPS, в которой рассчитываются параметры MNSQ INFIT и MNSQ OUTFIT, характеризующие соответствие данных модели Раша. Они определяются на основе средних сумм квадратов отклонений теоретических значений от эмпирических (*mean square statistics*). Значения этих параметров характеризуют степень «случайности» результатов или несоответствие данных используемой модели измерения. «Ожидаемые» значения MNSQ находятся вблизи 1,0[20]. Высокие значения MNSQ OUTFIT могут быть связаны со «случайными» откликами респондентов. Высокие значения MNSQ INTFIT обычно интерпретируются как индикатор низкой валидности измерений. Например, если в результате тестирования обнаруживаются

[19] Bond, T.G. and Fox, C.M. (2001) Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Lawrence Erlbaum.

[20] Wright, B.D. and Masters, G.N. (1982) Rating Scale Analysis, MESA Press.

высокие значения MNSQ INTFIT, то это свидетельствует о том, что данный тест непригоден для группы испытуемых, в которой он предъявлялся. Более важными с точки зрения характеристики качества результатов, как указывалось, являются значения MNSQ INTFIT.

Более критичными для измерений являются высокие значения MNSQ. Измерения со значениями MNSQ>2,0 рассматриваются как несоответствующие модели измерения, поэтому они не могут быть использованы при анализе результатов. Такие измерения рекомендуется исключать. Наиболее качественными и значимыми (*productive*) считаются измерения, для которых значения MNSQ лежат в диапазоне от 0,5 до 1,5. Более высокие значения (>1,5) указывают на неопределенность и «шум» в исходных данных. Слишком низкие значения (<0,5) также не очень желательны, поскольку они свидетельствуют об избыточности, «информационной перегруженности» инструмента.

Рассчитываемая статистика соответствия зависит от объема данных. Если количество испытуемых меньше 30, модель может оказаться не очень чувствительной («подходит все»). В случае, если количество испытуемых больше 300, модель, напротив, может оказаться слишком чувствительной («ничто не подходит»).

Оценка качества измерительных инструментов

Как отмечалось во введении к данной статье, основные характеристики качества измерений (валидность, надежность и точность) не могут «напрямую» относиться к измерительным инструментам, т.е. методологически некорректно говорить, например, о «валидном» тесте. Однако на практике качество измерительного инструмента оценивается через анализ результатов, полученных с помощью этого инструмента. Многократно проводя измерения, особенно при проектировании нового инструмента, мы оцениваем их качество. При этом, изменяя инструмент, например, исключая или добавляя задания теста, мы получаем измерения лучшего или худшего качества. Если мы стабильно получаем качественные измерения, то и используемый инструмент вызывает у нас больше доверия с точки зрения его качества. Например, если мы используем один и тот же инструмент для одной и той же выборки и получаем при этом аналогичные результаты при условии, что измеряемая характеристика не изменилась, это дает нам возможность предполагать, что мы используем качественный инструмент. То есть, принимая во внимание условия проведения измерений и оценивая качество результатов измерений по показателям их валидности и надежности, мы можем судить о том, способен ли инструмент обеспечивать валидные и надежные измерения, но при этом не можем говорить о том, что этот инструмент «надежен и валиден».

Однако мы можем говорить об «*эффективности*» инструмента как комплексной характеристике его качества[21]. Эффективным мы можем называть измерительный инструмент, обеспечивающий качественные измерения с точки зрения их валидности, надежности и точности.

Кроме того, в понятие эффективности измерительного инструмента входит такая характеристика, как *практичность* (*practicality*), т.е. экономичность применения (низкая затратность), удобство и простота использования. Например, мы можем говорить о том, что один тест эффективнее другого, если он обеспечивает более качественные измерения при тех же затратах времени или, например, более удобен в использовании, обеспечивая измерения примерно того же качества, что и другой тест.

Понятие эффективности может относиться не только ко всему инструменту, но и к отдельным его составляющим, например, заданиям теста[22]. Выше мы уже обсуждали, как качество отдельных заданий теста влияет на надежность-согласованность измерений. Исключение из теста «плохих» заданий делает его более эффективным. Еще одни показателем качества задания является коэффициент корреляции между ответами испытуемых на это задание и общей суммарной шкалой (суммарным показателем по всем заданиям). Считается, что этот показатель не должен быть меньше 0,2-0,3. Задания с меньшим коэффициентом «загромождают» тест и делают его менее эффективным. С другой стороны, если этот показатель очень высок (например, 0,95), это будет означать, что изменчивость ответов на одно это задание на 90% повторяет (или даже определяет) изменчивость откликов по всему тесту. Это означает *избыточность* теста (либо это задание «лишнее», либо все остальные задания не несут никакой дополнительной информации). Очевидно, что избыточность теста снижает его эффективность.

Эффективность теста зависит также от его *дифференцирующей способности*, связанной с отражением изменчивости измеряемых характеристик. Проверка дифференцирующей способности проводится для выделения и исключения заданий, не обеспечивающих достаточную степень «уверенного» разделения ответов. Например, если на одно задание *все* испытуемые отвечают правильно, а на другой вопрос *все* отвечают неправильно, то такие задания никакой информации фактически не несут, поэтому они не вносят никакой вклад в изучение того концепта, который интересует исследователя. Следовательно, такие задания не нужны в разрабатываемом тесте. Или другой пример. Предположим, мы проводим экзамен в студенческой группе и предлагаем такой тест, по которому все студенты выполняют все задания и получают «отлично», затем предлагаем другой тест, и в результате никто не выполняет ни

[21] Аванесов В.С. Проблема качества педагогических измерений // Педагогические измерения, №2, 2004, с. 3-27.

[22] Там же.

одного задания, и все получают «неудовлетворительно». Способны ли такие «тесты» дать представление о знаниях студентов? Их нельзя считать эффективными.

Для оценки дифференцирующей способности заданий используются более или менее сложные математические процедуры, как правило, связанные с методами проверки статистических гипотез. Рассмотрим одну из таких процедур.

После тестирования ответы всех испытуемых по каждому заданию суммируются. Например, 1-й испытуемый по ответам на все задания теста имеет 35 баллов, 2-й – 43, 3-й – 12 и т.д. Затем суммарные баллы ранжируются по величине. В итоге мы можем отобрать 20–25% испытуемых с наименьшим суммарным баллом и столько же с наибольшим суммарным баллом. Первая группа соответствует испытуемым с наихудшими результатами, вторая группа соответствует испытуемым с наилучшими результатами.

Таким образом, сформированы две группы по n человек: группа с низким суммарным баллом (группа L) и группа с высоким суммарным баллом (группа H). Оставшиеся испытуемые (50%) со «средним» баллом не рассматриваются. Далее *для каждого* задания теста определяются следующие величины:

f – число испытуемых, получивших определенную оценку (например, по 5-балльной системе оценки это 1, 2, 3, 4 или 5);

$fX = f * X$;

$fX^2 = f * X * X$;

$$\overline{X} = \frac{\sum fX}{n},$$

где X – оценка (например, 1, 2, 3, 4 или 5); $n = \Sigma f$ – число респондентов в группах L и H (в каждой группе это число должно быть одним и тем же).

Далее для каждого задания теста определяется модифицированный t-критерий.

$$t = \frac{\overline{X}_H - \overline{X}_L}{\sqrt{\frac{(\sum fX_L^2 - \frac{(\sum fX_L)^2}{n}) + (\sum fX_H^2 - \frac{(\sum fX_H)^2}{n})}{n(n-1)}}}.$$

В этой формуле индексы L (*low*) и H (*high*) соответствуют первой и второй группам соответственно.

После определения t-критерия задания ранжируются по его величине. Большее значение t-критерия соответствует лучшей дифференцирующей (разделяющей) способности задания. В качестве критерия пригодности вопросов шкалы по степени различения принимается $t_{критическое}$ = 1,75 для $n \geq 25$[23]. Задания с t <1,75 должны быть исключены. При n <25 критическое значение

[23] Cooper, D.R. and Shindler, P.S. (1995) Business Research Methods. Irwin/McGraw-Hill. p. 196.

t можно взять из стандартной таблицы *t*-распределения для соответствующего числа степеней свободы и выбранного уровня значимости.

Если тестируемая группа испытуемых состоит из нечетного числа респондентов (например, 71), то при формировании *L*- и *H*-групп не обязательно добавлять или удалять испытуемых, чтобы получить четное число, также как не нужно включать в группы одних и тех же испытуемых. Соотношение 25–25–50% – условное и может варьироваться. После ранжирования суммарных баллов всех испытуемых отбирается равное количество испытуемых «сверху» и «снизу» (приблизительно по 25% от численности группы); какое именно количество остается в средней группе (четное или нечетное) – не принципиально. Отбираются *относительно* «высокие» и «низкие» суммарные баллы, безотносительно к их абсолютным значениям.

Если «плохое» (неэффективное) задание представляется исследователю особо важным, то его необходимо переформулировать. Но после замены или переформулировки хотя бы одного из заданий необходимо вновь проводить тестирование и затем снова оценивать качество измерений. На практике исследователь часто несколько раз проходит через этапы создания, удаления и переформулировки заданий теста до тех пор, пока не придет к окончательному набору заданий, обеспечивающих эффективный измерительный инструмент.

Заключение

В заключение кратко остановимся на сопоставимости показателей качества измерений, получаемых на основе классической теории измерений и модели Раша. Характеристики качества измерений, определяемые «классическими» методами (в том числе альфа Кронбаха), и индикаторы качества измерений в модели Раша имеют разные смыслы, разную внутреннюю логику, и разные вычислительные процедуры. Например, коэффициент альфа Кронбаха основан на идее о согласованности результатов измерения, а модель Раша предлагает инструментарий для оценки соответствия данных модели (*fit statistics*). Поэтому прямо сопоставлять эти подходы нельзя, как нельзя напрямую сопоставить результаты их применения. В частности, нельзя сравнить согласованность измерений (например, вычислением коэффициента Кронбаха) по исходным данным и после их преобразования в шкалу Раша.

Дело в том, что применение модели Раша дает интегрированные результаты и по заданиям, и по испытуемым; в итоге осуществляется переход к вероятностным оценкам (например, может быть оценена вероятность правильного ответа на определенное задание определенным испытуемым). Коэффициент альфа Кронбаха рассчитывается по фиксированным баллам каждого респондента по каждому заданию. Поэтому при оценке

надежности измерений следует для полноты анализа не заменять, а дополнять одни подходы другими[24].

[24] Такая несопоставимость в моделях в определенном смысле отражает нестыковку, несопоставимость научных парадигм, проанализированную Т. Куном в его знаменитой книге (Кун Т. Структура научных революций. М.: Прогресс, 1975).