Кафедра педагогических измерений

ПРОБЛЕМЫ ОЦЕНКИ КАЧЕСТВА ИЗМЕРЕНИЙ

Игорь Дубина
Алтайский государственный университет
igor_dubina@yahoo.com

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 надёжность. 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 зару-

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 XX в.

бежной 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 *надёжностью* (*reliability*). Качество полученных результатов зависит также от *точности* (*accuracy*) проводимых измерений. В некоторых работах эти понятия (особенно надёжность) используются не только для характеристики качества *измерений*, но и для характеристики *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, посвящённых вопросам социальных измерений, также можно встретить определения надёжности 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 «надёжный» относятся к инструменту в том смысле, что этот инструмент обеспечивает валидные и надёжные измерения.

В.С. Аванесов полагает методологически неправомерным использование характеристик качества измерений для характеристики качества инструментов измерения. Он считает, что правильнее обсуждать вопрос не надёжности или валидности педагогических тестов, а надёжности 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, надёжности и точности должны относиться к результатам, а не к методам и инструментам изме-

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.

3

Примерами такого подхода являются работы по измерению характеристик организационного климата и когнитивных стилей решения проблем: Kirton M. Adaptors and Innovators: A Description and Measure, Journal of Applied Psychology, No. 61, 1976, pp. 622–629; Mathisen G.E., Einarsen S.A. Review of Instruments Assessing Creative and Innovative Environments Within Organizations, Creative Research Journal, Vol. 16, No. 1, 2004, pp. 119–140; Amabile T.M., Burnside R.M., Gryskiewicz S.S. User's manual for assessing the climate for creativity. Greensboro, NC: Center for Creative Leadership 1999; Basadur M., Graen G., Wakabayashi M.

рения. 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

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 ожидаётся 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 — объём 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

Identifying individual differences in creative problem solving style, Journal of Creative Behavior, No. 24, 1990, pp. 111–131.

4 Litwin M.S. How to measure survey reliability and validity. SAGE Publications, 1995; Shuman H. The random probe: A technique for evaluating the validity of closed questions», in Bulmer, M. (Ed.) Questionnaries. Vol. 3. SAGE Publications, 2004, pp. 389–396; Black T.R. Doing Quantitative Research in the Social Sciences: An Integrated Approach to Research Design, Measurement and Statistics. SAGE Publications, 1999; Cooper D.R., Shindler P.S. Business Research Methods. Irwin/McGraw-Hill, 1995.

5 *Аванесов В.С.* Проблема качества педагогических измерений // Педагогические измерения, № 2, 2004, с. 3–27.

6 Термин был впервые введен Юлом (Yule) в 1897 г.

объём выборки может считаться «достаточно большим». Известно эмпирическое правило, согласно которому принимается, что если объём 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 объёма 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 стандартная ошибка среднего арифметического рассчитывается как $\sigma_x = \frac{\sigma}{\sqrt{n}}$. 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 её объёма. 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

[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.

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 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 ($\bar{X}$) 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 $\Delta = \bar{X} \pm z\sigma_x$. Величина Δ называется *доверительным интервалом* (*confidence interval*), а величина $\delta = \pm z \cdot \sigma_x$ называется *предельной ошибкой среднего*. Доверительный интервал фактически характеризует *точность оценки* измеряемой величины. Таким образом, для оценки точности выборочных измерений достаточно опре-

делить 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 неопределённее 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 объёма выборки и задаваемого уровня вероятности).

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 чёткая 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 русско-

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

10 *Traub R.E.* Reliability for the social sciences: theory and applications. SAGE Publications. 1994. p. 42.

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

язычная 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:

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*

12 *Litwin M.S.* How to measure survey reliability and validity. SAGE Publications. 1995. p. 45.

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.

Оценка надёжности измерений и измерительных инструментов

Надёжность (*reliability*) — это характеристика, отражающая устойчивость и согласованность получаемых результатов измерения. В повседневном общении мы очень часто используем слово «надёжность» (надёжный человек, надёжный компьютер, надёжный 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 «надёжность» как характеристики измерения.

Надёжность характеризует, насколько измерения свободны от случайных ошибок. В отличие от оценки валидности, оценка надёжности измерительного инструмента всегда осуществляется с помощью математических операций. Общий подход к оценке надёжности 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 надёжности используются три основных

13 *Аванесов В.С.* Проблема качества педагогических измерений // Педагогические измерения, № 2, 2004, с. 3–27.

подхода, основанных на трёх разных аспектах понимания надёжности:

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

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

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

Рассмотрим эти подходы подробнее. Для оценки надёжности в смысле устойчивости результатов во времени проводится повторное измерение тем же инструментом по той же выборке через определённый 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 надёжности 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 определённую надёжность измерений. На практике статистически значимый коэффициент корреляции выше 0,8 считается свидетельством достаточной надёжности 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 надёжности.

Достоинство этого метода заключается в сравнительной простоте его использования, ясности основных посылок, лежащих в определении надёжности и простоте расчётов. 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 из-

1 Например, *Litwin M.S.* How to measure survey reliability and validity. SAGE Publications. 1995. p. 8.

мениться (например знания, способности, опыт испытуемых, отношения респондентов и т.д.). Приемлемым считается интервал между тестированиями от 2 недель до 2 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 надёжности 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 надёжности-эквивалентности используется метод параллельного тестирования или альтернативных тестов (*parallel forms*), проводимых либо одновременно, либо с небольшим интервалом. Данный метод оценки надёжности 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 надёжности тестируемого инструмента. Однако вычисление коэффициента корреляции может оказаться недостаточным в случае больших различий в средних значениях и дисперсиях по параллельным тестам. Ещё 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 надёжность 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 определённым образом интегрируем косвенную информацию, которую несёт каждый из индикаторов, наши выводы становятся более надёжными и обоснованными. Но при этом надёжное 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. Надёжность-согласованность как раз и показывает, в какой степени результаты измерений внутренне согласованы.

Для оценки надёжности-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 расчёте коэффициента корреляции по соответствующим двум наборам данных — суммарным баллам по каждому заданию. Суммирование баллов в двух сформированных группах даёт два набора данных, корреляция между которыми и характери-

зует надёжность-согласованность измерений. Если результаты измерений совершенно надёжны, то следует ожидать, что обе части абсолютно коррелируют (т.е. r = 1,0). Впрочем, такая «абсолютная надёжность» 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 надёжность при однократном тестировании. Однако использование этого метода предполагает допущение об эквивалентности не только отдельных форм, но и заданий теста. Ещё одна сложность использования метода раздельного коррелирования заключается в том, что два набора тестовых заданий можно получить разными способами, причём 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 нечётными номерами, а во вторую — с чётными (это наиболее распространённая 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 надёжности-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).

Ещё 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.:

$$r_{SB2} = \frac{2r}{1+r},$$

где r_{SB2} — скорректированный показатель надёжности-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 (*обобщённая 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) (см. рисунок).

Вычисляется корреляционная матрица (матрица коэффициентов корреляции). Затем определяется среднее арифметическое трёх коэффициентов корреляции — усредненный коэффициент корреляции (для нашего случая 0,594). Далее по обобщённой 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]:

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

15 HR-Лаборатория Human Technologies — www.ht.ru.

16 *Traub R.E.* Reliability for the social sciences: theory and applications. SAGE Publications. 1994, pp. 80–85.

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

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 KR_{20} *Кадера-Ричардсон*, *альфа Кронбаха* и *лямбда Гутмана*. Рассмотрим 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 учётом систематической ошибки (B), имеем: $O = T + E + B$.

Изменчивость измеряемого признака может быть связана с «естественной» изменчивостью самого признака (например, различие в подготовленности), но определённый 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 её изменчивость равна 0).

Тогда надёжность измерения (с) 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 надёжности измерений воплотилась в нескольких конкретных вариантах расчётных 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»):

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

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.

(*Продолжение статьи в следующем номере «ПИ»*)