Проблемы оценки качества измерений. Статья вторая

Статью первую см. ПИ, №2, 2007г.

Игорь Дубина

Алтайский государственный университет

igor_dubina@yahoo.com

Опубликовано в ж. «Педагогические Измерения» №3, 2007 г.

Аннотация

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

Например, получены ответы 5 испытуемых по трем заданиям теста, с дихотомической шкалой оценок. 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 1	2	3	Сумма	
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 испытуемых, отвечающих на вопросы анкеты, то можно вычислить дисперсию для каждого вопроса и для 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 измеряется не 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 вопросов. Поэтому коэффициент альфа будет равен нулю.

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 сравнить полученное значение с коэффициентом согласованности, определенным по обобщенной формуле Спирмена-Брауна, то мы 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 $\alpha > 0,7$[1]. 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 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

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

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 надо 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)[2]. 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 расчёт 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 какой-то из этих коэффициентов будет иметь отрицательные значения (это произойдет в случае, если сумма ковариаций между ответами на задания теста отрицательна). В

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

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.

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*)[3]. 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[4]. Высокие значения MNSQ OUTFIT могут быть связаны со «случайными» откликами

[3] Bond, T.G. and Fox, C.M. (2001) Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Lawrence Erlbaum.
[4] Wright, B.D. and Masters, G.N. (1982) Rating Scale Analysis, MESA Press.

респондентов. Высокие значения MNSQ INTFIT обычно интерпретируются как индикатор низкой валидности измерений. Например, если в результате тестирования обнаруживаются 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[5]. 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[6]. Выше мы уже

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

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 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 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 $n \geq 25$[7]. Задания с $t < 1{,}75$ должны быть исключены. При $n < 25$ критическое значение *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. После чего необходимо вновь проводить тестирование и затем снова оценивать качество измерений. На практике исследователь несколько раз проходит через этапы создания, удаления

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

и изменения 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 согласованности результатов измерения. 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[8].

[8] 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).