Проблема надежности психологических шкал и ее значение для психодиагностики в научных и практических целях[1]

С. А. Корнилов

МГУ им. М. В. Ломоносова (Москва)
sa.kornilov@gmail.com

Рассматриваются современные представления о понимании надежности психологических шкал. Описываются основные подходы к пониманию одного из наиболее распространенных коэффициентов надежности – альфа Кронбаха; выделяются переменные, от которых он зависит; приводятся основные формулы. Обосновывается, что надежность играет важную роль в установлении валидности психологических шкал; раскрываются методы корректировки коэффициентов валидности. Делается вывод о необходимости более тщательного анализа надежности в психологических исследованиях и практике.

Ключевые слова: надежность, согласованность, валидность, психодиагностика, коррекция аттенюации.

Ключевыми понятиями для психодиагностики являются понятия *надежности* и *валидности* психодиагностических методик как инструментов, разрабатываемых в целях косвенной оценки гипотетических конструктов на основании наблюдаемых данных (ответов на задания и вопросы; Cronbach, Meehl, 1955). В классической теории тестов (КТТ) задания (пункты теста) предположительно представляют собой случайную выборку всех возможных заданий, измеряющих заданный конструкт, и выступают индикаторами эффектов (effect indicators), связанных между собой через общую для них латентную переменную. Разработка психологических шкал, измеряющих гипотетические латентные переменные, осуществляется путем создания определенного количества заданий с целью надежного измерения конструкта; при этом надежность понимается как «степень, в которой [измерения] являются повторяемыми» (Nunnally, 1967, с. 206), т. е. реплицируемыми.

Ни один психологический инструмент не обладает идеальной надежностью, поэтому классической формулой для обозначения балла по любой шкале в рамках КТТ является

$$Score_{Total} = Score_{True} + Score_{Error},$$

где общий балл состоит из двух частей – *истинного* балла (как среднего балла при прохождении теста бесконечное количество раз) и случайной *ошибки измерения,* которая признается *несистематической* в том смысле, что средняя ошибка измерения для группы = 0. Таким образом, включенность в общий балл ошибки измерения иногда ведет к повышению, а иногда – к понижению *индивидуальных* тестовых показателей по сравнению с истинными баллом. Таким образом, надежность (r) может пониматься как отношение «истинной» дисперсии к общей дисперсии:

$$r = \frac{\sigma^2_{True}}{\sigma^2_{Total}}.$$

Эта формула применима только к группе тестовых показателей, поскольку для отдельного индивида истинная дисперсия всегда равна нулю (есть только один истинный балл).

1 Работа выполнена при поддержке гранта РГНФ 10-06-00416а.

Включение источников *систематических смешений* модифицирует обе формулы, что иллюстрируется следующими взаимоотношения между надежностью и валидностью (v) (Judd, Smith, Kidder, 1991):

$$Score_{Total} = Score_{CI} + Score_{SE} + Score_{RE},$$

$$r = \frac{\sigma^2_{CI} + \sigma^2_{SE}}{\sigma^2_O},$$

$$v = \frac{\sigma^2_{CI}}{\sigma^2_O},$$

где CI – интересующий нас конструкт, SE – систематическая ошибка, RE – случайная ошибка измерения. Обсуждение систематической ошибки затрагивает вопросы валидности, но не надежности. Увеличение же случайной ошибки ведет к понижению как показателя надежности, так и валидности. Таким образом, возможно создание надежного, но не валидного инструмента, однако без обеспечения надежности валидизация психологической методики как измерительного инструмента невозможна.

Низкая надежность психологических шкал имеет критические последствия при их использовании как в исследовательских, так и в практических целях. К примеру, точность заключения о высоком уровне развития способностей ребенка, получившего балл 115 (т. е. находящегося в верхней 1/6 популяции), сделанное на основе теста интеллекта с надежностью 0,75, на самом деле ограничивается стандартной ошибкой измерения (standard error or measurement, SEM), вычисляемой по формуле:

$$SEM = \sqrt{1 - r} \times SD_{observed},$$

где r – надежность, а $SD_{observed}$ – стандартное отклонение наблюдаемых показателей. Поскольку распределение случайных ошибок принимается нормальным (при M = 0, SD = SEM), можно сделать вывод о том, что в реальности для 96% испытуемых полученный тестовый балл будет в пределах двух стандартных ошибок измерений от истинного балла, т. е. для указанного выше ребенка его истинный балл может быть как 100, так и 130 при полученном балле в 115 по шкале IQ (SEM = 7,5; William, 2000). Данный пример иллюстрирует, что даже тест с приемлемым уровнем внутренней согласованности может крайне неаккуратно измерять диагностируемые свойства на уровне отдельных испытуемых. Это имеет важные следствия для стратегий обора людей в те или иные группы. Именно поэтому как решения о распределении в программы для одаренных, так и иные решения, связанные с «низким» полюсом шкалы IQ, например, никогда не должны делаться на основании единственный раз проведенных тестов.

При использовании психодиагностических шкал в рамках конкретных исследований, как и при валидизации методик, психологи сталкиваются с несколько иным ограничением: наличие случайной ошибки измерения в наблюдаемых переменных x и y накладывает ограничение на максимальный размер корреляции между ними (r_{xy}), которая будет меньше, чем корреляция между соответствующими x и y конструктами («истинными баллами») X и Y (r_{XY}). Игнорирование этого феномена, названного *аттенюацией корреляции* (correlation attenuation), может привести к ошибочным заключениям о взаимосвязях между гипотетическими конструктами,

в частности, при валидизации методик. Традиционным способом преодоления этого ограничения является использование уравнений коррекций аттенюации, наиболее частым из которых в рамках КТТ является следующее (Fan, 2003):

$$r_{XY} = \frac{r_{xy}}{\sqrt{r_{xx} \times r_{yy}}},$$

где r_{xx} и r_{yy} – коэффициенты надежности для переменных x и y соответственно.

Нетрудно увидеть, что если истинная корреляция между интересующими нас конструктами равна 0,60, но обе измеренные переменные имеют надежность 0,50, то корреляция между измеренными переменными составит всего 0,30 (см. рисунок 1). Применение методов коррекции аттенюации хотя вызывает множество споров в литературе, но является вполне обоснованным подходом к преодолению указанного ограничения (вторым подходом является использование методов структурного моделирования для установления связей между латентными переменными при их автоматической коррекции в так называемых моделях измерения, задающих наблюдаемые переменные через латентные переменные и ошибки).

Традиционные источники *ненадежности* данных – факторы времени, неэквивалентности тестовых форм, эффекты наблюдателя, а также гетерогенность тестовых заданий. Последнее критично для КТТ, поскольку она постулирует, что случайно отобранные тестовые задания должны быть высоко связаны между собой, если предназначены для измерения одного и того же конструкта. На этом основано вычисление одного из самых популярных коэффициентов надежности как внутренней согласованности заданий – коэффициента альфа Кронбаха, определяемого по формуле (Cronbach, 1951):

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum \sigma^2_k}{\sigma^2_{Total}}\right],$$

где k – количество заданий, во второй дроби в числителе находится сумма дисперсий всех заданий, а в знаменателе – общая дисперсия. В случае равенства дисперсий всех заданий альфа равна среднему всех коэффициентов надежности, посчитанных по методу расщепления (split-half), в противном случае альфа меньше этого среднего. Таким образом, альфа Кронбаха является функцией общности (communalities) тестовых заданий или, наоборот, их «уникальности» (uniqueness). Это лишь самое

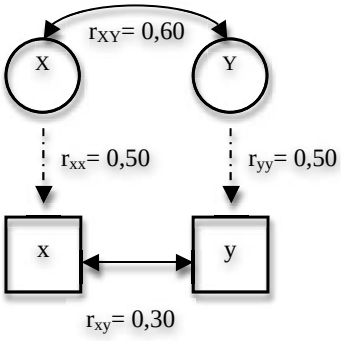

Рис. 1. Эффекты ненадежности психологических шкал при использовании корреляции r Пирсона

общее описание надежности как внутренней согласованности и альфы как способа ее измерения. Однако обратимся к использованию этого коэффициента в практике разработки психодиагностических методик.

В настоящее время в публикациях, посвященных созданию, апробации, валидизации, стандартизации и в целом разработке различных психологических инструментов (от личностных опросников до тестов интеллекта), альфа Кронбаха занимает особое почетное место и считается золотым стандартом описания психометрических свойств методики. При этом чаще всего не учитывается ряд ключевых ограничений, часть из которых прямо вытекает из приведенных выше уравнений (Cortina, 1993; Streiner, 2003).

Во-первых, как отмечает Страйнер (Streiner, 2003), альфа *не является фиксированным свойством шкалы,* это свойство конкретных полученных баллов: один и тот же инструмент может демонстрировать различные (высокие и низкие) альфы на разных выборках, поэтому не имеет смысла обращение к некому ранее установленному уровню согласованности шкалы. Связано это как с тем, что сама оценка согласованности несвободна от ошибки (ввиду того, что мы никогда не знаем истинный балл), так и с тем, что надежность зависит от дисперсии общего балла по шкале, которая варьирует от выборки к выборке. В связи с этим в публикациях важно указывать не ту величину альфы Кронбаха, которая была установлена при изначальной разработке методики (вне зависимости от разработчика), а то значение, которое получено на представленной в конкретной публикуемой работе выборке.

Во-вторых, не имеет смысла обсуждение альфы без *учета количества тестовых заданий:* Кортина (1993) показал, что увеличение длины теста с 6 до 18 заданий при константной средней интеркорреляции между заданиями (0,30) увеличивает альфу с 0,72 до 0,88. Таким образом, необходимо учитывать как длину теста, так и средний уровень связанности заданий. Увеличение количество тестовых заданий: первый способ повышения внутренней согласованности (William, 2000). При имеющейся же согласованности r, если мы хотим достичь согласованности R, мы должны умножить количество заданий на n, где

$$n = \frac{R \times (1 - r)}{r \times (1 - R)}.$$

В-третьих, альфа является мерой внутренней согласованности, но не дает информации о *количестве измеряемых факторов:* высокая альфа не означает наличие единого общего для заданий фактора, поскольку может быть получена для заданий, которые являются индикаторами нескольких ортогональных факторов – главное, чтобы внутри этих факторов задания были хотя бы на среднем уровне связаны между собой. Использование же высокой альфы в качестве аргументации в пользу наличия единого фактора недопустимо.

Четвертая проблема связана с конвенциями в отношении уровней, которые принято считать «приемлемыми». Разные авторы рекомендовали от 0,50–0,60 для ранних стадий исследований, 0,80 для исследовательских методов, и 0,90 для клинических (Nunnally, 1967) до меньших значений при учете содержания измеряемого конструкта: принцип «чем больше альфа, тем лучше» Страйнер (Streiner, 2003) называет одним из «мифов об альфе», поскольку альфа связана не только с гомогенностью заданий, но и с гомогенностью конструкта. Даже одномерные и однофакторные конструкты могут быть концептуализированы как имеющие множество различных

аспектов, что в итоге приведет к появлению определенной гетерогенности заданий разрабатываемой методики. Увеличение гомогенности заданий через уменьшение рассматриваемых в методике сторон гипотетического конструкта – второй способ повышения ее внутренней согласованности.

Таким образом, учет надежности (и, в частности, внутренней согласованности как одной из ее форм) психологических методик как средств операционализации тех или иных гипотетических конструктов, представляется одинаково важным как при проведении исследований, так и в практике психодиагностики. В первом случае недостаточный учет надежности (или, в крайнем, но распространенном случае полного ее игнорирования при публикации результатов исследований) потенциально ведет к серьезным ошибкам при проверке теоретических гипотез о взаимоотношениях между гипотетическими конструктами, стоящими за измеренными переменными. И даже частый случай инконсистентности получаемых в различных исследованиях результатов может быть функцией различий в показателях надежности примененных исследователями инструментов на конкретных выборках, тогда как скорректированные результаты могут быть схожими. Отдельной проблемой является проблема генерализации надежности, в частности, при использовании ее для корректировки показателей, получаемых в рамках метаанализа (Корнилов, Корнилова, 2010). Во втором случае результаты прикладной диагностики оказываются зависимыми от точности методики на индивидуальном уровне, поэтому любые заключения и выводы на этом уровне должны учитывать надежность используемого психодиагностического инструментария, как минимум полученную на схожих выборках. Важным шагом на пути к преодолению указанных ограничений является выработка «привычки» к рутинной проверке и сообщению в результатах исследованиях и справочных материалах к методикам подробных характеристик надежности разработанных и использованных психологических шкал.

Литература

Корнилов С. А., Корнилова Т. В. Мета-аналитические исследования в психологии // Психологический журнал. 2010. Т. 31. № 5. С. 5–17.

Cortina J. M. What is coefficient alpha? An examination of theory and practice // Journal of applied psychology. 1993. V. 78. № 1. P. 98–104.

Cronbach L. J. Coefficient alpha and the internal structure of tests // Psychometrika. 1951. № 16. P. 297–334.

Cronbach L. J., Meehl P. E. Construct validity in psychological tests // Psychological Bulletin. 1955. № 52. P. 281–302.

Fan X. Two approaches for correction correlation attenuation caused by measurement error: implications for research practice // Educational and Psychological Measurement. 2003. № 63. P. 915–930.

Judd C. M., Smith E. R., Kidder L. H. Research methods in social relations. New York: Harcourt Brace Jovanovich, 1991.

Nunnally J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Streiner D. L. Starting at the beginning: an introduction to coefficient alpha and internal consistency // Journal of Personality Assessment. 2003. V. 80. № 1. P. 99–103.

William D. Reliability, validity, and all that jazz // Education. 2000. V. 29. № 3. P. 9–13.