

ИЗМЕРИТЕЛЬНЫЕ ШКАЛЫ И ИХ ПРИМЕНЕНИЕ В ПЕДАГОГИЧЕСКИХ ИЗМЕРЕНИЯХ

Игорь Дубина

Алтайский государственный университет
igor_dubina@yahoo.com

В статье рассматриваются основные шкалы, в которых могут быть представлены результаты измерений. Исследованы свойства каждой шкалы, классификация шкал, основные методы обработки данных и проблема выбора измерительной шкалы. В заключительном разделе статьи исследованы свойства шкалы Лайкерта.

Введение

Умелое применение статистических методов при обработке эмпирических данных базируется на понимании сути измерительных шкал. К сожалению, тема измерительных шкал в российских вузовских учебниках не рассматривается или, в лучшем случае, рассматривается как вспомогательная тема, на уровне поверхностной классификации. Для сравнения: многие американские и европейские университеты предлагают специальные курсы по основам теории измерений. В этих курсах вопросам измерительных шкал в гуманитарной и социальной сферах отводится значительное место. Студенты же российской высшей школы, аспиранты и даже многие исследователи часто имеют лишь самое общее представление о возможностях и границах применения тех или иных шкал.

Для начинающих исследователей число обладает некоторой «магической силой», оно как бы «само по себе» *a priori* означает и

даже подчеркивает объективность исследования, а иногда и направляет порядок действия исследователя. Ведь всем знакомы элементарные арифметические операции, и у начинающих исследователей часто возникает соблазн получить некий результат, «что-нибудь сложив, разделив или умножив». В результате очень часто исследователи используют методы, совершенно не применимые к данным. Отсутствие необходимых знаний и навыков по применению измерительных шкал приводит к тому, что результаты исследований, представляемые в некоторых дипломных и даже диссертационных работах и публикациях, часто являются не только малообоснованными, но даже полностью лишёнными смысла с математической точки зрения.

В данной статье сделана попытка системно, но в то же время на достаточно популярном уровне, изложить ту часть теории измерений, в которой рассматриваются вопросы использования измерительных шкал. Автор надеется, что представленный здесь материал в некоторой степени восполнит пробелы, имеющиеся в русскоязычной литературе, и окажется полезным не только для начинающих исследователей, но и для практиков, работающих в области педагогических измерений.

Атрибуты измерительных шкал

Измерительная шкала — это правило соотнесения степени выражённости признаков исследуемых объектов или событий на числовой шкале. Измерительная шкала определяет тип получаемых данных и множество операций, которые можно с этими данными осуществлять; то есть фактически выбор измерительной шкалы влияет как на результат измерения, так и на исследовательский вывод. Используемая измерительная шкала является основным критерием при выборе и обосновании возможных математико-статистических методов обработки данных и интерпретации результатов. Прежде чем использовать ту или иную математическую операцию (пусть даже из области элементарной арифметики, не говоря уже о сложных статистических методах), очень важно понимать, позволяет ли выбранная шкала осуществлять такую операцию, проведено ли измерение таким способом, для которого возможно *корректное* использование данной операции.

Определение шкалы измерений является первым этапом любого эмпирического исследования. К сожалению, многие исследователи проходят этот этап «автоматически», совсем не задумываясь об этом или руководствуясь лишь «здравым

смыслом». Вопрос о типе используемой шкалы особо важен для педагогических измерений, которые в большинстве случаев являются косвенными. Непосредственно мы фиксируем некоторые признаки (время решения задачи, ответы на определённые вопросы и т.д.) и на этой основе делаем выводы о некотором скрытом (*латентном*) свойстве, недоступном прямому измерению (например, подготовленность по предмету).

Выделяют три основных атрибута измерительных шкал, наличие или отсутствие которых определяет принадлежность шкалы к той или иной категории. *Упорядоченность* шкалы означает, что одна позиция шкалы, определяемая числом и соответствующая выражённости измеряемого свойства, больше, меньше или равна другой позиции. *Интервальность* шкалы означает, что интервалы между позициями шкалы равны между собой. *Нулевая точка* (или точка отсчёта) шкалы означает, что набор чисел, соответствующих выражённости измеряемого признака, имеет точку отсчёта, обозначаемую за 0, которая соответствует полному отсутствию измеряемого свойства.

Классификация шкал

Наиболее общей классификацией измерительных шкал яв-

ляется разделение на *метрические* и *неметрические шкалы*. В первых может быть установлена единица измерения (например, килограмм, градус, рубль и т.д.). В *неметрических шкалах* единица измерения отсутствует. Метрические шкалы являются более *мощными*, т.е. они лучше дифференцируют измеряемые объекты (отражают больше информации о различных измеряемых объектах). К неметрическим относят номинативные и порядковые шкалы. В *номинативной шкале (nominal scale)* отсутствуют все главные атрибуты измерительных шкал, а именно упорядоченность, интервальность, нулевая точка. Для обозначения такой шкалы также используются термины *шкала наименований* и *номинальная шкала*.

Номинальная шкала используется для классификации или идентификации объектов (группировки по классам, каждому из которых приписывается число). Объекты группируются по классам таким образом, чтобы внутри класса они были идентичны по измеряемому свойству. Это самая простая шкала из тех, что могут рассматриваться как измерительные, хотя фактически эта шкала не ассоциируется с измерением и не связана с понятиями «величина» и «число». Она используется только с целью отличить один объект от другого.

Примером свойства, измеряемого в такой шкале, может являться пол. Эта переменная имеет две категории (наименьшее количество категорий, которое может иметь измеряемая переменная). Это свойство может быть измерено отнесением человека (объекта) в соответствующую категорию при подсчёте количества мужчин и женщин. Категория «мужчины» может, например, кодироваться цифрой «1», а категория «женщины» — цифрой «2». В результате можно сделать заключение о том, к какой категории отнесено больше объектов, а к какой меньше. Однако на основании этих данных нельзя произвести логические или математические операции над самими категориями, т.е. для такой шкалы нельзя, например, сказать, что 1 меньше, чем 2, или что $1 + 2 = 3$. Другими примерами свойств, измеряемых в номинальной шкале, являются семейное положение, национальность, религиозная принадлежность, место рождения.

Категории переменных, измеряемых в номинативной шкале, должны быть взаимоисключающими, чтобы не было неопределённости относительно классификации исследуемых объектов. Эти категории должны также быть исчерпывающими, т.е. каждому возможному значению должна однозначно соответствовать определён-

ная категория (хотя бы категория *другие* или *прочие*). Категории данных в номинативной шкале должны включать сопоставимые экземпляры.

При использовании номинативных шкал цифры, кодирующие соответствующие категории, являются лишь метками (именами) и не несут в себе численного значения. Поскольку категории измеряемой переменной фактически не являются численными, то их можно сравнивать друг с другом только на основе количества наблюдений, относящихся к ним. К данным, представленным в такой шкале, арифметические операции не применимы: их нельзя складывать, вычитать, умножать или делить. Однако разработаны специальные процедуры обработки и анализа данных в номинативных шкалах (например критерий χ^2 для статистической проверки гипотез). Они основаны на определении частоты (подсчёте количества появлений переменной в различных категориях).

В *порядковой шкале* (*ordinal scale*) присутствует упорядоченность, но отсутствуют атрибуты интервальности и нулевой точки. Для обозначения такой шкалы также используются термины *ранговая шкала* и *шкала рангов*. Результатом измерений в порядковой шкале является упорядочение объектов. Шкала ранжирует

объекты, приписывает им числа в зависимости от выражённости измеряемого свойства по некоторому признаку (в порядке убывания или возрастания). В отличие от номинативной шкалы здесь можно не просто определить, что один объект отличен от другого, но и то, что по определённому признаку один объект больше или меньше другого. Следовательно, шкала показывает, больше или меньше выражено свойство (измеряемая величина), но не насколько больше или насколько меньше оно выражено, а тем более — во сколько раз больше или меньше.

Порядковая шкала является наиболее распространённой в педагогических измерениях. Характерным примером измерений в порядковой шкале является 5-балльная система оценки знаний (1, 2, 3, 4, 5). Можем ли мы ответить, *насколько* 5 «лучше», чем 4? Настолько же, насколько 4 «лучше», чем 3? Возможно, для одного студента разница между 5 и 4 невелика, а разница между 4 и 3 значительна. Возможно, эта разница изменится от курса к курсу. Если считать, что оценки учащихся или студентов отражают их знания, то в случае применения к этой шкале арифметических действий может получиться, что знания отличника равны сумме знаний двоечника и троечника ($5 = 2 + 3$). А между зна-

ниями отличника и троечника такая же разница, как между знаниями «ударника» и двоечника ($5 - 3 = 4 - 2$). Таким образом, мы не можем сказать, что в такой шкале сохраняется равенство интервалов между пунктами.

Сложение, как и большинство других математических операций, предполагает, что интервалы между значениями равны. Если интервалы не равны, то $2 + 2$ может означать 5 или 555. Поэтому широко распространённая процедура усреднения оценок (баллов), т.е. вычисления среднего арифметического — это операция, лишённая смысла с математической точки зрения. Отсюда видно, что возможности статистического анализа для данных в порядковых шкалах ограничены. Строго говоря, упорядочение — это единственная математическая операция, применимая к порядковой шкале. Использование многих операций с порядковыми данными (например, вычисление среднего арифметического) математически некорректно, но широко распространено на практике.

В то же время шкала может вполне корректно использоваться в экспериментальных исследованиях, но для этого необходимо применять адекватные методы обработки данных, например, конвертирование порядковых шкал в метрические

шкалы или использование непараметрических методов для обработки данных.

В *интервальной шкале (interval scale)* присутствуют упорядоченность и интервальность, но нет нулевой точки. В этой шкале исследуемому объекту присваивается число единиц измерения, пропорциональное выраженности измеряемого свойства. Соответствующие интервалы разных участков шкалы имеют одно и то же значение, поэтому измерения в интервальной шкале допускают не только классификацию и ранжирование, но и точное определение различий между категориями.

Примеры интервальной шкалы: фиксация времени и даты, температурные шкалы Цельсия и Фаренгейта. Например, для известных шкал измерения времени можно сказать, что интервал между 3 и 6 ч равен интервалу между 4 и 7 ч. Но можно ли сказать, что 6 ч в два раза больше, чем 3 ч? Очевидно, нет. Аналогично, если сегодня уличный термометр фиксирует температуру -10°C , а вчера было -20°C , мы можем сказать, что сегодня теплее на 10 градусов. Но можно ли сказать, что сегодня теплее в 2 раза? Это кажется нелепым. Но можно сказать, что температура в 300°C отличается от температуры в 200°C в два раза больше, чем температура в 150°C отлича-

ется от температуры в 100°C , что кажется вполне логичным.

Подобные ответы связаны с тем, что интервальные шкалы не имеют нулевой точки отсчёта. Точнее, выбор нулевой точки в интервальной шкале условен. Имея данные, представленные в интервальной шкале, мы можем судить о том, *насколько* больше или *насколько* меньше выражено измеряемое свойство, но не о том, *во сколько раз* больше или меньше. Для интервальных шкал ноль не соответствует полному отсутствию измеряемого свойства, поэтому и измерения не соответствуют абсолютному количеству измеряемого свойства. Например, нулевое значение температуры по шкале Цельсия не соответствует полному отсутствию тепла. Для интервальных шкал мы можем корректно использовать большинство математических операций, поэтому если, например, имеются данные в порядковой шкале, то с точки зрения математической корректности целесообразно их конвертировать в интервальную шкалу, используя специальные методы, которые будут рассмотрены далее. На практике иногда возникают сложности, как при преобразовании исходных данных в метрические шкалы, так и при интерпретации и анализе полученных результатов.

В *относительной шкале (ratio scale)* присутствуют все

атрибуты измерительных шкал: упорядоченность, интервальность, нулевая точка. Для обозначения такой шкалы также используются термины *шкала отношений* и *абсолютная шкала*. Последний термин подчеркивает абсолютный характер нулевой точки.

Относительная шкала позволяет оценивать, во сколько раз свойство одного объекта больше или меньше аналогичного свойства другого объекта, принимаемого за эталон, единицу. Эта шкала характеризуется всеми атрибутами интервальной шкалы и, кроме того, имеет фиксированную нулевую точку (0), которая не является произвольной, так как она соответствует полному отсутствию измеряемого свойства. Например, переменная *количество студентов* имеет фактическое начало отсчёта, так как нулевое значение соответствует отсутствию студентов вообще. Аналогично, нулевое значение может соответствовать отсутствию образования, дохода и т.п. Другими примерами измерения в относительной шкале являются определение физических характеристик (например веса, высоты, площади), запись возраста респондента, временных периодов (например времени выполнения задания) и т.п.

В силу фиксированности нулевой точки, при сравнении исследуемых объектов

можно судить не только о том, насколько больше или меньше выражено свойство, но и во сколько раз больше или меньше оно выражено. Для относительной шкалы возможно корректное использование любых математико-статистических операций и процедур обработки данных. Впрочем, стоит отметить, что существование нулевой точки редко является критичным для статистического анализа. Например, для выполнения большинства известных параметрических методов (работающих с метрическими шкалами) для проверки статистических гипотез достаточно, чтобы данные были представлены в интервальной шкале.

Рассмотренная классификация измерительных шкал не является единственной. Некоторые специалисты выделяют лишь три категории измерительных шкал, не проводя существенных различий между интервальной и относительной шкалами и объединяя эти шкалы в одну. Авторы некоторых учебных и специальных изданий, напротив, выделяют большее количество шкал. Например, профессор А.И. Орлов выделяет шесть типов шкал (шкалу наименований, порядковую шкалу, шкалу интервалов, шкалу отношений, шкалу разностей, абсолютную шкалу).

Основные методы обработки данных и проблема выбора измерительной шкалы

Основными методами обработки данных практически любого экспериментального исследования являются процедуры получения агрегированных и комплексных оценок. Например, в эксперименте, основанном на выборочном исследовании, измеряются показатели отдельных элементов выборки. Набор таких измерений по каждому показателю может быть значительным (десятки, сотни, тысячи). Для анализа показателя по всей выборке необходимо *агрегировать* измерения, т.е. получить *агрегированные* (коллективные, групповые, обобщённые) оценки. Получение агрегированных оценок должно выполняться корректно. Для величин, измеренных в интервальной и относительной шкалах, агрегирование, как правило, осуществляется путем вычисления среднего арифметического по группе. Для метрических шкал эта простая процедура вполне корректна. Наиболее распространённая ошибка (с точки зрения математической корректности) — использование арифметических действий для агрегирования данных в неметрических шкалах (чаще всего порядковых), например вычисление среднего балла успеваемости по

отдельному предмету по студенческой группе. Довольно часто встречаются ситуации, когда необходимо синтезировать отдельные показатели в некоторый интегрированный (*комплексный*) показатель. Примером является усреднение баллов успеваемости студента (или студенческой группы) по нескольким дисциплинам.

Как и в случае с агрегированными оценками, исследователи нередко используют арифметические действия для интегрирования показателей, не задумываясь о содержательной и математической строгости используемых процедур. Ещё раз подчеркнем, что для неметрических шкал использование операций сложения, вычитания, умножения или деления для получения агрегированных или комплексных оценок с математической точки зрения некорректно и бессмысленно.

Таким образом, при выборе методов статистической обработки полученных данных в первую очередь следует обращать внимание на тип используемой измерительной шкалы. Идеальным с точки зрения корректности использования математико-статистического аппарата является выбор только тех методов, которые могут применяться для обработки данных, представленных в соответствующих шкалах. Однако на практике многие матема-

тические операции используются для данных, измеренных в таких шкалах, для которых эти операции в принципе не применимы. В частности, такие основные статистические показатели, как среднее арифметическое, стандартное отклонение или коэффициент корреляции, являются настолько привычными для представления результатов, что исследователь редко задумывается о корректности использования этих показателей для данных разных типов.

Дифференциация номинальной шкалы от порядковой, а тем более от метрической шкалы, обычно не вызывает сложностей. Гораздо сложнее определить различие между другими шкалами (особенно порядковой и интервальной). Весьма распространённой является ошибка, когда исследователь интерпретирует данные так, как если бы они были измерены по шкале, допускающей более сложные операции, чем эта шкала в действительности позволяет. Наиболее часто встречается ситуация, когда данные порядковой шкалы интерпретируются как измерения в интервальной и даже относительной шкале, поскольку статистические процедуры для этих шкал являются более информативными. Это создаёт трудности для обоснования корректности результатов.

Оценки и измерения

Многие специалисты разделяют подход, согласно которому *измерение* и *оценка* — разные способы количественной характеристики величин, поэтому эти характеристики следует различать. Различие измерения и оценки основано на двух факторах: а) используемая измерительная модель (включающая в себя измерительную шкалу) и б) получаемые результаты. *Оценка* даёт приближённое значение величины или параметра, причём в отличие от измерения нет однозначного соответствия между величиной и её количественным выражением в определённых единицах. Считается, что по исходным данным, представленным в неметрических шкалах, получаются *оценки* свойств исследуемого объекта, тогда как *измерения* могут быть получены при использовании либо метрических шкал, либо определённых процедур преобразования исходных данных в метрические шкалы.

Принято считать, что измерение, в отличие от оценки, характеризуется научной обоснованностью, корректностью использования математико-статистического аппарата, большей точностью, надёжностью и валидностью, стандартизованностью и универсальностью, меньшей чувствительностью к смене наборов исходных дан-

ных и т.д. Хотя чёткое и однозначное различие между оценками и измерениями провести достаточно сложно. Далее, с целью демонстрации такого различия, более подробно рассмотрены два подхода: оценки, получаемые на основе шкалы Лайкерта, и измерения с использованием модели Раша.

Шкала Лайкерта

Шкала Лайкерта (*Likert scale*) впервые была предложена в 1932 г. американским специалистом в областях организационной психологии и управления Р. Лайкертом (Likert). Эта шкала была разработана им во время обучения в аспирантуре в Колумбийском университете, при подготовке докторской диссертации. С тех пор эта шкала широко применяется для измерения отношения респондентов к предмету исследования.

Метод предполагает, что респондент выражает своё согласие или несогласие с каждым суждением (из предложенного набора) по шкале оценок (из 3, 4, 5, 6, 7 или более пунктов). Место респондента на итоговой шкале определяется суммой оценок каждого отдельного суждения. Лайкерт обосновал предположение о самом существовании измеряемой латентной переменной и дал обоснование того, что предложенная шкала является порядковой.

Хотя «классическая» шкала Лайкерта включает 5 пунктов, сходные шкалы с другим количеством позиций шкалы (3, 4, 6, 7...) также называют шкалами Лайкерта; чаще всего используется 5- или 7-позиционная шкала. Шкала может содержать или не содержать «нейтральную точку» (при нечётном числе позиций). Частным случаем шкалы Лайкерта является дихотомическая шкала (2 позиции, при которой выставляются оценки 1 или ноль).

Оценки по шкале Лайкерта могут быть выражены в категориях согласия, частоты, важности, качества и т.п., например:

- (5) Полностью согласен —
- (4) Согласен — (3) Нейтральное отношение — (2) Не согласен — (1) Полностью не согласен;
- (5) Всегда — (4) Очень часто — (3) Иногда — (2) Редко — (1) Никогда;
- (4) Очень важно — (3) Важно — (2) В некоторой степени важно — (1) Неважно;
- (5) Очень хорошо — (4) Хорошо — (3) Приемлемо — (2) Плохо — (1) Очень плохо;

Позиции шкалы Лайкерта обычно кодируются числами от 1 до 5 (для пятипозиционной шкалы), от 1 до 6 (для шестипозиционной шкалы) и т.д. Кодирование может включать отрицательные и нулевое значение (например для нейтральной позиции), что, впрочем, используется нечасто.

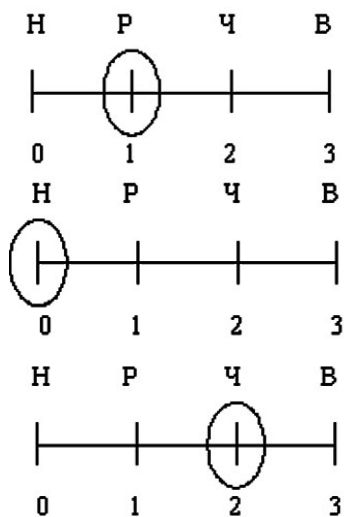
Направление кодирования зависит от формулировки вопроса. Для позитивно сформулированных утверждений (например «Нестандартные ответы студентов поощряются преподавателем») позитивно выражённое отношение (например «согласен») кодируется, как правило, большим числом (например 5), а негативно выражённое («не согласен») — меньшим. При негативной, по смыслу, формулировке — например, «Преподаватель даёт слишком много заданий на дом», позитивно выражённое отношение может кодироваться меньшим числом, а негативно выражённое — большим.

В случае, если анкета содержит как прямые, так и обратные утверждения, отклики на обратные утверждения перекодируются по формуле

$$r = (min - s) + max,$$

где r — отклик на прямые утверждения (вопросы); s — отклик на обратные утверждения (вопросы); min , max — минимальное и максимальное значения шкалы.

Рассмотрим пример агрегирования измерений в шкале Лайкерта (без нейтральной позиции). Предположим, что три студента отвечают на утверждение «Нестандартные ответы студентов поощряются преподавателем», и их отклики распределяются следующим образом



Н — Никогда Р — Редко Ч — Часто В — Всегда

Если использовать традиционный подход для агрегирования оценок, то мы получим $1 + 0 + 2 = 3$. Разделив полученное значение на количество респондентов, получим значение 1. Если изменить кодировку позиций шкалы, например, сдвинув на 1 (от 1 до 4), то мы получим

$$2 + 1 + 3 = 6; \quad \frac{6}{3} = 2.$$

Агрегированная оценка в численном виде изменилась, но средний «словесный» результат («редко»), казалось бы, остаётся тем же. Аналогичный результат мы получим при сдвиге кодировки шкалы на 2, 3, 4 и т.д., а также при использовании шкалы с другим количеством позиций и с наличием нейтральной пози-

ции. Средний «словесный» результат сохранится при любом линейном преобразовании численных значений пунктов шкалы. Является ли эта «инвариантность» результатом к подобной перекодировке свидетельством корректности используемого подхода к получению агрегированной оценки? Выше мы уже обсуждали, что агрегирование оценок на основе вычисления среднего арифметического корректно лишь для метрических шкал. Может ли шкала Лайкерта рассматриваться как метрическая? Другими словами, могут ли измерения в шкале Лайкерта, по крайней мере, интерпретироваться как интервальные?

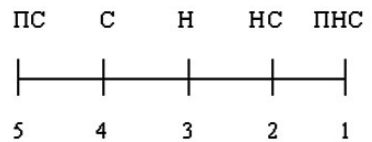
В шкале Лайкерта очень наглядно проявляется дилемма между простотой измерительных процедур и математической строгостью. В литературе встречаются и аргументируются различные точки зрения на возможности использования шкалы Лайкерта. Эта шкала, как показал ещё её автор, является порядковой (неметрической) шкалой, однако многие исследователи считают возможным использовать её как интервальную шкалу и применять процедуры для обработки данных, приемлемые для метрических шкал, так как, по их мнению, это не приводит к серьёзным ошибкам в результатах. Основными аргу-

ментами в пользу такого утверждения являются следующие *пред-суждения*:

- интервалы между позициями шкалы Лайкерта в отдельном пункте (вопросе или утверждении) анкеты приблизительно одинаковы;
- относительные интервалы между позициями шкалы Лайкерта изменяются незначительно при переходе от вопроса к заданию (от утверждения к утверждению).

Если эти два условия можно считать выполненными, то к шкале Лайкерта могут быть применены процедуры, приемлемые для интервальных шкал.

Посмотрим, что означает предпосылка о равенстве интервалов. Пусть мы имеем следующую шкалу:



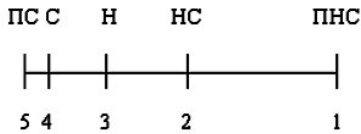
ПС — полностью согласен; С — согласен; Н — нейтральное отношение; НС — не согласен; ПНС — полностью не согласен

Равенство межпозиционных интервалов должно означать следующее:

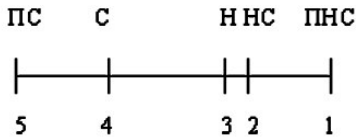
$$\begin{aligned}
 \text{ПС} - \text{С} &= \text{С} - \text{Н} = \text{Н} - \text{НС} = \\
 &= \text{НС} - \text{ПНС} = 5 - 4 = 4 - 3 = \\
 &= 3 - 2 = 2 - 1 = 1.
 \end{aligned}$$

Пусть мы имеем в качестве пункта анкеты следующее утверждение: «Нестандартные от-

веты студентов поощряются преподавателем». Для респондента А, который в целом согласен с этим утверждением, интервалы между позициями шкалы могут быть такими:



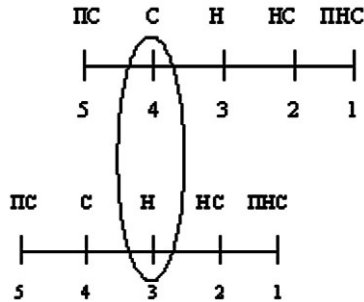
В этом случае респондент А с большей вероятностью выберет одну из позиций — ПС или С, нежели Н, а тем более ПНС, т.е. в этом случае говорить о равенстве интервалов (даже приблизительном) не приходится. Предположим, что некий респондент В скорее не согласен с предложенным утверждением. Для него интервалы между позициями шкалы могут быть следующими:



Интервал между Н и НС для респондента В много меньше, чем интервалы между другими позициями. Очевидно, что вычисление арифметического среднего для данного случая некорректно.

Аналогично складывается ситуация и с получением комплексных оценок в шкале Лайкерта. Предположим, в анкете есть два пункта: «Нестандартные ответы студентов при-

нимаются преподавателем» и «Нестандартные ответы студентов поощряются преподавателем». Очевидно, что эти пункты имеют разный «вес». (Одно дело — просто принимать нестандартные ответы, другое — стимулировать их поиск.) Поэтому отклик «согласен» при ответе на первый пункт может не соответствовать позиции «согласен» при отклике на второй пункт. Возможно, «вес» позиции «согласен» (4) в первом пункте соответствует нейтральной позиции (3) во втором пункте.



Нестандартные ответы студентов принимают преподавателем.

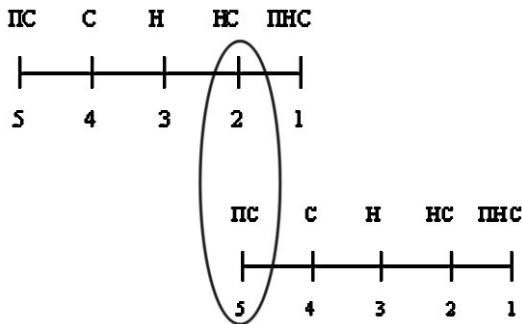
Нестандартные ответы студентов поощряются преподавателем.

Ещё более рельефно неэквивалентность значений позиций шкалы при переходе от пункта к пункту при получении комплексных оценок в шкале Лайкерта демонстрирует пример, приведённый в книге Т. Бонда и К. Фокса¹. Оценивается отношение школьников к работе за компьютером. В анкете используется несколько пунктов, в том числе: «Я настолько боюсь компьютера, что стараюсь избегать пользоваться им» и «Я опасаясь, что сделаю

1

Bond T.G., Fox C.M. (2001) Applying the Rasch Model: Fundamental Measurement in the Human Science, Lawrence Erlbaum.

ошибку при работе за компьютером». Школьник, выбирающий «полностью согласен» по второму пункту, может выбрать «не согласен» по первому, и комплексная оценка, получаемая традиционным способом, будет равна 7, хотя очевидно, что «вклад» пунктов в общую оценку не равноценен. Школьник, выбравший «полностью согласен» по первому пункту и «не согласен» по второму, получит те же 7 баллов, что кажется нелогичным. Более реалистичную оценку можно было бы получить, «выровняв» «вес» этих пунктов, например, следующим образом: 2 в первом пункте равно 5 во втором.



Таким образом, в реальности достаточно сложно обеспечить выполнение двух основных условий (равенство межпозиционных интервалов шкалы и эквивалентность позиций шкалы при переходе от пункта к пункту) для того, чтобы шкалу Лайкерта можно было бы считать интервальной. Даже для приблизительного выполнения этих

условий «веса» пунктов анкеты должны быть приблизительно равны, а отношения респондентов примерно одинаковы.

Возможности преобразования данных из неметрических шкал в метрические

При планировании проводимых измерений целесообразно выбирать такие способы сбора данных, которые позволяют применить метрические шкалы, отражающие больше информации о различии исследуемых объектов. Однако это не всегда возможно, но при применении статистических методов требуется обосновать, что выбранная шкала позволяет эти методы использовать, поэтому исследователь часто стремится найти свидетельства того, что его шкала может рас-

считываться как метрическая.

Почти повсеместно используемые в педагогических измерениях подходы к получению агрегированных и комплексных оценок по шкале Лайкерта путем вычисления арифметического среднего работают с практической точки зрения, но лишены смысла с точки зрения математической, поскольку

Я настолько боюсь компьютера, что стараюсь избегать пользоваться им

Я опасюсь, что сделаю ошибку при работе за компьютером

эта шкала не является интервальной. В то же время шкала Лайкерта является простым и удобным инструментом для сбора данных. Возможен ли некий компромисс между практическим удобством и математической строгостью? Возможно ли значимое обоснование данных в шкале Лайкерта как данных интервальной шкалы? Возможно ли некое преобразование данных из шкалы Лайкерта в метрическую шкалу?

Одним из самых «сильных» аргументов в пользу метрического характера шкалы является распределение измерений по нормальному закону. Исходным принципом здесь является предположение о том, что измеряемое свойство распределено в генеральной совокупности в соответствии с нормальным законом. Следовательно, измерение данного свойства также должно обеспечивать нормальное распределение. Если это так, то шкала может считаться по крайней мере интервальной. Если это не так, то измеряемое свойство отражается в не метрической шкале, а, например, в порядковой. Таким образом, один из подходов к разработке и анализу измерительной шкалы заключается в обеспечении и проверке нормального распределения полученных измерений. Если выборочное распределение не отличается от нормального, то, по

мнению, многих специалистов, можно считать, что измеряемое свойство отражается в интервальной шкале². Однако такой подход к обоснованию метрического характера измерений принимается не всеми исследователями. Например, А.И. Орлов в одной из своих работ³ показал, что во многих случаях распределения результатов измерения существенно отличаются от нормальных.

Причиной отклонения формы выборочного распределения признака от нормального вида чаще всего является особенность процедуры измерения: используемая шкала может обладать неравномерной чувствительностью к измеряемому свойству в разных частях диапазона его изменчивости. Для проверки нормальности используются различные процедуры, позволяющие выяснить, отличается ли от нормального выборочное распределение измеренной переменной. Необходимость такого сравнения особенно актуальна, когда есть сомнения: в какой шкале представлен признак — порядковой или интервальной? Это важно, поскольку от этого зависит возможность использования методов для дальнейшего анализа. А метрические шкалы позволяют использовать гораздо более широкий диапазон методов, чем порядковые, поэтому и исследовательские выводы могут быть

2

Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных. СПб.: Речь, 2006.

3

Орлов А.И. Часто ли распределение результатов наблюдений является нормальным? // Западская лаборатория. 1991. Т. 57. No. 7.

более глубокими и информативными. При значительном отклонении эмпирического распределения от нормального следует отказаться от предположения о том, что признак измерен в метрической шкале.

Методы проверки нормальности распределения включены в статистические пакеты SPSS, STATISTICA и др. Удобный инструментарий для оценки вида распределения предлагает программный продукт @Risk компании Palisade, специализирующейся на разработке программного обеспечения для принятия решений в условиях риска и неопределённости⁴.

Если исходные данные (*raw data*) не проходят проверку на соответствие нормальному распределению, можно их нормализовать, т.е. привести к нормальному виду. Наиболее простым и распространённым нормализующим преобразованием исходных данных является приведение данных к так называемому стандартному z -распределению со средним, равным 0, и стандартным отклонением, равным 1:

$$z_i = \frac{X_i - M}{s},$$

где X_i — исходный тестовый балл испытуемого i ; M — среднее значение по тесту; s — стандартное отклонение результатов теста.

В психологических и педагогических исследованиях также широко используется *стандартная T-шкала*, связанная с z -шкалой простым преобразованием:

$$T_i = 10z_i + 50.$$

T -распределение имеет среднее, равное 50, и стандартное отклонение, равное 10.

Описанные методы стандартизации данных имеют принципиальный недостаток с точки зрения математической строгости. Они изначально используют методы, неприменимые для порядковых шкал (определение среднего значения и стандартного отклонения) по причинам, которые подробно обсуждались выше. Однако исследователи «забывают» о такой «мелочи», и обычно считается, что нормализованные данные, полученные в порядковых шкалах, можно рассматривать как интервальные со всеми вытекающими последствиями по применению методов обработки данных.

Модель Раша

Процедуру более корректного преобразования данных из шкалы Лайкерта в интервальную шкалу можно осуществить на основе *модели Раша* (*Rasch model*), предложенной датским математиком Г. Рашем ещё в конце 1950-х гг., хотя понима-

ние и признание среди исследователей и специалистов эта модель получила лишь в 1970–1980-х гг. Первоначально модель Раша была разработана для оценки способностей (знаний, умений и т.п.). Позднее её также начали использовать для оценки отношений и восприятий. В настоящее время модель Раша наиболее активно применяется в образовательно-педагогической практике (в инструментах тестирования), но имеет серьёзные перспективы для социально-экономических и гуманитарных исследований.

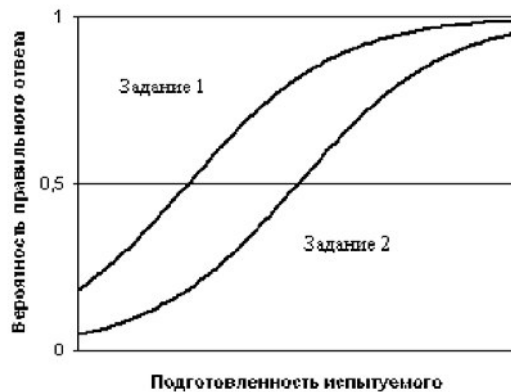
Разработка этой модели была связана с обнаружившимися проблемами представления данных в порядковых шкалах. Во-первых, эти шкалы не обеспечивают линейности (пропорциональности) в представлении данных, а во-вторых, получаемые данные субъективны и ситуативны в том смысле, что они зависят от конкретных респондентов (испытуемых) и конкретных заданий теста. Изменение выборки респондентов и/или системы заданий теста приводит к другим результатам. Г. Раш предложил сравнительно простую модель, которая отчасти снимает эти две проблемы.

В контексте измерения способностей ключевая идея модели Раша может быть сформулирована следующим образом: вероятность правильного ответа на задание теста зависит

от способностей (уровня подготовленности испытуемого (B) и трудности задания (D):

$$P_i(B) = \frac{e^{B-D}}{1 + e^{B-D}}.$$

Эта базовая идея может быть проиллюстрирована с помощью *характеристической кривой* тестового задания. Пусть в тесте Задание 2 сложнее Задания 1. Тогда при той же подготовленности испытуемого (или при тех же самых знаниях) вероятность правильного ответа на Задание 2 будет меньше, чем вероятность правильного ответа на Задание 1.



Единицей шкалы Раша является «логит» (d):

$$d = \ln \left(\frac{P}{1-P} \right).$$

Здесь P — вероятность правильного ответа на вопрос. Чем больше вероятность правильного ответа на вопрос испытуемым, тем большее значе-

ние по шкале Раша он имеет. На практике при получении данных в шкале Раша вместо вероятностей используют частоты откликов, т.е. чем чаще тестируемый отвечает правильно, тем выше его показатель в шкале Раша.

Если подставить в эту формулу значения вероятности успешного (правильного) ответа на вопрос, как она определяется в модели Раша, то получим $d = B - D$. Этот результат показывает, что измерения в шкале Раша теоретически зависят только от способности испытуемого и трудности задания, т.е. мы получаем измерение, независимое от выборки испытуемых (*person free measurement*) и набора тестовых заданий (*item free calibration*). Это одно из важнейших, наряду с интервальностью, свойство шкалы Раша. Независимость измерения от конкретного набора заданий теста и от выборки респондентов (в пределах ошибки измерения) позволяет специалистам в области теории измерений называть шкалу Раша «объективной» измерительной шкалой и сравнивать со шкалами, используемыми в точных и естественных науках, например в физике.

Для определения трудности задания используется следующая формула:

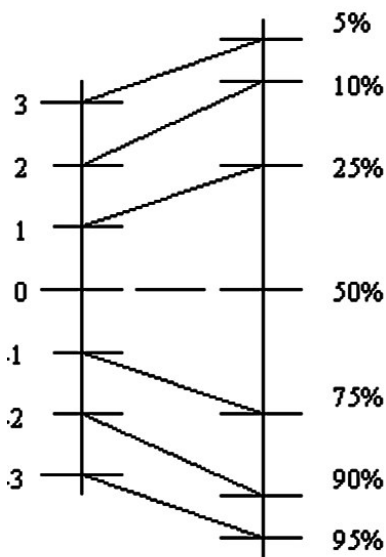
$$d = \ln \left(\frac{1-P}{P} \right).$$

Следовательно, чем легче задание, тем меньшее значение оно имеет по шкале Раша.

Некоторую сложность может вызывать интерпретация данных в шкале Раша. Значение ноль по шкале Раша отражает ситуацию, когда знания (способности) тестируемого соответствуют трудности задания, а вероятность правильного ответа на него составляет 50%. При уменьшении трудности задания вопроса на 1 логит вероятность правильного ответа возрастает приблизительно до 75%. При увеличении трудности задания на 1 логит вероятность правильного ответа уменьшается примерно до 25%.

Иначе значение «ноль логитов» можно интерпретировать следующим образом: 50% тестируемых правильно отвечают на задание, 50% — неправильно. Значение $-2,94$ соответствует «простому» заданию, на который 95% тестируемых отвечают правильно, и лишь 5% — неправильно. Значение $+2,94$, напротив, соответствует «трудному» заданию, на который лишь 5% тестируемых отвечают правильно (см. рис.).

Таким образом, модель Раша оценивает «сложность» каждого задания и «способность» тестируемого в логитах (единицах интервальной шкалы Раша).



Преимущества модели Раша

Модель Раша обладает следующими преимуществами в сравнении с порядковыми шкалами, в частности шкалой Лайкерта:

- ◆ Модель обеспечивает шкалирование заданий в соответствии с их трудностью.
- ◆ Единица измерения шкалы Раша, представляющая собой результат преобразования исходных данных, имеет стандартизованный характер, что даёт возможность объективно сравнить результаты, полученные по разным методам и выборкам.
- ◆ Данные представляются в интервальной шкале.
- ◆ Модель обеспечивает статистику соответствия (*fit statistics*) каждого задания и от-

ветов каждого испытуемого, что является чрезвычайно важным инструментом оценки точности результатов измерений, существенно дополняющим классические методы (коэффициент альфа Кронбаха, коэффициенты лямбда Гутмана и др.).

◆ Задания, которые каким-либо испытуемым оставлены без ответа, не представляют трудностей при использовании модели Раша, т.е. пропуски не критичны для обработки данных.

◆ Измерения по модели Раша не зависят от выборки (*sample free*) и набора заданий (*item free*), т.е. теоретически они не зависят от состава выборочной совокупности и состава теста. Это позволяет получать объективированные результаты, что, во-первых, положительно отражается на качестве измерений, а во-вторых, следует принципу объективности как одного из критериев научного исследования. По этой причине измерения, основанные на модели Раша, называют «объективными измерениями» (*objective measurement*) и даже «фундаментальными измерениями» (*fundamental measurement*).

Разработаны специальные программы для конвертирования исходных данных в шкалу Раша и для выполнения всех необходимых расчётов (например WINSTEPS, RUMM, BiLog, MultiLog). Процессы, ре-

ализованные в этих программах, достаточно сложны и громоздки. Чтобы проиллюстрировать общий подход к реализации подобных процедур, рассмотрим простейший алгоритм преобразования исходных данных в шкалу Раша. Для простоты начнём с дихотомической шкалы, в которой измерение представлено двумя оценками, например 1 и 0. Примером такой шкалы может являться тест для проверки знания учебного предмета, в котором испытуемому требуется выбрать один правильный ответ из двух ответов каждого задания. Правильный ответ кодируется 1, неправильный — 0.

Предположим, что в тест включено 7 заданий. Допустим, что на задания ответили 7 испытуемых, которых обозначим латинскими буквами А, В, ... G. Для конвертирования исходных данных в шкалу Раша необходимо определить общую сумму баллов (S_1), набранных испытуемыми, и общую сумму баллов по каждому заданию теста (S_2). Далее определяется максимальное значение, которое *мог бы* набрать испытуемый ($Max1$): $Max1 = 1 \times 7 = 7$. Для дихотомической шкалы это значение соответствует количеству заданий, на каждое из которых может быть дан правильный ответ. Отношение S_1 к $Max1$ может быть интерпретировано как вероятность правильного отве-

та (P_1). Это значение определяется для каждого испытуемого. В итоге для каждого испытуемого определяется его балл по

$$\text{шкале Раша } D_1 = \ln \left(\frac{P_1}{1 - P_1} \right).$$

Например, при проверке знаний или подготовленности более высокие значения D_1 соответствуют лучшей подготовленности и наоборот.

Аналогично вычисляется максимальное количество баллов, которое может быть приписано каждому заданию ($Max2$). В данном случае, максимальное значение по каждому заданию равно 1, поэтому $Max2 = 1 \times 7 = 7$. Далее для каждого задания вычисляется

отношение $P_2 = \frac{S_2}{Max2}$ и значение

$$D_2 = \ln \left(\frac{1 - P_2}{P_2} \right).$$

В последней формуле числитель и знаменатель поменялись местами (что эквивалентно смене знака перед логарифмом). В классическом варианте модели Раша, которая разрабатывалась в контексте именно тестовых измерений, это было сделано для более «естественной» интерпретации результатов. В таком случае более высокие значения D_2 соответствуют большей трудности тестового задания и наоборот. Следовательно, чем более трудным яв-

ляется задание, тем меньше вероятность правильного ответа на него и, следовательно, тем меньшее количество испытуемых отвечает на него правильно.

ния, позиции шкалы Лайкерта могут соответствовать оценкам, полученным испытуемым, например по трёх- или пятибалльной шкале.

Испытуемые	Задания теста									
	1	2	3	4	5	6	7	S_1	P_1	D_1
A	1	0	1	1	0	1	0	4	0,57	0,29
B	0	1	0	1	1	0	0	3	0,43	-0,29
C	0	1	1	1	1	1	0	5	0,71	0,92
D	0	1	1	1	1	1	1	6	0,86	1,79
E	1	0	1	0	1	1	0	4	0,57	0,29
F	0	0	1	1	0	1	1	4	0,57	0,29
G	1	1	1	1	1	0	0	5	0,71	0,92
S_2	3	4	6	6	5	5	2			
P_2	0,43	0,57	0,86	0,86	0,71	0,71	0,29			
D_2	0,29	-0,29	-1,79	-1,79	-0,92	-0,92	0,92			

Такой подход к конвертированию данных может быть расширен и на многопозиционные порядковые шкалы, в частности, на шкалу Лайкерта. В отличие от дихотомической шкалы, эта шкала имеет не один пороговый переход «правильно-неправильно», а несколько. Например переходы между позициями «всегда» и «часто», «часто» и «иногда», «иногда» и «никогда». Переход от одной позиции к другой может рассматриваться как более или менее «сложный» для испытуемого, что отражает степень соответствия (*agreeability*) позиции действительному положению вещей. В контексте тестирова-

Например, первый пороговый переход моделируется как вероятность выбора 1 («всегда») вместо 0 («часто») по каждой градации оценки за выполнение задания теста:

$$P_{ni1} \left(x = \frac{1}{B_n, D_i, F} \right) = \frac{e^{(B-D_i+F_i)}}{1 + e^{(B-D_i+F_i)}}$$

где P_{ni1} – вероятность выбора испытуемым n позиции «всегда» (1) вместо «часто» (0) по i -му заданию; F_1 – сложность первого порогового перехода; D_i – трудность задания теста.

Рассмотрим пример. Имеется 7 заданий теста. Для оценки ответов используется 5-позиционная шкала Лайкерта; позиции кодируются от 1 до 5. Пусть на задания ответили 40

испытуемых. Исходные данные (частоты оценок) представлены в таблице.

Из таблицы видно, что при ответе на первое задание 5 испытуемых получили оценку 1, 8 — 2, 11 — 3, 10 — 4 и 6 чел. — 5. В наиболее простом варианте ал-

Аналогично вычисляются агрегированные оценки всех заданий теста.

Здесь мы вновь сталкиваемся с вопросом об интерпретации результатов. В данном случае задания, оцененные большим числом, имеют меньшее

№№ заданий теста	1	2	3	4	5	S	P	D	D_1
1	5	8	11	10	6	124	0,62	-0,49	0,49
2	9	7	10	9	5	114	0,57	-0,28	0,28
3	9	8	9	10	4	112	0,56	-0,24	0,24
4	7	9	9	8	7	119	0,595	-0,38	0,38
5	3	4	11	12	10	142	0,71	-0,90	0,90
6	5	9	9	9	8	126	0,63	-0,53	0,53
7	12	13	8	4	3	93	0,465	0,14	-0,14

горитм перевода исходных данных в шкалу Раша следующий. Рассчитывается балл соответствующей градации ответа (*item counts*) как сумма произведений количества испытуемых, получивших определённую оценку, и её балла (S). Например, для первого задания: $5 \times 1 + 8 \times 2 + 11 \times 3 + 10 \times 4 + 6 \times 5 = 124$. Максимальный балл для каждого задания составляет $40 \times 5 = 200$. Тогда «вероятность» правильного ответа на первое задание составит

$$P = \frac{124}{200} = 0,62. \text{ Первоначальный}$$

(до коррекции) шкальный балл этого задания у Раша (в классическом варианте) определяется как

$$D = \ln\left(\frac{1-P}{P}\right) = -0,49.$$

значение, что полностью соответствует классической («тестовой») интерпретации измерений на основе модели Раша. В ситуациях измерения через отношения респондентов более удобным и понятным может быть использование формулы

$$D = \ln\left(\frac{P}{1-P}\right). \text{ Очевидно, что ре-}$$

зультаты измерения будут одинаковыми по абсолютной величине и отличаться лишь знаком (значения D_1 в таблице).

Ещё раз отметим, что при использовании описанного метода конвертирования данных заданий теста, на которые респонденты не ответили по каким-то причинам, не являются препятствием для использова-

ния модели Раша. Задания, по которым не получены ответы, не учитываются при подсчёте максимально возможного балла для каждого пункта. Например, если по первому вопросу один из испытуемых не ответил, то значение максимального балла для этого пункта составляет не $40 \times 5 = 200$, а $39 \times 5 = 195$. В основном процесс конвертирования данных не отличается от описанного выше.

Приведённые примеры иллюстрируют лишь базовую и самую простую процедуру преобразования данных в шкалу Раша. В более сложных алгоритмах, которые реализованы в WINSTEPS и других специальных программных продуктах по *Rasch Measurement*, учитываются распределение и размах данных, делается поправка на размер выборки, рассчитываются ошибки измерений, оценивается согласованность исходных

данных и измерительной модели, вычисляются показатели надёжности измерений.

Примечания

При подготовке статьи частично использованы результаты исследований, выполненных при поддержке Российского гуманитарного научного фонда (проект №06-02-60201 а/г).

Автор благодарен профессору В.С. Аванесову за возможность участвовать в Московской Летней школе (2-8 июля 2006 г.), что позволило существенно обогатить содержание материала, подготовленного для данной статьи.

Более подробно материал данной статьи изложен в учебном пособии: И.Н. Дубина. Математические основы эмпирических социально-экономических исследований. Барнаул: Изд-во Алт. гос. ун-та, 2006.