РАЗЛИЧИЯ МЕЖДУ ИСХОДНЫМИ ТЕСТОВЫМИ БАЛЛАМИ И ИЗМЕРЕНИЯМИ[1]

Б.Д. Райт,
Дж.М. Линека
rmt@rasch.org

Сокращённый перевод
Г.И.Смирновой и *А.В.Смирнова*
smirnova_g_i@mail.ru

Статья посвящена ключевым различиям между полученными в научном эксперименте исходными баллами и итоговыми результатами измерений. Приводятся наглядные примеры таких различий. Анализируются значения делений используемых в педагогических измерениях шкал. Определяется историческая необходимость перехода от результатов подсчета исходных баллов к измерениям. Аргументируется выбор точки отсчета шкалы. Рассматриваются способы измерения явлений, считающихся не измеряемыми.

Данные наблюдения представляются в порядковой шкале

Все результаты научных наблюдений являются порядковыми, если, конечно, они даны не в номинальной шкале. Точные науки начинаются с определения тех событий и условий, которые при более детальном рассмотрении должны быть подсчитаны. Результаты подсчета иногда называют «исходными баллами» (raw scores) для того, чтобы отличить их от статистически обработанных данных — например, взвешенных или шкалированных посредством моделей измерения баллов. Последние, собственно, и называются измерениями. Обычно исходные тестовые баллы называют короче — эти просто «баллы» (scores), которые являются ни чем иным, как следствием подсчёта числа правильных ответов испытуемых на задания теста, Это только начало измерений, но ещё не сами измерения. Если исходные баллы ошибочно назвать «измерениями», то далее возникает трудный вопрос — где исходные тестовые баллы, а где подлинные результаты измерения.

1 *Wright B.D., Linacre J.M.* Differences between scores and measures. Rasch Measurement Transactions, 1989, 3:3 p. 63. Эту небольшую статью было решено печатать ввиду несомненной актуальности поднимаемых в ней вопросов.

В некоторых работах это называлось шкалой упорядоченной классификации (Прим. ред.)

Счет является начальным этапом квантификации результатов, лежащих в основе получаемых затем измерений. Истоки измерения относятся к счету. Самый элементарный уровень — это подсчет случаев наличия интересующего признака или отсутствия. Информации может быть получено гораздо больше, если интересующие признаки градуированы и упорядочены на интересующем континууме. Что открывает возможность не только посчитать число проявляемого признака, но и число уровней, по которым количество признаков может быть ранжировано на каждом фиксированном уровне.

Когда шкала состоит из градаций: «ни одного» (none), «много» (*plenty*), «все» (*all*), то очевидный порядок этих категорий позволяет использовать их в качестве уровней оценивания. Уровню «ни одного» может быть присвоено число 0 (ноль), уровень «много» — число 1, а уровень «все» — числом 2. Эти числа не соотносятся, напрямую, с числами, которые могут стать оценками для каждого уровня проявления интересующего свойства. Например, уровень «много» может быть выражен числом 20 или 40, и это никак не меняет того факта, что градация «много» на этой шкале отстоит от значения «ни одного» всего на одну ступень (единицу).

Все классификации являются качественными. Некоторые классификации могут быть упорядочены, и потому представляются в шкале более высокого уровня, чем в номинальной[2]. Другие классификационные признаки — такие как, например, пол испытуемых — обычно не поддаётся осмысленному упорядочению, хотя иногда это и может быть сделано. Но это не значит, что номинальные данные не могут иметь «исследовательской» ценности. Это лишь означает, что номинальные данные не являются измерениями в общепринятом смысле этого слова.

Числа, представляющие каждый уровень классификации объектов, ничего не говорят ни о расстоянии между этими уровнями (категориями), ни об условии, которому должны удовлетворять все задания, включенные в категорию. Для «счетного ряда» не играет никакой роли тот факт, что для одного вопроса деления шкалы могут иметь значения «ни одного» (*none*), «много» и «все», а для другого — «ни одного» (*none*), «почти ни одного» (*almost none*) и «все» (*all*).

Измерения всегда представляются на интервальной или пропорциональной шкалах. Интервальная шкала

Результаты наблюдения в любой науке никогда не являются «измерениями» ещё и потому что из-

мерение подразумевает и требует наличия концепции измерения, выделение измеряемого свойства, и необходимых элементов измерительной системы, целесообразность использования которых оказывается доказуемой.

Счёт и измерения часто путают

Счёт лежит в основе примитивно построенной пропорциональной шкалы. Он начинается с нуля и продолжаются со значением «плюс один» (*one more*). Но свойства, которые подсчитываются, являются специфическими, а не общими, конкретными, поэтому измерения предпочтительней производить по степени сходства данных. Иногда следующее значение «плюс один» означает небольшое возрастание, (как в примере шага со ступени «ни одного» (*none*) на ступень «практически ни одного» (*almost none*). Иногда следующее значение включает в себя большое возрастание, как, например, из ступени «ни одного» (*none*) к ступени «все» (*all*). Поэтому всё, что мы можем сделать — это прибавить еще одну ступень. Чтобы получить значение ступени опытным путем, мы должны создать измерительную систему, основанную на систематизации множества результатов наблюдений. Для этого требуется провести измерительный анализ результатов исследования первичных наблюдений, обобщающих исходные данные.

От результатов наблюдений — к измерениям

Необходимость перехода от результатов подсчетов к измерениям отмечал Э. Торндайк ещё в 1904 году. Л.Л. Тэрстон частично решил эту проблему в 1920 году. А в 1953 году Г. Раш разработал изобрел модель, которая является необходимой составляющей измерений в любой науке. Г. Раш полагал что задание теста только тогда полезно для измерения знаний испытуемых, когда для всех испытуемых оно одинаково сложно, а испытуемый пригоден для отбора заданий, когда он показал стабильный уровень знаний. Так же Г. Раш отметил, что результаты взаимодействия испытуемого и задания не могут быть полностью предопределены, но всегда должны содержать вероятностные элементы. Это приводит к следующему условию (стохастическое распределение Л.Л. Гуттмана):

- чем выше уровень знаний у испытуемого, тем на большее количество заданий он отвечает;
- чем сложнее задание, тем меньшее количество испытуемых на него отвечает;
- модель Раша предназначена для перевода числовых данных в измерения с учетом этих необходимых условий.

Выбор точки отсчета

«Измерения» (measurement) включают в себя подсчет стандартных единиц измерения. Графически это описывается расстоянием между точками на линии. В такой шкале нет абсолютного нуля. В качестве нуля обычно принимается среднее значение шкалы. По-видимому, выбор интервальной шкалы является скорее теоретическим, чем практическим. Логарифмы преобразуют любую относительную шкалу в интервальную, а экспонента — наоборот, преобразует любую интервальную шкалу в относительную.

Одномерность

Свойство одномерности придаёт смысл процессу измерения какого-либо одного свойства личности. Особенность модели Раша заключается в том, что эта модель может использоваться для измерения, казалось бы, неизмеряемых данных. Такое сочетание слов иногда вызывает недоумение, хотя работа с такими данными является подлинным смыслом измерений. Ценность модели Раша заключается в том, что с ее помощью можно построить одномерную модель интересующего свойства, поскольку этого требует сама модель. Одномерность является концепцией измерения, а не реальным свойством данных.

Ни один из существующих тестов не может считаться совершенно одномерным. Так же , как и не бывает эмпирических данных, которые абсолютно отвечают требованиям модели Г. Раша. Хотя в практике часто принимается решение пробовать применить одномерную модель измерения. Пригодность данных к которой оценивается эмпирически, специальным методом, основанным на применении метода c^2-квадрат. На английском языке это называется *fit statistics*. Если данные соответствую модели, то это является некоторым указанием на валидность результатов. Процесс проверки свойства одномерности не имеет конца. С каждым новым применением теста получаются новые результаты, подтверждающие или отрицающие искомое свойство одномерности данных.

Если тест включает в себя задания по медицине и по правоведению, то экзаменатор вправе считать этот тест непригодным для измерения знаний того и другого. И это вполне оправдано, потому что любой человек может привести количественные или качественные аргументы, доказывающие невозможность совмещения в одном гомогенном тесте двух таких разных учебных дисциплин. После получения интервальной шкалы целесообразно провести статистический анализ пригодности теста для измерения интересующего свойства.