УДК 159.9.01
DOI: DOI: 10.26140/anip-2020-0904-0095

В ЧЕМ ОШИБОЧНОСТЬ ПРИМЕНЕНИЯ ФАКТОРНОГО АНАЛИЗА ТЕСТОВ, СООТВЕТСТВУЮЩИХ ТЕОРИИ ОТВЕТОВ НА ТЕСТОВЫЕ ЗАДАНИЯ (IRT)?

SPIN-код: 6174-3930
AuthorID: 947445

Равен Джон, почетный профессор университета Эдинбурга, Печского университета (Венгрия), Католического Люблинского университета (Польша)
Эдинбургский университет
(Шотландия, Эдинбург, EH3 6QH, Грейт Кинг-стрит, 30, e-mail: jraven@ravenfamily.co.uk)
Фугард Энди, Ph.D в области психологии, старший преподаватель кафедры Исследовательских методов социальных наук
Лондонский университет
(Великобритания, Лондон, Тауэр-Хамлетс, кафедра психологии, e-mail: a.fugard@bbk.ac.uk)

Аннотация. Многие исследователи, которые знакомы с теорией ответов на тестовые задания (ТОТЗ, Item Response Theory - IRT), (или со шкалами Раша и Гутмана), знают, что применение факторного анализа в попытке оценить внутреннюю согласованность или одномерность, таких тестов, имеют тенденцию давать недостоверные результаты. К сожалению, это известно немногим из тех, кто работал только с тестами, разработанными с использованием классической теории тестирования (КТТ). Такое положение дел привело к тому, что многие исследователи пришли к серьезным ошибочным выводам, применив факторный анализ к матрицам корреляций между заданиями, составляющими тесты, построенные на основе ТОТЗ (IRT). Данная статья иллюстрирует проблему, на примере факторного анализа данных, генерируемых компьютером и имитирующих то, что было бы получено при использовании этой архетипической формы ТОТЗ-теста как портновский метр или метровая линейка для измерения роста или способности прыгать в высоту.

Ключевые слова: теорией ответов на тестовые задания (ТОТЗ, Item Response Theory - IRT), задания тестов, корреляция трудности заданий, одномерность и многомерность теста, измерительная рулетка, однофакторный и многофакторный анализ, ошибочные результаты факторного анализа

WHAT'S WRONG WITH FACTOR-ANALYSING TESTS CONFORMING TO THE REQUIREMENTS OF ITEM RESPONSE THEORY?

Raven John, Honorary Professor University of Edinburgh, University of Pecs (Hungary), Catholic University of Lublin (Poland)
University of Edinburgh
(Scotland, Edinburgh EH3 6QH, 30 Great King Street, e-mail: jraven@ravenfamily.co.uk)
Fugard Andy, Senior Lecturer in Social Science Research Methods BEng, MSc PhD
University of London
(Great Britain, Birkbeck College, Department of Psychology, e-mail: a.fugard@bbk.ac.uk)

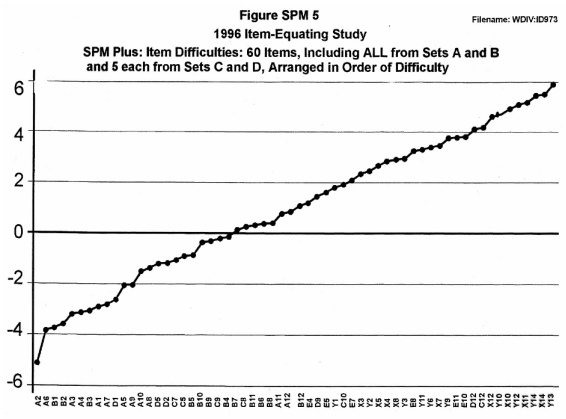
Abstract. Many researchers who are familiar with Item Response Theory (IRT) (or variants such as Rasch or Guttman scales) know that applying factor analysis in an attempt to assess the internal consistency, or unidimensionality, of such tests tends to yield misleading results. Unfortunately, few of those who have worked only with tests developed using Classical Test Theory are aware of this. This has resulted in many researchers coming to seriously misleading conclusions when they have applied factor analysis to the matrices of correlations between the items constituting IRT-based tests. The current paper illustrates the problem by factor-analysing computer-generated data simulating that which would be obtained from using that archetypical form of an IRT test – a tape measure or meter stick – to measure height or the ability to make high jumps.

Keywords: Item Response Theory, test items, correlation of difficulty of items, unidimensionality and multidimensionality of the test, measuring tape, one-factor and multifactor analysis, inadequate results of factor analysis

Цель этой статьи - проиллюстрировать как можно более наглядно то, что относительно хорошо известно исследователям, знакомым с прикладной теорией ответов на тестовые задания, но неизвестно большинству других. Речь идет о том, что приложение процедур факторного анализа, которые рутинно используются для установления «одномерности» (или наоборот неодномерности) тестов, сконструированных согласно классической теории тестирования (КТТ), приводят к «бессмыслице», в случае применения для проверки тестов, построенных в соответствии с ТОТЗ (IRT) (или с шкалами Раша и Гуттмана).

Процедуры IRT нацелены на создание теста, состоящего из набора заданий, будь то задания «проверки способностей» или задания на «лайкертовы отношения» или «личностных» заданий, так что респонденты пройдут или одобрят все задания (пункты) вплоть до такого, которое указывает максимум их способности (или максимальную силу их эмоций), и не пройдут все последующие задания (пункты) данного теста.

Данные на рисунке ниже, относящиеся к тесту «Стандартные прогрессивные матрицы Равена Плюс», могут служить примером теста, который почти соответствует этому критерию. По сути, респонденты дают правильные ответы на все вопросы (задания), вплоть до самых сложных, которые они могут решить, и не отвечают на остальные, хотя для полной демонстрации этого утверждения требуются дополнительные данные [1].



Исследование равномерного нарастания трудности заданий, 1996 г.
«Стандартные прогрессивные матрицы Равена Плюс».

Трудности заданий: 60 заданий, включающие все из блоков A и B, и по 5 из блоков C и D, расположенных в порядке возрастания трудности.

Такой тест будет похож на линейку для размера ноги или портновский метр.

В идеале, «итоговым баллом» будет уровень или сложность последнего пройденного или подтвержденного задания. Аналогом может служить наибольшая высота планки, которую смог преодолеть прыгун в высоту.

Такие итоговые баллы заметно отличаются от баллов, полученных с помощью сомнительной процедуры подсчета количества пройденных респондентом заданий среди тех, которые составляют один «фактор» или «размерность» «личностного» теста. Сомнительны и заключения о том, что тот, кто пройдет больше пунктов, имеет более высокую уровень рассматриваемого признака. Классическая теория тестирования рекомендует эту процедуру, призывая исследователя установить, что все задания, входящие в то, что представлено как конкретная область «способностей» или «черт» личности, умеренно коррелируют друг с другом, но мало коррелируют, если вообще коррелируют, с теми, которые, как говорят, отражают другое измерение или область.

К сожалению, идеал единственной надежной цифры как показателя результата теста IRT обычно не может быть достигнут. Это привело к созданию баллов, основанных на подсчете количества правильно решенных или пройденных заданий. Эти оценки внешне аналогичны оценкам, полученным в тестах, разработанных в соответствии с классической теорией тестирования, но их теоретическая основа сильно отличается.

Все это послужило источником бесконечной путаницы.

В этой статье мы показываем, что, когда процедуры, обычно применяемые для оценки внутренней согласованности тестов, разработанных в соответствии с классической теорией тестирования, применяются к матрицам корреляций между заданиями тестов, соответствующими требованиям ТОТЗ, они *всегда* и обязательно декларируют, что эти тесты многомерны.

Подчеркнем это еще раз:

- применение процедур, рекомендуемых классической теорией тестирования, к тестам на основе ТОТЗ (IRT) *всегда* приводит к выводу о том, что для учета наблюдаемой картины корреляций между заданиями теста необходимы три или более факторов. А из этого делается заключение, что исследуемый тест является многомерным.

Первая часть этого наблюдения верна. Но это никоим образом не подтверждает вывод о многомерности теста. «Факторы», которые процедура правильно указывает как необходимые для вычисления максимума объяснимой дисперсии в корреляционной матрице, на самом деле являются «силовыми» факторами, каждый из которых характеризуется заданиями аналогичной трудности и отличается от тех факторов, которые характеризуются группами более простых или более трудных по этому единственному фактору заданий.

Наша демонстрация производит то, что, как подчеркнула Целевая группа АПА (Американская психологическая ассоциация) по Статистическому выводу, *делается слишком редко*. Она основана на возвращении к матрице корреляций между заданиями (пунктами), составляющими любой тест [Отчет АПА не публиковался, но о нем можно прочесть в работе: L. Wilkinson and Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604].

Мы ожидаем, что многим исследователям будет трудно полностью оценить актуальность того, что мы делаем.

Многие, если не большинство, исследователи довольствуются тем, что применяют *готовые компьютерные пакеты факторного анализа* к своим наборам данных и считывают такие показатели, как количество и характер факторов, полученных в различных условиях, долю дисперсии, приходящуюся на каждый фактор, нагрузки по каждому фактору, набор факторных оценок и т.д.

Несмотря на рекомендации Целевой группы APA по статистическому выводу, эти исследователи редко изучают корреляционные матрицы вида задание-задание или задание-тест, которые лежат в основе результатов, выдаваемых применяемыми статистическими пакетами.

В этом и состоит проблема. Как говорится, дьявол кроется в деталях.

Итак, в этой статье мы сосредоточимся именно на этих скрытых матрицах корреляций (фактически ковариаций) между заданиями и их решениями.

Мы хотим проиллюстрировать, насколько неуместно стремиться к демонстрации «одномерности» (или наоборот многомерности) тестов, которые действительно удовлетворяют требованиям ТОТЗ (или шкалам Раша или Гуттмана), с помощью применения критериев классической теории тестирования матрицам корреляции между заданиями, составляющими тест.

Рекомендации Целевой группы АПА по Статистическому выводу, требуют от исследователей внимательно изучить *корреляционную матрицу*, лежащую в основе их факторного анализа, и задаться вопросом, на основании какой модели эта матрица составлена, прежде чем приступить к факторному анализу. Нарушая эти требования, такие исследователи обычно затем выделяют явное содержание заданий с высокими «нагрузками» на каждый фактор и маркируют факторы на этом основании ... хотя на самом деле группы заданий состоят в основном из заданий одинаковой трудности, различаемых всего лишь как «менее *трудные*» или «более *трудные*» задания.

Повторим: мы начинаем с рассмотрения матриц корреляций, а не результатов, выдаваемых статистическими компьютерными программами.

Более конкретно, мы проиллюстрируем неуместность попытки использовать факторный анализ для установления одномерности шкал IRT в целом, используя со ссылкой на общедоступную измерительную рулетку или метр, используемые для измерения роста или способности совершать прыжки в высоту.

Обычная измерительная рулетка представляет собой идеальную шкалу IRT. «Люди» «проходят» (правильно решают задание) все сантиметровые отметки («задания») до той, которая регистрирует их рост (или высоту самой высокой планки, которую они могут перепрыгнуть) и «терпят неудачу» (не могут достигнуть) до всех сантиметровых отметок («заданий») выше этой отметки.

Мы представим две иллюстрации того, что происходит, когда кто-то пытается проанализировать матрицы корреляций между заданиями теста, которые возникают в результате взаимной корреляции этих «заданий» (сантиметровые отметки).

Одна основана на сгенерированных компьютером данных, приближенных к тем, которые были бы получены, если бы 36-сантиметровая рулетка использовалась для измерения высоты случайной выборки из тысячи особей одного вида или линии животных, имеющих средний рост 18 см. То есть компьютер был запрограммирован на создание набора данных, в котором среднее значение будет 18, а «баллы» распределены по всей шкале из 36 пунктов в соответствии с «гауссовым» распределением (авторы используют дискретное распределение, поэтому называют его «гауссовым» условно, т. к. нормальное распределение является непрерывным. В данном случае, видимо, используется биномиальное распределение. – *прим. перев.*). Естественно, каждая «оценка» (то есть «рост») предполагала, что конкретное животное «прошло» каждую сантиметровую отметку (задание) до этой точки и не смогло достичь каждой сантиметровой отметки над ней.

Вторая имитация использовала прямолинейное (равномерное) распределение, то есть предполагалось, что одинаковое количество животных должно получить каждую «оценку» от 0 до 36. Это будет соответствовать распределению, которое могло бы быть получено, если бы рулетка использовалась для измерения роста всех животных и объектов в определенном месте (или если психологический тест был построен для получения характеристической кривой заданий теста (ICC), которая показала бы, что тест имеет одинаковую различительную способность во всем рабочем диапазоне).

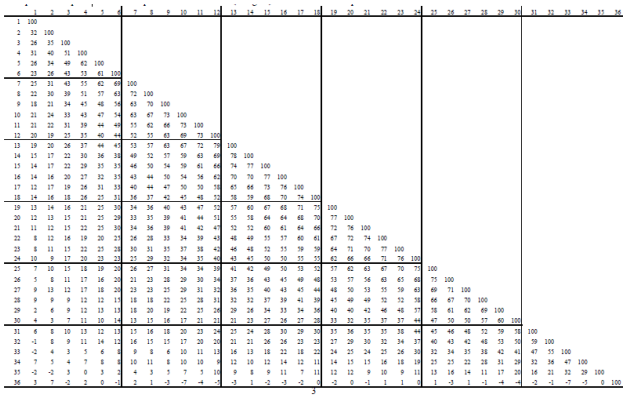
Фактически, хотя это и признается немногими из тех, кто утверждает, что использует IRT, на самом деле необходимо составить такую выборку, если нужно получить надежную статистику по элементу, попадающему в хвост нормального распределения (то есть задание, которое выполняют очень немногие люди) [2].

Первая имитация дает данные, приближенные к тем, которые обычно получают путем проведения типичного теста, основанного на ТОТЗ (IRT), для широкого круга респондентов и последующего факторного анализа результатов.

Но результаты второй имитации еще яснее иллюстрируют основную мысль, которую здесь необходимо высказать.

В таблице 1 показана корреляционная матрица, полученная путем генерации данных прохождения/непрохождения («итоговых баллов») в соответствии с 36 сантиметровой рулеткой со средним значением 18 и распределением Гаусса между верхним и нижним значениями рулетки.

Таблица 1. Корреляции между данными «прошел/не прошел» для отметок 36 сантиметровой рулетки. (Компьютерно генерированные данные. Гауссово распределение общего «балла» со средним значением 18. N= 1000. Десятичная точка опущена. «Респонденты» «проходят» все «задания», которые «ниже» их итогового «балла» (роста) и «не проходят» все последующие «задания».)



Корреляции, прилегающие к главной диагонали, стремятся к 0.99, потому что это отражает весьма высокую вероятность того, что в нашей случайной выборке из 1000 животных, если животное, например, более 1 см ростом, оно почти наверняка будет и выше, чем 2 см. Корреляция между прохождением отметки 1 см и отметки 2 см будет высокой, но не идеальной. С другой стороны, мало что можно сказать о шансах прохождения отметки 35 см, зная, что пройдена отметка 2 см. Прогностическая достоверность от одного бита информации к другому будет низкой, то есть, как показано в дистальных углах матрицы (далеких от главной диагонали), корреляция между дистальными «заданиями» (сантиметровые отметки) будет почти нулевой. Иначе говоря, доля дисперсии «балла» по второму заданию, которая объясняется дисперсией по первому заданию, мала (Отклонения от такой «идеальной» матрицы будут объяснены позже).

Подбор «однофакторного» («главного компонента») факторно-аналитического решения к этим данным дает результаты, представленные в таблице 2.

Таблица 2. Факторные нагрузки (Factor Loadings), полученные с помощью Однофакторного решения с помощью применения компьютерной программы факторного анализа к корреляционной матрице, представленной в Табл. 1 (“SS loadings” - sum of squared loadings – сумма квадратов нагрузок; Proportion Var – доля объясненной дисперсии.)

Final Score/ cm. mark	Loadings on Factor 1
1	0.19
2	0.23
3	0.29
4	0.38
5	0.43
6	0.47
7	0.54
8	0.57
9	0.62
10	0.65
11	0.67
12	0.72
13	0.77
14	0.76
15	0.81
16	0.81
17	0.81
18	0.79
19	0.81
20	0.81
21	0.79
22	0.76
23	0.76
24	0.74
25	0.72
26	0.67
27	0.63
28	0.59
29	0.53
30	0.44
31	0.46
32	0.40
33	0.31
34	0.22
35	0.14
36	
SS Loadings	13.24
Proportion Var	0.37

Большинство исследователей, хорошо разбирающихся в факторном анализе, но не знакомых с теорией ТОТЗ, интерпретируют эти результаты как значимые, а не просто то, что корреляционная матрица не может быть статистически оценена (или «объяснена») с помощью «оценок» по одному основному фактору (она составляет только 13% дисперсии), но также и то, что тест не является «одномерным».

Стоит немного отвлечься, чтобы сказать об этой таблице еще кое-что очень важное.

Хотя таблица фактически показывает нагрузку каждого элемента (галочка) на первый главный компонент, она, по сути, дает корреляции между заданиями и «итоговым баллом». Теперь, согласно классической теории тестов, между каждым заданием и итоговым баллом должна быть высокая корреляция (индекс дискриминации).

Итак, согласно этой теории, задания (галочки) в центре измерительной линейки - это «хорошие» задания, а те, что на концах, - «плохие». Это, конечно, чепуха, опять же из-за применения неадекватной модели измерения.

Первый вывод сделан правильно. Второй – нет (хотя многое зависит от понимания термина «одномерный»).

Затем они приступают к выявлению дополнительных факторов. Результаты трехфакторного решения показаны в таблице 3.

Возвращаясь к таблице 1, можно увидеть, что же произошло.

Говоря простым языком и не обращая внимания на технические детали, которые отличают факторный анализ от кластерного, многофакторный анализ «пытается» определить группы элементов, которые имеют высокую корреляцию друг с другом, но низкую корреляцию с элементами из других групп (так называемые «факторы»

или «кластеры»).

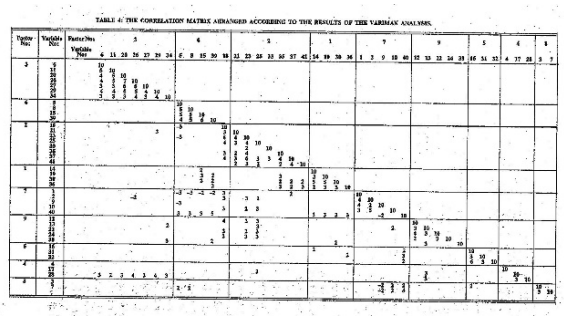
Таблица 3. 3-факторное решение на основе корреляционной матрицы из таблицы 1. (Cumulative Var – накопленная объясненная дисперсия)

Table 3

3-factor solution from factor analysing the correlation matrix in Table 1.

Final Score/ cm. mark	Loadings on: Factor 1	Factor 2	Factor 3
1		0.30	
2		0.37	
3		0.51	
4		0.64	0.12
5	0.10	0.69	
6	0.14	0.71	0.10
7	0.20	0.79	
8	0.25	0.78	
9	0.35	0.72	
10	0.39	0.72	
11	0.45	0.65	
12	0.56	0.58	
13	0.63	0.57	
14	0.68	0.46	
15	0.74	0.40	0.15
16	0.75	0.33	0.18
17	0.77	0.28	0.21
18	0.76	0.21	0.23
19	0.77	0.17	0.31
20	0.76	0.14	0.37
21	0.69	0.15	0.41
22	0.66		0.48
23	0.61	0.12	0.53
24	0.54	0.10	0.62
25	0.48	0.10	0.68
26	0.41		0.72
27	0.33	0.12	0.72
28	0.28		0.75
29	0.19	0.10	0.76
30	0.12		0.66
31	0.14	0.12	0.64
32		0.13	0.61
33			0.51
34			0.39
35			0.21
36			
SS Loadings	7.47	6.01	5.85
Proportion Var	0.21	0.17	0.16
Cumulative Var	0.21	0.37	0.54

Для полноты и контраста, вот корреляционная матрица, полученная в результате изучения ответов 4000 подростков на анкету о стремлении к карьере после того, как она была перестроена в соответствии с результатами многофакторного факторного анализа [3].



Ясно, что элементы делятся на несколько групп или кластеров, которые, как правило, мало пересекаются (хотя фактор 4 явно может быть объединен с фактором 3).

Таким образом, программа фактически сказала: «Смотрите. Здесь, в середине, находится группа элементов, которые сильно коррелируют друг с другом и относительно меньше с элементами в двух других группах элементов на нижнем и верхнем концах шкалы. Итак, ребята, вам нужно как минимум 3 фактора, чтобы учесть эти данные».

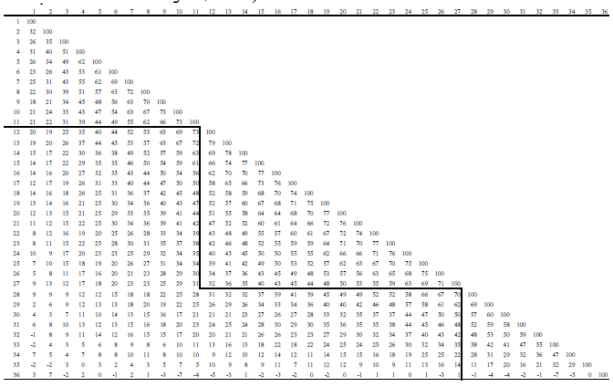
Способ, которым программа «сгруппировала» элементы, показан в Таблице 4.

Конечно, мы могли бы продолжать и действительно *продолжили,* то есть извлекли 5 факторов. (Заметим, что, если бы мы извлекли 36 факторов, мы фактически смогли бы использовать факторные нагрузки для точного воспроизведения исходной матрицы корреляций.)

Но мы сделали уже достаточно, чтобы доказать выдвинутое положение: мы же *знаем*, что рулетка одномерная.

Применение процедур, основанных на классической теории тестирования, к данным, полученным с ее же помощью, чтобы установить, являются ли эти данные одномерными, ошибочно. Коротко, факторный анализ группирует вместе предметы аналогичной *трудности* и заявляет, что они представляют собой основные факторы или «размерности» в пределах тест. Со времен Гуттмана (наиболее известного своей работой по анализу «Шкалограмм», которые, по сути, является вариантом IRT) и позже эти факторы были известны как факторы «мощности».

Таблица 4. Корреляции в таблице 1 сгруппированные в кластеры процедурой 3-факторного анализа. (Десятичная точка опущена.)



Но, не заметив этого, тысячи исследователей, *не* знакомых с целями и моделью измерения, лежащими в основе IRT-тестов, и *не* выполнявших рекомендации рабочей группы APA по статистическим выводам (которые могли быть еще не опубликованы, на момент проведения исследований), требующей «сначала взглянуть на исходные данные», из-за этого совершили ужасное преступление, которое повлияло на мышление целых поколений исследователей.

Как указывалось ранее, они исследовали явное содержание заданий с высокими нагрузками на эти 3 или 5 факторов и, исходя из этого, делали вывод, что в тесте было 3 или 5 (или более) «типов» заданий ... другими словами, тест был беспорядочным и объединял 3, 5 или более «независимых» размерностей 1, 2, 3.

На самом деле, использование слова «независимый» само по себе свидетельствует о более чем небольшом незнании того, как работает факторный анализ, потому что последовательные факторы в действительности не являются независимыми от тех, что были ранее, но фактически представляют собой следующую лучшую попытку исправить с помощью еще одного единственного фактора, ошибки, которые были сделаны при предположении, что матрица корреляции [фактически ковариационная] может быть «объяснена» в терминах ранее извлеченных факторов.

Термин «одномерный» на самом деле крайне неоднозначен. Подробное обсуждение термина см. в [4].

Эта ошибка привела многих исследователей к выводу, что Прогрессивные матрицы Равена измеряют несколько разных характеристик.

Хотя в каком-то смысле это может быть правдой, суть в том, что это не демонстрируется путем факторизации элементов. Здесь не место для подробного обсуждения этих вопросов, но, поскольку интересом к ним привел авторов к подготовке этой статьи, то стоит отметить, что обзор многочисленных демонстраций на основе ТОТЗ (IRT), которые различают качественно разные типы заданий, составляющих Прогрессивные матрицы Равена, измеряют один и тот же базовый континуум способности (по аналогии с тем, что качественно разные типы заданий, составляющих геологическую шкалу, используемую для измерения «твердости», все измеряют одну и ту же основную переменную «твердость»), можно найти в главе 1 книги Raven J. & Raven, C.J. (2008) Uses and Abuses of Intelligence.(Opus Cit.) [5] (http://eyeonsociety.co.uk/resources/UAIChapter1.pdf), а примером исследова-

ния, которое может выявить другие «измерения» трудности (аналогичные дополнительным измерениям твердости кирпичей) может служить статья [6].

Второй подход

Приведенная выше история во всех существенных деталях соответствует тому, что на самом деле происходит, когда исследователи применяют факторный анализ к корреляционным матрицам, полученным из взаимных корреляций заданий теста, основанного на ТОТЗ (IRT).

Таким образом, большинство читателей уже получили все, что они могут извлечь из этой статьи.

Таблица 5. Корреляция между данными прохождения/непрохождения для сантиметровых отметок на 36 см рулетке. Компьютерно сгенерированные данные для равного количества объектов, достигших каждой отметки (и не прошедших выше) от 1 до 36. n=3,700 («Респонденты» «проходят» все «задания» (сантиметровые отметки) до того, которое указывает их рост, и «не проходят» все последующие «задания». Десятичная точка опущена.)

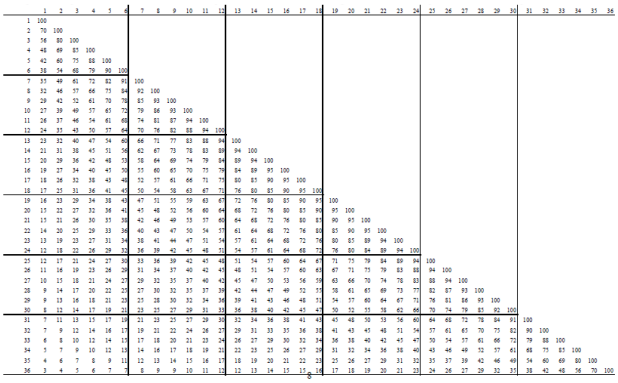


Таблица 6. 5-факторное решение факторного анализа данных Таблицы 5.

cm. mark	Factor 1	Factor 2	Factor 3	Factor 4
1				0.46
2				0.64
3		0.13		0.77
4		0.19		0.85
5	0.11	0.25		0.89
6	0.12	0.34		0.87
7	0.13	0.44		0.81
8	0.15	0.54		0.72
9	0.16	0.64		0.63
10	0.18	0.73		0.53
11	0.20	0.80		0.44
12	0.23	0.85	0.11	0.36
13	0.27	0.86	0.13	0.30
14	0.31	0.85	0.15	0.26
15	0.37	0.81	0.16	0.23
16	0.43	0.76	0.17	0.21
17	0.49	0.70	0.18	0.20
18	0.56	0.64	0.19	0.20
19	0.64	0.56	0.20	0.19
20	0.70	0.49	0.20	0.18
21	0.76	0.43	0.21	0.17
22	0.81	0.37	0.23	0.16
23	0.85	0.31	0.26	0.15
24	0.86	0.27	0.30	0.13
25	0.85	0.23	0.36	0.11
26	0.80	0.20	0.44	
27	0.73	0.18	0.53	
28	0.64	0.16	0.63	
29	0.54	0.15	0.72	
30	0.44	0.13	0.81	
31	0.34	0.12	0.87	
32	0.25	0.11	0.89	
33	0.19		0.85	
34	0.13		0.77	
35			0.64	
36			0.46	

Однако в корреляционной матрице, представленной в таблице 1, есть что-то непонятное. А именно, почему корреляции, близкие к диагонали, так далеки от 0,99 на верхнем и нижнем концах шкалы?

Ответ таков: поскольку данные моделирования были сгенерированы для получения гауссова распределения по длине рулетки, то получено всего несколько «респондентов» выполнивших «самое легкое» и «самый трудное» задания.

Поэтому, анализы были повторены с выборкой, дающей одинаковое количество «животных», для каждого «роста» от 1 до 36 см.

Результаты представлены в таблицах 5 и 6.

История похожа на ту, что мы получили ранее, за исключением того, что корреляции между «самыми легкими» и «самыми трудными» «заданиями» намного выше.

Можно подумать, что эти результаты не имеют отношения к основным положениям данной статьи.

Но это не так.

На самом деле исследователи чрезвычайно часто проводят анализ заданий, будь то на основе классической теории тестирования или теории ТОТЗ, используя данные, полученные из тестируемых популяций (часто ошибочно называемых «выборками»), которые дают лишь очень узкий диапазон оценок. Это означает, что слишком мало людей (а часто и вовсе ни одного) с оценками в хвостах распределения, чтобы можно было вычислить значимую статистику по элементам (см. врезку 2).

Это значительно усугубляет ошибки, которые допускаются, когда исследователи пытаются интерпретировать свои результаты, особенно полученные с помощью рутинного применения факторного анализа.

СПИСОК ЛИТЕРАТУРЫ:

1. Raven, J., Prieler, J. & Benesch, M. (2008). Using the Romanian data to replicate the IRT-based Item Analysis of the SPM+: Striking achievements, pitfalls, and lessons. Chapter 5 in J. Raven & J. Raven (Eds.) Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metrics. Unionville, New York: Royal Fireworks Press; Edinburgh, Scotland: Competency Motivation Project; Budapest, Hungary: EDGE 2000; Cluj Napoca, Romania: Romanian Psychological Testing Services SRL. См. также http://eyeonsociety.co.uk/resources/UAIChapter5.pdf

2. Appendix to Chapter 3 The need for, and development of, the SPM Plus in J. Raven & J. Raven (Eds.) (2008) Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metrics. Opus Cit http://eyeonsociety.co.uk/resources/UAIChapter3.pdf

3. Raven, J., Ritchie, J., & Baxter, D. (1971). Factor analysis and cluster analysis: Their value and stability in social survey research. Economic and Social Review, Vol. 2, 367-391.

4. Hattie, J. (1985) Methodology Review: Assessing Unidimensionality of Tests and Items. Applied Psychological Measurement 9 139-164.

5. Raven J. & Raven, C.J. (2008) Uses and Abuses of Intelligence. (Opus Cit.) (http://eyeonsociety.co.uk/resources/UAIChapter1.pdf)

6. DeShon, R.P., Chan, D., & Weissbein, D.A. (1995). Verbal Overshadowing Effects on Raven's Advanced Progressive Matrices: Evidence for Multidimensional Performance Determinants. Intelligence, 21, 135-155.

Статья поступила в редакцию 02.10.2020
Статья принята к публикации 27.11.2020