ВВЕДЕНИЕ В RASCH MEASUREMENT[1]

Everett V. Smith Jr., Ph.D.
Karen M. Conrad, Ph.D.
Karen Chang, Ph.D.
Jo Piazza, R.N.
evsmith@uic.edu.

Сокращённый перевод с английского Надежды Пракиной
ГОУ ВПО «Уральский государственный университет
им. А.М. Горького», г. Екатеринбург

В статье рассматриваются ограничения статистической (классической) теории тестов (CTT), некоторые преимущества Rasch Measurement (RM), сравниваются результаты факторного анализа и Rasch Measurement.
206 женщин заполнили анкеты почтового опроса членов национальной ассоциации медсестёр по тесту Рикмана (Ryckman, 1982). Факторный анализ выявил наличие только одного фактора. Коэффициент надёжности тестовых результатов среди испытуемых оказался равным 0,78. Оценка адекватности этих данных модели Г. Раша подтвердила одномерный характер распределения результатов, с коэффициентом надёжности 0,76. Метод Раша использовался также для оценки качества шкалы, оценки валидности результатов и диагностики затруднений испытуемых. Рассмотрены различия оценок коэффициентов т.н. внутренней состоятельности шкалы, а также вопросы практической полезности методологии Раша и факторного анализа.

Ключевые слова: *теории тестов, модели Раша, тест, Winsteps.*

1

Correspondence concerning this paper should be addressed to Everett Smith, University of Illinois at Chicago, 1040 W. Harrison Street, M/C 147, Chicago, IL, 60607. Email evsmith@uic.edu.
This manuscript was originally published as Smith, Jr., E.V., Conrad, J.M., Chang, K. & Piazza, J. (2002). An introduction to Rasch measurement for scale development and person assessment. Journal of Nursing Measurement, 10, 189-206. Used by permission from Springer Publishing.

Введение

Модели Г. Раша (Rasch Measurement, RM) и математической теории тестов (Item Response Theory, IRT) стали популярными во всём мире и в различных сферах. Эта популярность касается не только сфер педагогики и психологии, но и медицины (Beck & Gable, 2001). Она связана также и с некоторыми ограничениями статистической (классической) теории тестов CTT. Чтобы проиллюстрировать полезность моделей и теории, результаты Rasch Measurement сравниваются с результатами факторного анализа, полученными на одних и тех же данных. Мы надеемся, что наш пример поможет донести до более широкой аудитории медиков-исследователей информацию о том, как модель измерения Раша может быть использована для повышения качества результатов.

Ограничения классической теории тестов

Хотя в этой статье рассматриваются только некоторые ограничения статистической (классической) теории тестов (CTT), этот обзор является важным для понимания причин полезности и других теорий педагогических измерений. Одно из ограничений CTT — зависимость статистических характеристик заданий и теста — напр., бисериальный, точечно-бисериальный и классический коэффициенты корреляции заданий с суммой баллов, доли правильных ответов и значения коэффициентов надёжности — от уровня подготовленности испытуемых. Другое ограничение CTT — невозможность определить теоретическую вероятность правильного ответа испытуемого на то или иное задание теста. Различающиеся единицы измерения меры трудности заданий и уровня подготовленности испытуемых препятствуют прогнозированию результатов ответов множества испытуемых на множество заданий.

Кроме того, многие статистические модели, основанные на CTT, предполагают интервальной уровень шкалы измерений. Хотя на самом деле исходные тестовые баллы представляются, как обычно полагают, в порядковой шкале. В порядковой шкале не имеют заметного смысла математические операции, необходимые для вычисления средних арифметических и стандартных отклонений. Анализ данных, представляемых на порядковой шкале, может привести к ошибочным выводам о влиянии или невлиянии факторов (Merbitz, Morris, & Grip, 1989; Wright & Linacre, 1989; См. также Michell, 1990; 1999; and 2002).

Другие ограничения статистической (классической) теории тестов:

1) трудность в сравнении достижений у тех испытуемых, кто тестируется по параллельным вариантам теста;

2) трудности определения ошибки измерения, дифференцированной для испытуемых различного уровня подготовленности. Как известно, в классической теории ошибка измерения принимается одинаковой для всех испытуемых, хотя известно, что высокие и низкие значения тестовых баллов измеряются с меньшей точностью;

3) невозможность сравнивать оценки, полученные одним и тем же тестом на различных выборках испытуемых;

4) трудность сравнения результатов испытуемых, набравших одинаковые баллы, но имеющих различающиеся профили (векторы строки). Например, сравнимы ли по уровню знаний те, кто получили одинаковое число баллов, но один правильно ответил на первые пять лёгких заданий, а второй — на такое же число трудных заданий? Исходный тестовый балл у них будет один и тот же, но качественно это совсем другой балл;

5) невозможность внести поправки в различные источники погрешностей измерений, например, в случаях учёта погрешности, привносимой каждым экспертом в свои оценки (хотя на этот случай есть другая теория, так называемая generalizability theory).

Несмотря на упомянутые ограничения CTT, эта теория измерений все ещё доминирует в литературе. Мы предполагаем две причины, которые могут быть рассмотрены как достоинства или недостатки, в зависимости от точки зрения. Во-первых, CTT основана на так называемых «слабых» предположениях. Слово «слабые» надо понимать в позитивном смысле, так, что предположения, лежащие в основе CTT, могут сравнительно легко подтверждаться в большинстве практических случаев; следовательно, методы классической теории тестов оказываются пригодными во многих ситуациях. Во-вторых, CTT — теория, преподаваемая во многих вводных курсах педагогических измерений.

Модели Rasch Measurement: краткий обзор

Для преодоления отмеченных ограничений в процессе измерения могут использоваться модели датского математика Георга Раша. (См.: Wright, 1977.) Эти математические модели применимы при свойствах одномерности и свойства аддитивности данных. Эти свойства вытекают

из свойства матрицы, где представлены оценки множества испытуемых, полученные ими по множеству заданий. Взаимодействие этих двух множеств образует данные, обладающие свойством т.н. «совместной аддитивности» (conjoint additivity).

Одномерность означает, что все задания теста измеряют один и тот же общий концепт (концептуально образуемое свойство личности). Если тест содержит субтесты, то требование одномерности шалы относится к каждому субтесту.

Свойство conjoint additivity касается единицы измерения. Эта единица называется «логит». Данные отвечают этому свойству при условии, если уровень трудности заданий измеряется сходными единицами: это логит трудности заданий и логит уровня подготовленности испытуемых. Данные, представленные логитами, отвечают требованиям интервальной шкалы измерений и могут подвергаться методам так называемой параметрической статистики. По этому поводу есть некоторые разногласия. (См. Merbitz, Morris, & Grip, 1989.)

Если данные подходят для модели, то параметры испытуемых и параметры заданий измеряются на общей шкале логитов (Rasch, 1960; Perline, Wright, & Wainer, 1979). Общая шкала способствует вычислению вероятности правильного ответа испытуемого как функций от взаимодействия между испытуемым и заданием. В результате у нас есть ясная картина того, что можно ожидать от испытуемого в выполнении отдельных заданий и теста в целом.

Правильное использование метода Г. Раша позволяет отделить оценки испытуемых от оценок трудности заданий, и наоборот. Это известное свойство Rasch Measurement называется по-английски separability of parameter estimates (Hambleton, Swaminathan, & Rogers, 1991; Van der Linden & Hambleton, 1997), что можно перевести как «независимость оценок заданий от испытуемых и оценок испытуемых от параметров заданий».

Методы оценки пригодности измерения по модели Раша

Когда параметры модели Раша определены, например, методом максимальной правдоподобности, эти параметры далее используются для вычисления вероятности правильного ответа на каждое задание теста. Чем больше эмпирические данные совпадают с теоретическими, тем лучше соответствие данных модели. Критерием соответствия является значение хи-квадрат. Задания, не соответствующие этому критерию, рассматриваются как не соответствующие гипотезе одномерности шкалы

и обычно подлежат переработке или удалению.

Свойства модели Раша проявляются только при условии соответствия данных этой модели.

Рассматриваются два варианта оценки пригодности, данных для Rasch Measurement. Первый вариант называется «Item fit» statistics. Эта статистика позволяет проверить пригодность заданий для одномерной модели измерения по Рашу. Она применяется для выбраковки заданий, которые не отвечают гипотезе одномерности. Оставляемые для разрабатываемого теста задания имеют различную вероятность правильного решения хорошо подготовленными испытуемыми. Задания, не отвечающие этому требованию, либо улучшаются, либо удаляются. Второй вариант оценки пригодности данных — «Person fit statistics» оценивает соответствие множества имеющихся испытуемых уровню трудности заданий, используемых для разработки теста.

В соответствии с проверяемой гипотезой испытуемые с более высоким уровнем подготовленности должны с высокой вероятностью давать правильные ответы на сравнительно лёгкие для них задания. В этом варианте проверяется соответствие реально полученных векторов — строк (профилей) испытуемых ожидаемым профилям. Подходящим считается такой профиль, где эмпирическая вероятность правильного ответа у сильного испытуемого выше при ответах на лёгкие задания и снижается по мере роста трудности задания. Испытуемые, отвечающие иным образом, считаются не соответствующими данной задаче измерения. Несоответствующими считаются также испытуемые, подготовленность которых не соответствует уровню трудности заданий теста.

Компьютерная программа WINSTEPS

WINSTEPS — компьютерная программа (J.M. Linacre, 2000), обеспечивающая полный цикл измерений по модели Раша. Эта программа даёт возможность вычисления обоих упомянутых вариантов статистики для оценки пригодности результатов измерения. В этой программе использованы два вида статистик для проверки гипотез пригодности заданий для разрабатываемого теста. Первая статистика называется по-английски «Infit», которая считается менее чувствительной к менее вероятным ответам испытуемых на задания с уровнем трудности, заметно отличающихся от уровня их подготовленности. Вторая статистика называется «Outfit», которая, напротив, чувствительна к несоответствию ответов испытуемых на задания. Программа считает значения хи-квадрат, делённые на число

так называемых степеней свободы. Значения обеих статистик, меньше единицы, рассматриваются как указание на недостаточную вариацию данных (Wright & Linacre, 1994). Значения больше единицы указывают на значительную вариацию. Значения в пределах от 0,6 до 1,4 считаются приемлемыми для интересующей нас шкалы (Wright & Linacre, 1994). Правда, исследователям надо быть готовым к тому, чтобы признать, что выбираемые значения пригодности статистик достоверности зависят от цели и от значимости измерения. В социально значимых случаях принимаются более строгие требования. (В нашем исследовании для решения вопроса пригодности заданий для теста мы использовали значения от 0,7 до 1,3, которые несколько консервативнее значений, рекомендованных Б. Райтом и М. Линека (B.D. Wright and J.M. Linacre (1994). Если значения соответствия заданий уровню подготовленности испытуемых связаны с валидностью результатов измерения по модели Rasch Measurement, то значения стандартных ошибок указывают на меру точности измерения. Эти значения стандартных ошибок используются для вычисления доверительных интервалов, в пределах которых находятся интересующие значения параметров испытуемых и параметров заданий. (Wright & Stone 1988).

Преимущества Rasch Measurement

Преимущества Rasch Measurement представлены в табл. 1.

До начала изложения результатов нашего исследования важно обсудить свойства использованной здесь шкалы градуированных оценок, именуемой по-английски rating scale, и показать, как применяются измерения по модели Раша для данных такой шкалы. Значения шкал градуированных оценок относятся к данным, получаемым на порядковой шкале. И, соответственно, они не применимы для методов параметрической статистики.

Например, на десятибалльной шкале можно экспертно оценить профессиональный уровень выполняемой работы. В этом случае крайние значения шкалы будут эквивалентны таким характеристикам, как «абсолютно непрофессионально» и «высоко профессионально». Тем самым все остальные уровни будут располагаться между этими крайними градациями, с предположительно равными интервалами.

1	2	3	4	5	6	7	8	9	10
абсолютно непрофессионально							высоко профессионально		

Таблица 1

- Если данные соответствуют модели, то в результате процесса измерения данные представляются на интервальной шкале
- Шкала устойчива к потере некоторых исходных данных
- Обеспечивает получение валидных результатов посредством применения статистик адекватности (fit statistics), диагностической информации, карты сравнения уровня трудности заданий с уровнем подготовленности испытуемых. Также даёт информацию о надёжности измерений посредством расчёта стандартных ошибок измерений, оценок параметров заданий и параметров подготовленности испытуемых на одной шкале
- Даёт возможности оценить параметры уровня подготовленности испытуемых независимо от уровня трудности заданий имеющейся выборки заданий
- Даёт возможности оценить параметры уровня трудности заданий независимо от уровня подготовленности выборки испытуемых
- Представляет параметры испытуемых и заданий на одной общей линейной шкале, что помогает критериально-ориентированной и нормативно-ориентированной интерпретации данных. Ставит в фокус исследования отдельные задания и результат отдельных испытуемых, в отличие от CTT, где исследователь имеет дело с обобщённой статистикой
- Возможность уравнивания баллов испытуемых, полученных на параллельных вариантах заданий, измеряющих одно и то же интересующее свойство [2]

Но реально шкала может дифференцировать испытуемых по данному признаку неравномерно. Например, вот так:

1 2 3 5 4 6 7 8 9 10

из чего видно, что получаемые значения данного признака необязательно соответствуют свойству монотонности числовой шкалы. К счастью, модель Раша помогает и в таких случаях неизоморфного отображения.

Применение Rasch Measurement основано на некоторых предположениях:
1. Каждое последующее число отображает большее наличие интересующего свойства. Для проверки правильности такого предположения считается средний арифметический балл. В случае нарушения этого предположения больший уровень латентного свойства не получает адекватного отображения, соответственно, большим числом. (Linacre, 1999). Для устранения таких случаев рекомендуется уменьшить число градаций. Тогда ошибок отображения будет меньше (Linacre, 1997).
2. Чем выше балл, тем большей должна быть вероятность правильного ответа (Linacre, 1997, 1999a; Zhu, Updyke, & Lewandowski, 1997).

2

Более точную информацию о Rasch Measurement можно получить в работах Wright and Stone (1979), Wright and Masters (1982), and Bond and Fox (2001).

3. Адекватности лексики и терминологии изучаемой сферы. Часто появляются неудачные определения, своеобразные названия, которые неоднозначно понимаются испытуемыми. Неадекватность лексем можно выявить в процессе Rasch Measurement посредством анализа результатов Outfit статистики. На такого рода явления могут указывать значения хи-квадрат, большие единицы. Значения хи-квадрат, большие единицы указывают на большие различия в восприятии словесного названия данной градации.

Основываясь на этих предположениях, используемые методы позволяют обнаружить и увеличить долю дисперсии истинных компонентов в общей дисперсии результатов испытуемых, что, в свою очередь, повышает надёжность измерений, корреляцию между результатами Rasch Measurement и внешними критериями.

Методы исследования

Выборка и организация

Была разработана анкета, которая на условиях анонимности ответов была разослана 300 членам ассоциации медсестёр, живущих в одном штате. Число заполненных и возвращённых анкет равно 225. Полностью заполненных и пригодных для исследования оказалось 206, что составляет 69 процентов. Возраст респондентов варьировал от 27 до 71 года, средний возраст 48 лет. По уровню образования большинство медсестёр имели диплом бакалавра.

Инструментарий

Для измерения самооценки уровня физического развития использовалась шкала, разработанная Ryckman, Robbins, Thornton и Cantrell (1982). Эта шкала гетерогенного типа включает в себе две другие шкалы — шкалу самооценки уровня развития физической подготовленности (Perceived Physical Ability (PPA) и шкалу самооценки готовности демонстрировать уровень своей физической подготовленности (Perceived Self Presentation Confidence). В этой статье полученные данные проанализированы при использовании десятибалльной шкалы.

Шкала PPA свидетельствует о том, как респонденты воспринимают свою физическую подготовленность. Коэффициент надёжности результатов по Cronbach's равнялся 0,78. У других авторов значения этого коэффициента варьировали между 0,60 и 0,89 (Bosscher, Van Der, Van Dasler, Deeg, & Smit, 1995; Desmond, 1993; McAuley, 1992; McAuley, Bane, & Mihalko, 1995). Ryckman et all. (1982). Коэффициент надёжности, по-

лученный методом повторного анкетирования для PPA был равен 0,80.

Факторный анализ (ФА)

Факторный анализ (ФА) представляет собой метод выделения таких переменных величин, у которых общая вариация (ковариация) больше, чем вариация с остальными переменными. (Nunnally & Bernsteim, 1994). Цель факторного анализа — уменьшить число переменных, объясняющих (измеряющих) интересующее свойство. Использовался метод главных компонент. Факторизуемость корреляционной матрицы проверялась по критерию Kaiser-Meyer-Olkin (KMO). Для проверки размерности матрицы использовались рекомендации Tabachnick и Fidell (2001). Для определения числа извлекаемых факторов использовался метод сравнительного анализа (Horn, 1965; Thompson & Daniel, 1996).

Этот метод требует генерации матрицы случайных чисел такого же размера. Программный синтаксис для генерации такой матрицы можно найти в работе Thompson & Daniel, (1996). Обе матрицы подвергаются анализу методом главных компонент. Для исследования оставляют только те значения характеристических корней содержательной матрицы, которые больше, чем в матрице случайных чисел. В качестве коэффициенты общности (communalities) использовались квадраты коэффициентов множественной корреляции. Для оценки коэффициента надёжности результатов использовался коэффициент альфа (alpha) Cronbach.

Применение модели Раша

Для оценки качества шкалы — надёжности и валидности получаемых результатов, а также меры соответствия множества испытуемых множеству вопросов анкеты (person and item misfit) использовалась программа WINSTEPS (Linacre, 2000), позволяющая вычислять интересующие параметры данных (RSM; Andrich, 1978; Wright & Masters, 1982). Для анализа по модели Раша могут использоваться данные, собранные на другой шкале, и данные, имеющие распределения, отличные от нормального (Wright & Mok, 2000).

Формула 1 содержит три параметра: воспринимаемая физическая мера подготовленности испытуемого (n), трудность задания () и установка шага-меры (). Для RSM допускается, что расстояние между каждым шагом-мерой — константа во всех пунктах. Если это предположение неадекватно, то можно принять гипотезу о неравномерности интервалов (Wright & Masters, 1982) между градациями шкалы.

$$P(X_{ni}=x)=\frac{\exp\sum_{j=0}^{x}\left[\beta_n-\left(\delta_i+\tau_j\right)\right]}{\sum_{x=0}^{m}\exp\sum_{j=0}^{x}\left[\beta_n-\left(\delta_i+\tau_j\right)\right]},\ x=0,1,\ldots,m, \qquad (1)$$

где $P(X_{ni} = x)$, — вероятность того, что испытуемый n с результатом по градации x по заданию под номером j находится на уровне $m + 1$, и что каждая градация шкалы должна содержать не менее десяти результатов наблюдений. Это требуется для получения статистически устойчивых результатов (Linacre, 1997).

Средние значения по каждой градации располагаются в определённом порядке, средний квадрат Outfit statistic для каждой градации меньше двух (Linacre, 1997).

Результаты факторного анализа

Значение KMO равнялось 0,82, что является указанием на достаточность информации, имеющейся в матрице, составленной по ответам на десять вопросов анкеты. Зна-

Таблица 2. **Упорядоченные значения коэффициентов корреляции заданий с основным (первым) фактором «самооценка» физической подготовленности**

Задание	Суждения шкалы	Значения коэффициента корреляции заданий с фактором
22	Благодаря быстроте, я была способна сделать то, что другие не могли	0,77
6	Я могу быстро бегать	0,66
12	У меня хороший мышечный тонус	0,62
8	Я не чувствую дискомфорта, когда прохожу тесты, включающие физическую ловкость	0,55
2	Я — быстрая и изящная (ловкая)	0,53
19	Моя подвижность помогла мне в некоторых трудных ситуациях	0,46
4	Мое телосложение довольно крепкое	0,45
13	Я испытываю небольшую гордость в моих спортивных успехах	0,44
1	У меня отличные рефлексы	0,31
21	У меня хорошая хватка	0,31

чения характеристических корней имели значения 3,44, 1,16, и 1,01. На одномерность результатов указывают различия между значениями первого корня и остальными, что объясняет 78,47% всей дисперсии. Коэффициенты корреляции вопросов с фактором принимали значения от 0,31 до 0,77 (см. табл. 2.)

Значение коэфициента альфа было 0,78 для этих 10 заданий. Результаты факторного анализа подтверждают гипотезу одномерности шкалы PPA.

Применение модели Раша для оптимизации шкалы

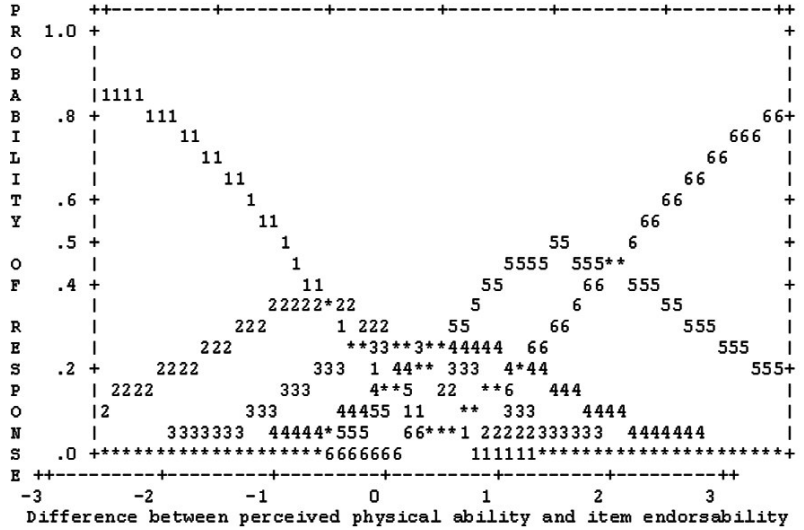
Для вычисления частот каждой градации шкалы, средних арифметических значений и статистики соответствия двух множеств — испытуемых и заданий использовалась программа WINSTEPS. Именно эта программа помогает исследователям выявить затруднения и несоответствия в процессе использования шкалы. Табл. 3 (приводится в оригинале из-за отсутствия подходящей лексической системы на русском языке) содержит информацию о результатах применения шестибалльной шкалы.

Такую же информацию даёт рис. 1, где представлены характеристические кривые ответов испытуемых на каждую градацию шестибалльной шкалы.

Результаты табл. 3 и рис. 1 указывают на необходимость изменения числа градаций шкалы. Вычисления по изменённой шкале данных приведены в табл. 4 и на рис. 2. Эти данные также приведены здесь на языке оригинала. Сокращение числа градаций шкалы не имеет принципиального значения, так как они представляют концептуальные чёткие понятия (напр., сумма соответствия в сравнении с суммой несоответствия).

Таблица 3. **Rating Scale Category Counts, Average Measures, Outfit Mean-Square Statistics and Step Measures for the 6-Point Scale**

Category label	Observed count	Average measure	Outfit MNSQ	Step measures	Category meaning
1	184	–0,66	1,82		Strongly agree
2	285	–0,38	1,08	–1,01	Moderately agree
3	332	–0,14	0,86	–0,41	Slightly agree
4	360	0,17	0,88	–0,02	Slightly disagree
5	585	0,61	0,87	–0,07	Moderately disagree
6	314	1,30	0,94	1,51	Strongly Disagree



Note. 1 represents Strongly Agree; 2 Moderately Agree; 3 Slightly Agree; 4 Slightly Disagree; 5 Moderately Disagree; 6 Strongly Disagree.

Рис. 1. Вероятные кривые для шкалы, состоящей из 6 градаций

Была испробована 5-балльная и 4-балльная шкала. Результаты анализа 4-бальной шкалы показаны в табл. 4 и изображены наглядно на рис. 2. Результаты удовлетворяют нашим требованиям, а потому они послужили основой для последующего анализа.

Таблица 4. **Rating Scale Category Counts, Average Measures, Outfit Mean-Square Statistics and Step Measures for the 4-Point Scale (лексика на русском языке не сформирована)**

Category label	Observed count	Average measure	Outfit MNSQ	Step measures	Category meaning
1	184	–1,20	1,40		Strongly Agree
2	617	–0,44	0,88	–2,07	Moderately/Slightly Agree
3	945	0,65	0,87	–0,30	Moderately/Slightly Disagree
4	314	2,06	0,96	2,37	Strongly Disagree

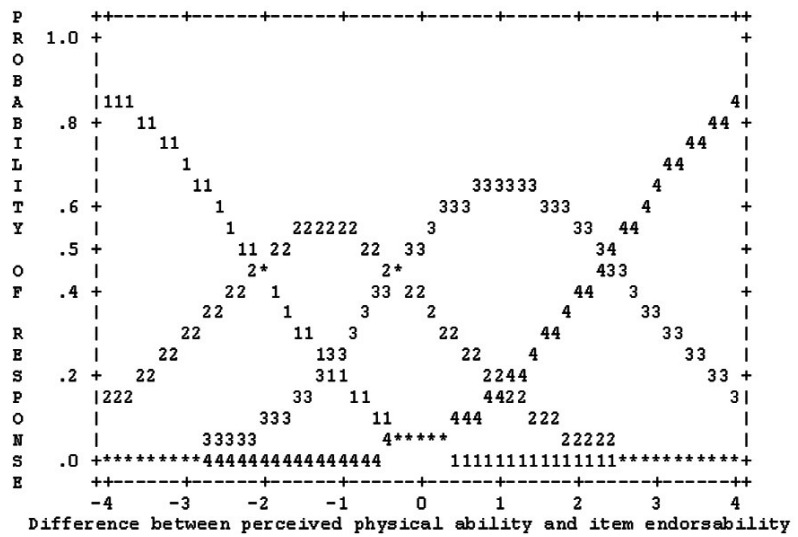
Оценка пригодности и надёжности заданий по модели Раша

Используя оптимальную 4-балльную шкалу оценки ответов респондентов, мы приступили к исследованию размерности данных и пригодности заданий данной шкале оценивания. С этой точки зрения оказалось проблемным задание №22 шкалы. Оно имело средний квадрат (значение дисперсии) менее единицы, а потому признано неинформативным. Удаление этого задания с последующим пересчётом параметров испытуемых и параметров заданий дало наиболее подходящие значения статистик остальных заданий, в пределах от 0,7 до 1,3, что и обеспечивает одномерность шкалы.

Оценка валидности шкалы на основе методологии Раша

Проблему валидности может прояснить рис. 3. На нём справа представлены задания, слева — гистограмма ответов респондентов. Анализ соответствия показывает на значимость отдельных заданий для исследуемого концепта.

С применением остальных 9 заданий был найден коэффициент надёжности ответов испытуемых, равный 0,76 (эквивалент коэффициенту альфа



Note. 1 represents Strongly Agree; 2 Moderately/Slightly Agree; 3 Moderately/Slightly Disagree; 4 Strongly Disagree.

Рис. 2. Характеристические кривые для шкалы с 4-мя градациями

Кронбаха). Это показывает, что задания дифференцируют испытуемых. Коэффициент надёжности, рассчитанный для заданий, оказался равным 0,98. Возможности оценки концептуальной валидности данной иерархии заданий шкалы в рамках

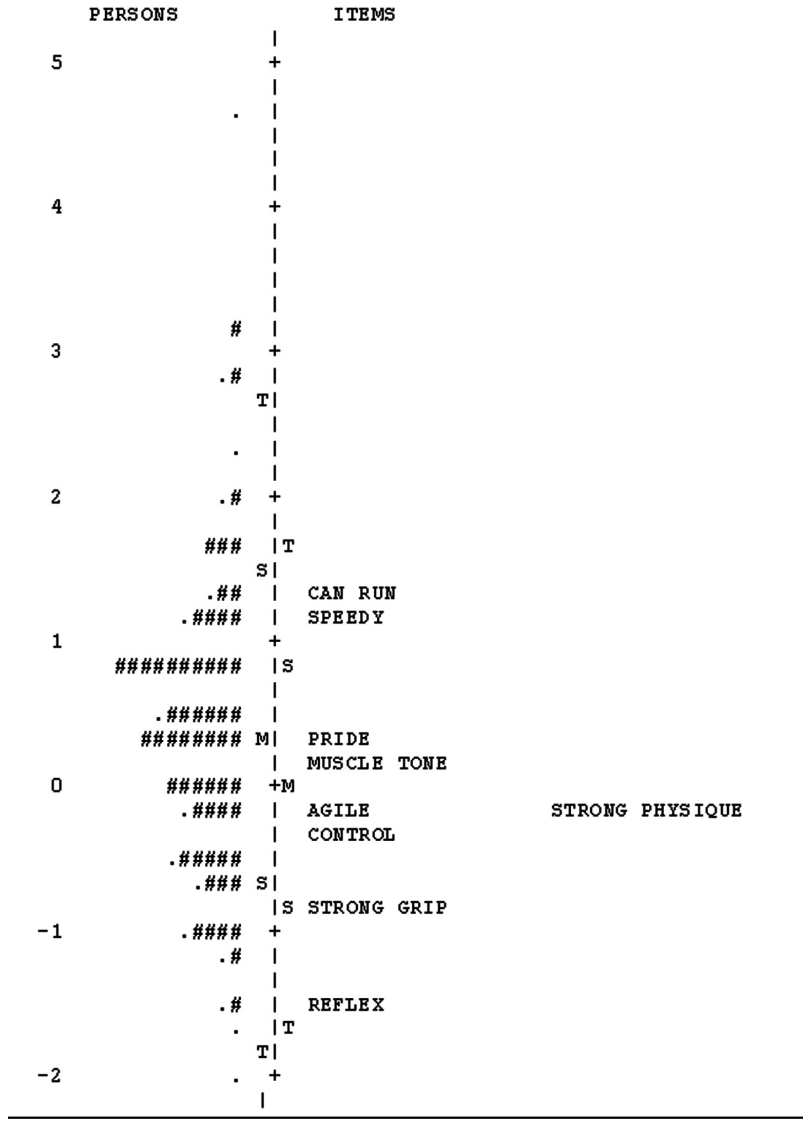


Рис. 3. Сравнительное положение испытуемых и заданий на измеряемом континууме

Таблица 5. **Examples of Person Diagnostic Information from WINSTEPS**

Person A					
response:	4 4 0 1 0	4 0 4 0 0	0 4 4 0 0	0 0 0 4 0	4 0
Z-residual:	X–5 X	X XX	X XX	XXX X	XXX X
Person B					
response:	4 4 0 4 0	4 0 1 0 0	0 3 4 0 0	0 0 0 4 0	4 0
Z-residual:	X X	X–5 XX	X XX	XXX X	X

методологии Rasch — Validity Evaluation дают и другие методы (см. Смит, 2001).

Оценка соответствия испытуемых модели Раша

Табл. 5 отображает информацию о проверке соответствия двух испытуемых применяемой модели измерения.

Обсуждение результатов

Мы показали некоторые ограничения CTT и преимущества использования измерения для проведения надёжного и достоверного исследования. Общий набор данных был проанализирован с использованием модели Раша и факторно-аналитических методов.

Обоими методами установлена одномерность шкалы (читатели могут обратиться к Smith [1996] и Wright [1996] для выявления условий, при которых эти две методологии отличаются (Linacre 1998a; 1998b, Smith, 2002, и Silver, Smith, и Greene, 2001), также условий, при которых эти две методологии могут дополнить друг друга в определении размерности данных). Как отмечено Banjeri, Smith, и Dedrick (1997), выбор между факторным анализом и методами Раша должен производиться на основе целей измерения и методов операционализации интересующих понятий. Оба метода эффективны для удаления заданий, не образующих латентные переменные. Метод Раша помогает лучше определить место задания на континууме, располагать их в определённой иерархии, как это видно, например, из расположения заданий на рис. 3. Ликвидация разрывов в расположении заданий на континууме увеличивает точность измерения при условии, что добавляемые задания измеряют то же самое свойство личности, которое измеряют и остальные задания.

Дополнительные преимущества Rasch Measurement связаны с валидизацией выво-

дов о соответствии заданий, шкал и испытуемых (относительно) цели измерения. Например, в нашем исследовании было показано преимущество четырёхбалльной шкалы по сравнению с шестибалльной шкалой. В последующих исследованиях было бы полезно проверить правильность такого суждения на других выборках испытуемых, тем самым генерализовать полученный вывод (Smith, Wakely, de Kruif, & Swartz, in press).

Что касается подготовленности испытуемых, то Rasch Measurement даёт возможность выявить и оценить необычные профили ответов испытуемых на задания. Дополнительную информацию по вопросам валидности читатели могут получить в работах Smith (2001), где показана связь между Rasch Measurement и идеями Messick (1989; 1995) относительно валидности исследовательских выводов в процессе измерений; это можно видеть на примере т.н. Medical Outcomes Trust (1995).

В большинстве случаев применения факторного анализ

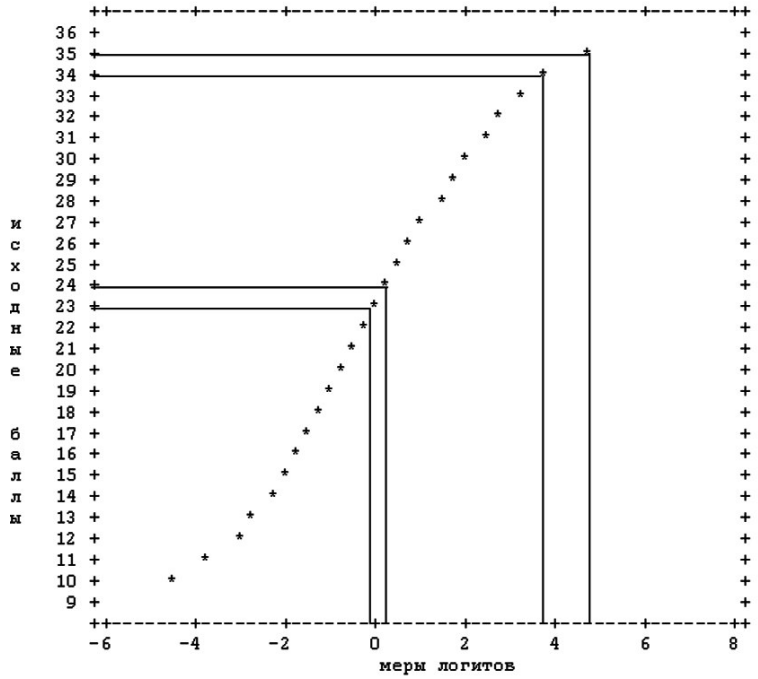


Рис. 4. Соотношение исходных баллов и шкалы логитов

предполагается наличие данных, полученных на интервальной шкале. Однако, как видно из рис. 4, полученного из распечатки расчётов по программе WINSTEPS, исходные результаты, полученные в анкетном опросе на градуированной шкале, соответствуют требованиям только порядковый шкалы. То есть различия в сумме исходных баллов на интервале от 34 до 35 отображаются на шкале логитов большим интервалом, чем различия между 23-мя и 24-мя баллами.

Некоторые читатели могут также заинтересоваться, почему отличаются коэффициент альфа Кронбаха и коэффициент надёжности, считаемый среди испытуемых по методу Г.Раша. Первая причина — расчёт коэффициента альфа производится на данных шкалы исходных данных. А эти данные не образуют линейную шкалу, в то время как сам метод предполагает наличие линейной интервальной шкалы.

Во-вторых, в обычных исходных данных иногда используются минимальные и максимальные значения данных; это случаи, когда у испытуемого ответы на все задания оценены нулём или, наоборот, все ответы правильны. В таких случаях в ответах испытуемого общая и ошибочная дисперсии равны нулю. Включение этих нулевых и максимальных баллов в общую сумму исходных баллов снижает стандартную ошибку измерения исходных данных. Исходя из структуры формулы расчёта коэффициента надёжности, уменьшение ошибочной компоненты измерений означает увеличение коэффициента надёжности измерений. В методике Раша полностью нулевые и единичные вектор-строки испытуемых исключаются из обработки. Linacre [1996, 1999b] объясняет это с метрической точки зрения: стандартные ошибки измерения у таких испытуемых принимают бесконечно большие значения, а потому содержат очень мало информации о реальном уровне подготовленности испытуемого. Очевидно, такое истолкование прямо противоречит содержательно ориентированной интерпретации результатов: чем выше балл, тем выше уровень подготовленности.

Наконец, мы подчёркиваем, что эта статья только слегка затрагивает потенциальные преимущества использования измерения Раша. Мы поддерживаем заинтересованных исследователей приведёнными ссылками для приобретения более точной информации о потенциальных преимуществах применения моделей Раша в своих собственных научно-исследовательских стремлениях.