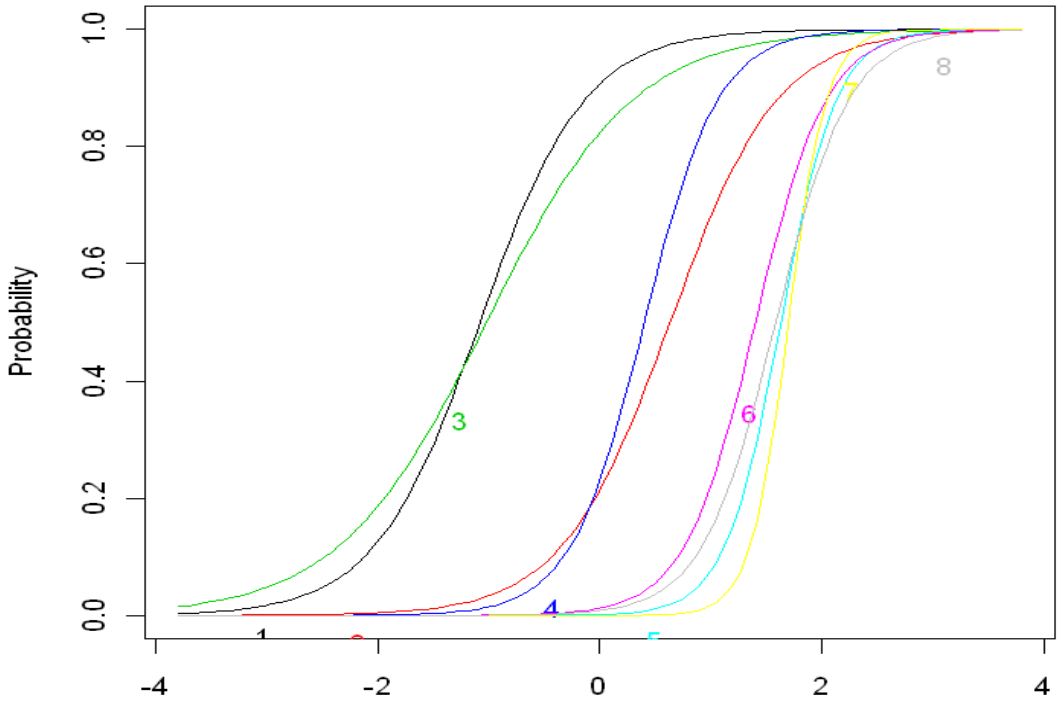
Теория тестовых пунктов

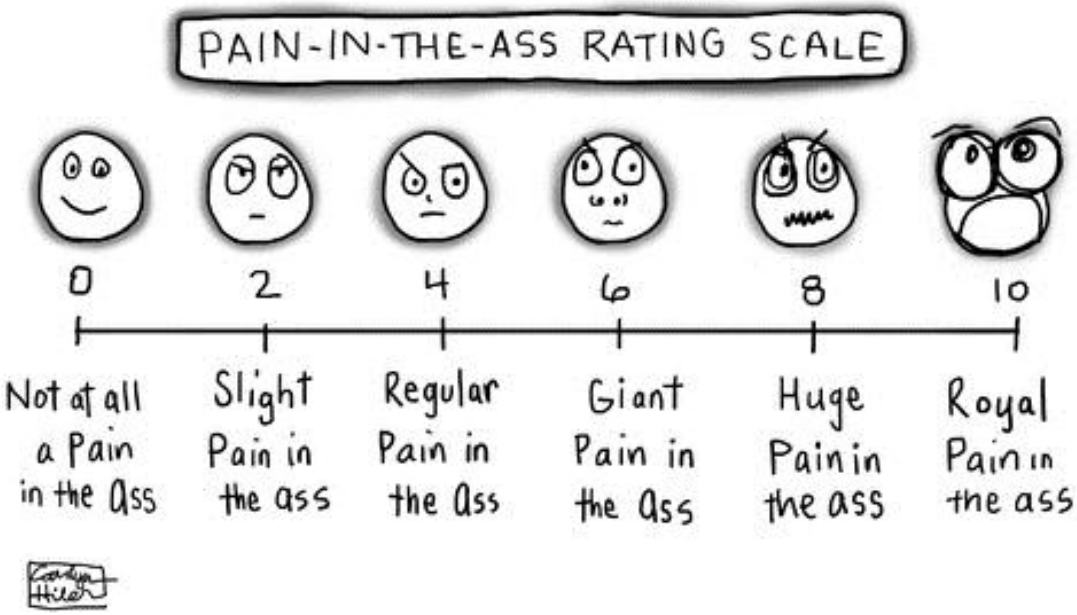
и ее применение в современном тестировании

Анна Браун



Измерения в психологии

- Твердая научная база для физических измерений (температура, масса, давление, и т.д).
- Какова научная база для измерений интеллекта, личности, или психического здоровья?



Оцениваемое качество и Тестовый балл

- Оцениваемое качество и Тестовый балл – это не одно и то же.
- Отношения между ними описывают используя понятия:
  - Валидность – насколько тест измеряет заявленное качество
  - Надежность – точность с которой тест измеряет это качество
  - Объективность – насколько тест справедлив к различным группам в популяции

Классическая теория тестов

$$\mathbf{Y} = \mathbf{T} + \mathbf{E}$$

$\mathbf{Y}$ = тестовый балл
$\mathbf{T}$ = истинный балл (истинная величина измеряемого качества)
$\mathbf{E}$ = погрешность измерения

- В этом уравнении два неизвестных – T и E
  - Чтобы решить это уравнение, нужно сделать некие предположения

Предположения в классической теории

1. Погрешность измерения распределена нормально со средним 0

$$\overline{E} = 0$$

- Из этого следует что

$$\overline{Y} = T$$

  – “Истинный балл – это математическое ожидание в длинной серии повторных измерений с промежуточным промыванием мозгов испытуемым” (Lord & Novick, 1968; *перевод мой*)

Предположения в классической теории

2. Погрешность измерения и истинный балл независимы друг от друга

$$\operatorname{cov}(\mathrm{T}, \mathrm{E}) = 0$$

3. Погрешности измерения в двух тестированиях независимы друг от друга

$$\operatorname{cov}(\mathrm{E}_i, \mathrm{E}_k) = 0$$

Важные результаты

- С помощью базовой формулы Y=T+E и трех предположений были получены важные результаты в классической теории:
  - Формулы для вычисления надежности
  - Формула стандартной погрешности измерения (одна величина независимо от тетстового балла)
  - Коррекция для оценки корреляции между теоретическими качествами (а не ненадежными тестовыми баллами)
  - Формула Келли для предсказания истинного балла на основе тестового балла, среднего бала в популяции и надежности

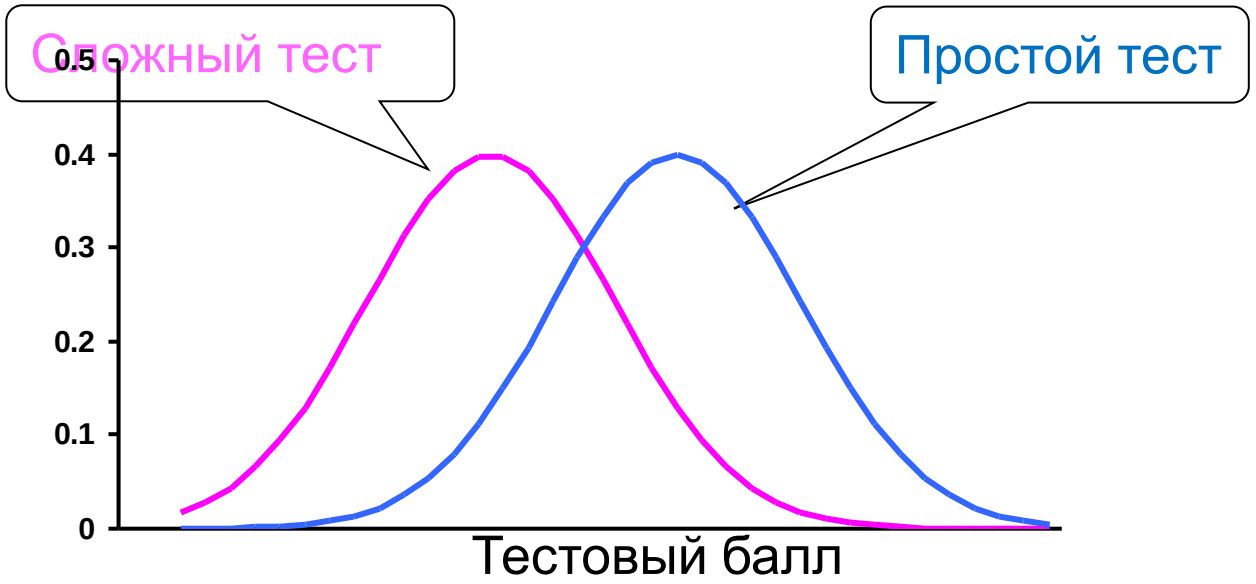
Преимущества классической теории

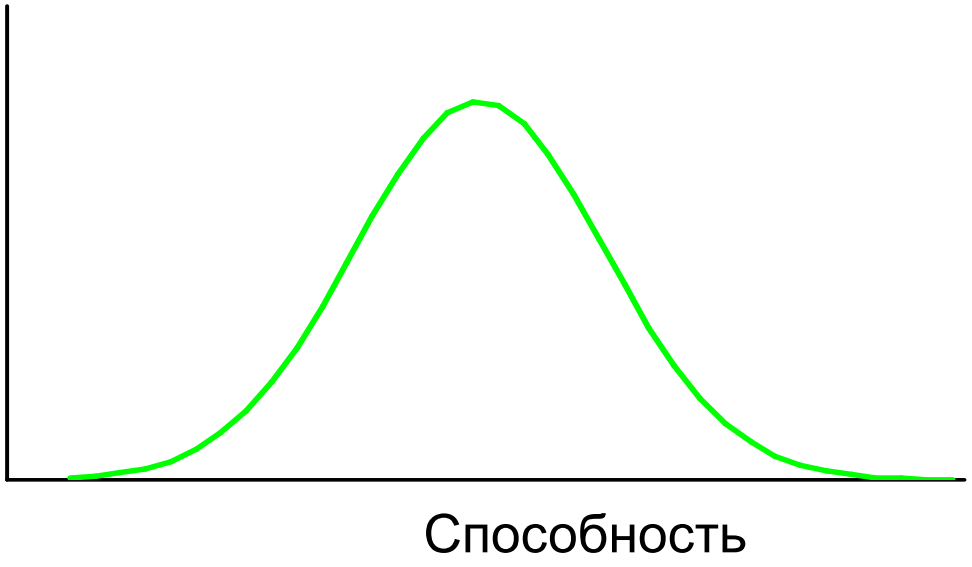
- Простота
- Построена на слабых предположениях (которые легко удовлетворить во многих тестовых выборках), поэтому теория имеет широкое применение

Ограничения классической теории

- Статистические показатели тестовых пунктов зависят от выборки
- Стандартная погрешность измерения не зависит от тестового балла (все баллы предполагают одинаковую погрешность)
- Баллы испытуемых зависят от сложности (и других характеристик) тестовых пунктов
- Моделирование ведется на уровне теста, но моделирование на уровне пунктов необходимо для гибкости использования тестов

Сложность тестовых пунктов и тестовый балл





Чего хотят разработчики тестов?

- Чтобы тестовый балл не зависел от характеристик конкретных пунктов
- Чтобы характеристики тестовых пунктов не зависели от конкретных выборок
- Определить стандартную погрешность для конкретного испытуемого
- Моделирования на уровне пунктов для гибкости утилизации пунктов
- Размещения испытуемых и пунктов на одной шкале

Введение в Теорию Тестовых Пунктов

- Может быть прослежена в работах 1940х годов (Лоули, Ричардсон, Такер).
- В 1950-х, Лорд, Бирнбаум, и Раш заложили формальные основы этой теории.
- В 1960-х и 1970-х, работы Бока, Лорда, Макдональда, Самедзимы, Раша, Райта, Андрича, Гольдштейна и многих других.
- Интерес к компьютерному адаптивному тестированию был главной силой развития в 1960-х (но не было компьютерных мощностей).
- С мощными компьютерами и программным обеспечением, теория быстро развивается.

Основополагающие принципы

- «Тест представляет собой серию коротких экспериментов, из которых выводится измерение» (van der Linden & Hambleton, 1997)

- Ответы на тестовые пункты моделируются в соответствии с теорией об их отношениях с психологическим качеством
  - «Связь между качеством (то, что тест измеряет) и ответами на пункты описана нелинейными моделями, основанными на предположениях, которые всегда могут быть проверены» (Hambleton)

Латентное качество

- Измеряемое психологическое (*латентное*) качество
  - Может описывать узкий или широкий домен; может быть стабильным или быстро меняющимся...
    - память, внимание, скорость реакции, способности, знания, черты личности, социальные установки и т.д.
  - Предполагает интервальную шкалу
  - Обозначается буквой θ («тета») – для простоты, тета измеряется на стандартной z шкале

Ответы на тестовые пункты

- Ответы на тестовые пункты чаще всего в категориях
  - Двоичные (*да/нет, правильно/неправильно*)
  - Порядковые категории (*никогда-иногда-часто-всегда*)
  - Номинальные категории (три или более категорий которые не могут быть выстроены по порядку)
- Линейные модели предполагают интервальные шкалы и не подходят к пунктам выраженным в категориях
- Как моделировать процесс ответов на тестовые пункты?

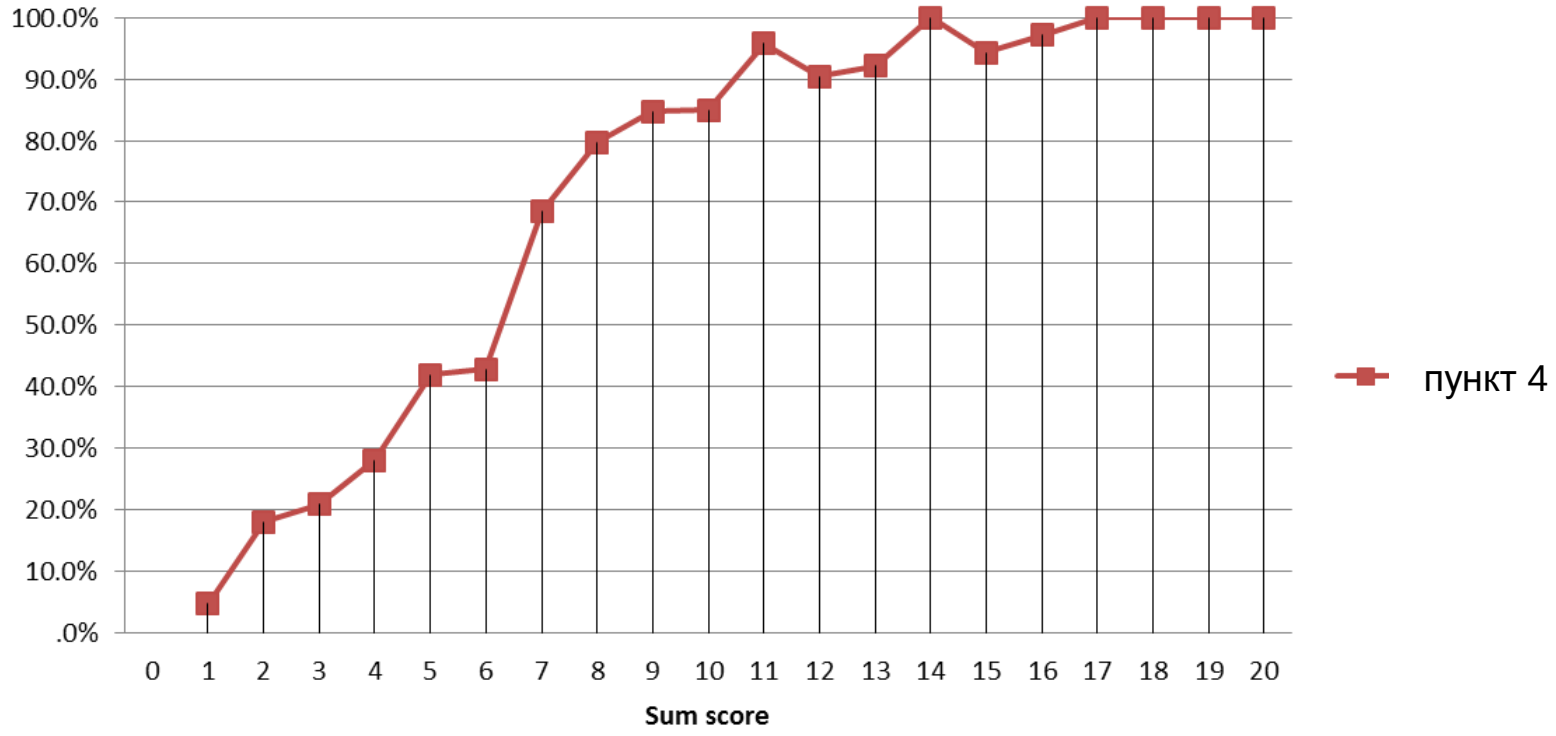
Пример

- Тест способности состоящий из 20 пунктов
  - Каждый пункт предполагает измерять определенный аспект этой способности
  - На каждый пункт, испытуемый может ответить правильно (код 1) или неправильно (код 0)

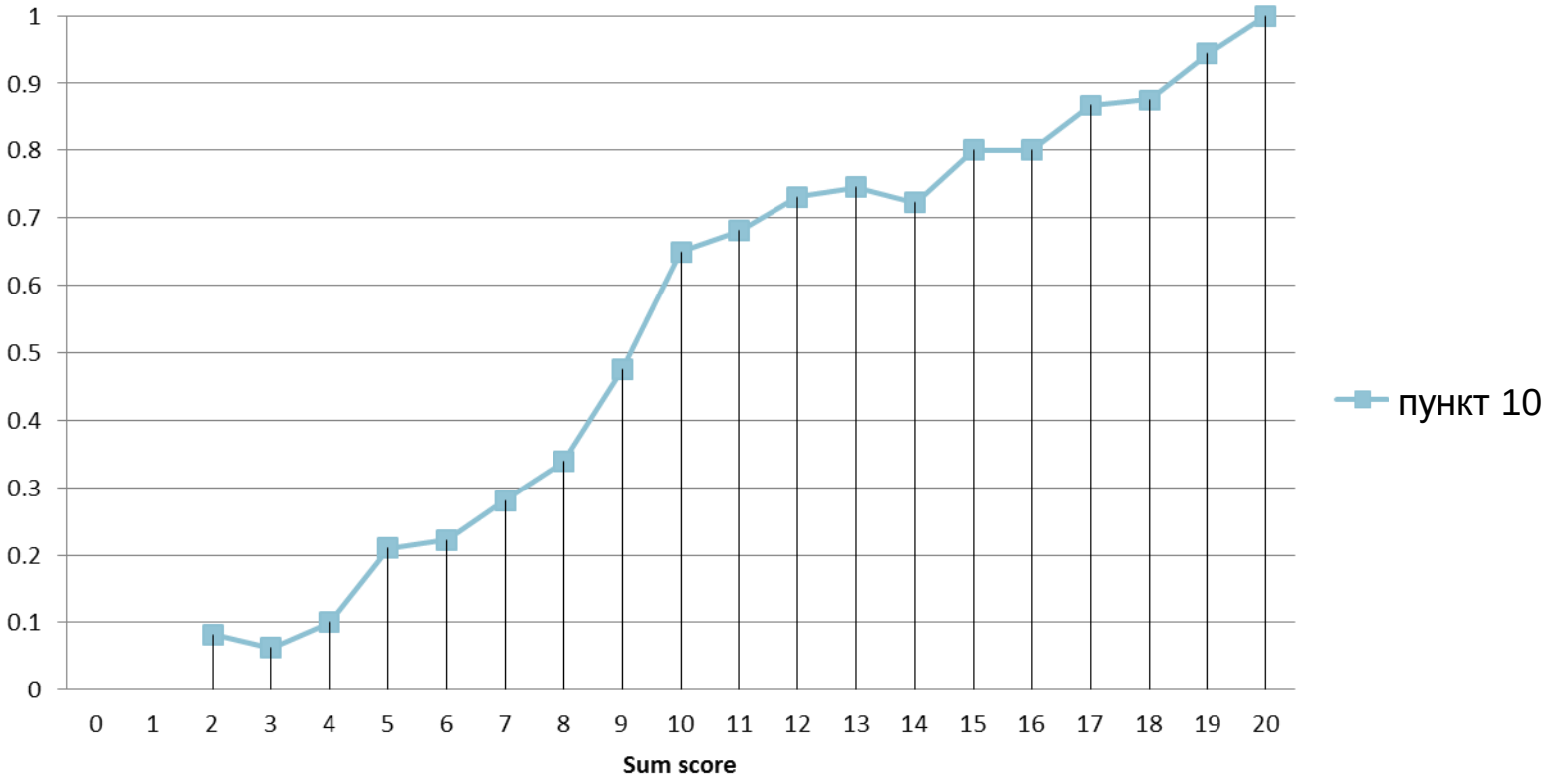
испытуемые	пункты									
	1	2	3	…	…	…	….	….	20
1	1	0	0	…	…	…	…	…	1
2	1	1	0	…	…	…	…	…	0
3	0	1	1	…	…	…	…	…	1
:	:	:	:						:
:	:	:	:						:
:	:	:	:						:
:	:	:	:						:
:	:	:	:						:
N	1	1	0	…	…	…	…	…	1

Процент испытуемых ответивших правильно как функция способности

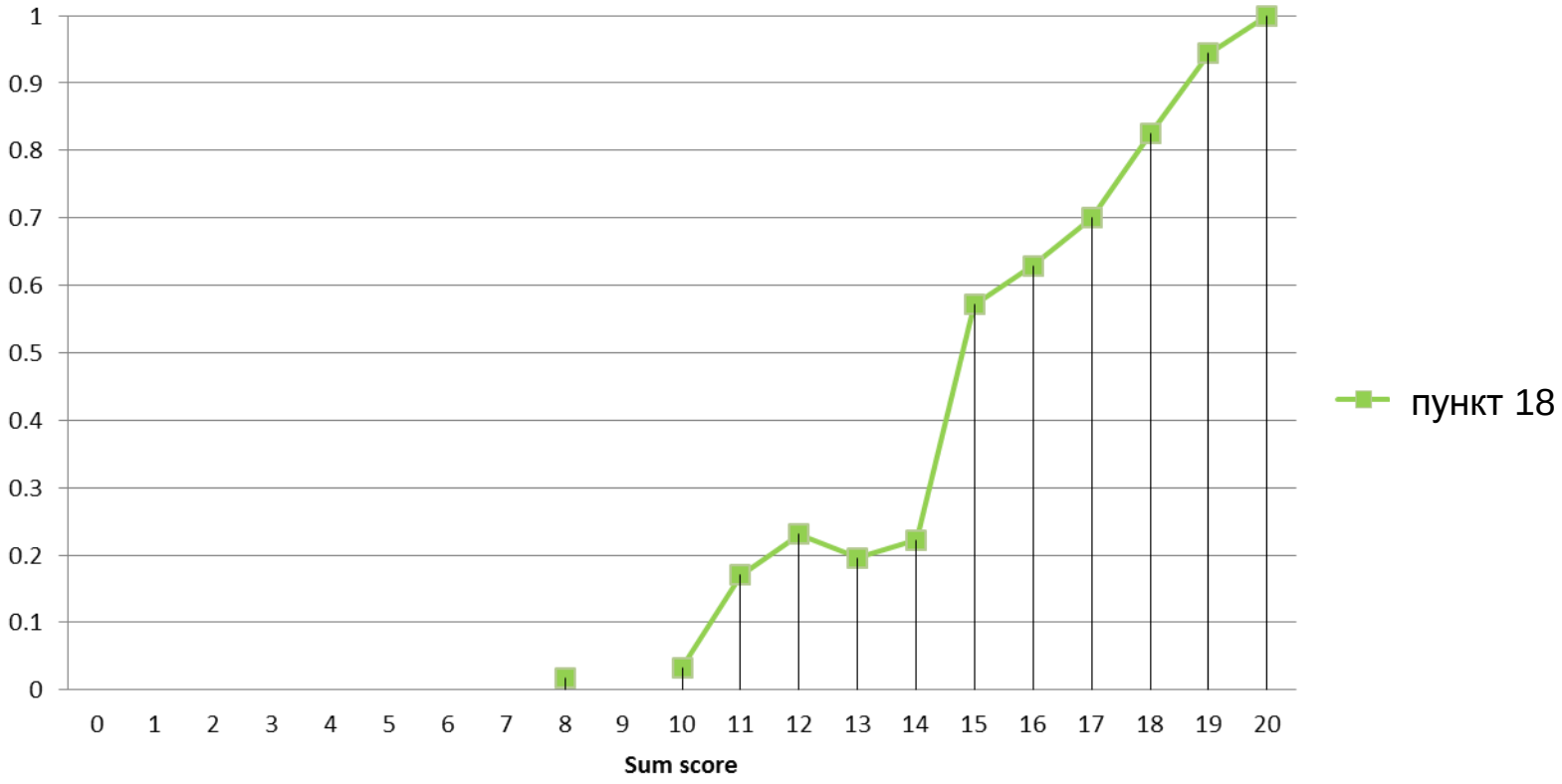
- Мы используем количество правильных ответов как индикатор способности (пока)



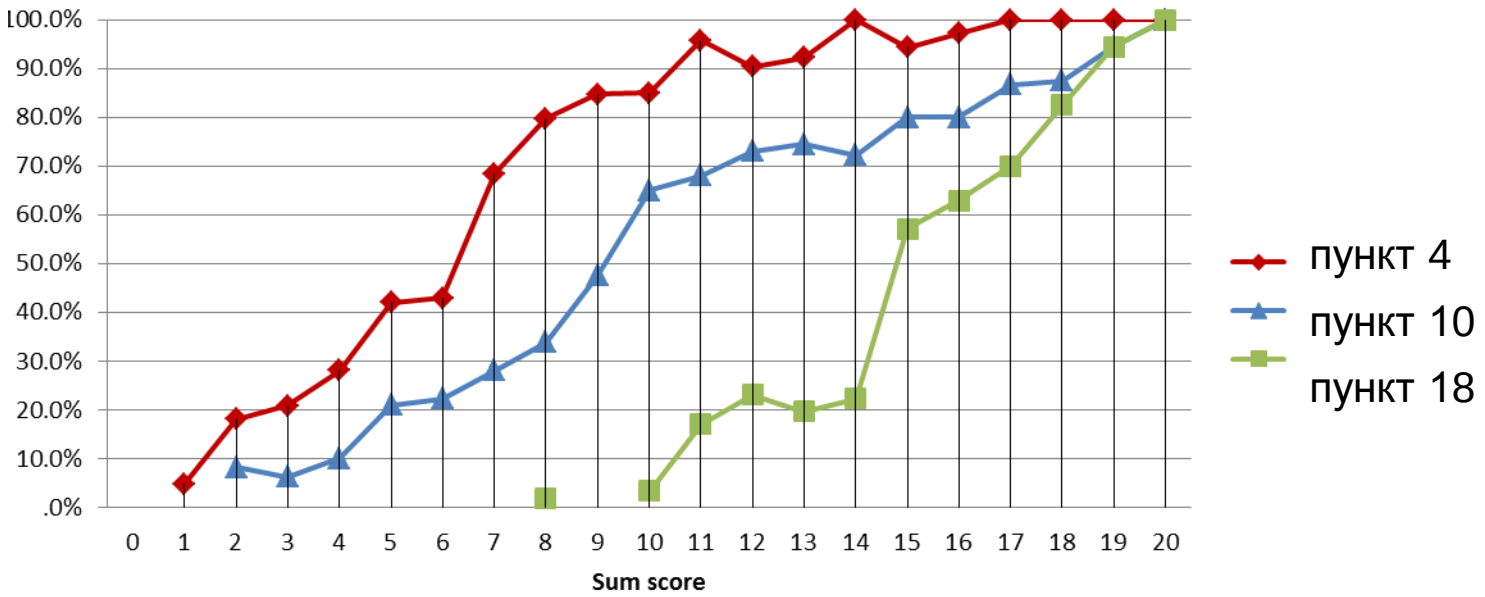
…и для другого пункта



…и для еще одного пункта

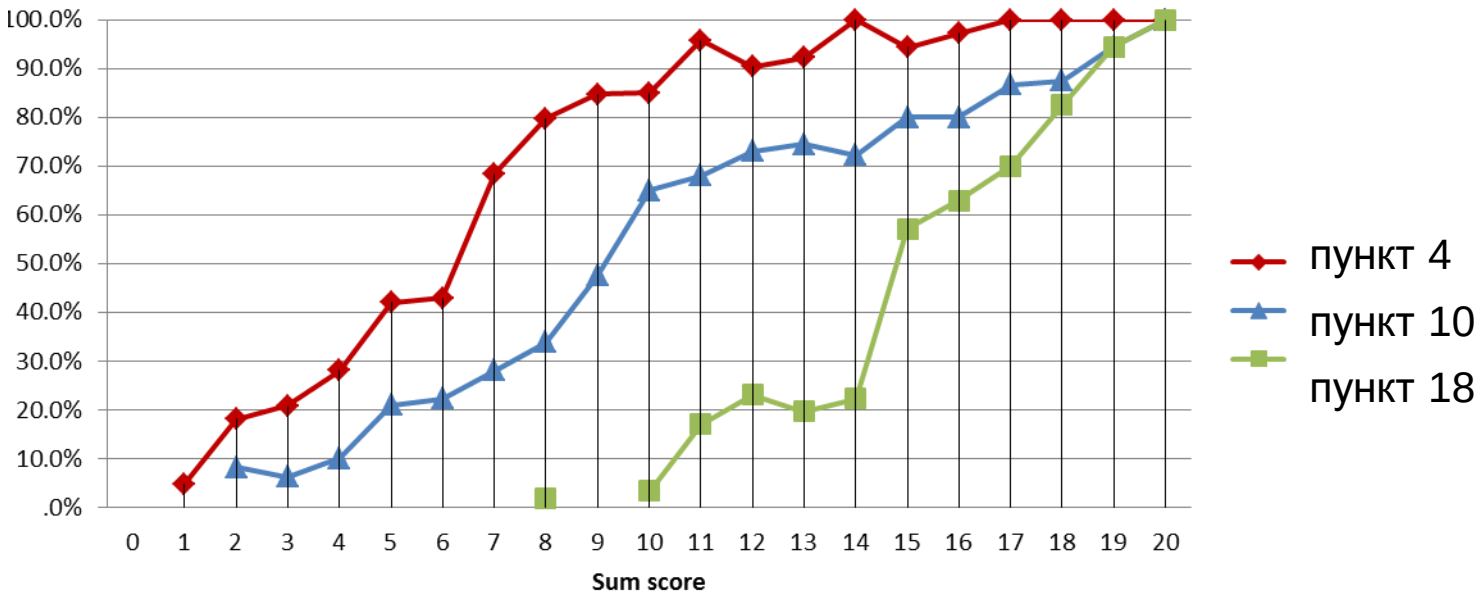


Пункты отличаются друг от друга (1)



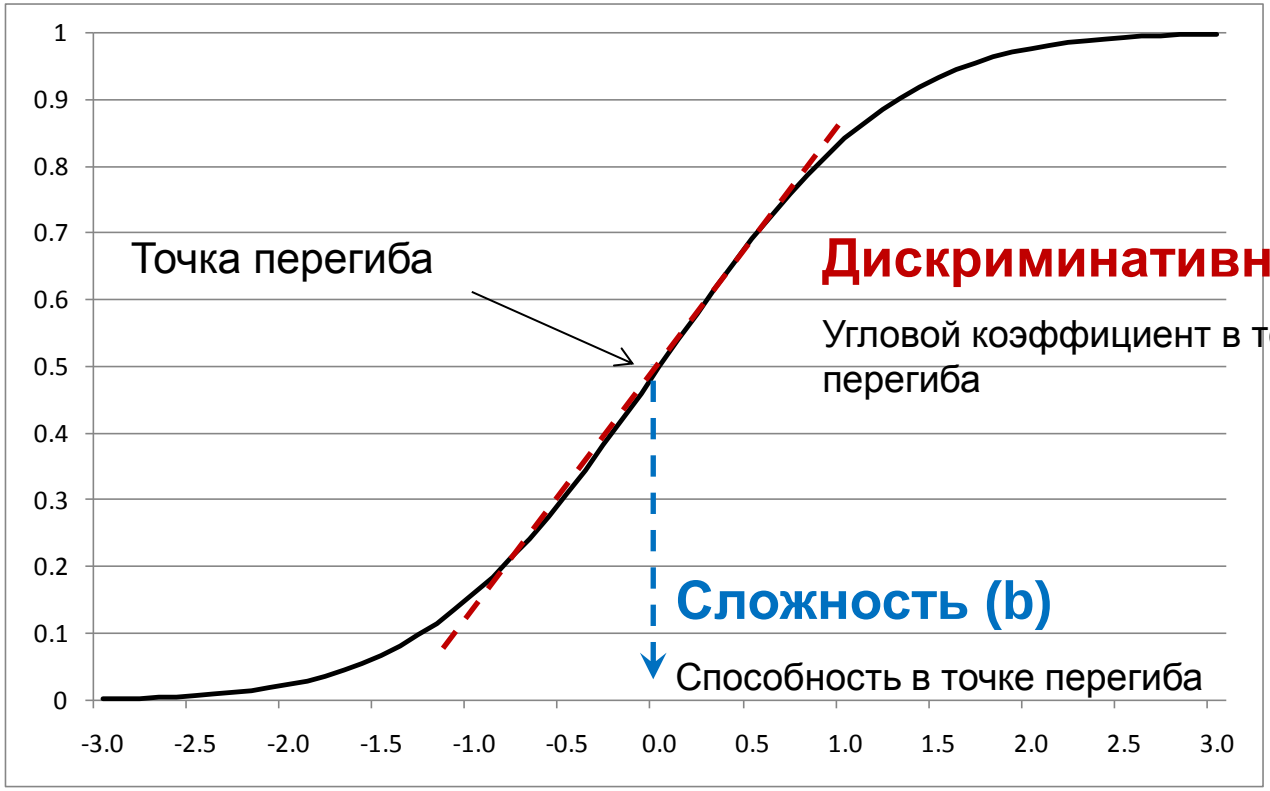
- Пункты различаются вероятностью правильного решения на каждом уровне способности (сложность)

Пункты отличаются друг от друга (2)



- Пункты различаются скоростью увеличения вероятности правильного решения с увеличением способности (дискриминация)

Характеристическая кривая пункта (ICC)



- С параметрами ***a*** = 1, ***b*** = 0
- С другими параметрами, положение и форма функции будут другими

Основные свойства

- Пункт наиболее дискриминативен для уровня способности равному сложности пункта
  - Задания со сложностью много **ниже** чем способность будут пройдены с очень высокой вероятностью, и не дадут информации чтобы различить между способными испытуемыми
  - Задания со сложностью много **выше** чем способность будут провалены с очень высокой вероятностью, и не дадут информации чтобы различить между испытуемыми с низкой способностью
- Пункты должны примерно соответствовать уровню способности в популяции

Нормальная огива

- Знакомая всем нормальная огива
  - Самая первая модель в теории тестовых пунктов (описана Фредериком Лордом в 1952)
- Вероятность правильного ответа зависит от способности и двух параметров пункта, *a* и *b*

$$P_i = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = N\left(a_i\left(\theta - b_i\right)\right)$$

  - Математика непроста, так что нужна была модель с более легкой формулой (сегодня это не проблема с приходом мощных компьютеров)

Логистическая модель

- Аллан Бирнбаум в конце 1950х поставил цель разработки модели с формулой, которая имела бы аналитическое решение
- Предложил заменить нормальную огиву на логистическую функцию
  - $|N(x)-L(1.7x)|<0.01$

$$P_i = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}, \text{ где } D \approx 1.7$$

  - Отсюда следует очень простая формула для шансов

$$\log_e\left[P_i / (1 - P_i)\right] = Da_i(\theta - b_i)$$

Логистическая модель с 2 параметрами (2PL)

- Вероятность правильного ответа на пункт i зависит от способности испытуемого и 2х параметров пункта
  - сложности b_i
  - дискриминативности a_i

$$P(u_i = 1 | \theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

Логистическая модель с 1 параметром (1PL), или модель Раша

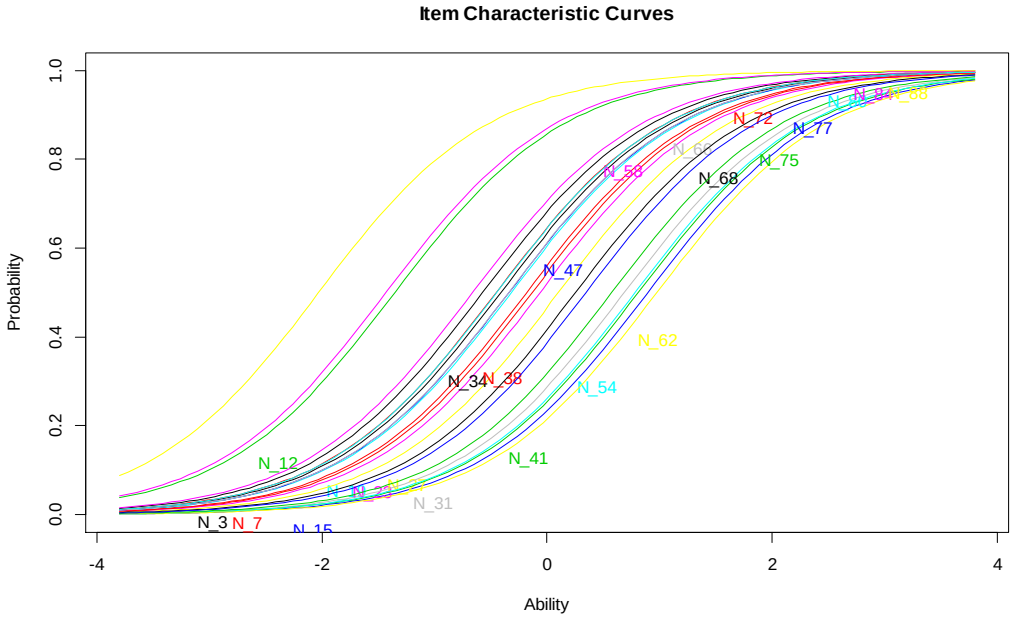
- Вероятность правильного ответа на пункт i зависит от способности испытуемого и 1го параметра пункта
  - сложности b_i

$$P(u_i = 1 | \theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

- Эта модель делает наиболее очевидным тот факт что испытуемые и пункты размещены на одной шкале!
  - Если тета выше сложности, выражение в скобках положительно и вероятность >.05
  - Если тета ниже сложности, выражение в скобках отрицательно и вероятность <.05

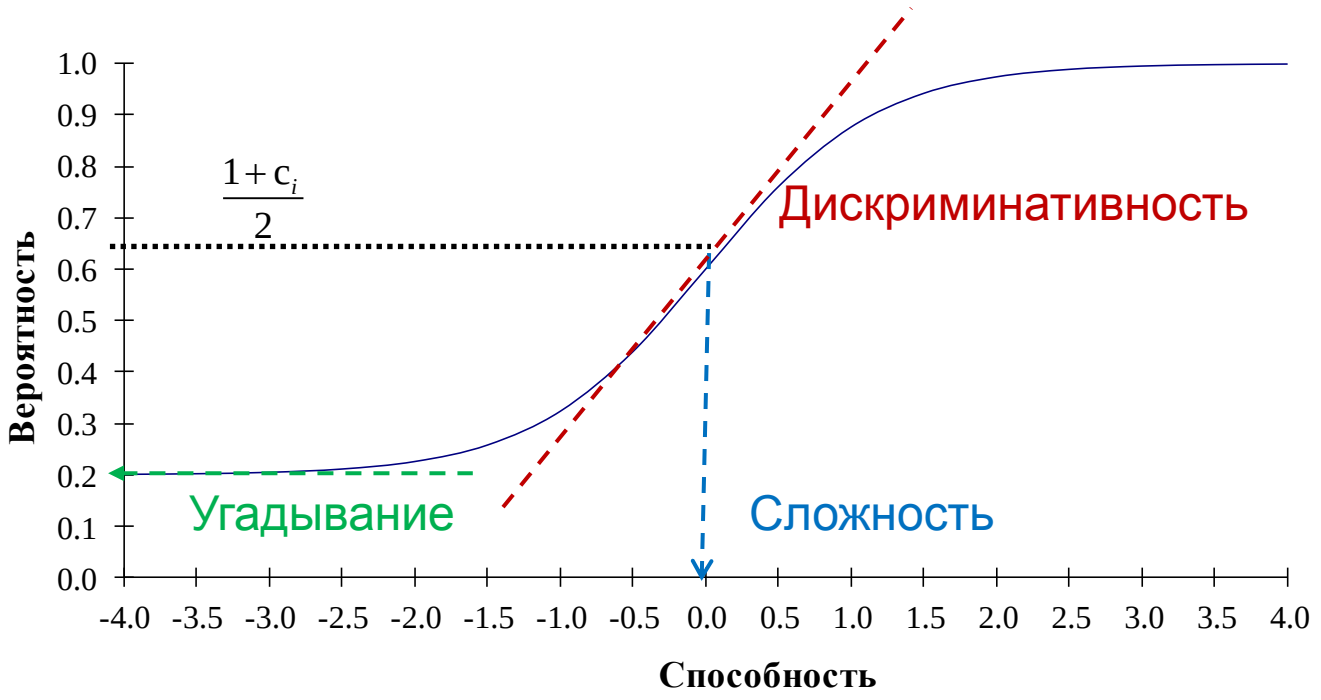
“Специфическая объективность”

- Когда ответы на пункты следуют модели Раша
  - Независимо от выбора пунктов, испытуемые расположены на шкале в том же порядке
  - Независимо от выбора испытуемого, пункты всегда упорядочены по сложности



Функция ответа на пункт с угадыванием

- Что если испытуемый может просто угадать правильный ответ?
  - В вопросе с 5 альтернативами, вероятность угадать ответ можеть достигать 0.2



Логистическая модель с 3 параметрами (3PL)

- Вероятность правильного ответа на пункт i зависит от способности испытуемого и 3х параметров пункта
  - сложности b_i
  - дискриминативности a_i
  - вероятности угадывания c_i

$$P\left(u_i = 1 \mid \theta\right) = c_i + \left(1 - c_i\right)\frac{e^{Da_i\left(\theta - b_i\right)}}{1 + e^{Da_i\left(\theta - b_i\right)}}$$

Пример: Опросник мобильности

- Выборка 8445 сельских жительниц из Бангладеш (Huq & Cleland, 1990).
  - Данные описаны в книге Bartholomew, D., Steel, F., Moustaki, I. and Galbraith, J. (2002) *The Analysis and Interpretation of Multivariate Data for Social Scientists.* London: Chapman and Hall.
  - Данные доступны в стат. программе R, пакет “ltm” (аббревиатура «*latent trait modelling*»)
- Необходимые команды для R могут быть предоставлены любому желающему повторить анализ

Опросник

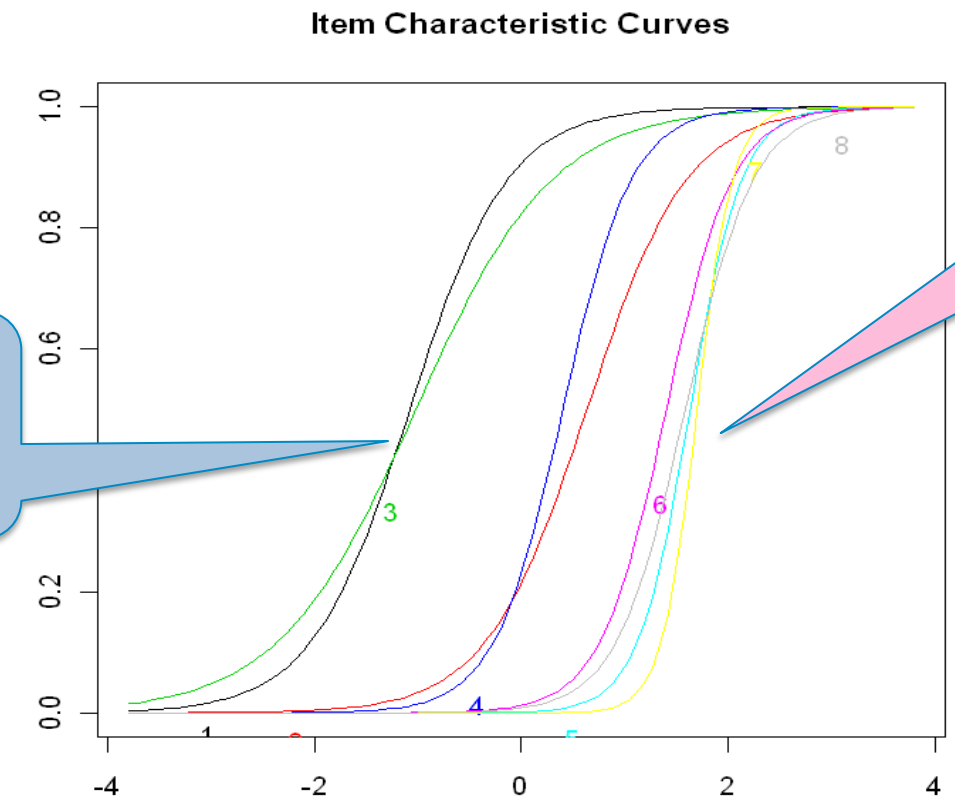
- Мы заинтересованы в измерении социальной мобильности и свободы сельских женщин
- Женщин спросили, могут ли они выполнить следующие действия самостоятельно (ДА или НЕТ):
  1. Сходить в любую часть деревни/села / города.
  2. Выйти за пределы деревни/села / города.
  3. Поговорить с незнакомым мужчиной.
  4. Сходить в кино / посетить культурное мероприятие.
  5. Пройтись по магазинам.
  6. Сходить в кооператив / клуб / клуб матерей.
  7. Принять участие в политическом собрании.
  8. Сходить в поликлинику / больницу.

Результаты моделирования (2PL)

- Мы попробуем 2PL модель

	Сложность	Дискрим.
Item 1	-1.084	2.109
Item 2	0.631	2.058
Item 3	-1.025	1.509
Item 4	0.400	3.010
Item 5	1.630	3.976
Item 6	1.402	3.138
Item 7	1.699	5.816
Item 8	1.585	3.022

Функции пунктов в опроснике мобильности

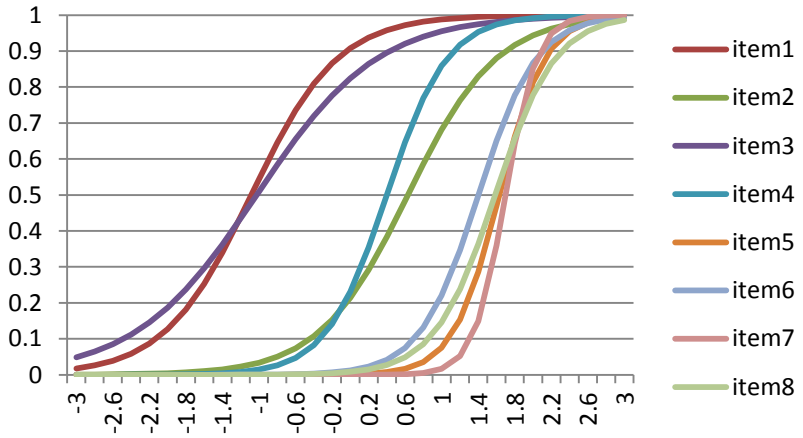


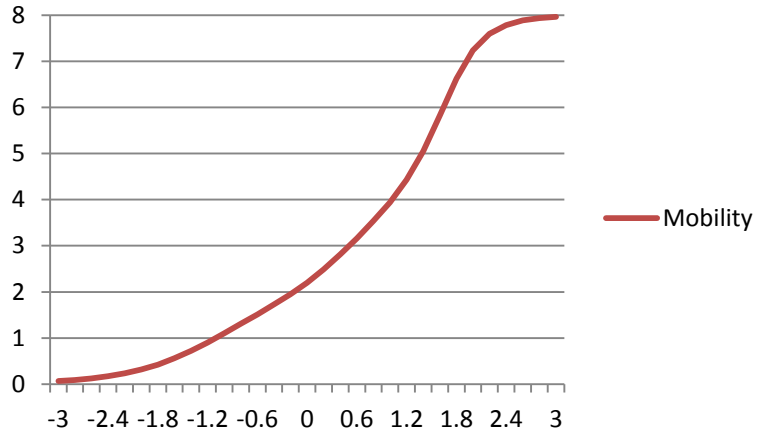
Эти пункты различают между женщинами с **низкой** мобильностью

Эти пункты различают между женщинами с **высокой** мобильностью

Характеристическая Кривая Теста (TCC)

- ICC описывает не только вероятность правильного ответа, но и ожидаемый средний балл по пункту в зависимости от латентного качества
- TCC это сумма ICC всех пуктов





- TCC описывает отношения между латентным качеством и тестовым баллом (заметьте, что это **нелинейная** функция!)

Два допущения ТТП

- Локальная независимость / одномерность
  - Ответы на пункты являются независимыми, после учета латентного качества
    - или, что то же самое
  - Существует только одно латентное качество, объясняющее дисперсию в ответах на пункты
- Форма функции ответа на пункт
  - Вероятность ответа на пункт описывается заявленной функцией (например, нормальной огивой)

Подсчет баллов в ТТП

- В приложениях, параметры пунктов известны (откалиброваны во время стандартизации)
- Вероятность ответов будет зависеть только от латентных качеств испытуемых
- Предполагая локальную независимость,
  - вероятность полученных ответов на все пункты равна произведению вероятностей ответов на отдельные пункты (для конкретного значения θ)

$$P(u_1\ u_{2\ldots}\ u_m|\theta) = P(u_1|\theta) \cdot P(u_2|\theta)\ \ldots\ \cdot P(u_m|\theta)$$

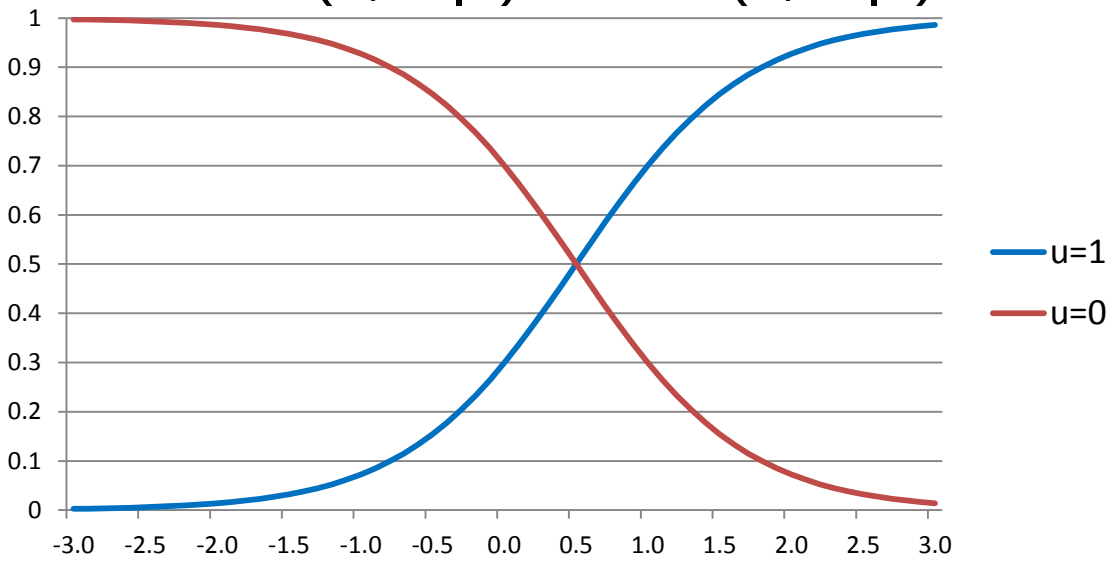
Вероятность правильного и неправильного ответов

- Вероятность правильного ответа описана соответствующей формулой для ICC

$$P(u_i=1|\theta)$$

- Тогда вероятность неправильного ответа

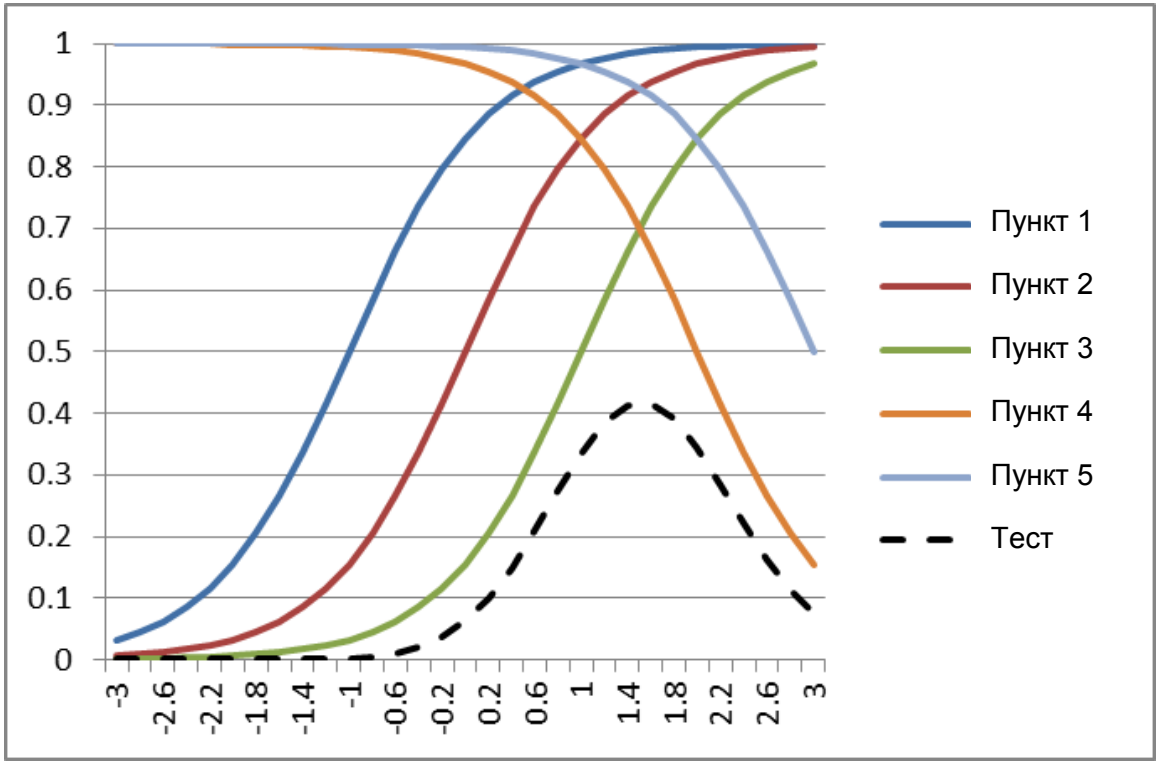
$$P(u_i=0|\theta)= 1 - P(u_i=1|\theta)$$



Вероятность ответов на целый тест

Предположим, что наш испытуемый ответил правильно на пункты 1, 2 и 3; и неправильно на пункты 4 и 5

$$\mathrm{P}(u_1,u_2,\ldots,u_5) = \mathrm{P}(u_1=1)\cdot\mathrm{P}(u_2=1)\cdot\mathrm{P}(u_3=1)\cdot\mathrm{P}(u_4=0)\cdot\mathrm{P}(u_5=0)$$

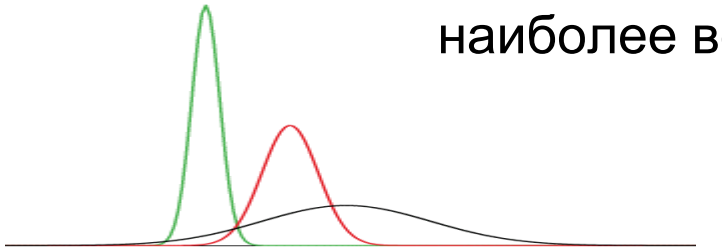


«Метод максимального правдоподобия»

- Поиск тестового балла, который максимизирует вероятность наблюдаемых ответов на все пункты
- плюсы
  - Метод находит оптимальный балл без искажений
  - Стандартная погрешность распределена нормально
- минусы
  - Нахождение оптимального балла не гарантированно в случае аберрантных ответов
  - Конечная оценка не существует когда **все** ответы правильные, или неправильные

Стандартная погрешность

- Вероятность набора ответов на тестовые пункты – это не одно значение, а огромное число возможных значений, каждое со своей вероятностью
  - Много возможных решений, но только одно наиболее вероятно



- Стандартное отклонение в распределении вероятности – это стандартная погрешность тестового балла (SE)
  - **Важно**: погрешность зависит от конкретных ответов на пункты, и значит, будет разной для разных испытуемых

Опросник мобильности: Подсчет баллов

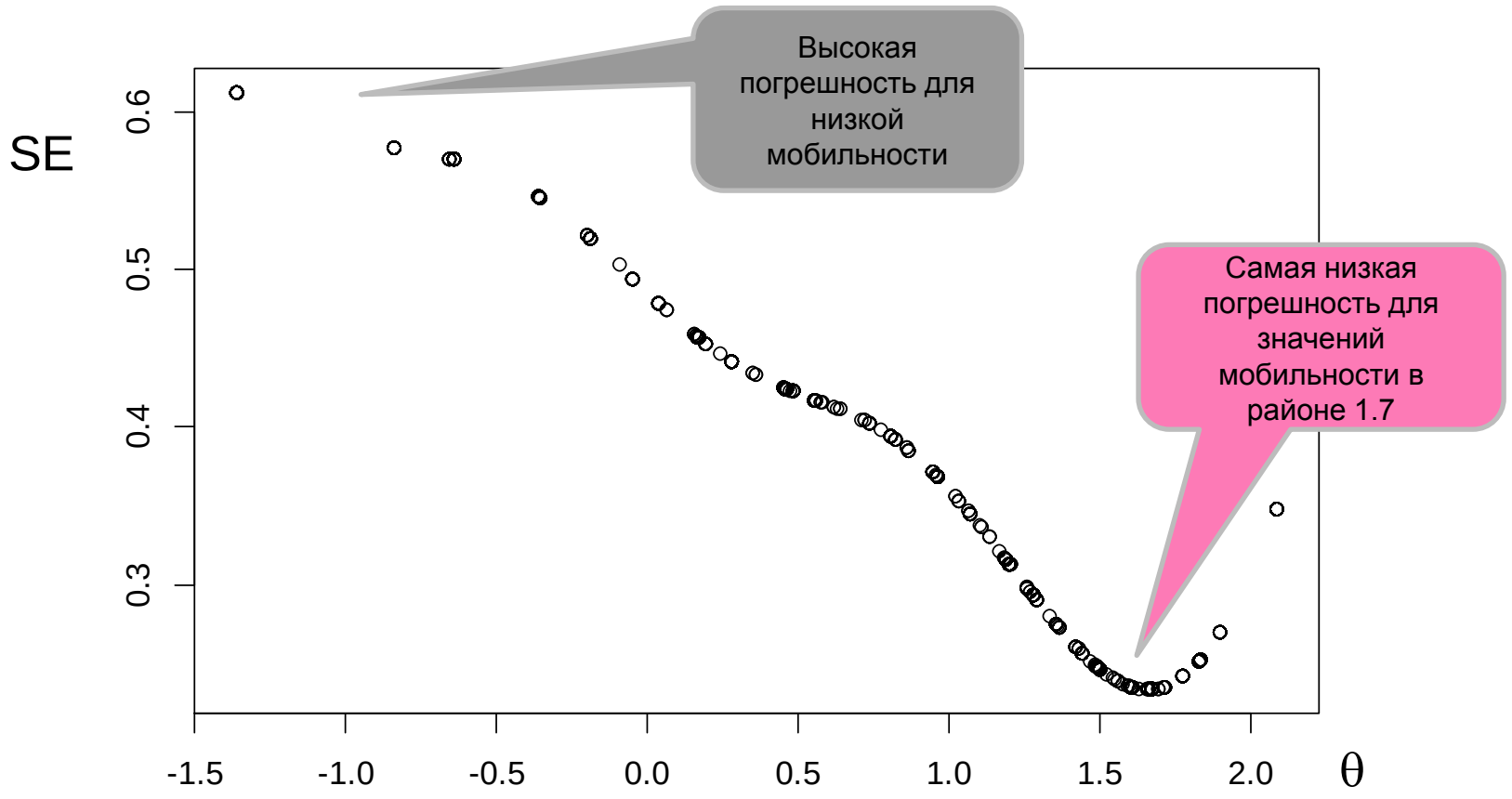
	п1	п2	п3	п4	п5	п6	п7	п8	θ	SE
1	1	1	1	1	0	0	0	0	0.805	0.394
2	0	0	0	0	0	0	0	0	-1.364	0.612
3	0	0	0	0	0	0	0	0	-1.364	0.612
...										
8440	1	1	1	1	0	0	1	0	1.420	0.261
8441	1	1	1	0	0	0	0	0	0.280	0.442
8442	1	1	1	1	0	0	1	1	1.603	0.236
8443	1	0	1	1	0	0	0	0	0.458	0.425
8444	1	1	1	1	0	0	1	1	1.603	0.236
8445	1	0	1	0	0	0	0	0	-0.188	0.520

Большая погрешность для низкой мобильности

Одинаковые ответы – одинаковый балл

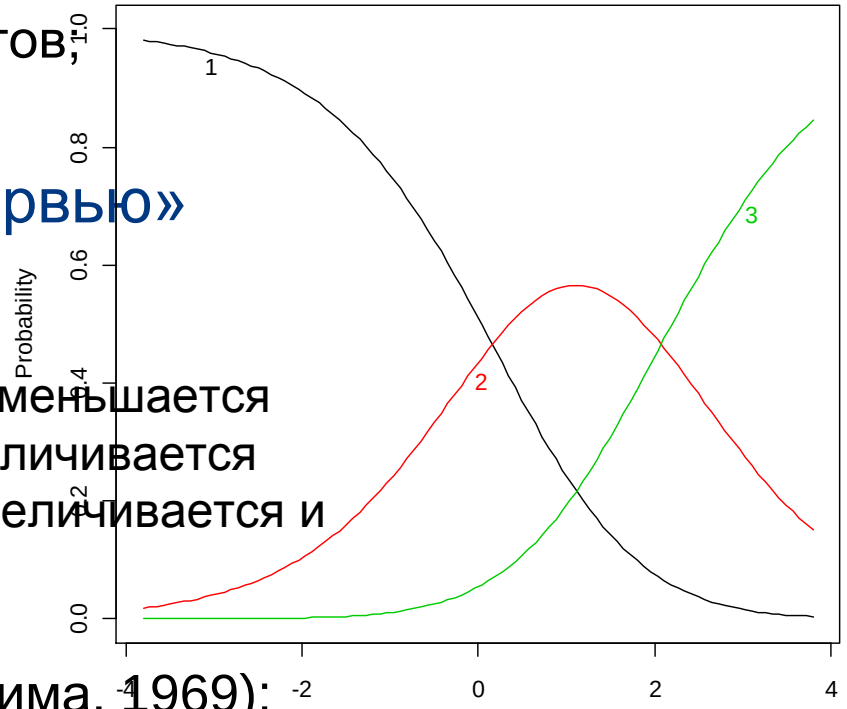
Маленькая погрешность для высокой мобильности

Опросник мобильности: Погрешность измерения



Модели для порядковых категорий

- Вероятности выбора каждой категории должны в сумме давать 1 для любой θ
  - как это было для двоичных ответов, P(u=1)+P(u=0)=1
- «Я не могу заснуть перед интервью»
  (1) никогда (2) иногда (3) часто
  - С ростом тревожности
    - Вероятность выбора «никогда» уменьшается
    - Вероятность выбора «часто» увеличивается
      Вероятность выбора «иногда» увеличивается и затем уменьшается

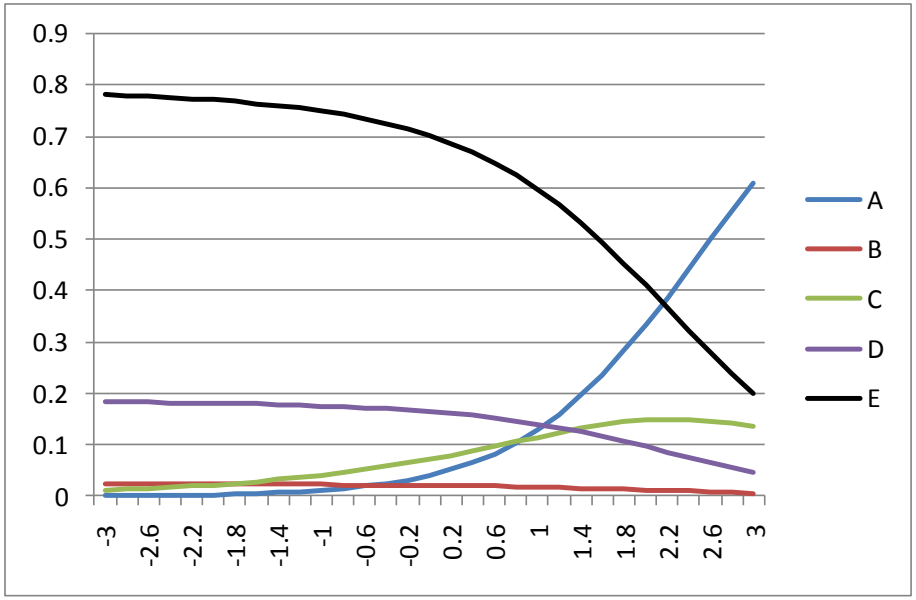


  - Graded Response Model (Самедзима, 1969);
  - Partial Credit Model (Мастерс, 1982)

Модель для номинальных категорий

- Пример: оценка булимии
- “Я предпочитаю есть”

  (a) дома одна (b) дома с семьей (c) в кафе
  (d) с друзьями (e) не важно



- Nominal Response Model (Bock, 1972)

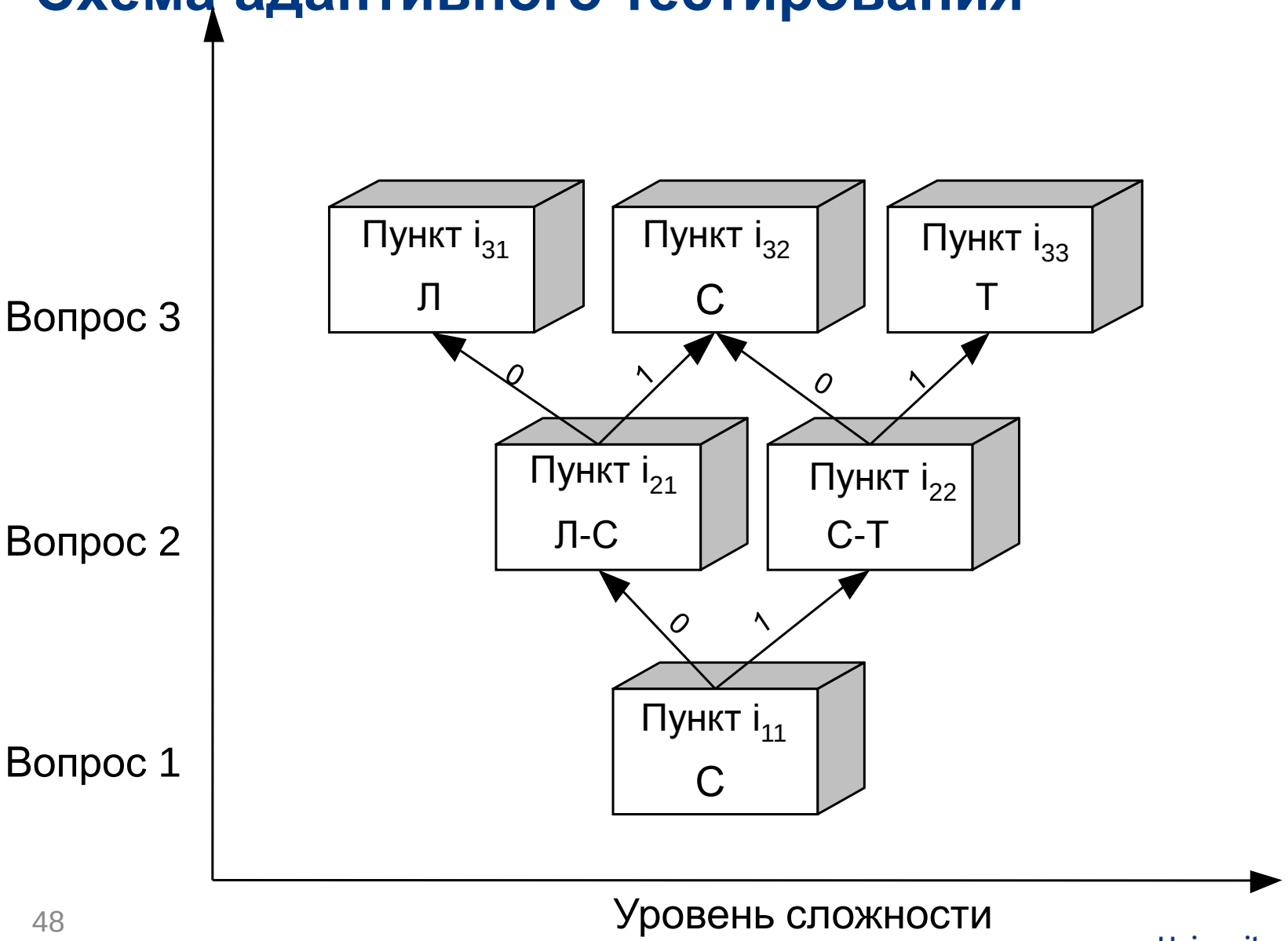
Приложения Теории Тестовых Пунктов

- Очень удобная база для решения многих задач в тестировании
  - Оценка необъективности пунктов по отношению к различным группам испытуемых, так как характеристические функции можно сравнить
  - Уравнивание тестовых форм, так как параметры пунктов полностью контролируются
  - Сравнение испытуемого с критерием, так как испытуемые и пункты на одной шкале
  - Компьютерное адаптивное тестирование (CAT), так как можно выбрать пункты наиболее эффективные для измерения конкретного уровня способностей

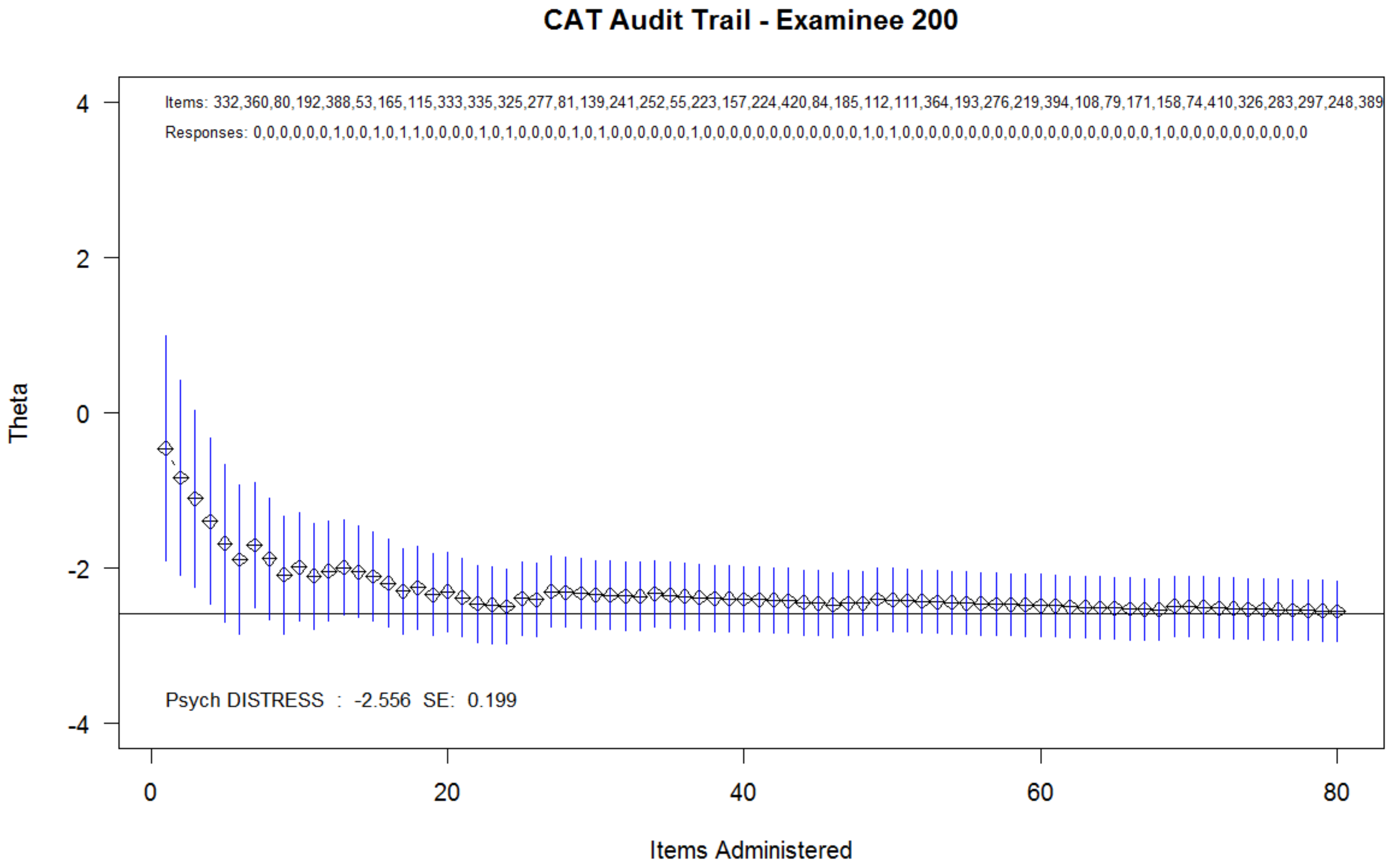
Компьютерное адаптивное тестирование

- Тест адаптируется к тому как испытуемый отвечает
  1. Представляется первый пункт
  2. Испытуемый отвечает, и подсчитывается промежуточный балл
  3. Выбирается следующий пункт
     – Если предыдущий пункт был решен правильно, выбирается более сложный пункт
     – Если предыдущий пункт был решен неправильно, выбирается более легкий пункт
  4. Шаги 1-3 повторяются до тех пор, пока не достигнуто условие остановки
     – Как правило, погрешность оценки достаточно мала
     – Или, достигнуто максимально позволенное количество пунктов
- Попробуйте сами – а лучше с другом
  http://www.nihpromis.org/software/demonstration

Схема адаптивного тестирования



Приближение к наиболее вероятному баллу



Преимущества Теории Тестовых Пунктов

- Теория делает возможным целевой отбор пунктов
  - Достижение наибольшей точности где она необходима
    - Отбор по проходному баллу
    - Отсев больных в зоне риска
  - Создание параллельных форм
  - Адаптивное тестирование
- Подсчет тестового балла контролируя свойства пунктов
  - Эти свойства учитываются в моделях, таким образом, естовый балл становится независим от них
- Точная оценка погрешности измерения для каждого испытуемого
  - Дает возможность оценивать значимость изменений, например в результате лечения

Рекомендуемые книги

- Для начинающих
  - Hambleton, Swaminathan & Rogers (1991). Fundamentals of Item Response Theory.
- Для тех кто любит прикладное изложение
  - Embretson, S. & Reise, S. (2000). Item Response Theory for psychologists.
  - R.J. de Ayala (2009). The theory and practice of Item Response Theory.
- Для тех кто любит элегантное и строгое математическое изложение
  - McDonald, R. (1999). Test Theory: A Unified Treatment.
- Для продвинутых
  - Van der Linden, W. & Hambleton, R. eds. (1997). Handbook of modern Item Response Theory.

СПАСИБО!

ВАШИ ВОПРОСЫ?

a.a.brown@www.kent.ac.uk

University of Kent

50 1965-2015 THE UK'S EUROPEAN UNIVERSITY