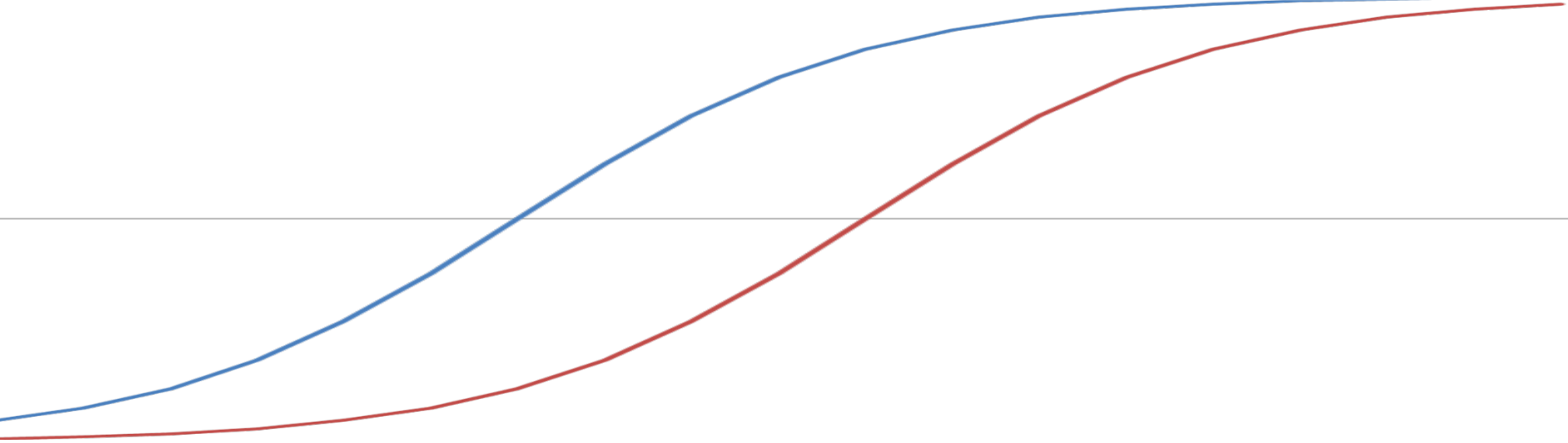


What is Item Response Theory?

Nick Shryane
Social Statistics Discipline Area
University of Manchester

nick.shryane@manchester.ac.uk

What is Item Response Theory?

1. It's a theory of measurement, more precisely a psychometric theory.
   - 'Psycho' – 'metric'.
     - From the Greek for '*mind/soul*' – '*measurement*'.
2. It's a family of statistical models.

Why is IRT important?

- It's one method for demonstrating reliability and validity of measurement.
- Justification, of the sort required for believing it when...
  - Someone puts a thermometer in your mouth then says you're ill...
  - Someone puts a questionnaire in your hand then says you're post-materialist
  - Someone interviews you then says you're self-actualized

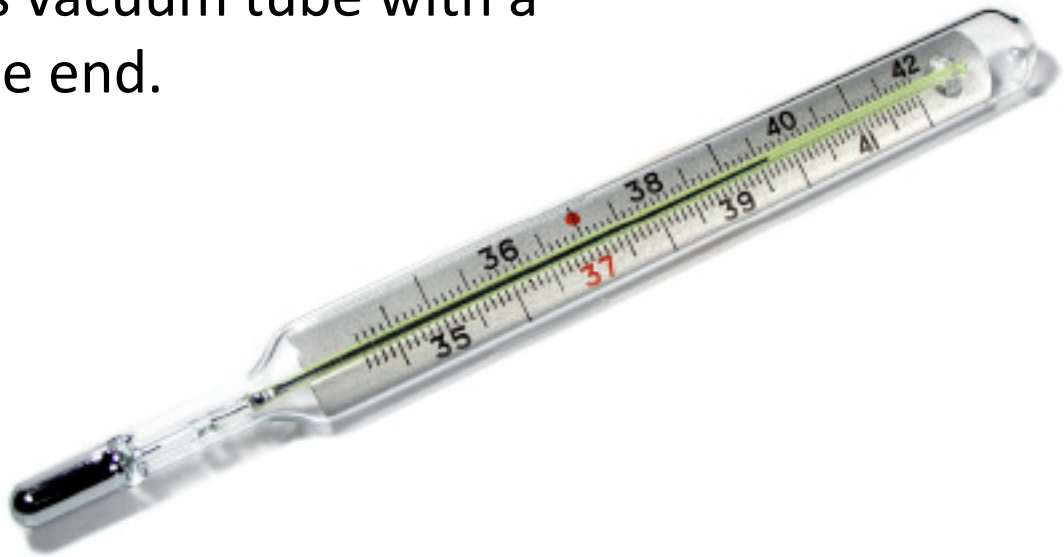
This talk will cover

- A familiar example of measuring people.
- IRT as a psychometric theory.
  - 'Rasch' measurement theory.
- IRT as a family of statistical models, particularly:
  - A 'one-parameter' or 'Rasch' model.
  - A 'two-parameter' IRT model.
- Resources for learning/using IRT

Measuring body temperature

Using temperature to indicate illness

Measurement tool: a mercury thermometer - a glass vacuum tube with a bulb of mercury at one end.



Measuring body temperature

Thermal equilibrium

Stick the bulb in your mouth, under your tongue.

The mercury slowly heats up, matching the temperature of your mouth.

Measuring body temperature

Density – temperature proportionality

Mercury expands on heating, pushing up into the tube. Marks on the tube show the relationship between mercury density and an abstract scale of temperature.

Measuring body temperature

Medical inference

Mouth temperature is assumed to reflect core body temperature, which is usually very stable. Temperature outside normal range may indicate illness.

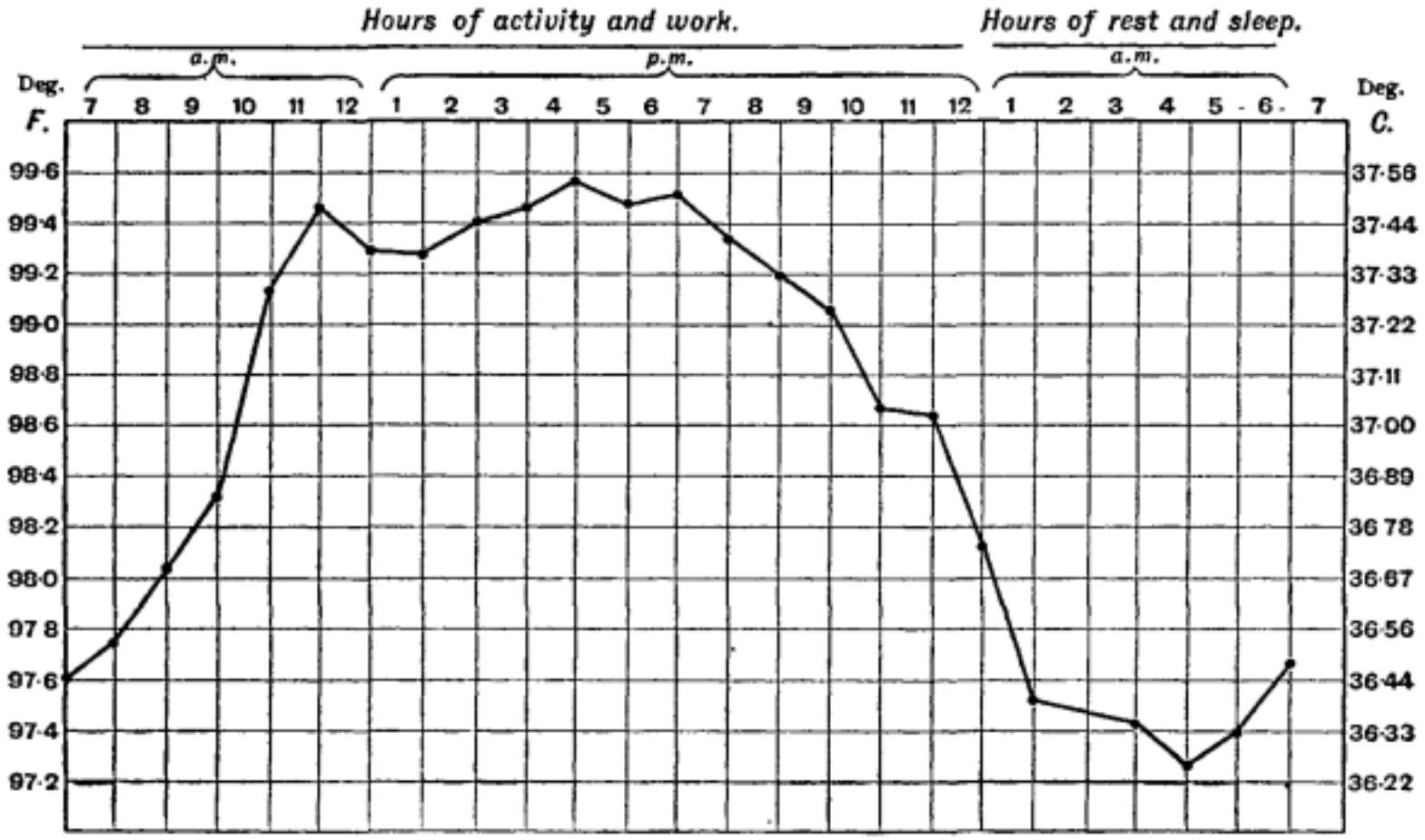
Measuring body temperature

- To make inference between taking temperature and illness rests upon **theory** regarding:
  - Thermal equilibrium via conduction.
  - The proportionality of mercury density with a conceptual temperature scale.
  - Relationship between mouth and core body temperature.
  - Relationship between core body temperature and illness.

Measuring body temperature

- At each stage, **error** may intrude:
  - Thermal equilibrium may not have been reached (e.g. thermometer removed too quickly).
  - Expansion of mercury also affected by other things (e.g. air pressure).
  - Mouth temperature may not reflect core body temperature (e.g. after a hot cup of tea).
  - Core body temperature does not vary with all illnesses, and is not even completely stable in health.

Daily variation in body temperature



Measurement: key features

- Rules for mapping observations onto conceptual structures
  - Level of mercury onto temperature, temperature onto health
- Scaling
  - What type of mapping? Quantitative, qualitative?
    - Density of mercury with a quantitative temperature scale.
    - Quantitative temperature scale with a qualitative health state (i.e. well/ill).
- Error
  - Where does the mapping break down? Bias vs. variance

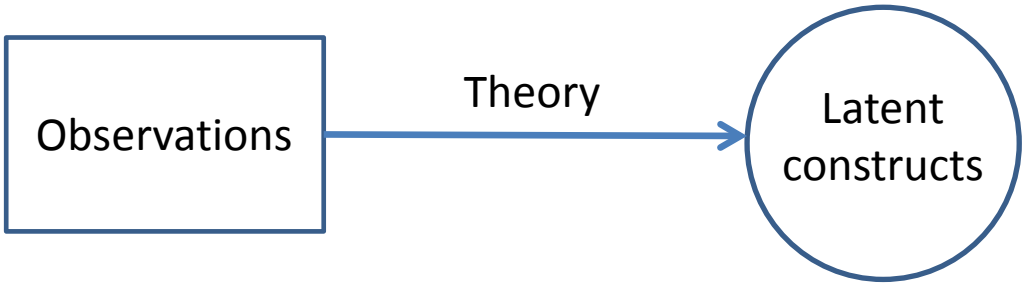
Measuring what people think

- We need to do the same thing when trying to infer what people...

  ...think/believe/know/feel

- based upon how they...

  ...behave/speak/write/interact



Psychometric measurement

- Mapping observations onto internal states/traits
  - Test scores onto knowledge/intelligence
  - Questionnaire item responses onto attitudes/beliefs
  - Interview transcripts into a narrative account

Psychometric measurement

- Measurement tool
  - Often a test / questionnaire consisting of several ‘items’.
  - Could be many things: facial recognition camera, accelerometer, an observer/rater/examiner, an inkblot plus a rater, etc.
- Measurement theory
  - Participant has an unobserved trait, e.g. Intelligence, knowledge, optimism, anger, etc.
  - The output of the measurement tool is mapped to the unobserved trait using some ‘scaling’ rules.
- Questionnaires often involve mapping discrete (e.g. binary) responses onto unobserved traits that are assumed to be continuous (i.e. you can have any ‘amount’ of it)
    - Popular method: Add up all the responses into a ‘score’
    - What’s the justification for this?

Example psychometric model

- Trait – Perceived disposable wealth
- Questionnaire items
  - "If I wanted to, I could probably afford to do the following this month:"

Example psychometric model

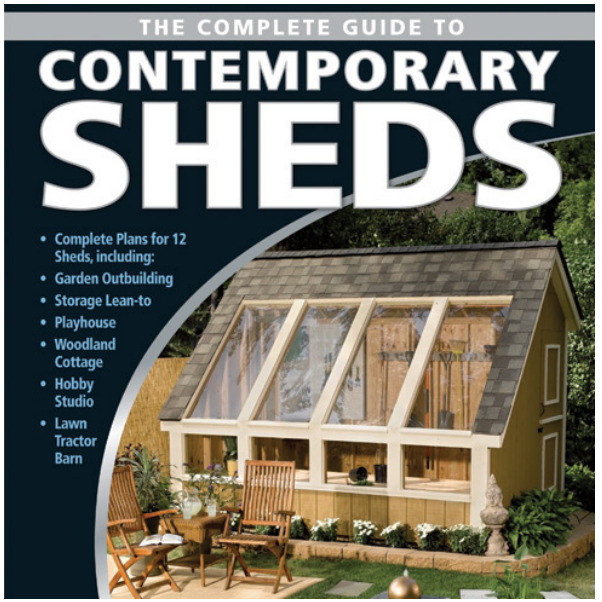
- Trait – Perceived disposable wealth
- Questionnaire items
  - "If I wanted to, I could probably afford to do the following this month:"
  - Buy a cup of coffee

Example psychometric model

- Trait – Perceived disposable wealth
- Questionnaire items
  - "If I wanted to, I could probably afford to do the following this month:"
  - Save £10

Example psychometric model

- Trait – Perceived disposable wealth
- Questionnaire items
  - "If I wanted to, I could probably afford to do the following this month:"
  - Buy a book about sheds



Example psychometric model

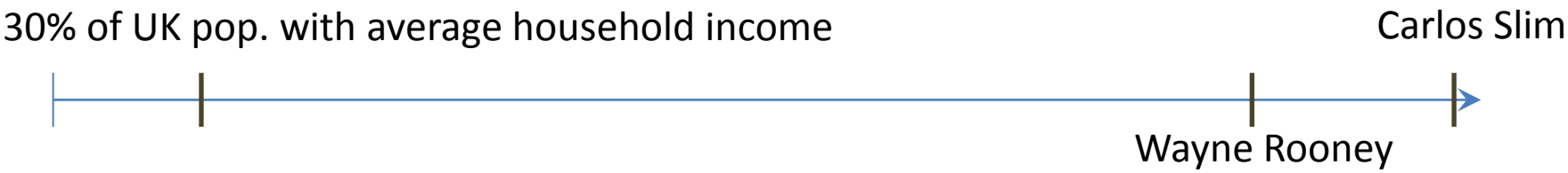
- Trait – Perceived disposable wealth
- Questionnaire items
  - "If I wanted to, I could probably afford to do the following this month:"
  - Buy a new fridge

Example psychometric model

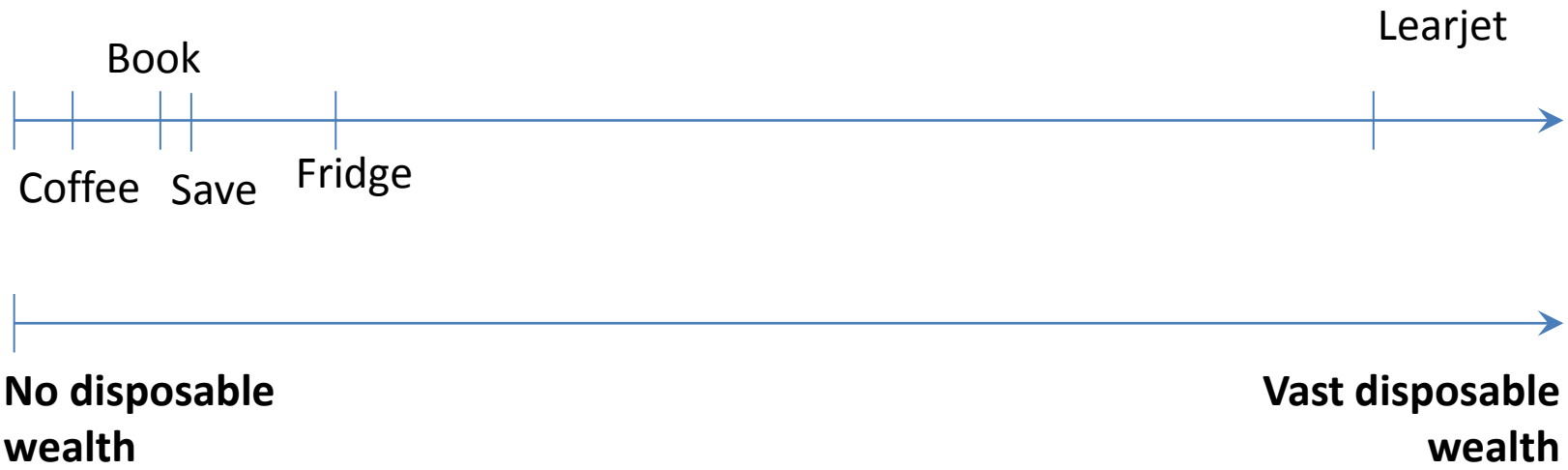
- Trait – Perceived disposable wealth
- Questionnaire items
  - "If I wanted to, I could probably afford to do the following this month:"
  - Buy a Learjet

Items and people on the same scale

Individuals



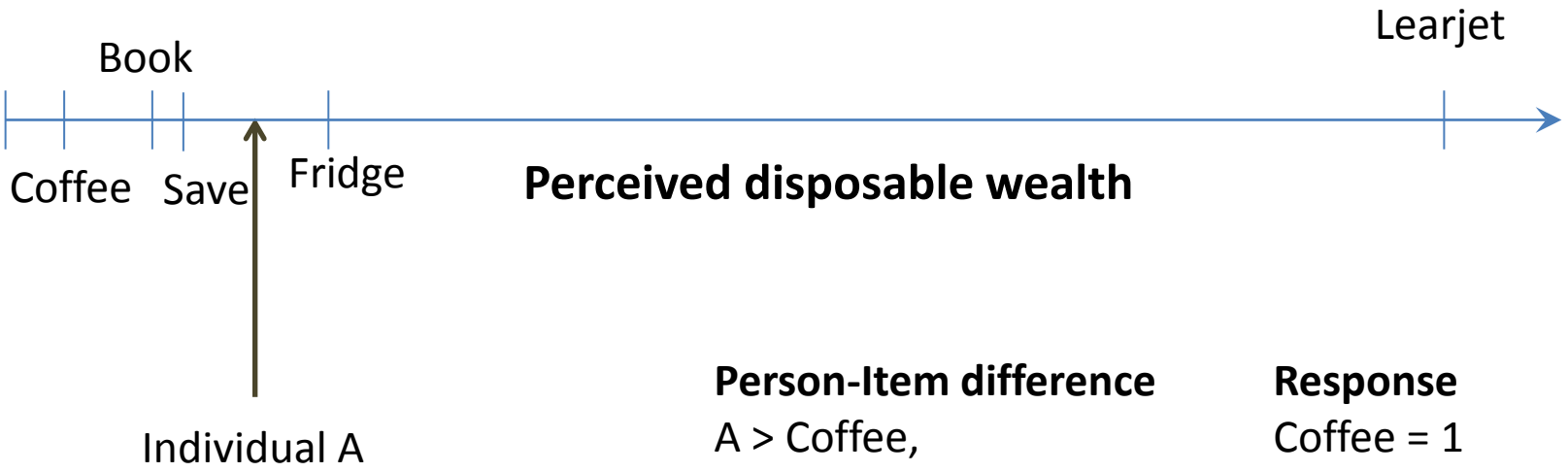
Items



Mapping binary responses to the scale

- Some items require greater disposable wealth to purchase than others – items *cheap/expensive*
- Some participants have greater disposable wealth than others – people *poor/wealthy*
  - If "participant wealth" > "item cost", we should see a positive item response
- 'Level' of positive item response tells us about where on the scale the participant lies, e.g.
  - No positive responses (i.e. can't afford even a coffee), very low disposable wealth
  - All positive responses (i.e. can afford a Learjet) – very high disposable wealth

Mapping binary responses to the scale



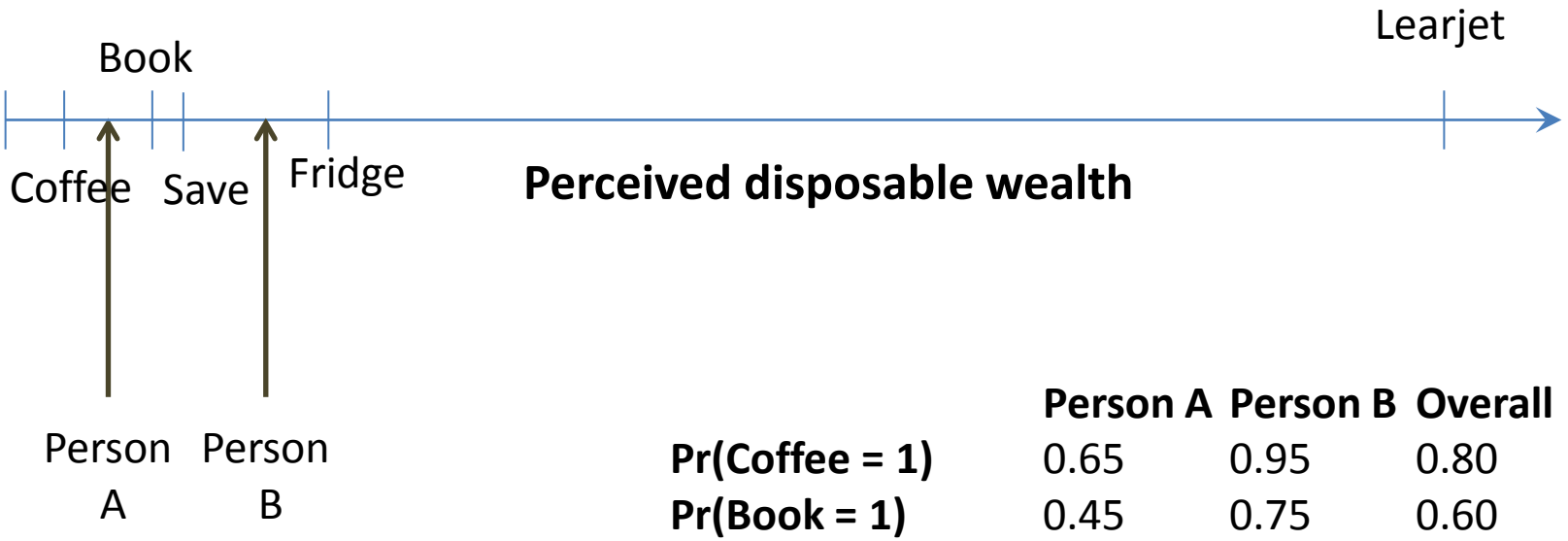
Person-Item difference	Response
A > Coffee,	Coffee = 1
A > Book,	Book = 1
A > Save,	Save = 1
A < Fridge,	Fridge = 0
A < LearJet,	LearJet = 0

Probabilistic mapping

- The mapping across and within individuals will not be completely consistent, e.g.
  - Different estimates of how much things cost
  - Different knowledge of how much money he or she has available (available = credit?)
  - Wishful thinking
  - Disposable wealth changes over time – not a fixed trait.
- The mapping will be probabilistic, contains error
  - It's probable that a rich person will be more able to afford a Learjet, not certain.

Probabilistic mapping

Probability of observing a positive response will vary by item and by a person's level on the scale.



	Person A	Person B	Overall
Pr(Coffee = 1)	0.65	0.95	0.80
Pr(Book = 1)	0.45	0.75	0.60
Pr(Save = 1)	0.40	0.70	0.55
Pr(Fridge = 1)	0.15	0.45	0.30
Pr(Learjet = 1)	0.00	0.00	0.00

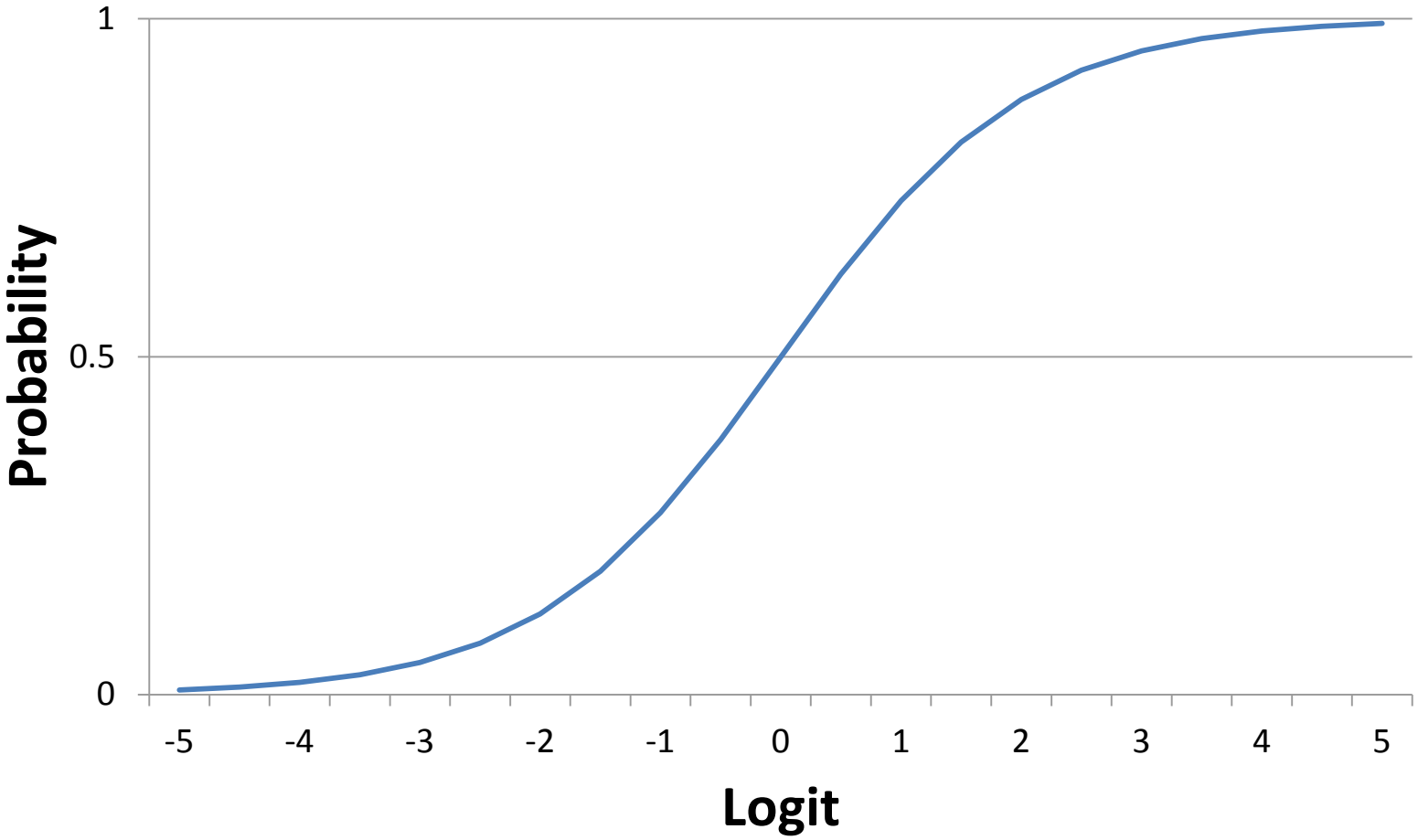
Transforming probability

- Probabilities are not convenient for statistical modelling
  - Bounded between [0, 1].
- Much easier to model a transformation of probability that ranges from [-∞, +∞]:
  - Natural log of the odds, a.k.a. logit:

$$\text{Logit} = \ln(\Pr / (1-\Pr))$$

$$\text{e.g., } 0 = \ln(0.5 / (1-0.5)).$$

Probability vs. logit



Statistical model

$$\text{Logit}_{person\_endorses\_item} = \text{Wealth}_{person} - \text{Cost}_{item}$$

$$Y_{ij} = \theta_j - b_i$$

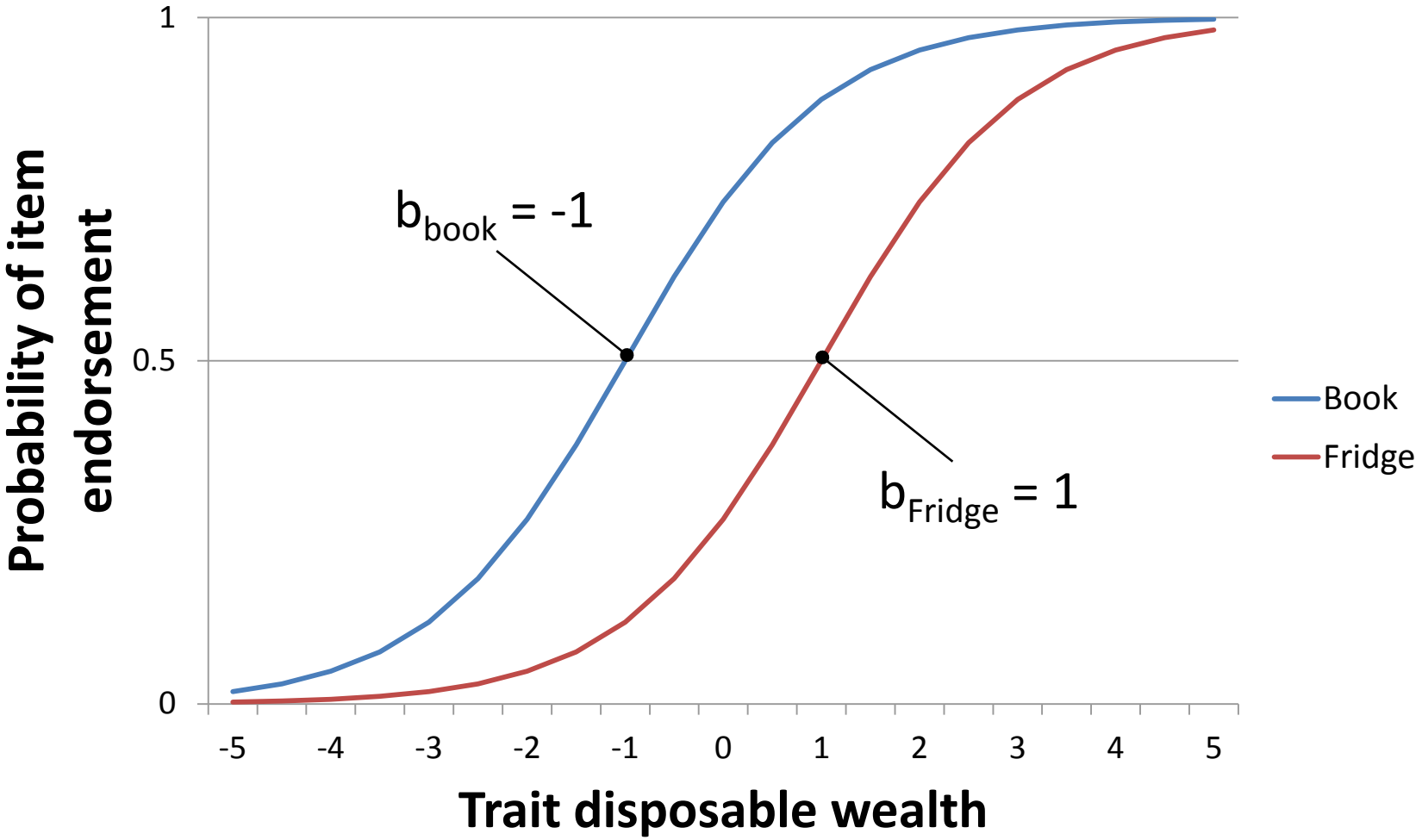
Y_{ij} = Logit that item *i* is endorsed by person *j*

θ_j = ***Trait*** level of person *j*

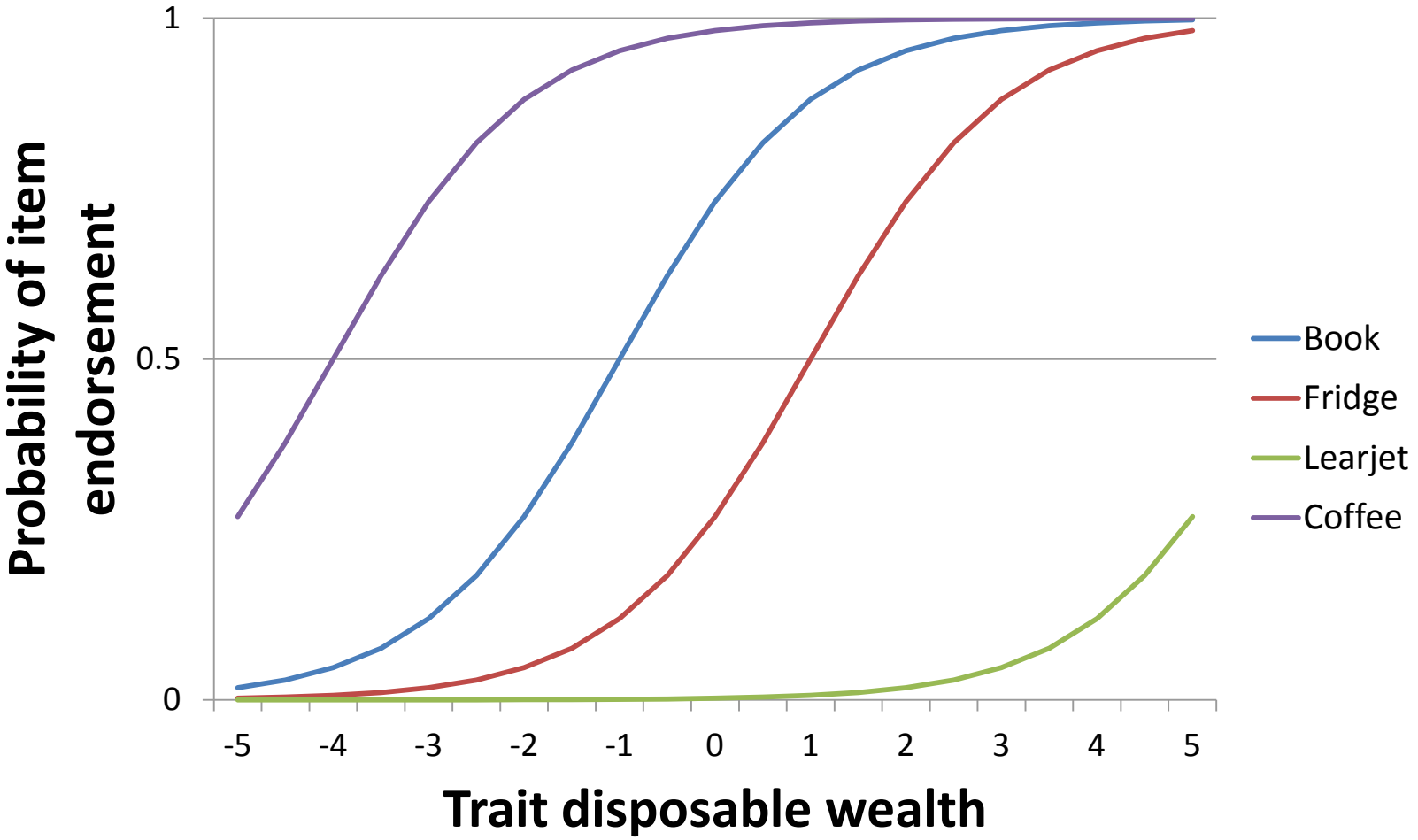
b_i = ***Difficulty*** of item *i* (a.k.a. item *Threshold*)

- This model called '1-parameter' or 'Rasch' model (Rasch, 1960).

Item characteristic curves



Items ‘informative’ about different trait levels



Rasch theory of measurement

- 'Rasch model' describes the theory of measurement as well as the statistical model just described.
- It has some desirable properties:
  - Specific objectivity
    - Each item should rank two individuals similarly.
    - Each person should rank two items similarly.

Rasch theory of measurement

- 'Rasch model' describes the theory of measurement as well as the statistical model just described.
- It has some desirable properties:
  - Sum-score sufficiency
    - Sum of item responses is an unbiased, sufficient statistic for estimating the latent trait.
    - The **number** of endorsements tells us about the trait, their **pattern** does not.

Specific objectivity violated

- Trait – Perceived disposable wealth.
- Additional questionnaire item:
  - "If I wanted to, I could probably afford to do the following this month:"

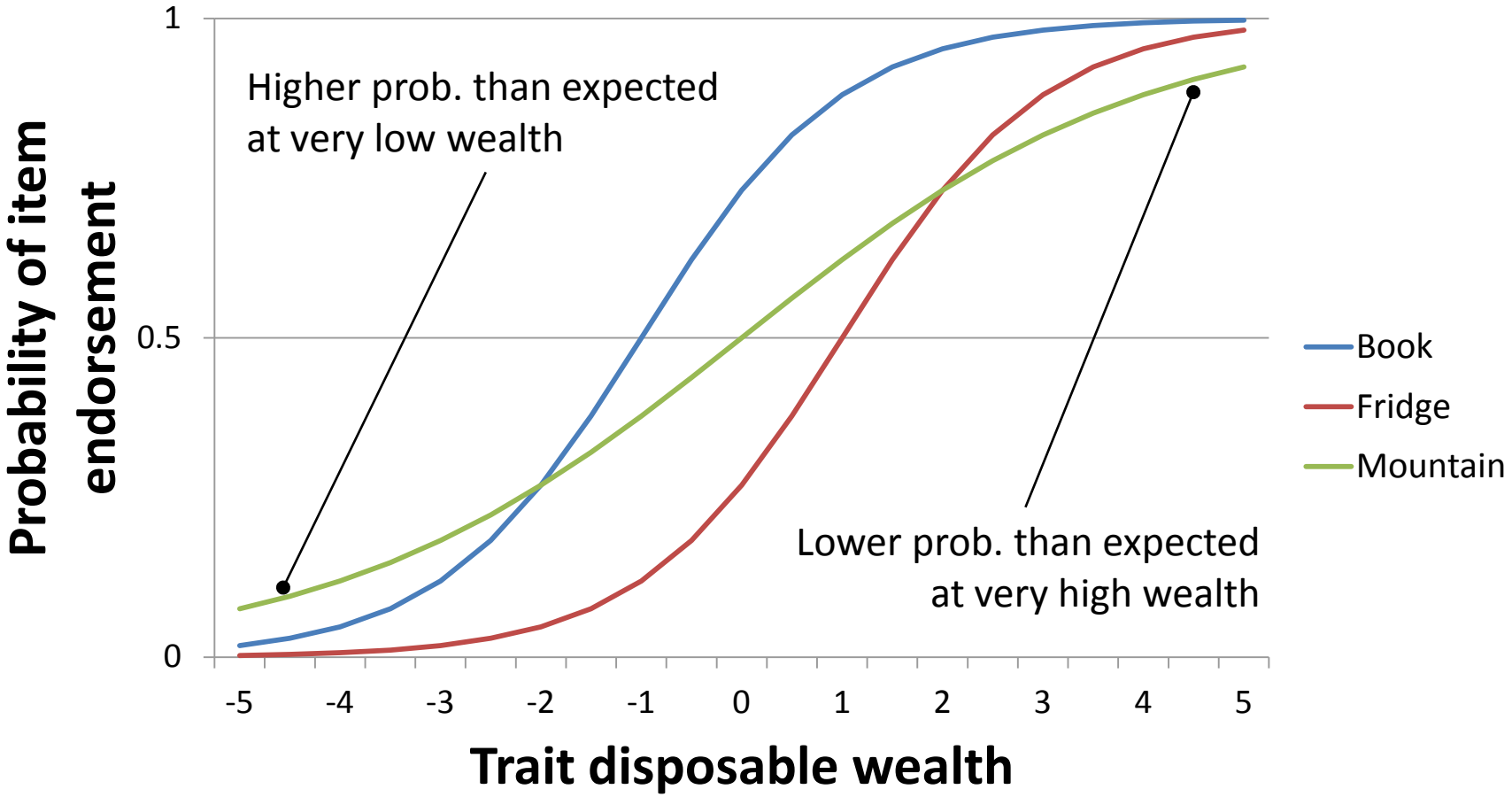
Specific objectivity violated

- Trait – Perceived disposable wealth
- Additional questionnaire item
  - "If I wanted to, I could probably afford to do the following this month:"
  - Climb up a mountain

Specific objectivity violated

- Trait – Perceived disposable wealth
- Additional questionnaire item
  - "If I wanted to, I could probably afford to do the following this month:"
  - Need money (travel, clothes)
    - Also need knowledge, ability
    - Not just asking about wealth

Specific objectivity violated



Specific objectivity violated

- At low levels of disposable wealth, people may be more able to climb a mountain than might be expected, because:
  - They might live nearby, no need to travel far.
  - They might be in a club, go with friends.
- At high levels, people might be less able to climb a mountain because
  - Too much champagne and foie gras, not very fit.

Revised statistical model

$$Y_{ij} = a_i \theta_j - b_i$$

Y_{ij} = Logit that item *i* is endorsed by person *j*

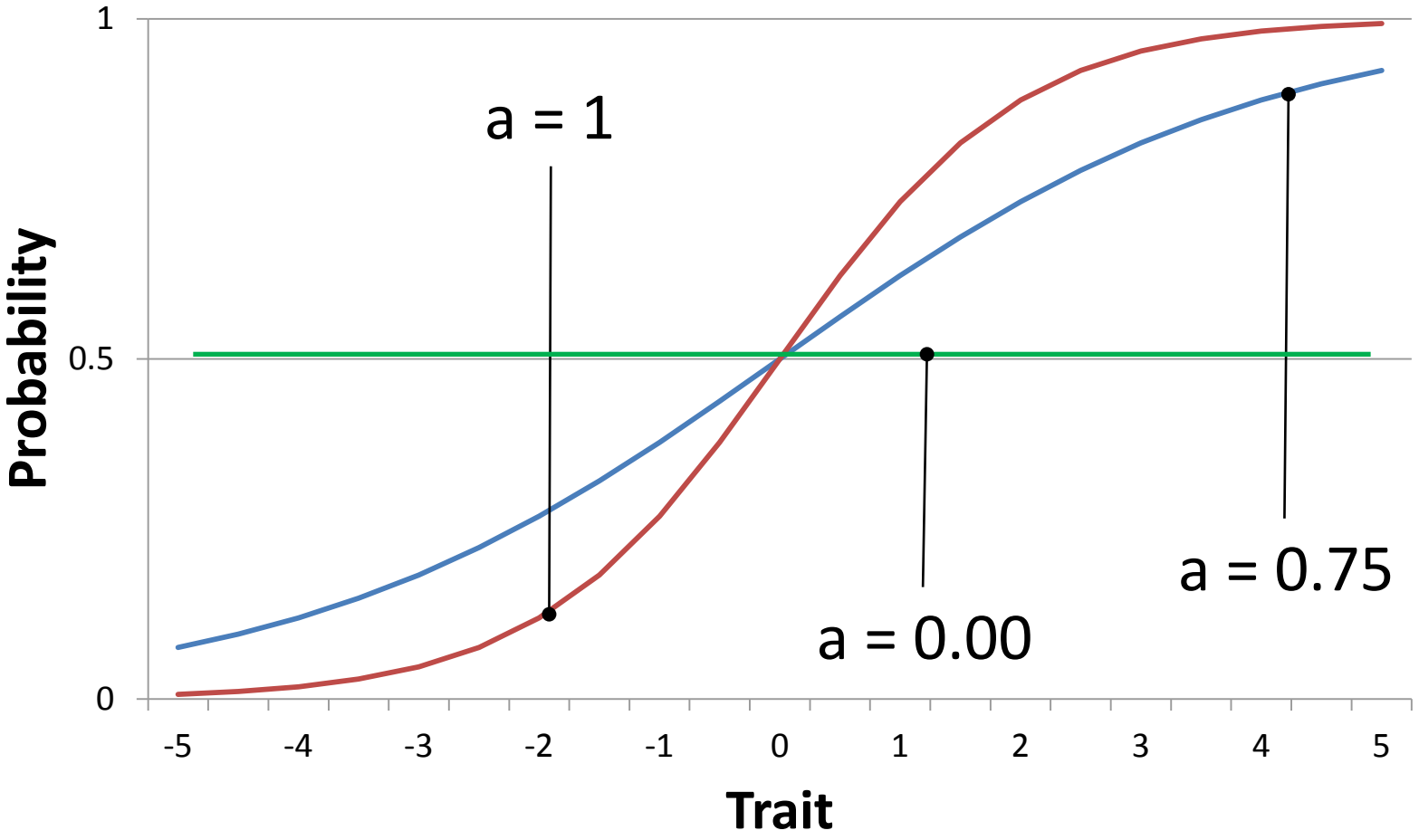
θ_j = ***Trait*** level of person *j*

b_i = ***Difficulty*** of item *i* (a.k.a. item *threshold*)

a_i = ***Discrimination*** of item *i* (a.k.a. item *slope*, or *loading*)

This model called '2-parameter' IRT model.

Same difficulty, different discriminations



British Social Attitudes Survey '09

- How do you think you would feel if a person with a mental health condition such as depression or a personality disorder...
  1. Had been appointed as your boss?
  2. Had joined your quiz team, community group or swimming club?
  3. Were to marry and have a family with one of your close relatives?
  - Very/somewhat comfortable vs. very/somewhat uncomfortable .

British Social Attitudes Survey '09

4. Generally speaking, do you think there is a lot of prejudice in Britain against disabled people in general?
   - A lot/little vs.
     hardly any/none?

British Social Attitudes Survey '09

- Modelling strategy

1. Fit a 1-parameter ('Rasch') model
2. Fit a 2-parameter model

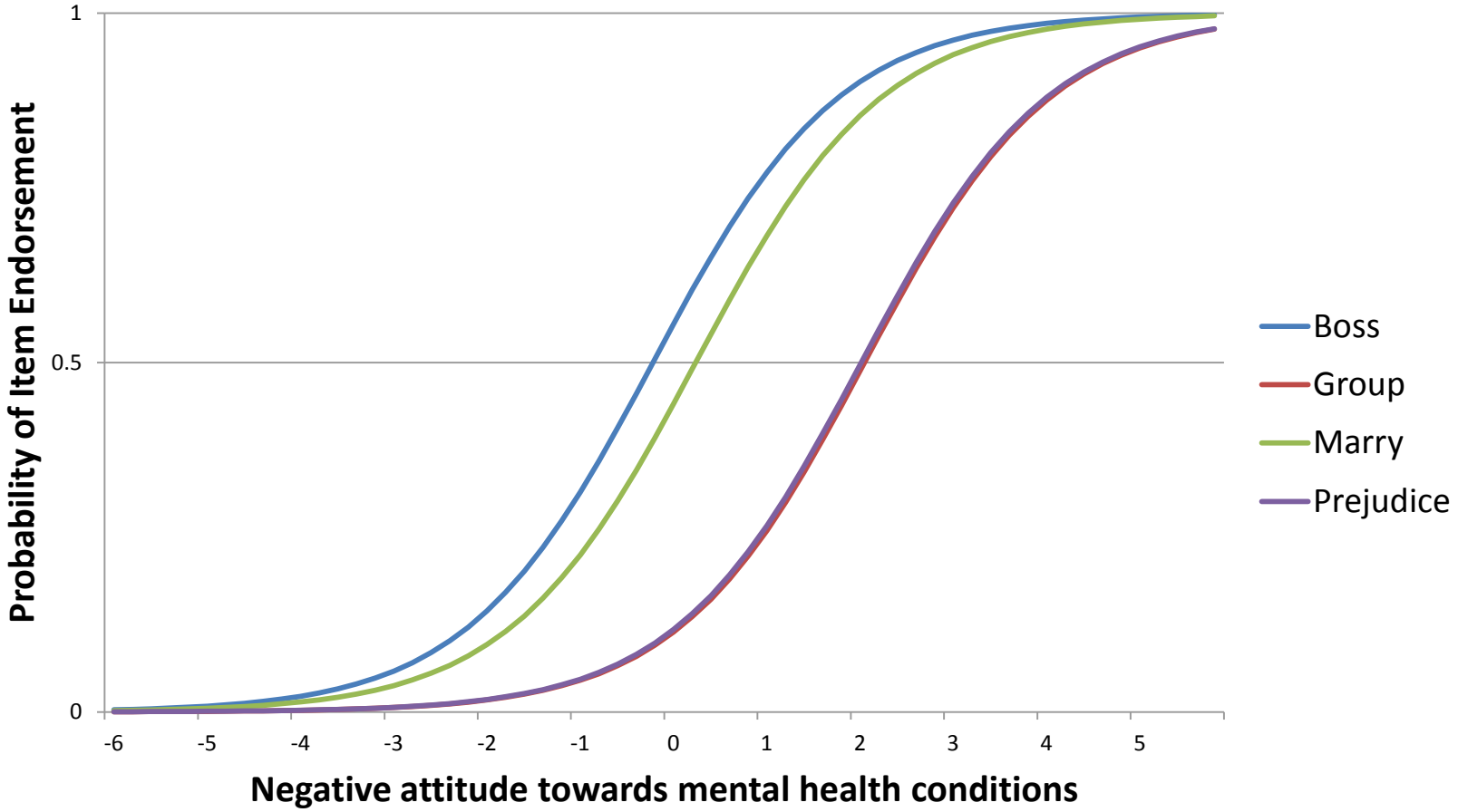
- Test if model 2. fits better than model 1.
  - If so, 'Rasch' measurement is rejected
    - May not be a uni-dimensional scale
  - Summing item responses may not be a good idea

British Social Attitudes Survey '09

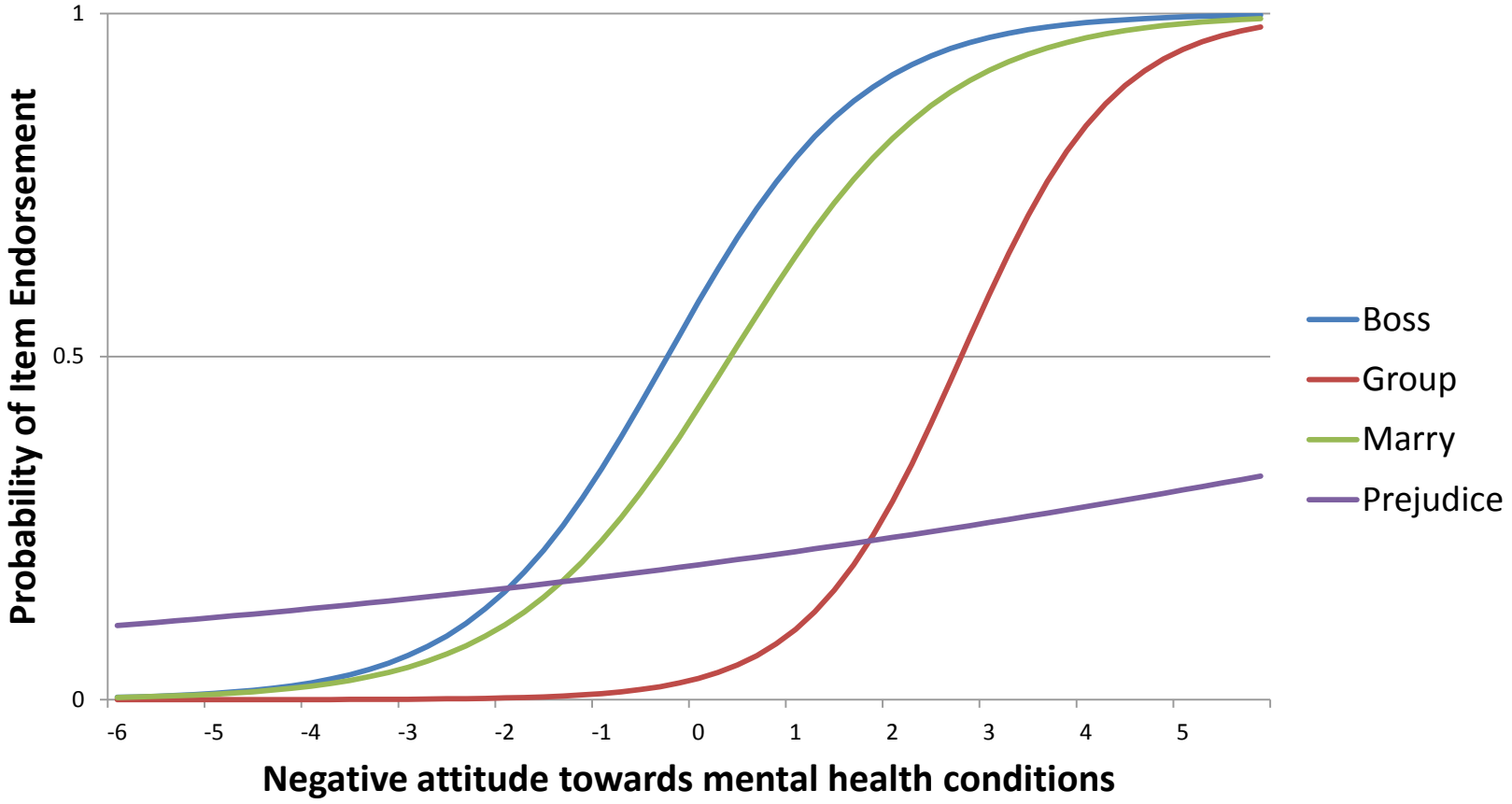
- Modelling strategy

3. Make some predictions, test some hypotheses:
   1. Social 'distance' or 'fixedness' will predict acceptability.
      - $b_{marry} < b_{boss} < b_{group}$. ($b_{prejudice}$?)
   2. 'Prejudice' question is about disability, not mental health per se.
      - $a_{prejudice} < (a_{boss} \mid a_{marry} \mid a_{group})$

1-parameter model of negativity towards mental health conditions



2-parameter model of negativity towards mental health conditions



Expanding IRT – including predictors

- IRT measurement model can form the basis of a model to test substantive hypotheses
  - Original model:

$$Y_{ij} = a_i \theta_j - b_i,$$

  - Attitudes to mental health **generally** less positive with age (period/cohort):

$$\theta_j = \gamma_1 AGE_j,$$

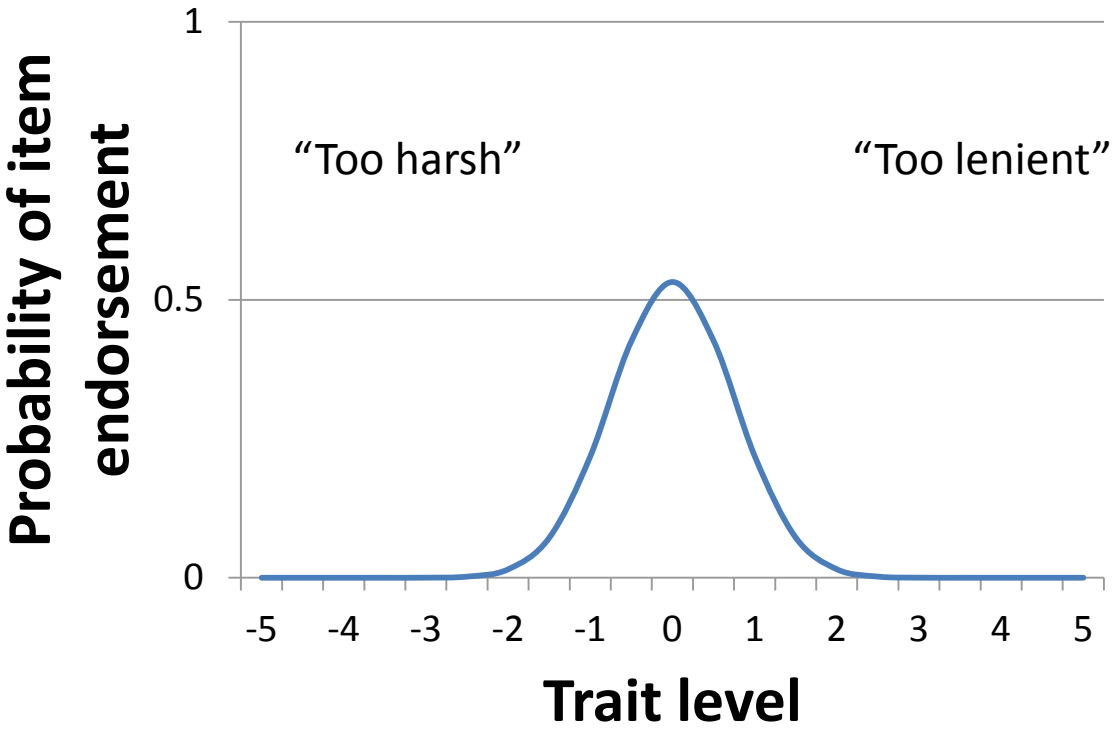
  - Attitudes to mental health in **marriage specifically** less positive with age (period/cohort):

$$b_{marry} = \gamma_2 AGE_j.$$

Other types of IRT model

- There are literally dozens of kinds of IRT model, each suitable for a particular measurement application.
  - For example, 1- and 2-parameter models assume a monotonic relationship between the latent trait and response probability.
    - This is not always the case.
  - Do you agree with the following?:
    - “A whole-of-life prison sentence gives the murderer what he deserves”

Other types of IRT model



"A whole-of-life prison sentence gives the murderer what he deserves"

Non-monotonic

- Response probability goes up then down with increasing trait level
- This requires an 'unfolding' model (e.g. Coombs, 1960; Andrich, 1988)

Summary

- IRT is a measurement theory that maps data observed on participants to the latent traits assumed to be causing the observations.
  - Data often comes from questionnaires, but could come from anywhere, as long as we have a substantive theory that links the two.
- IRT is a family of statistical models that can be used to assess the plausibility of the measurement theory

Summary

- IRT makes explicit the assumptions required to justify making inference about latent qualities based upon observations.
- IRT can be used to assess the reliability and validity of observations.
- IRT provides a method to specify and test detailed substantive hypotheses.

Guides and tutorials - theory

- Baker, F. B. (2001). The basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation. http://tinyurl.com/bakerIRT
- Reeve, B. B. (2002?). Modern Measurement Theory. Tutorial written for the Cancer Outcomes Measurement Working Group, National Cancer Institute, USA. http://tinyurl.com/reeveIRT
- Van der Linden, W. J. & Hambleton, R. K. (1997). Handbook of modern item response theory. New York: Springer

Guides and tutorials - practice

- Mplus
  - Uses a Structural Equation Modelling approach to fit exploratory and confirmatory IRT models.
  - Download free demo version of Mplus from:
    - www.statmodel.com
  - Download introductory tutorial from:
    - http://tinyurl.com/shryane-mplus-manual
    - http://tinyurl.com/shryane-mplus-examples
    - See section 9, IRT models

Guides and tutorials - practice

- Stata
  - The **gllamm** command uses a multilevel modelling approach to fit confirmatory IRT models.
  - Download the manual and lots of worked examples from
    - www.gllamm.org

Guides and tutorials - practice

- R
  - Download R for free from
    - www.r-project.org
  - The **ltm** (latent trait modelling) library allows you to fit a wide range of IRT models
    - Can't include predictors of the latent traits

Guides and tutorials - practice

- SPSS v.19
  - The GLMM (generalized linear mixed models) command allows you use a multilevel modelling approach to fit a 1-parameter (‘Rasch’) model.
  - Not possible to fit a 2-parameter or other models.

References

- Andrich, D. (1988). The Application of an Unfolding Model of the PIRT Type to the Measurement of Attitude. *Applied Psychological Measurement*, 12(1), 33-51. http://conservancy.umn.edu/bitstream/104143/1/v12n1p033.pdf
- Coombs, (1960). A theory of data. *The Psychological Review*, 67(3), 143-159.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA.