A gentle intro to item response theory

Alvin Tan & George Kachergis

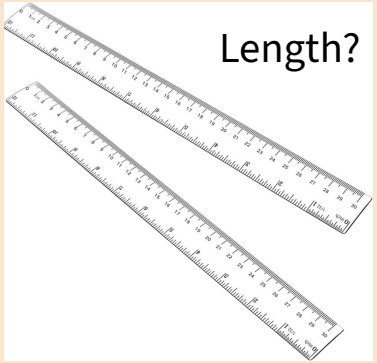
CogSci pre-workshop tutorial, 2024-07-01

Outline

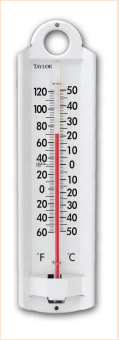
- Why IRT?
- What is IRT?
- How to use IRT?
- Other IRT variants
- Cool stuff you can do with IRT
- Practical

IRT motivation

How do we measure stuff?

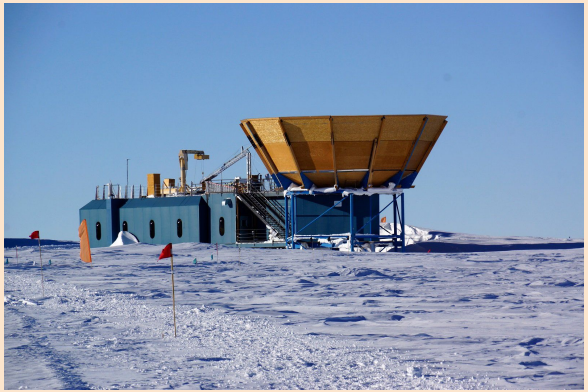


Length?



Temperature?

Cosmic microwave background radiation?



Personality?

Affect?

Lack of direct observational access

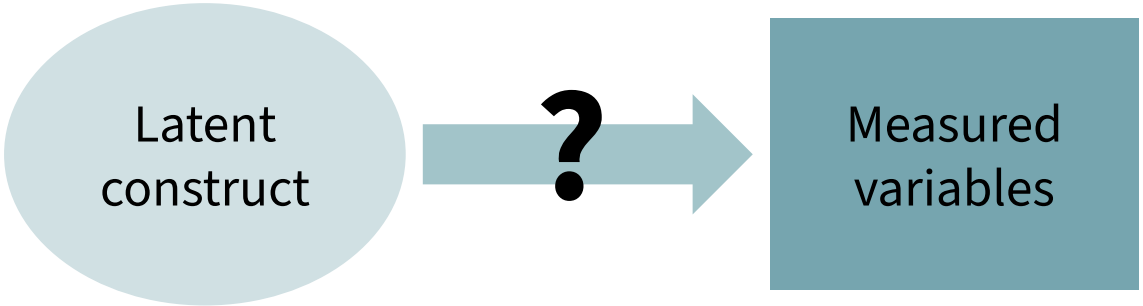
Memory?

Cognitive ability?

The place for psychometrics

What are the sources of variance?

Is my measurement consistent?



How are latent constructs organised?

How much error is there?

What is psychometrics?

Validity

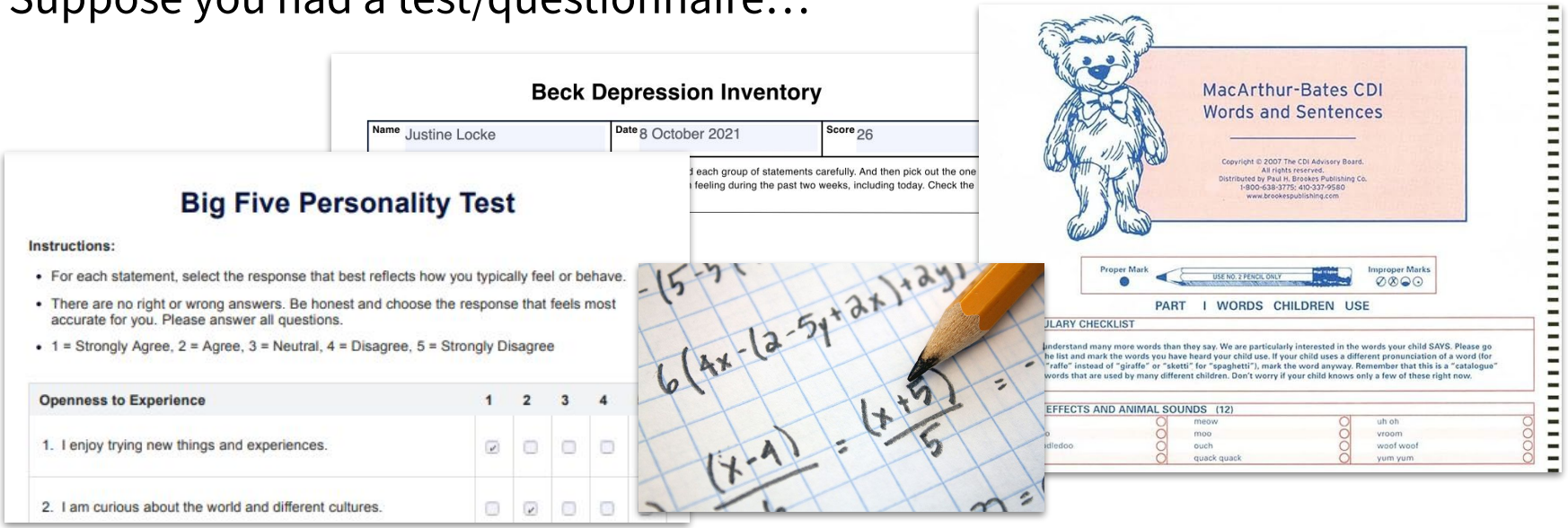
How do measurements relate to the underlying latent constructs?

Instrumentation	*Measurement model*	*Latent structure*
• Scaling • Reliability • Bias • Measurement invariance • Differential item functioning	• **Classical test theory** • **Item response theory** • Generalisability theory	• Factor analysis • Latent variable models • Network theory • Structural equation modelling

Borsboom, Mellenbergh, & van Heerden (2004)

Classical test theory

Suppose you had a test/questionnaire…



…how would you determine a test-taker's level?

Classical test theory

One obvious answer: just add it up (= sum scores/true scores)!

$Y_i = T_i + e_i$

Y_i: observed score
T_i: true score
e_i: error

With k items,

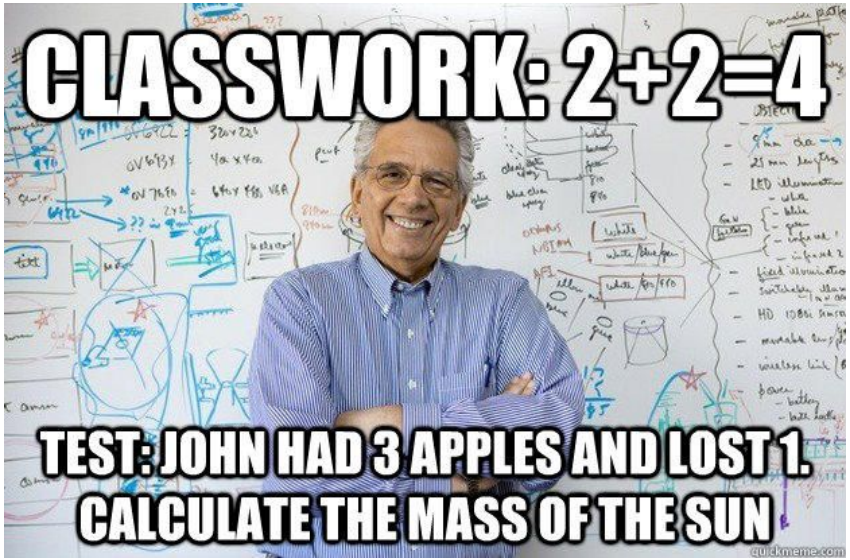
$(\Sigma Y_i)/k = (\Sigma T_i)/k + \underbrace{(\Sigma e_i)/k}$

Expected value = 0
Variance decreases as k increases

Classical test theory

One obvious answer: just add it up (= sum scores/true scores)!

BUT this assumes that **all items behave exactly the same**

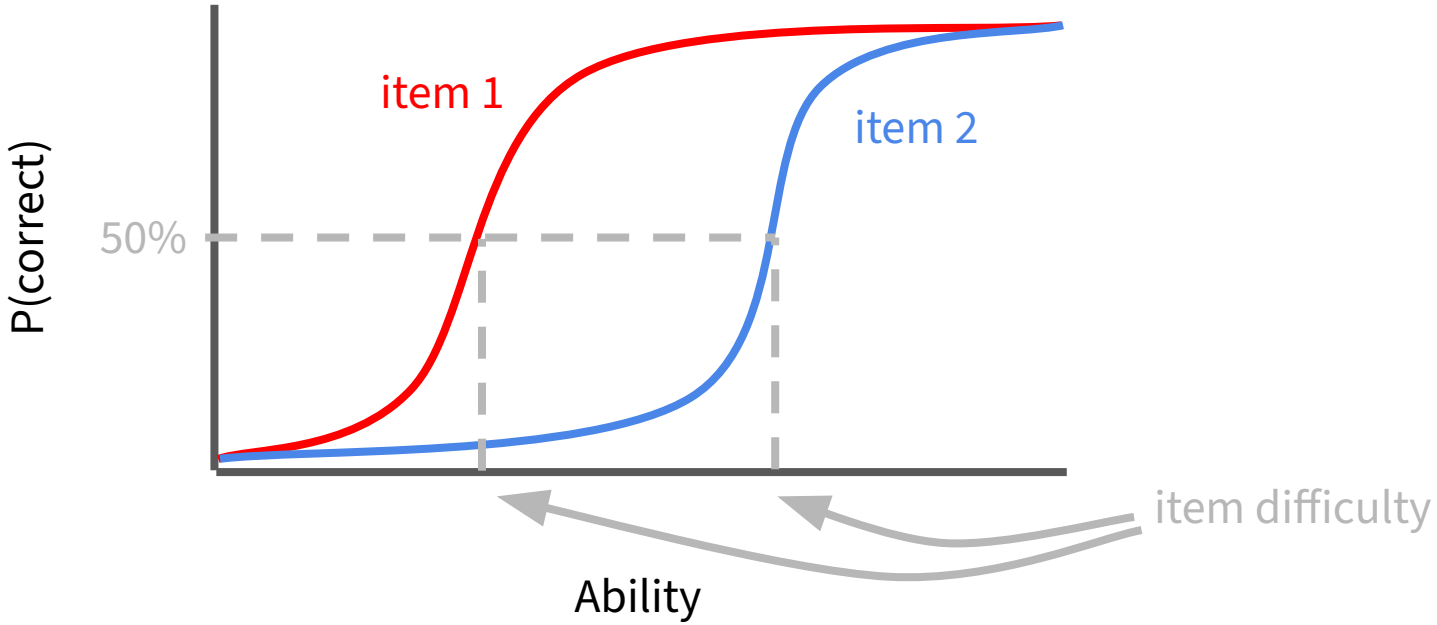


Classical test theory

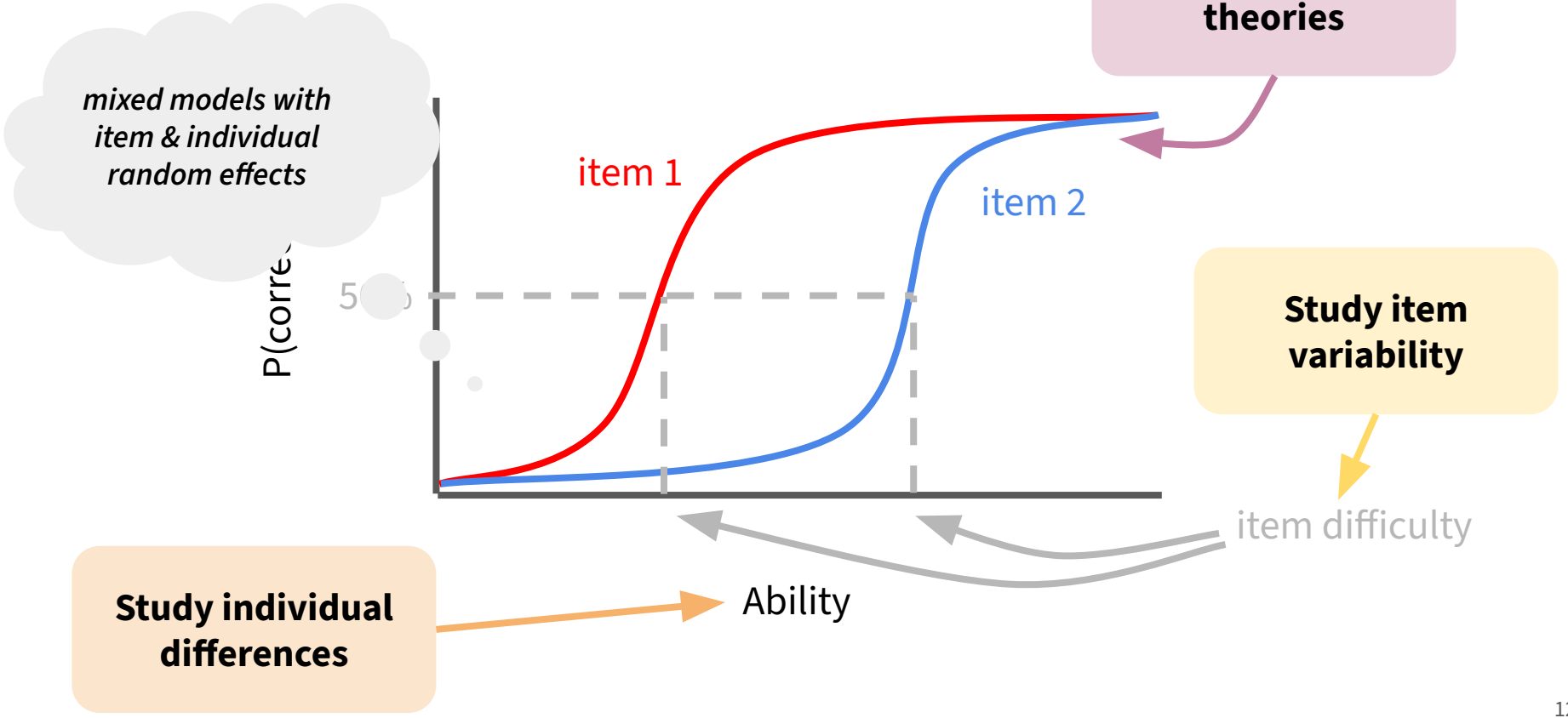
Consequences of this assumption:

- Only test-level information; no item-level information
- Reliability assumes parallel forms, but no way of testing
- Assumes measure is equally good over all levels of ability
- No generalisability to new tests → incentive to use the same ones

Item response theory

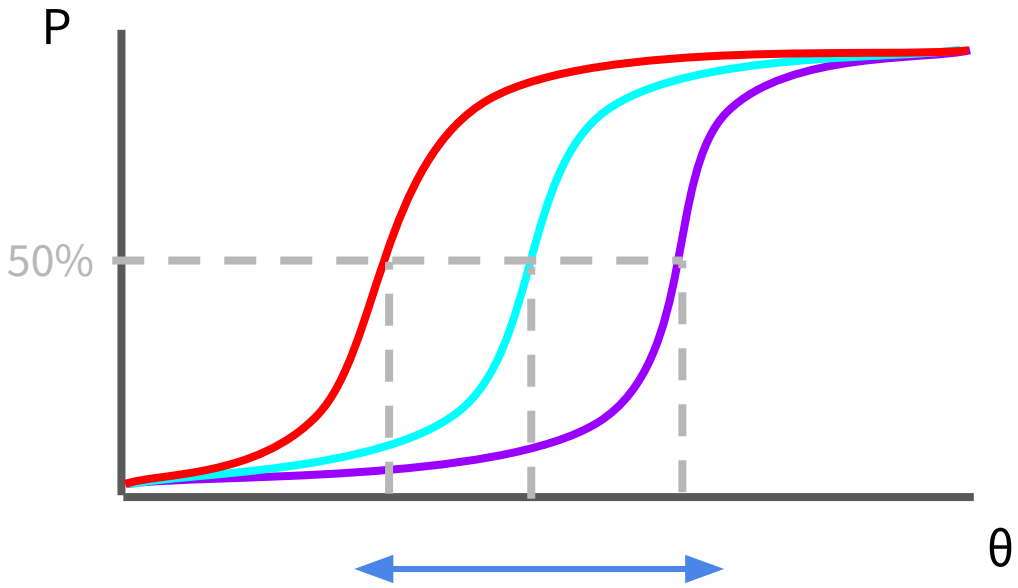


Item response theory



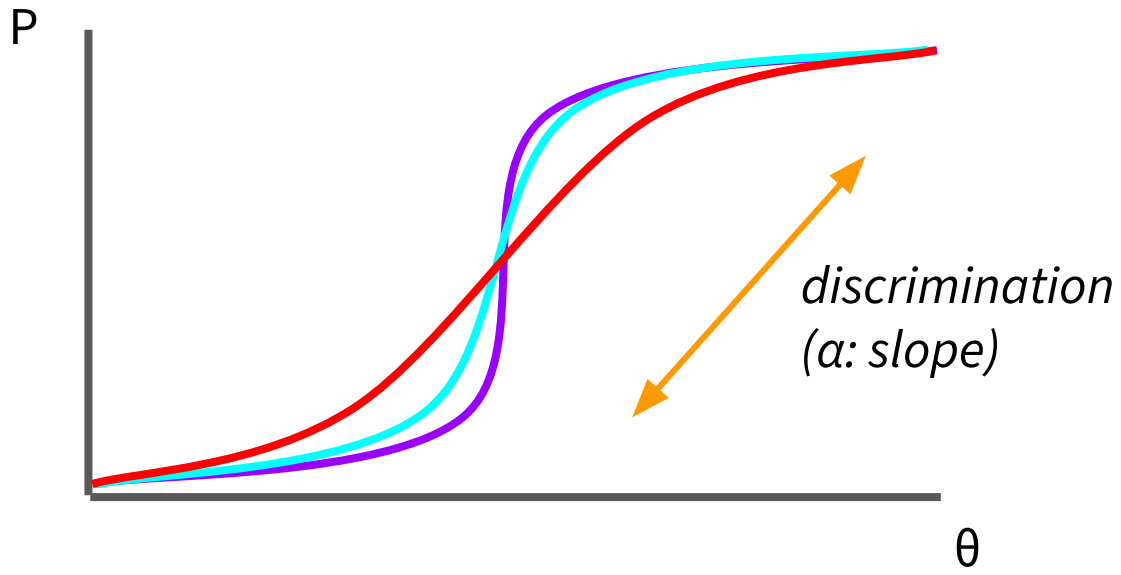
IRT basics

Item response theoretic model (2PL)

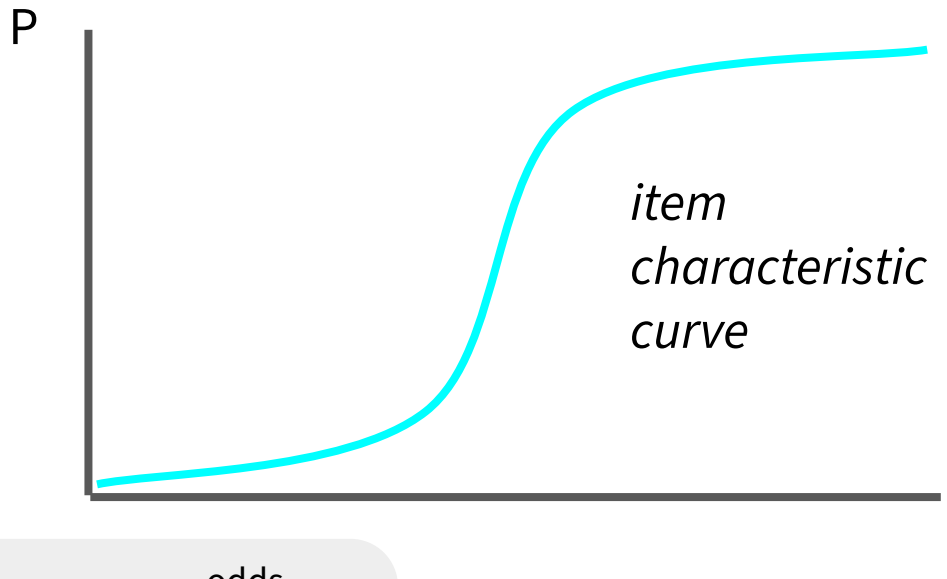


difficulty (β: location)

Item response theoretic model (2PL)



Putting the L in 2PL



$$\text{logit}(P) = \overbrace{\ln(P / (1 - P))}^{\text{odds}} = \boldsymbol{\alpha(\theta - \beta)}$$

$$P(X_{ip} = 1 | \theta_p, \beta_i) = \frac{\exp(\alpha_i(\theta_p - \beta_i))}{1 + \exp(\alpha_i(\theta_p - \beta_i))}$$

Item response theory

Consequences of this model:

- Item are on a meaningful scale (theta measures difficulty)
- Individuals are on a meaningful scale (latent trait scores)
- These two dimensions lie on the same scale
- Can calculate probability of an individual getting an item correct directly

$$P(X_{ip} = 1|\theta_p, \beta_i) = \frac{\exp(\alpha_i(\theta_p - \beta_i))}{1 + \exp(\alpha_i(\theta_p - \beta_i))}$$

Ex: Communicative Development Inventories (CDIs)

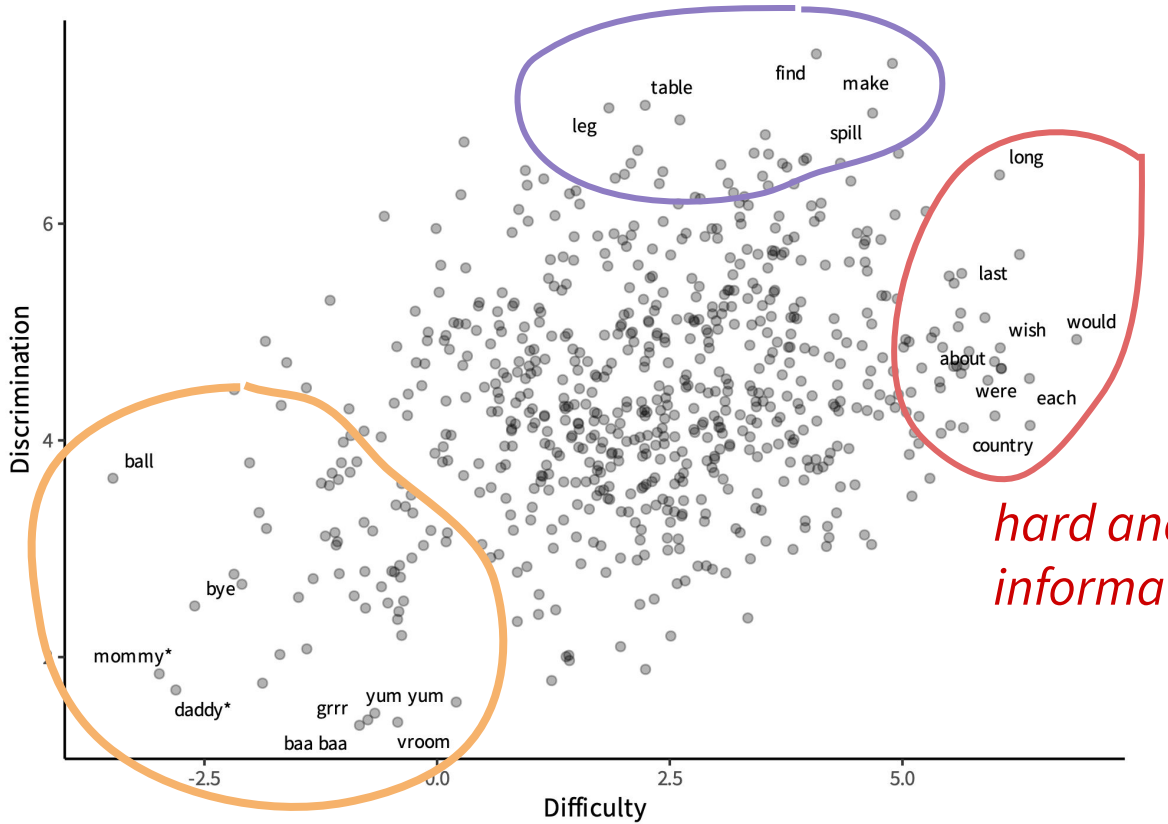
1. SOUND EFFECTS AND ANIMAL SOUNDS (12)		
baa baa	meow	uh oh
choo choo	moo	vroom
cockadoodledoo	ouch	woof woof
grrr	quack quack	yum yum

2. ANIMALS (Real or Toy) (43)		
alligator	duck	penguin
animal	elephant	pig
ant	fish	pony
bear	frog	puppy
bee	giraffe	rooster
bird	goose	sheep
bug	hen	squirrel
bunny	horse	teddybear
butterfly	kitty	tiger
cat	lamb	turkey
chicken	lion	turtle
cow	monkey	wolf
deer	moose	zebra
dog	mouse	
donkey	owl	

Frank et al. (2017)

Ex: CDI parameter space (2PL)

fairly hard and informative



easy and not informative

hard and somewhat informative

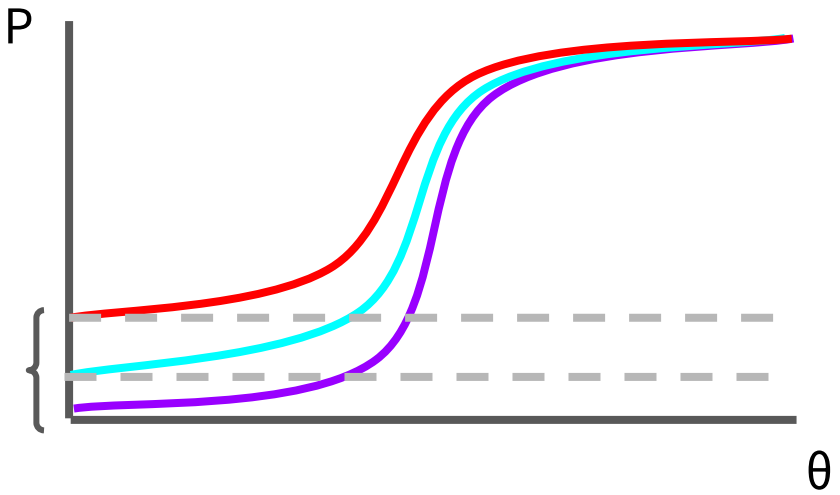
slide credit: Mike Frank

IRT variants

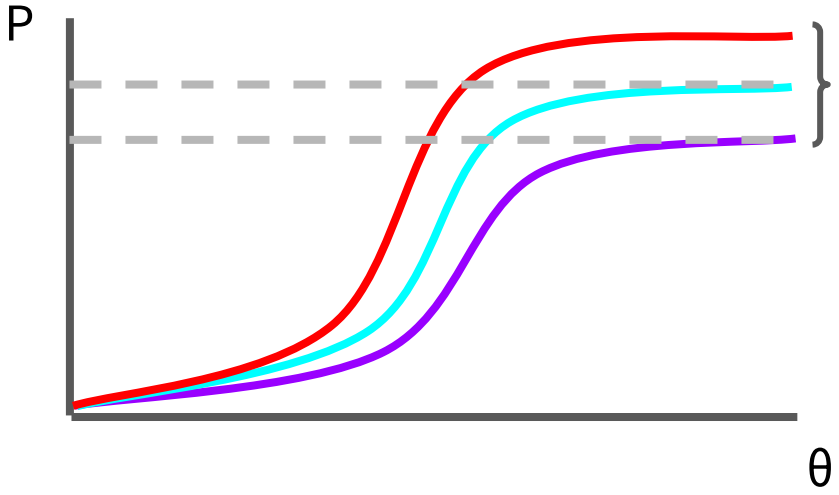
IRT variants: Number of parameters

- 1PL: location (difficulty)
- 2PL: slope (discrimination)
- 3PL: lower bound (guessing)
- 4PL: upper bound (errors)

Para para parameter



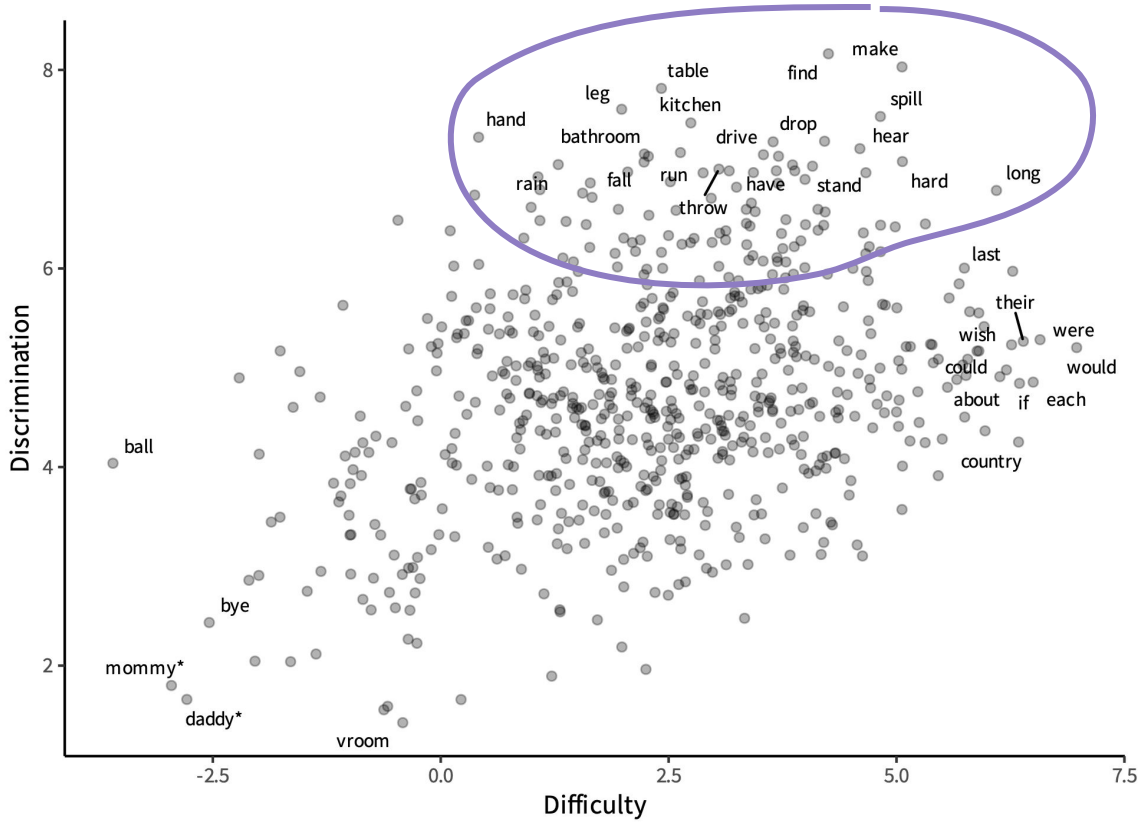
3PL: *lower asymptote*



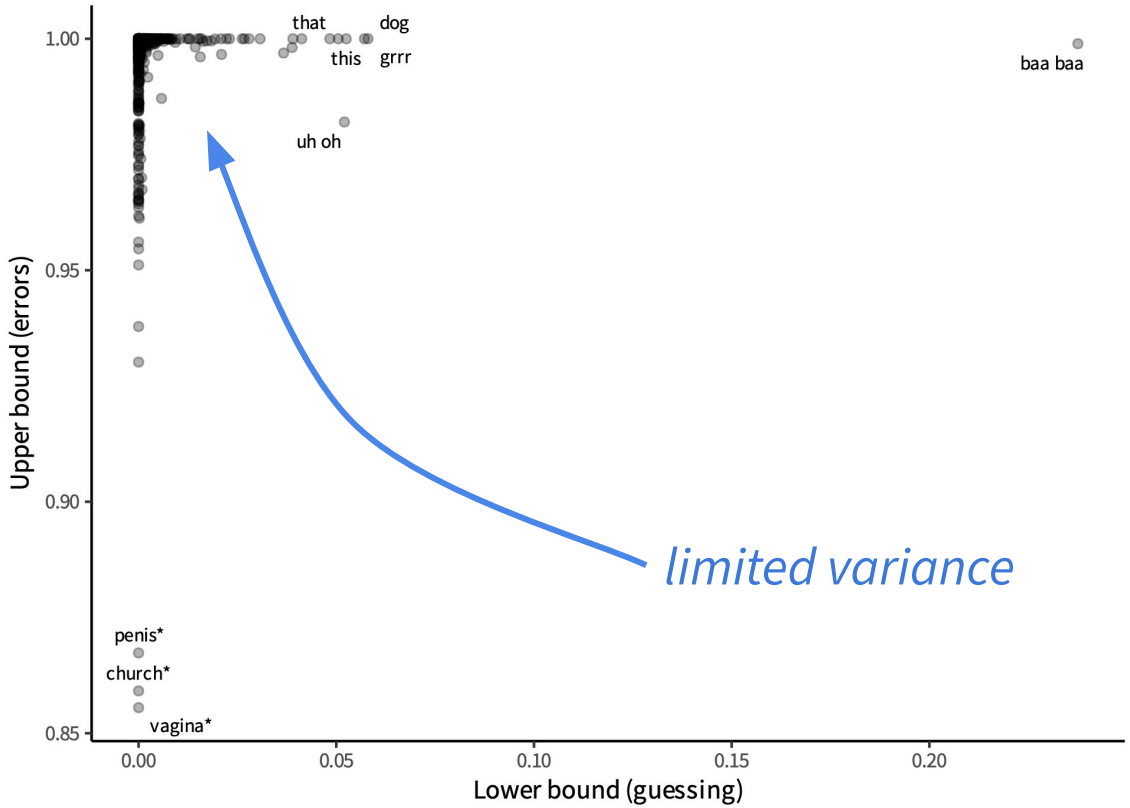
4PL: *upper asymptote*

Ex: CDI parameter space (4PL)

more informative items



Ex: CDI parameter space (4PL)



Ex: CDI model fits

Model	BIC
1PL	4788209
2PL	**4668081**
4PL	4676954

see Kachergis et al. (2022)

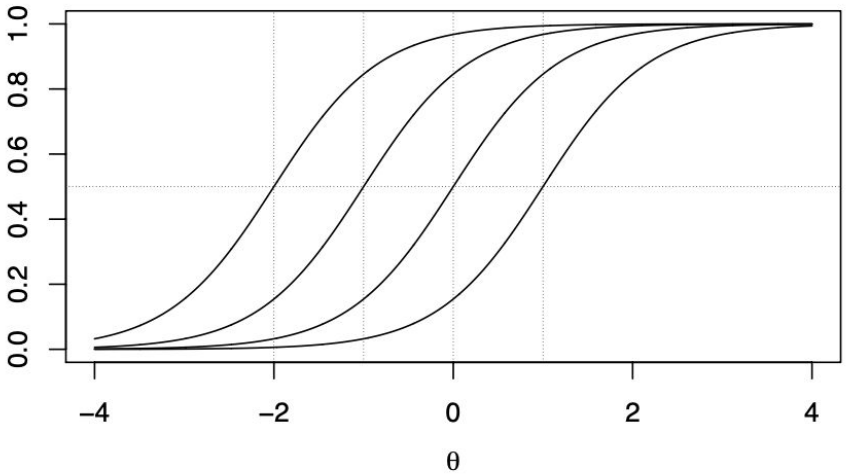
IRT variants: Item type

- Vanilla IRT operates over binary observations
- How do we move from dichotomous to polytomous measures?
- Solution: decompose polytomous items into sets of dichotomous comparisons
- Two main classes: difference models (e.g., GRM) and divide-by-total models (e.g., PCM)

Difference models: Graded response model

Dichotomise based on being above/below a threshold; e.g., with $k = 5$

- $P_1 = Pr(y_{ip} > 0)$
- $P_2 = Pr(y_{ip} > 1)$
- $P_3 = Pr(y_{ip} > 2)$
- $P_4 = Pr(y_{ip} > 3)$



cumulative response curves

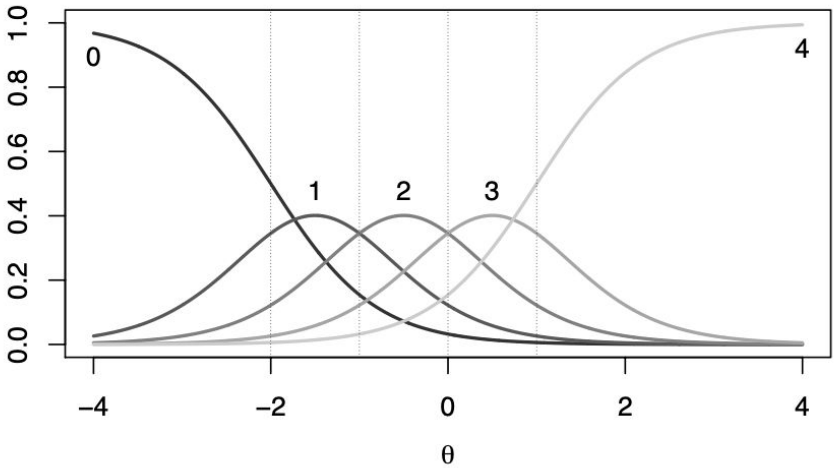
slide credit: Ben Domingue

Difference models: Graded response model

Dichotomise based on being above/below a threshold; e.g., with $k = 5$

- $P_1 = Pr(y_{ip} > 0)$
- $P_2 = Pr(y_{ip} > 1)$
- $P_3 = Pr(y_{ip} > 2)$
- $P_4 = Pr(y_{ip} > 3)$

$$Pr(y_{ip} = 2) = P_3 - P_2$$



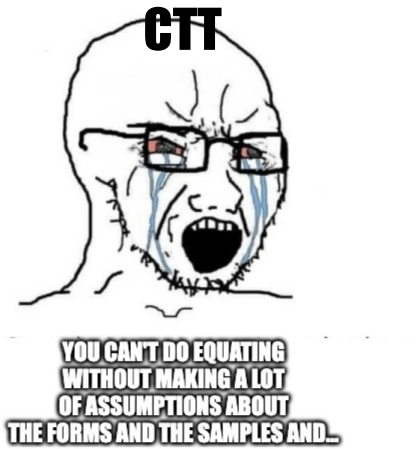
category response curves

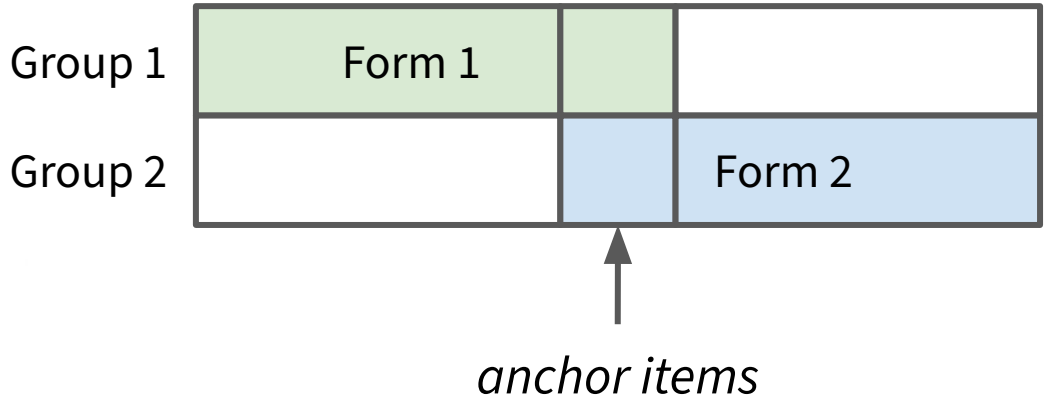
IRT techniques

IRT techniques: Equating

How do you handle multiple versions of an instrument?

- Avoid practice effects
- Prevent cheating
- Updating

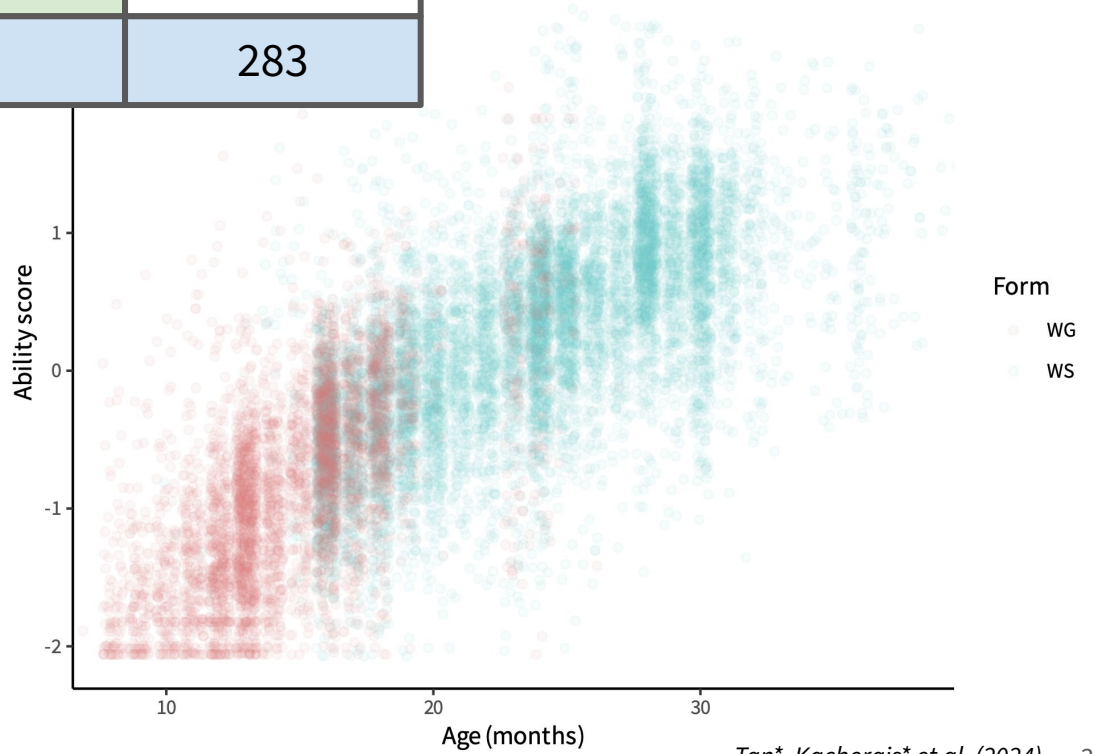




→ **concurrent calibration**

Ex: Stitching different CDI forms

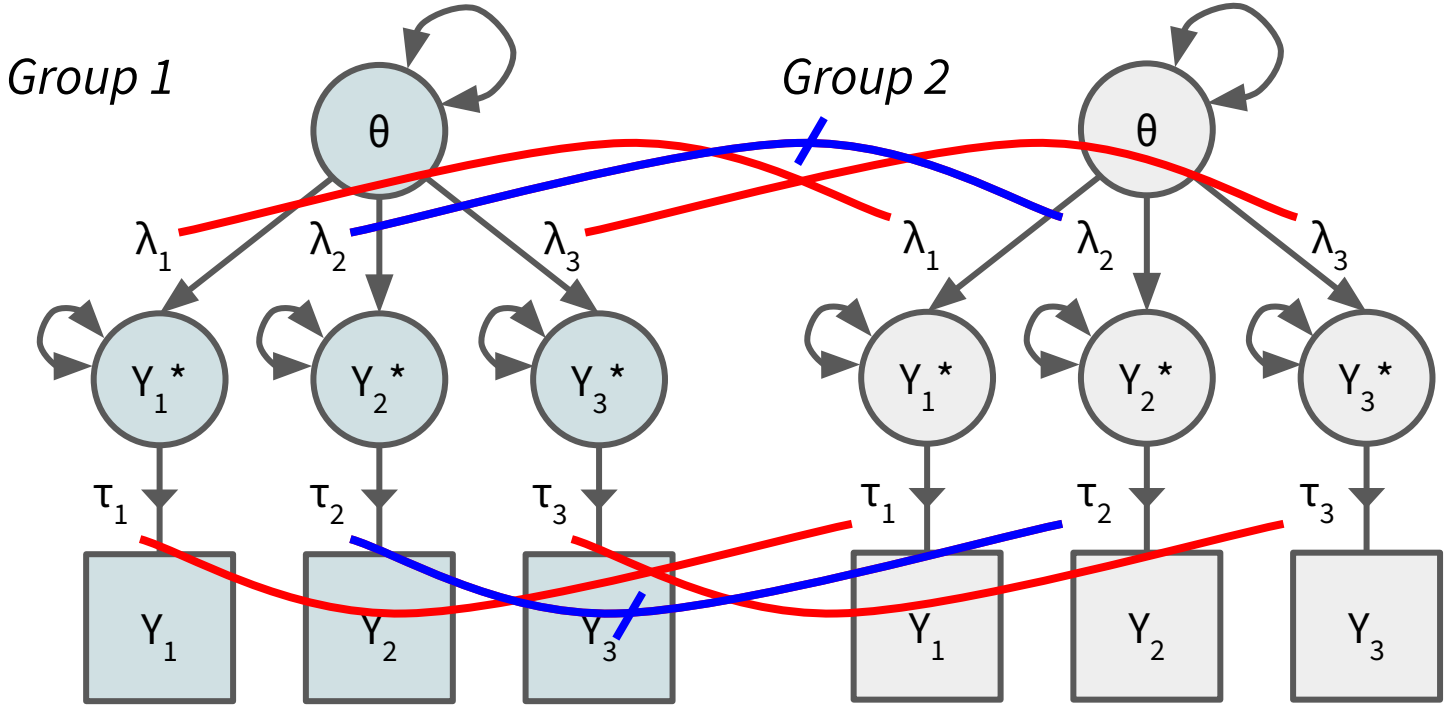
WG: 12–18mo	2	397	
WS: 16–36mo		397	283



Tan, Kachergis* et al. (2024)*

IRT techniques: Differential item functioning

How do we know if items have measurement invariance across groups?

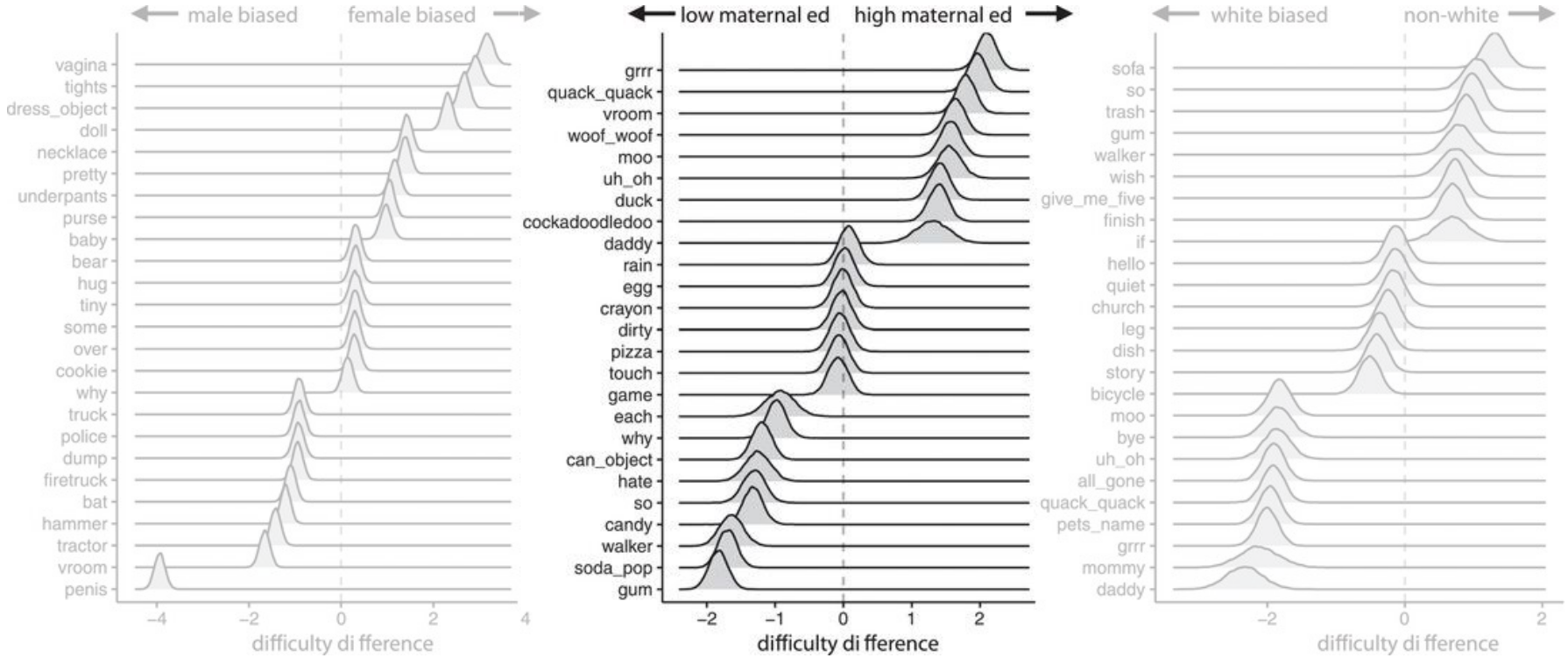


Constrained model

Free model

Likelihood ratio test

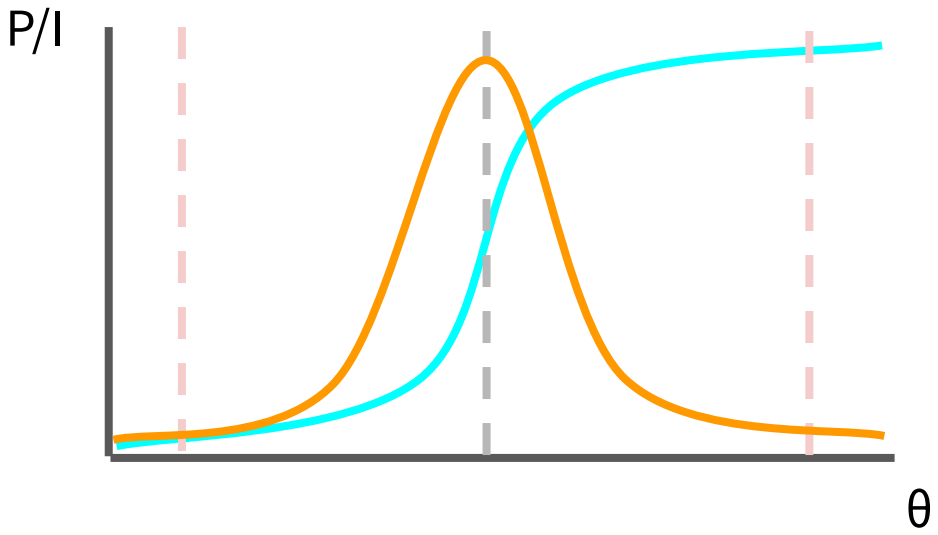
Ex: DIF in CDIs



Kachergis, Francis, & Frank (2022)

IRT techniques: Item information

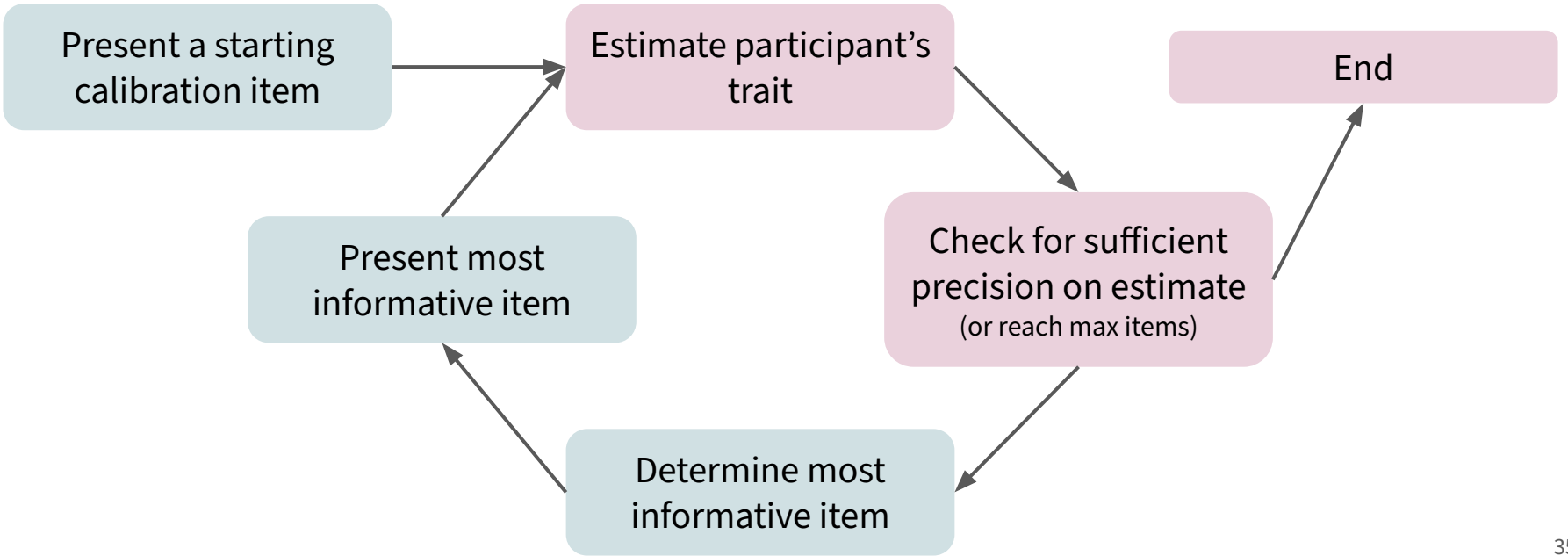
Because items are at different locations, they are differently informative about individuals of different trait levels



$$I(\theta) = dP(\theta) / d\theta$$
$$= P(\theta)\,(1 - P(\theta))$$

IRT techniques: Computerised adaptive testing

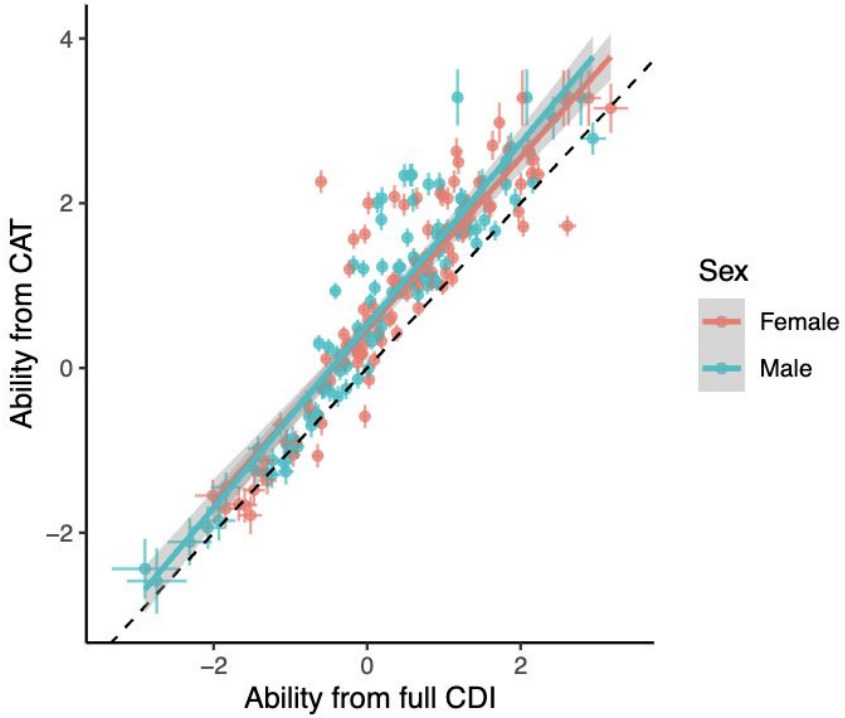
Instead of giving a full instrument, choose items that are most informative given your current estimate of the participant's trait



Ex: CDI-CAT

CDI:
~600–800 items

CDI–CAT:
~35 items



Kachergis et al. (2022)

Summary

- IRT helps us better handle inter-item variability
- IRT is effectively logistic regression with item-level parameters
- IRT models can have various numbers of parameters, and can handle different types of categorical variables
- IRT enables us to do lots of statistical techniques that are not possible with CTT (e.g., equating, DIF, CAT)

ty+q?

Slides, links, resources:
https://psychometrics-workshop.github.io/