# Item response theory for designing calibrated face ability tests

Geraldine Jeckeln+, Ying Hu+, Jacqueline Cavazos+, Amy N. Yates*,
Carina A. Hahn*, Larry Tang++, Alice J. O'Toole+, **P. Jonathon Phillips***

* NIST, + University of Texas at Dallas, ++ University of Central Florida

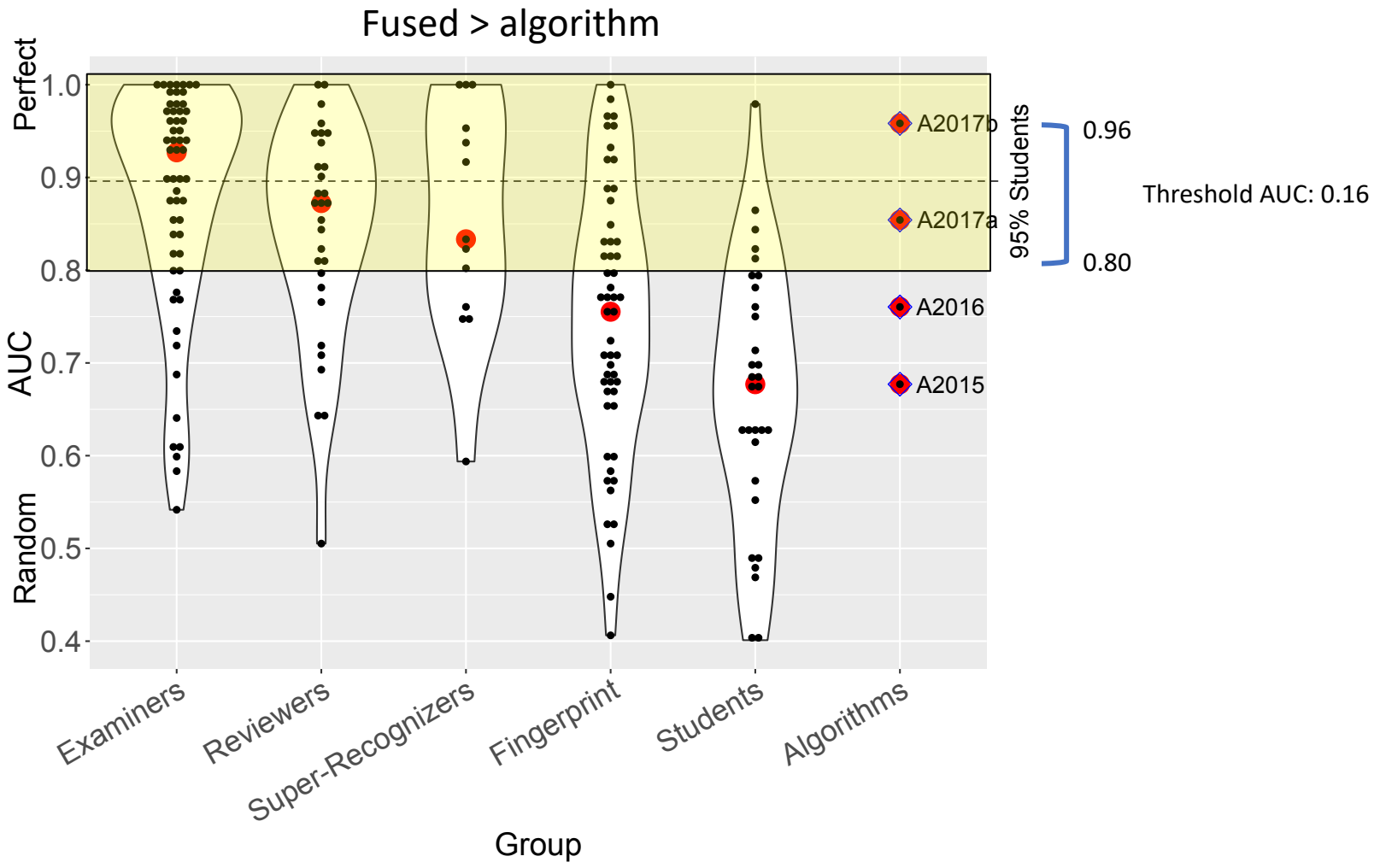
National Institute of Standards and Technology
U.S. Department of Commerce

INFORMATION TECHNOLOGY LABORATORY

# Outline

- Need for Calibrated Face Tests
- Introducing item response theory (IRT)
- Triads
- IRT experiments
  - Baseline
  - Extended

# Lesson from fusion



# Importance

- Over use of existing tests
- Screening for ability
  - Large range of performance for face expert groups (Phillips et al 2018)
  - Recruitment
- When to fuse, when not to fuse
- Proficiency for face identification professionals
- Consistency of performance
  - Day-to-day variation in ability

# Understanding Item Response Theory (IRT)

# What is IRT?

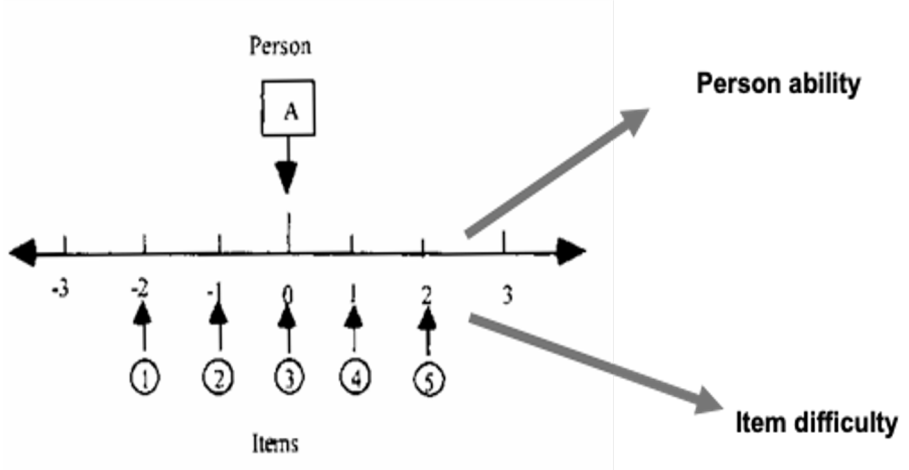
Model of a person's ability on a given test (e.g. comparing face images)

Think of the SAT. A large bank of questions (items) to pull from for each test. The difficulty of each item is known, so scores (a person's ability) between different tests are directly comparable.

# What is IRT?

Model of a person's response on a given test item (question, image pair, etc.)

Advantage: Subject's ability and item difficulty located on the same scale



# Advantages of IRT

- Measure subjects' ability based on a set of test items
- Measure the difficulty of an item
- Create a "item bank," with prior knowledge of the test items
- Design tests of same difficulty

# Triads

# Comparison / Identification / Matching

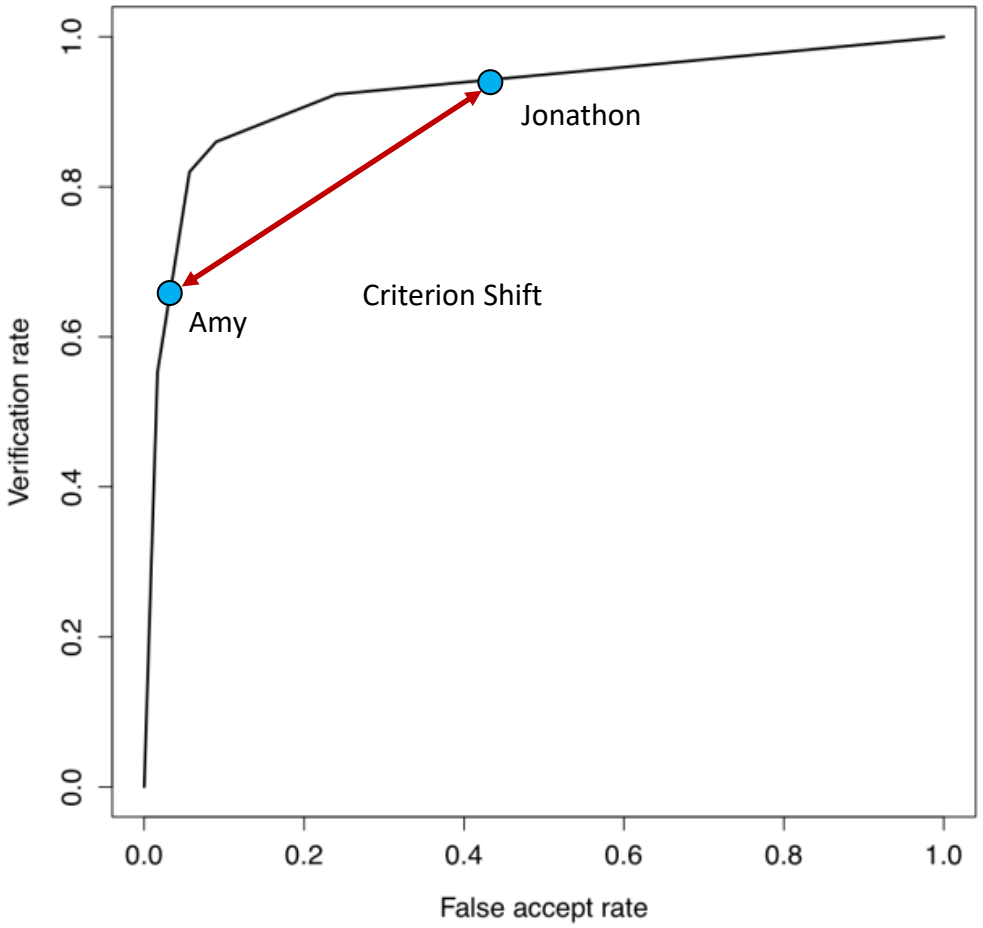

| | |
|---|---|
| +3 | The observations strongly support that it is the same person |
| +2 | The observations support that it is the same person |
| +1 | The observations support to some extent that it is the same person |
| 0 | The observations support neither that it is the same person nor that it is different persons |
| -1 | The observations support to some extent that it is not the same person |
| -2 | The observations support that it is not the same person |
| -3 | The observations strongly support that it is not the same person |

# Same/Different Paradigm


+1 The observations support to some extent that it is the same person

-1 The observations support to some extent that it is not the same person

# Criterion shift



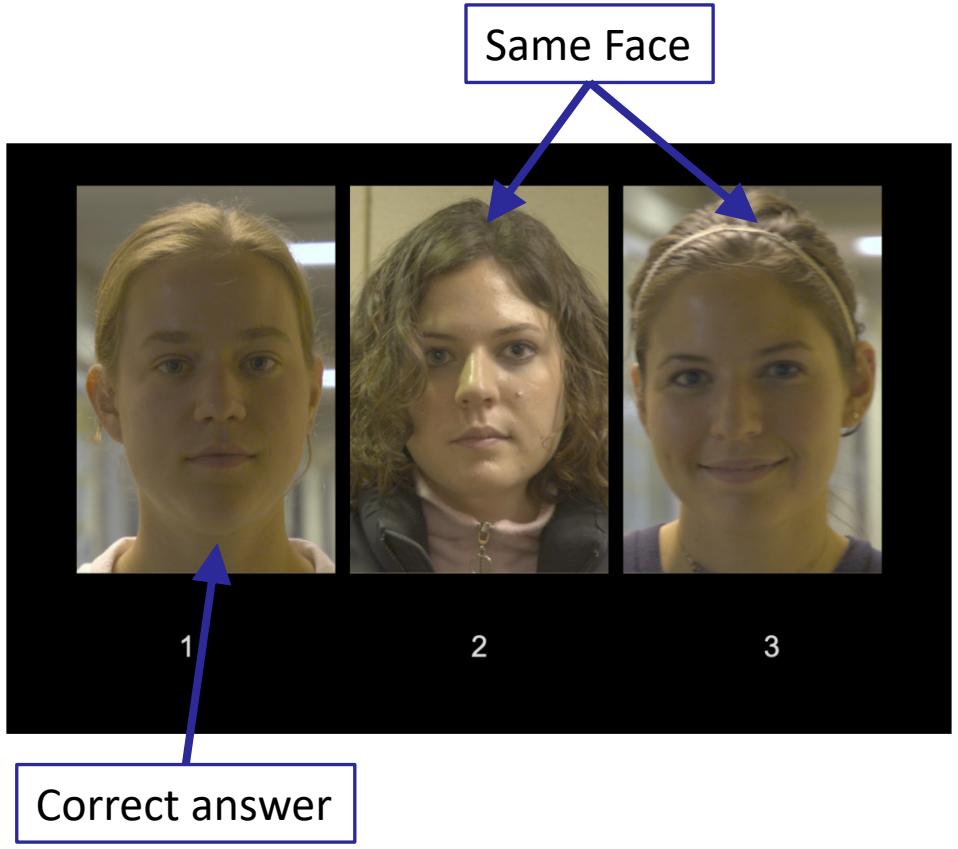
# Triads test

Three images

- Two images of same face
- One image of a different face

Choose the “odd” one out.

3 Alternative Forced Choice Task (3AFC)



# Why triads?

- Overcomes the criterion problem
  - Accuracy is not dependent on match/non-match decisions
- Note: cannot calculate false alarms response with triads

# Experiments—baselining

# Triad experiment

Goals of experiment

- Validate triad design for IRT
- Create item bank for future experiments

Participants

- 198 UT Dallas students

Stimuli

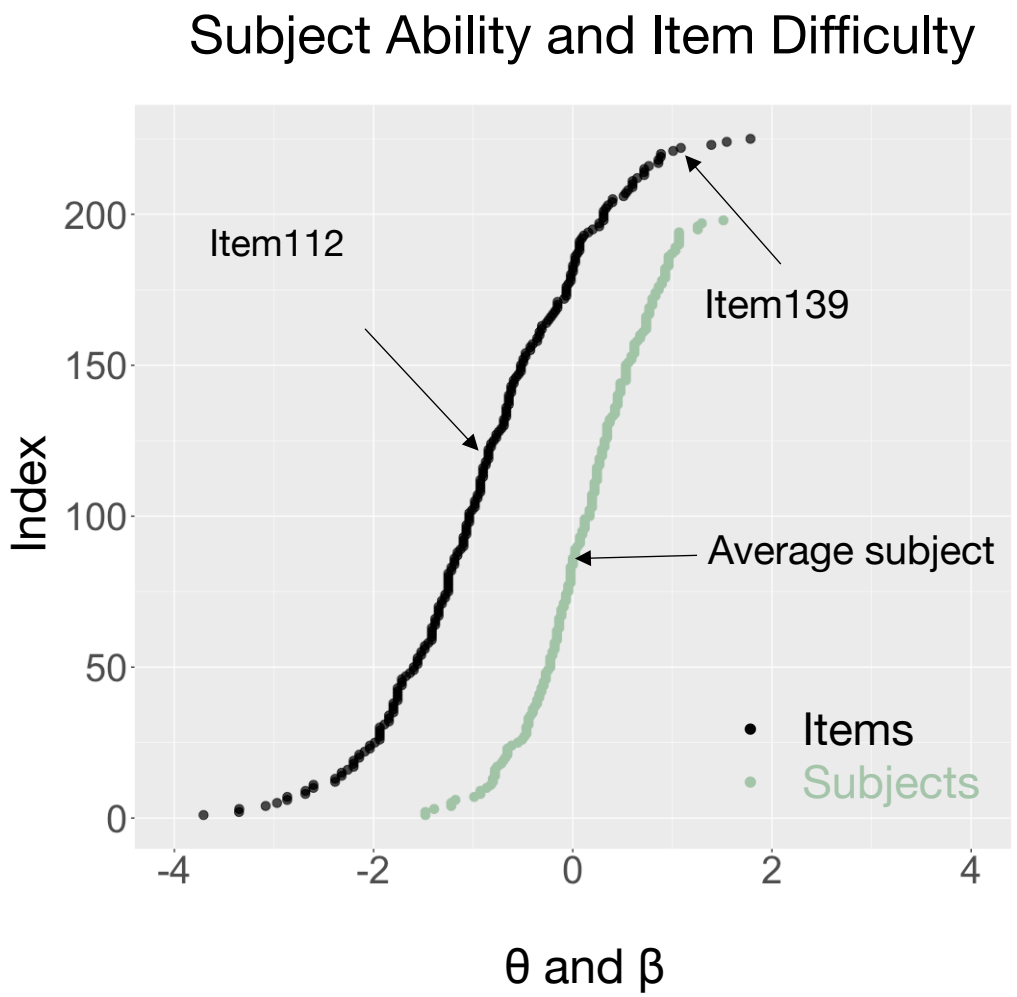
- 225 face-image triads

Three-alternative forced choice

- random order, image position
- 3.5 s exposure time,  unlimited RT
- accuracy free of decision bias

# Subjects and items on same scale

| Parameters | |
|---|---|
| Subject | ability $\theta_j$ |
| Item | difficulty $\beta_i$ |



# Rasch one-parameter logistic model

$$p(x_{ij} = 1 | \theta_i, \beta_j) = [e^{(\theta_i - \beta_j)}]/[1 + e^{(\theta_i - \beta_j)}]$$

answer on item j is correct

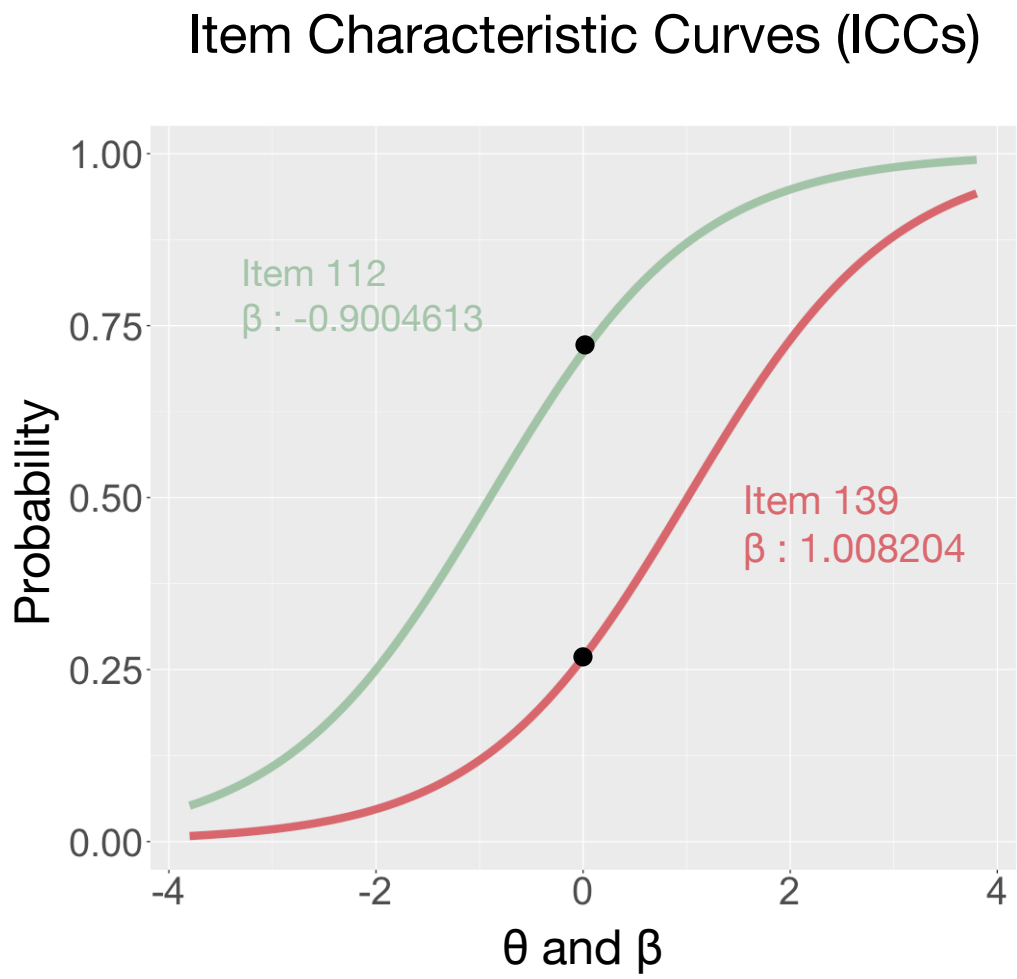
ability of subject *i*

difficulty of item *j*

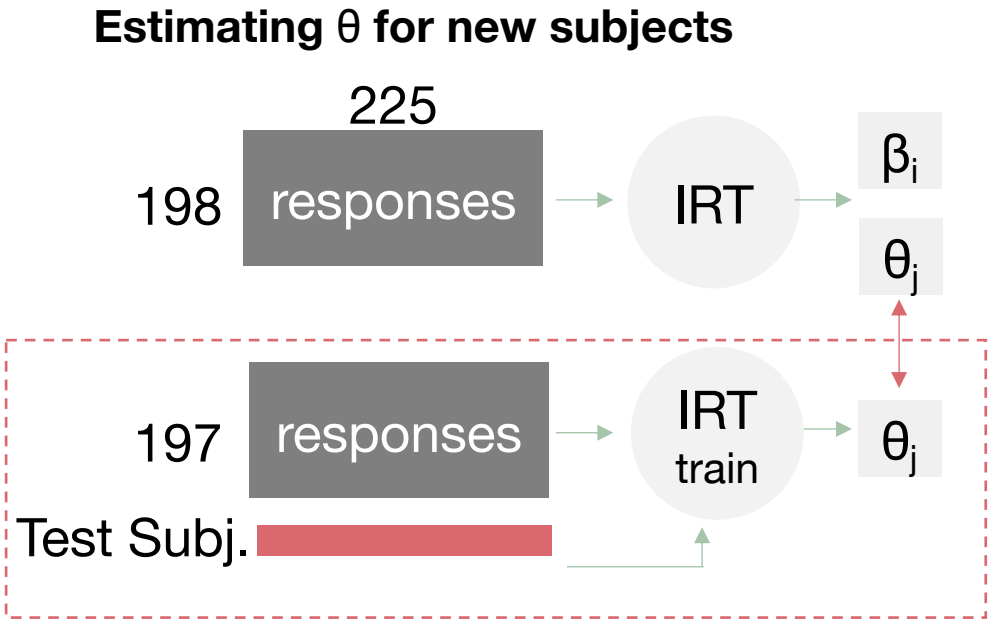
| Parameters | |
|---|---|
| Subject | ability $\theta_j$ |
| Item | difficulty $\beta_i$ |

# Item characteristic curves

- θ Subject ability
- β Item difficulty
- Average subject (θ=0) has a probability of ~.75 & ~.25 of responding to items 112 and 139 correctly



# Validating fit of model

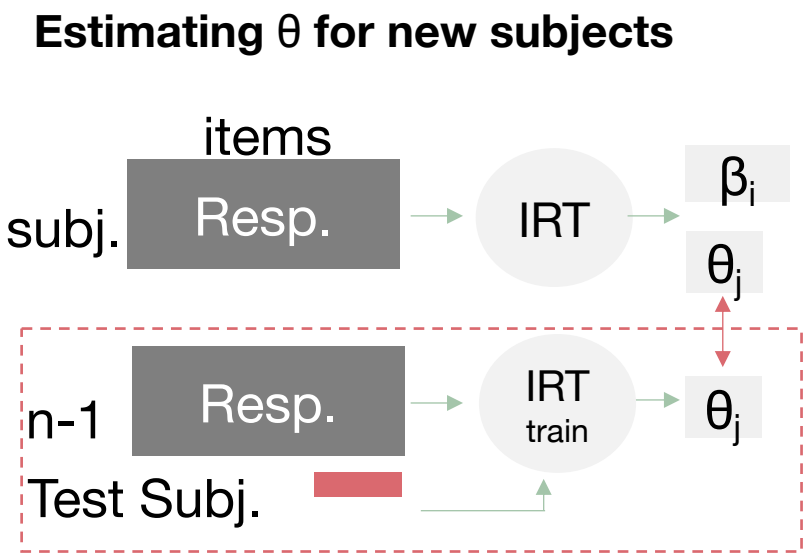


# Validating fit of model

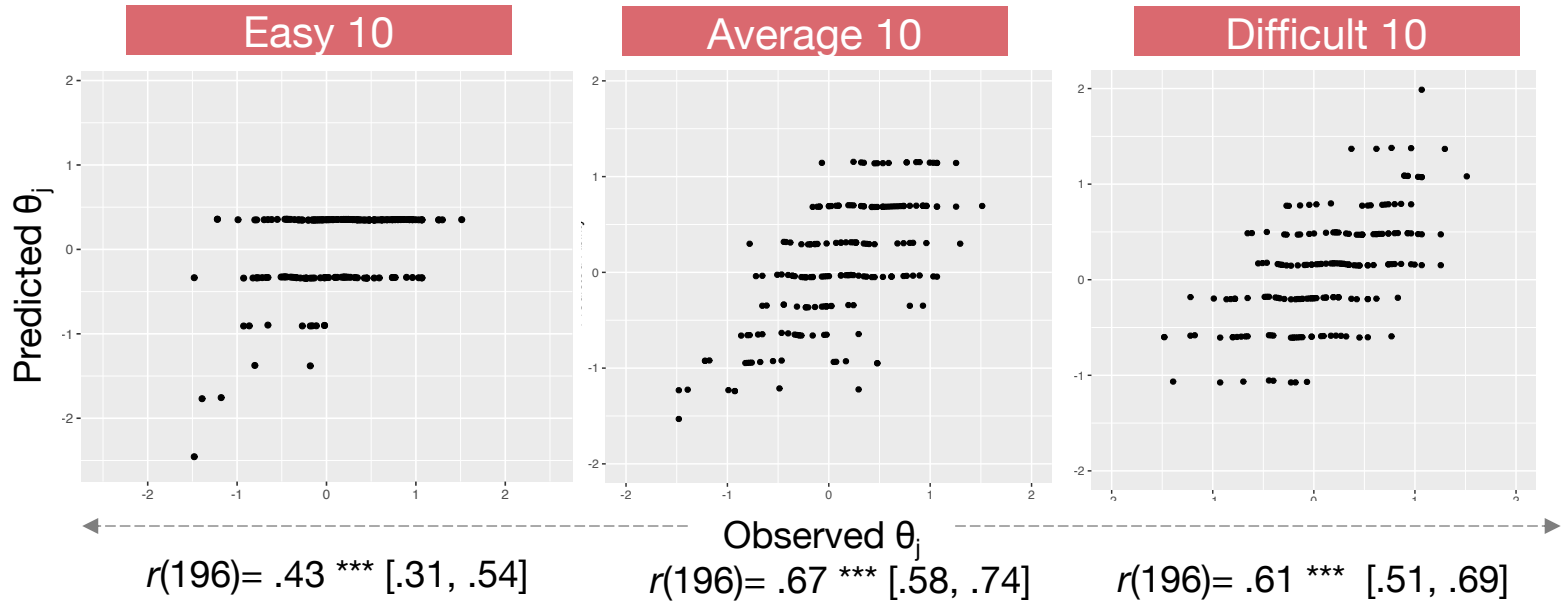
$r(196)=.99,\ p<.001$

- Rasch model estimated the ability of future subjects based on their responses to the full set of face-triad items

# Estimating subject ability $\theta_j$ from subsets



# Estimating subject ability $\theta_j$ from subsets



# Experiments—extended

# Two main goals

- Measure between day variance for subjects
  - Triad test

- How well does the triad test predict accuracy on comparison tests?
  - Glasgow Face Matching Test
  - Black-box Test

# Design of experiment

Participants

- 56 NIST staff

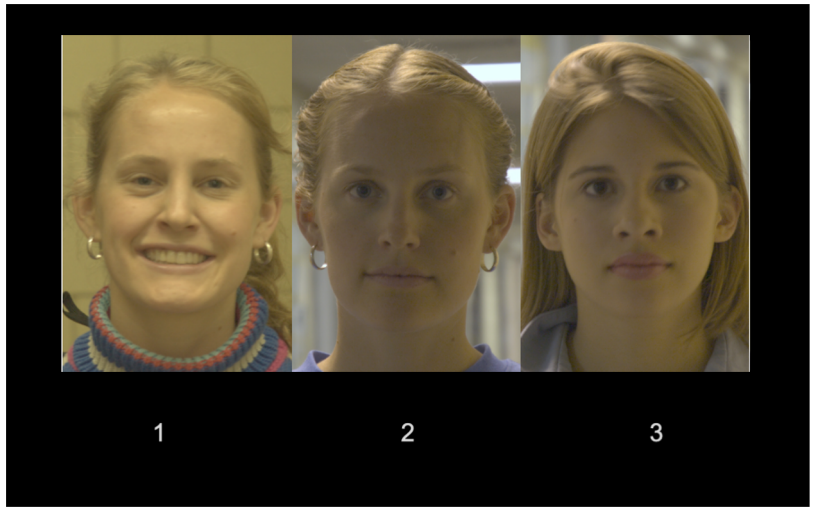
Triad subsets of 75

| Session 1 | Session 2 |
|---|---|
| Triad Subtest A | Triad Subtest B |
| GFMT | CFMT+ |
| Black-box | |

# Between Day Variance

Triad subset A


Triad subset B



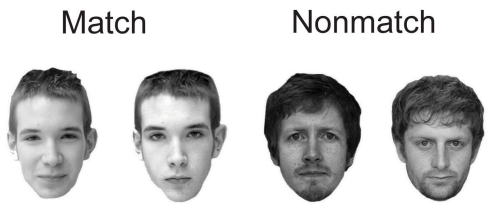

$r(56)=.66,\ p<.001$

# Accord between triad and signal detection

Triad subset A


$r(56)=.47, p<.001$

Glasgow Face Matching Test




$r(56)=.66, p<.001$

$r(56)=.48, p<.001$

PNAS Black-box Test

# Conclusions

- Item Response Theory (IRT) used to measure subject's ability as well as item difficulty
- IRT enables the construction of a reliable, flexible, and efficient face-identification test
- stablished a technique for creating an item bank (of triads) with known difficulty in order to create a set of tests of equal difficulty
- Account for day-to-day variation

# Thank you