Item response theory for designing calibrated face ability tests


* NIST, + University of Texas at Dallas, ++ University of Central Florida
• Need for Calibrated Face Tests
• Introducing item response theory (IRT)
• Triads
• IRT experiments
  • Baseline
  • Extended
Lesson from fusion

Fused > algorithm

Threshold AUC: 0.16

95% Students

0.80

0.96

Examiners

Reviewers

Super-Recognizers

Fingerprint

Students

Algorithms

Group

A2015

A2016

A2017a

A2017b

Random

Perfect

AUC
• 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
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
The observations strongly support that it is the same person

The observations support that it is the same person

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

The observations support neither that it is the same person nor that it is different persons

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

The observations support that it is not the same person

The observations strongly support that it is not the same person
Same/Different Paradigm

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

-1
Criterion shift

Amy

Jonathon

Criterion Shift

Verification rate

False accept rate
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
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

<table>
<thead>
<tr>
<th>Subject</th>
<th>ability $\theta_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>difficulty $\beta_i$</td>
</tr>
</tbody>
</table>

Subject Ability and Item Difficulty

- Item 112
- Item 139
- Average subject

$\theta$ and $\beta$
Rasch one-parameter logistic model

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

- Difficulty of item \( j \)
- Ability of subject \( i \)
- Answer on item \( j \) is correct

Parameters:
- **Subject**: ability \( \theta_j \)
- **Item**: difficulty \( \beta_i \)
Item characteristic curves

- $\theta$ Subject ability
- $\beta$ Item difficulty
- Average subject ($\theta=0$) has a probability of $\sim.75$ & $\sim.25$ of responding to items 112 and 139 correctly
Validating fit of model

Estimating $\theta$ for new subjects

- 225 responses
- 198 responses

Test Subj.

IRT

$\beta_i$

$\theta_j$

IRT train

$\theta_i$

$\theta_j$
Validating fit of model

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

$r(196) = .99, p < .001$
Estimating subject ability $\theta_j$ from subsets

Estimating $\theta$ for new subjects

- Items
- Resp. subj.
- IRT
- $\beta_i$
- $\theta_j$

- Resp. n-1
- IRT train
- $\theta_j$

Test Subj.
Estimating subject ability $\theta_j$ from subsets

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Predicted vs Observed Ability</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy 10</td>
<td>Predicted vs Observed Ability (cor = 0.67)</td>
<td>$r(196) = .67$ *** [0.58, 0.74]</td>
</tr>
<tr>
<td>Average 10</td>
<td>Predicted vs Observed Ability (cor = 0.61)</td>
<td>$r(196) = .61$ *** [0.51, 0.69]</td>
</tr>
<tr>
<td>Difficult 10</td>
<td>Predicted vs Observed Ability (cor = 0.43)</td>
<td>$r(196) = .43$ *** [0.31, 0.54]</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Session 1</th>
<th>Session 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triad Subtest A</td>
<td>Triad Subtest B</td>
</tr>
<tr>
<td>GFMT</td>
<td>CFMT+</td>
</tr>
<tr>
<td>Black-box</td>
<td></td>
</tr>
</tbody>
</table>
Between Day Variance

$r(56) = .66, p < .001$
Accord between triad and signal detection

Triad subset A

Glasgow Face Matching Test

PNAS Black-box Test

$r(56) = 0.47, p < 0.001$

$r(56) = 0.66, p < 0.001$

$r(56) = 0.48, p < 0.001$
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
- Established 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