Item Response Theory

Psych 818
DeShon
IRT

• Typically used for 0,1 data (yes, no; correct, incorrect)
  - Set of probabilistic models that...
  - Describes the relationship between a respondent’s level on a construct (a.k.a. latent trait; e.g., extraversion, cognitive ability, affective commitment)...
  - To his or her probability of a particular response to an individual item

• Samejima's Graded Response model
  - For polytomous data where options are ordered along a continuum (e.g., Likert scales)
Advantages of IRT?

• Provides more information than classical test theory (CTT)
  - Classical test statistics depend on the set of items and sample examined
  - IRT modelling not dependent on sample examined
• Can examine item bias/ measurement equivalence
• SEM’s vary at different levels of the trait (conditional standard errors of measurement)
• Used to estimate item parameters (e.g., difficulty and discrimination) and...
• Person true scores on the latent trait
Advantages of IRT?

• Invariance of IRT Parameters
  − Difficulty and Discrimination parameters for an item are invariant across populations
    • Within a linear transformation
  − That is no matter who you administer the test to, you should get the same item parameters
  − However, precision of estimates will differ
    • If there is little variance on an item in a sample with have unstable parameter estimates
IRT Assumptions

- Underlying trait or ability (continuous)
- Latent trait is normally distributed
- Items are unidimensional
- Local Independence
  - if remove common factor the items are uncorrelated
- Items can be allowed to vary with respect to:
  - Difficulty (one parameter model; Rasch model)
  - Discriminability (two parameter model)
  - Guessing (three parameter model)
Model Setup

- Consider a test with $p$ binary (correct/incorrect) responses.
- Each item is assumed to ‘reflect’ one underlying (latent) dimensions of ‘achievement’ or ‘ability’.
- Start with an assumed 1-dimensional test, say of sexual attitude mathematics with 10 items.
- How do we get a value (score) on the mathematics scale from a set of 40 (1/0) responses from each individual?

Set up a Model....
A simple model

• First some basic notation...

• $f_j$ is the latent (factor) score for individual $j$.

• $\pi_{ij}$ is the probability that individual $j$ responds correctly to item $i$.

• Then a simple item response model is: $\pi_{ij} = a_i + b_i f_j$

• Just like a simple regression but with an unobserved predictor
Classical item analysis

- Can be viewed as an item response model (IRM) –
  \[ \pi_{ij} = a_i + b_i f_j \]

- The maximum likelihood estimate of \( f_j \) (red used for a random variable) is given by the ‘raw test score’ - Mellenbergh (1996)

- \( a_i \) is the item difficulty and \( b_i \) is the item discrimination
- Instead of using a linear linkage between the latent variable and the observed total score, IRT uses a logit transformation
  \[ \log(\pi_{ij} / 1 - \pi_{ij}) = \text{logit}(\pi_{ij}) = a_i + b_i f_j \]
**Parameter Interpretation**

- **Difficulty**
  - Point on the theta continuum (x-axis) that corresponds to a 50% probability of endorsing the item
  - A more difficult item is located further to the right than an easier item
  - Values are interpreted almost the reverse of CTT
  - Difficulty is in a z-score metric

- **Discrimination**
  - the slope of the IRF
  - The steeper the slope, the greater the ability of the item to differentiate between people
  - Assessed at the difficulty of the item
Item Response Relations

For a single item in a test
The Rasch Model (1-parameter)

\[
\text{logit}(\pi_{ij}) = a_i + f_j
\]

- Notice the absence of a weight on the latent ability variable...item discrimination (roughly the correlation between the item response and the level of the factor) is assumed to be the same for each item and equal to 1.0
- The resulting (maximum likelihood) factor score estimates are then a 1 – 1 transformation of the raw scores.
2 Parameter IRT model

\[ \pi_{ij} = a_i + b_i f_j \]

• Allows both item difficulty and item discrimination to vary over items

• Lord (1980, p. 33) showed that the discrimination is a monotonic function of the point-biserial correlation
Rasch Model Mysticism

• There is great mysticism surrounding the Rasch model
  - Rasch acolytes emphasize the model over the data
  - To be a good measure, the Rasch model must fit

  - Allowing item discrimination to vary over items means that you don't have additive measurement
  - In other words, 1 foot + 1 foot doesn't equal 2 feet.

  - Rasch maintains that items can only differ in discrimination if the latent variable is not unidimensional
# Rasch Model – Useful Ruler Logic

<table>
<thead>
<tr>
<th>Word Recognition</th>
<th>Mastery Scale</th>
<th>Grade Scale 50% Mastery</th>
<th>Sample Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure</td>
<td></td>
<td></td>
<td>is</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>1.1</td>
<td>is</td>
</tr>
<tr>
<td></td>
<td>41</td>
<td>1.3</td>
<td>red</td>
</tr>
<tr>
<td></td>
<td>58</td>
<td>1.4</td>
<td>down</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>1.5</td>
<td>black</td>
</tr>
<tr>
<td></td>
<td>86</td>
<td>1.7</td>
<td>away</td>
</tr>
<tr>
<td></td>
<td>101</td>
<td>1.8</td>
<td>cold</td>
</tr>
<tr>
<td></td>
<td>114</td>
<td>2.0</td>
<td>drink</td>
</tr>
<tr>
<td></td>
<td>124</td>
<td>2.2</td>
<td>shallow</td>
</tr>
<tr>
<td></td>
<td>143</td>
<td>2.8</td>
<td>through</td>
</tr>
<tr>
<td></td>
<td>159</td>
<td>3.3</td>
<td>octopus</td>
</tr>
<tr>
<td></td>
<td>174</td>
<td>4.1</td>
<td>allowable</td>
</tr>
<tr>
<td></td>
<td>192</td>
<td>5.7</td>
<td>hinderance</td>
</tr>
<tr>
<td></td>
<td>211</td>
<td>9.3</td>
<td>equestrian</td>
</tr>
<tr>
<td></td>
<td>240</td>
<td>12.9</td>
<td>heterogeneous</td>
</tr>
</tbody>
</table>
Five Rasch Items

3rd Graders
2nd Graders
1st Graders

Red  Away  Drink  Octopus  Equestrian

Word Order stays the same!

Item Difficulty (relative to "Drink" item)

Logit Scale
Item Response Curves
With different item discriminations
Results in Chaos

Item Difficulty (relative to “Drink” item)
IRT Example: British Sexual Attitudes (n=1077, from Bartholomew et al. 2002)

11% Should male homosexuals be allowed to adopt?
13% Should divorce be easier?
13% Extra-marital sex? (not wrong)
19% Should female homosexuals be allowed to adopt?
29% Same sex relationships? (not wrong)
48% Should gays be allowed to teach in school?
55% Should gays be allowed to teach in higher education?
59% Should gays hold public office?
77% Pre-marital sex? (not wrong)
82% Support law against sex discrimination?
How Liberal are Brits with respect to sexual attitudes?
### Item correlations

<table>
<thead>
<tr>
<th></th>
<th>Div</th>
<th>Sd</th>
<th>pre</th>
<th>ex</th>
<th>gay</th>
<th>sch</th>
<th>hied</th>
<th>publ</th>
<th>fad</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEXDISC</td>
<td>-.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PREMAR</td>
<td>.08</td>
<td>.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXMAR</td>
<td>.09</td>
<td>.05</td>
<td>.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAYSEX</td>
<td>.02</td>
<td>.17</td>
<td>.28</td>
<td>.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAYSCHO</td>
<td>-.02</td>
<td>.16</td>
<td>.25</td>
<td>.11</td>
<td>.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAYHIED</td>
<td>-.01</td>
<td>.16</td>
<td>.28</td>
<td>.11</td>
<td>.41</td>
<td>.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAYPUBL</td>
<td>-.01</td>
<td>.09</td>
<td>.24</td>
<td>.13</td>
<td>.34</td>
<td>.66</td>
<td>.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAYFADOP</td>
<td>.06</td>
<td>.03</td>
<td>.18</td>
<td>.16</td>
<td>.34</td>
<td>.28</td>
<td>.26</td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td>GAYMADOP</td>
<td>.07</td>
<td>.06</td>
<td>.15</td>
<td>.17</td>
<td>.39</td>
<td>.27</td>
<td>.24</td>
<td>.22</td>
<td>.71</td>
</tr>
</tbody>
</table>

Div  Sd  pre  ex  gay  sch  hied  publ  fad
<table>
<thead>
<tr>
<th>Item</th>
<th>Scale Mean if Item Deleted</th>
<th>Scale Variance if Item Deleted</th>
<th>Corrected Item-Total Correlation</th>
<th>Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIVORCE</td>
<td>3.9136</td>
<td>5.2686</td>
<td>.0387</td>
<td>.7773</td>
</tr>
<tr>
<td>SEXDISC</td>
<td>3.2154</td>
<td>5.0223</td>
<td>.1588</td>
<td>.7679</td>
</tr>
<tr>
<td>PREMAR</td>
<td>3.2730</td>
<td>4.6224</td>
<td>.3500</td>
<td>.7453</td>
</tr>
<tr>
<td>EXMAR</td>
<td>3.9099</td>
<td>4.9947</td>
<td>.2195</td>
<td>.7588</td>
</tr>
<tr>
<td>GAYSEX</td>
<td>3.7512</td>
<td>4.2484</td>
<td>.5265</td>
<td>.7190</td>
</tr>
<tr>
<td>GAYSCHO</td>
<td>3.5645</td>
<td>3.8650</td>
<td>.6736</td>
<td>.6915</td>
</tr>
<tr>
<td>GAYHIED</td>
<td>3.4921</td>
<td>3.8691</td>
<td>.6743</td>
<td>.6915</td>
</tr>
<tr>
<td>GAYPUBL</td>
<td>3.4513</td>
<td>4.0453</td>
<td>.5809</td>
<td>.7088</td>
</tr>
<tr>
<td>GAYFADOP</td>
<td>3.8542</td>
<td>4.5652</td>
<td>.4340</td>
<td>.7338</td>
</tr>
<tr>
<td>GAYMADOP</td>
<td>3.9341</td>
<td>4.7419</td>
<td>.4486</td>
<td>.7349</td>
</tr>
</tbody>
</table>

N of Cases = 1077, N of Items = 10

Alpha = .7558
## Rasch Model (1 parameter) results

<table>
<thead>
<tr>
<th>Label</th>
<th>Item P</th>
<th>Diffic. (se)</th>
<th>Discrim. (se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIVORCE</td>
<td>0.126</td>
<td>1.573 (.057)</td>
<td>1.146 (.070)</td>
</tr>
<tr>
<td>SEXDISC</td>
<td>0.847</td>
<td>-1.409 (.051)</td>
<td>1.146 (.073)</td>
</tr>
<tr>
<td>PREMAR</td>
<td>0.787</td>
<td>-1.118 (.047)</td>
<td>1.146 (.072)</td>
</tr>
<tr>
<td>EXMAR</td>
<td>0.129</td>
<td>1.548 (.056)</td>
<td>1.146 (.070)</td>
</tr>
<tr>
<td>GAYSEX</td>
<td>0.293</td>
<td>0.753 (.044)</td>
<td>1.146 (.071)</td>
</tr>
<tr>
<td>GAYSCHO</td>
<td>0.486</td>
<td>0.040 (.042)</td>
<td>1.146 (.068)</td>
</tr>
<tr>
<td>GAYHIED</td>
<td>0.561</td>
<td>-0.229 (.042)</td>
<td>1.146 (.067)</td>
</tr>
<tr>
<td>GAYPUBL</td>
<td>0.603</td>
<td>-0.383 (.043)</td>
<td>1.146 (.068)</td>
</tr>
<tr>
<td>GAYFADOP</td>
<td>0.187</td>
<td>1.224 (.050)</td>
<td>1.146 (.070)</td>
</tr>
<tr>
<td>GAYMADOP</td>
<td>0.105</td>
<td>1.719 (.061)</td>
<td>1.146 (.071)</td>
</tr>
<tr>
<td>Mean</td>
<td>0.412</td>
<td>0.372 (.049)</td>
<td>1.146 (.070)</td>
</tr>
</tbody>
</table>
Latent Trait Model Item Plots

DIVORCE  SEXDISC  PREMAR  EXMAR

GAYSEX  GAYSCHO  GAYHIED  GAYPUBL

GAYFADOP  GAYMADOP
All together now...

Sex discrim.

Males adopting
What do you get from one parameter IRT?

• Items vary in difficulty (which you get from univariate statistics)?
• A measure of the fit.
• A nice graph, but no more information than table.

• Person scores on the latent trait
## Two Parameter IRT Model

<table>
<thead>
<tr>
<th>Label</th>
<th>Item P</th>
<th>Diffic.</th>
<th>(se)</th>
<th>Discrim</th>
<th>(se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIVORCE</td>
<td>0.126</td>
<td>2.417</td>
<td>(.110)</td>
<td>0.499</td>
<td>(.026)</td>
</tr>
<tr>
<td>SEXDISC</td>
<td>0.847</td>
<td>-3.085</td>
<td>(.154)</td>
<td>0.323</td>
<td>(.017)</td>
</tr>
<tr>
<td>PREMAR</td>
<td>0.787</td>
<td>-1.107</td>
<td>(.058)</td>
<td>0.935</td>
<td>(.050)</td>
</tr>
<tr>
<td>EXMAR</td>
<td>0.129</td>
<td>2.449</td>
<td>(.110)</td>
<td>0.503</td>
<td>(.026)</td>
</tr>
<tr>
<td>GAYSEX</td>
<td>0.293</td>
<td>0.552</td>
<td>(.024)</td>
<td>2.288</td>
<td>(.159)</td>
</tr>
<tr>
<td>GAYSCHO</td>
<td>0.486</td>
<td>0.138</td>
<td>(.020)</td>
<td>3.000</td>
<td>(.195)</td>
</tr>
<tr>
<td>GAYHIED</td>
<td>0.561</td>
<td>-0.028</td>
<td>(.021)</td>
<td>3.000</td>
<td>(.189)</td>
</tr>
<tr>
<td>GAYPUBL</td>
<td>0.603</td>
<td>-0.127</td>
<td>(.021)</td>
<td>2.995</td>
<td>(.189)</td>
</tr>
<tr>
<td>GAYFADOP</td>
<td>0.187</td>
<td>0.858</td>
<td>(.029)</td>
<td>2.239</td>
<td>(.146)</td>
</tr>
<tr>
<td>GAYMADOP</td>
<td>0.105</td>
<td>1.102</td>
<td>(.033)</td>
<td>2.995</td>
<td>(.220)</td>
</tr>
</tbody>
</table>

Mean 0.412 0.317 (.058) 1.878 (.122)
Threshold effects and the form of the latent variable.
2-p IRT

Latent Trait

P (liberal response)

-3 -1 1 3

sex discrim.
pre-marital
ex-marital divorce

r = .09
What do we get from the 2 parameter model?

- The graphs clearly show not all items equally discriminating. Perhaps get rid of a few items (or a two trait model).
- Getting some items right, more important than others, for total score.
**Item Information Function (IIF)**

\[ I(f) = b_i \cdot b_i \cdot p_i(f)q_i(f) \]

- \( p \) is the probability of correct response for a given true score and \( q \) is the probability of incorrect response
- Looks like a hill
- The higher the hill the more information
- The peak of the hill is located at the item difficulty
- The steepness of the hill is a function of the item discrimination
  - More discriminating items provide more information
IIF’s for Sexual Behavior Items

- date
- break up
- intercourse
- sex before 15
- afraid pregnant
- pregnant
**Item Standard Error of Measurement (SEM) Function**

- Estimate of measurement precision at a given theta value

- $\text{SEM} = \text{inverse of the square root of the item information}$

- SEM is smallest at the item difficulty

- Items with greater discrimination have smaller SEM, greater measurement precision
Test Information Function (TIF)

- Sum of all the item information functions
- Index of how much information a test is providing a given trait level
- The more items at a given trait level the more information
Test Standard Error of Measurement (SEM) function

- Inverse of the square root of the test information function
- Index of how well, i.e., precise a test measure’s the trait at a given difficulty level
Figure 1. Effect of Item Discrimination on Information
IRT Item and Test Functions

\( a_1 = 2.0, b_1 = -2.0, c_1 = 0.0; a_2 = 1.5, b_2 = -0.8, c_2 = 0.10; a_3 = 1.5, b_3 = -0.25, c_3 = 0.20; a_4 = 0.8, b_4 = 2.0, c_4 = 0.25 \)

a. Item Response Functions (IRFs)

b. Item Information Functions (IIFs)

c. Item Conditional SEM Functions

d. Test Response Function (TRF)

e. Test Information Function (TIF)

f. Test Conditional SEM Function
Check out this for more info

- [www2.uni-jena.de/svw/metheval/irt/VisualIRT.pdf](http://www2.uni-jena.de/svw/metheval/irt/VisualIRT.pdf)
  - Great site with graphical methods for demonstrating the concepts
Testing Model Assumptions

- Unidimensionality
- Model fit – More on this later when we get to CFA
Unidimensionality

- Perform a factor analysis on the tetrachoric correlations
- Or... Use HOMALS in SPSS for binary PCA
- Or CFA for dichotomous indicators

![Scree Plot](image)
IRT vs. CTT


  There is no obvious superiority between classical sum-score item and test parameters and the new item response theory item and test parameters, when a large, representative sample of individuals are used to calibrate items.
Computer Adaptive Testing (CAT)

- In IRT, a person’s estimate of true score is not dependent upon the number of items correct.
- Therefore, can use different items to measure different people and tailor a test to the individual.
- Provides greater:
  - Efficiency (fewer items)
  - Control of precision - given adequate items, every person can be measured with the same degree of precision.
- Example: GRE
Components of a CAT system

• A pre-calibrated bank of test items
  – Need to administer a large group of items to a large sample and estimate item parameters

• An entry point into the item bank
  – i.e., a rule for selecting the first item to be administered
  – Item difficulty, e.g., $b = 0, -3, \text{ or } +3$
  – Use prior information about examinee
Components of a CAT system

• An item selection or branching rule(s)
  – E.g., if correct to first item, go to a more difficult item
  – If incorrect go to a less difficult item
  – Always select the most informative item at the current estimate of the trait level
  – As responses accumulate more information is gained about the examinee’s trait level
Item Selection Rule -

Select the item with the most information at the current θ estimate of the latent trait.
Components of a CAT system

- A termination rule
  - Fixed number of items

- Equiprecision
  - End when SEM around the examinee’s trait score has reached a certain level of precision. The precision of test varies across individuals.
  - Examinee’s whose responses are consistent with the model will be easier to measure, i.e., require fewer items.

- Equiclassification
  - SEM around trait estimate is above or below a cutoff level.
Figure 1. Examinee #1 — Progress Through Exam