Workshop
Polytomous IRT models
(№ 144, Remo Ostini and Michael L. Nering)

Jorge Tendeiro

16 April 2014
Presentation based on the book:


I also used a classic book:

Overview

1 Introduction

2 (Some) Polytomous IRT models
   Nominal response model (NRM)
   Partial credit model (PCM)
   Generalized partial credit model (GPCM)
   Rating scale model (RSM)
   Graded response model (GRM)

3 Model selection

4 Software
Introduction
Item response theory (IRT): Main idea

Modeling the relationship item ↔ person by means of a mathematical function:

\[
P(X_i = c|\theta) = f(\theta)
\]

\( P_{ic}(\theta) \)

✓ \( X_i = \) Item \( i \) with discrete response categories.
✓ \( c = \) Coded response category:
  • If \( X \) is dichotomous, \( c = 0, 1 \);
  • If \( X \) is polytomous, \( c = 0, 1, \ldots, m \) (\( m > 1 \)).
✓ \( \theta = \) Person trait parameter.

This is the item response function (IRF).
IRT: Important property

Item location (to be defined shortly) and person trait are indexed on the same metric.

Example: Dichotomous item

- \( \theta > b \) \( \rightarrow \) person is more likely to answer \( X_i = 1 \).
- \( \theta < b \) \( \rightarrow \) person is more likely to answer \( X_i = 0 \).
IRT: Dichotomous models recap.

- Dichotomous items: 
  \( X_i = 0 \) (incorrect, false) or \( X_i = 1 \) (correct, true).
- Most common models (logistic): 1PLM, 2PLM, 3PLM
- These models typically relate \( \theta \) and \( P_{i1}(\theta) \):

\[
P_{i1}(\theta) = f(\theta).
\]

\[
[P_{i0}(\theta) \equiv 1 - P_{i1}(\theta)].
\]

We usually simplify notation in the dichotomous case:

\[
P_i(\theta) = P_{i1}(\theta).
\]
IRT: Dichotomous models recap.

1PLM

\[ P_i(\theta) = \frac{1}{1 + \exp[-(\theta - b_i)]} \]

- \( b_i \) = difficulty param.
2PLM

\[ P_i(\theta) = \frac{1}{1 + \exp[-a_i(\theta - b_i)]} \]

- \( b_i \) = difficulty param., \( a_i \) = discrimination param.
IRT: Dichotomous models recap.

3PLM

\[ P_i(\theta) = c_i + (1 - c_i) \frac{1}{1 + \exp[-a_i(\theta - b_i)]} \]

- \( b_i \) = difficulty param., \( a_i \) = discrimination param., \( c_i \) = guessing param.
IRT: Polytomous models

In this case $X_i = 0, 1, \ldots, m$, where $m > 1$.

Example of items with multiple response items:

- Rating scale
  (e.g., Likert-type items: ‘Strongly disagree’, ..., ‘Strongly agree’).
- Ability test items awarding partial credit.

Now we need to define models which allow estimating each $P_{ic}(\theta)$, $c = 0, 1, \ldots, m$:

$$
\begin{align*}
P_{i0}(\theta) &= f_1(\theta) \\
\vdots \\
P_{im}(\theta) &= f_m(\theta)
\end{align*}
$$

These are the item category response functions (ICRFs).
IRT: Polytomous models – Why?

Polytomous items... 

• are extensively used in applied psychological measurement.
• measure across a wider range of the trait continuum $\theta$.
• are related to an increase of statistical information when compared to dichotomous items.
• (in some settings) may help reducing test length (time $\downarrow$, costs $\downarrow$, respondents’ motivation $\uparrow$).
Nominal response model (NRM)
NRM (Bock, 1972)

- Type of items: Polytomous with two or more nominal categories.
- Here, nominal categories = unordered in terms of the trait being measured.
- E.g.: Multiple choice items (namely the distractors).

The NRM is a “divide-by-total”, or “direct” model: The ICRFs are modeled directly.
NRM (Bock, 1972)

The ICRF for category \( c \) \( (c = 0, 1, \ldots, m) \) is

\[
P_{ic}(\theta) = \frac{\exp(\lambda_{ic}\theta + \zeta_{ic})}{\sum_{h=0}^{m} \exp(\lambda_{ih}\theta + \zeta_{ih})}.
\]

- \( \lambda_{ih} = \) slope associated to category \( h \) of item \( i \).
- \( \zeta_{ih} = \) intercept associated to category \( h \) of item \( i \).

To identify the model (i.e., to estimate parameters), one of two constraints is typically imposed:

- \( \sum_{h=0}^{m} \lambda_{ih} = \sum_{h=0}^{m} \zeta_{ih} = 0 \), or
- \( \lambda_{i0} = \zeta_{i0} = 0 \).
NRM (Bock, 1972): Example

Item measuring student mathematical achievement ($N \approx 2,000$).

<table>
<thead>
<tr>
<th>Response options</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>$\sum$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_i$</td>
<td>-.30</td>
<td>.81</td>
<td>-.31</td>
<td>-.20</td>
<td>.000</td>
</tr>
<tr>
<td>$\zeta_i$</td>
<td>.21</td>
<td>.82</td>
<td>-.09</td>
<td>-.94</td>
<td>.000</td>
</tr>
</tbody>
</table>

![Graph showing response probabilities for different student abilities](Image)
NRM (Bock, 1972): Example

Interpretation:

- Response B is the most popular for the more able respondents.
- Response A is the most popular for the less able respondents (followed by Response C).
- Response D was not popular across the entire trait scale.

In general, for the NRM:

- The popularity of response categories across the entire trait scale is associated to the order of the intercepts $\zeta_{ic}$.

For the example, in increasing order of popularity:

Response D $<$ Response C $<$ Response A $<$ Response B.
Partial credit model (PCM)
PCM (Masters, 1982)

- Type of items: Polytomous with two or more ordinal categories.
- Ideal when the answer to an item consists of an ordered sequence of steps.
- Partial credit can be given if the respondents only answered correctly to the first (but not all) steps.
- Varying number of categories across items is possible.
- PCM = Applying the 1PLM to each pair of adjacent item response categories.
- The PCM is an extension of the 1PLM.

The PCM is a “divide-by-total”, or “direct” model: The ICRFs are modeled directly.
PCM (Masters, 1982)

The ICRF for category $c$ ($c = 0, 1, \ldots, m$) is

$$P_{ic}(\theta) = \frac{\exp \left[ \sum_{j=0}^{c}(\theta - \delta_{ij}) \right]}{\sum_{h=0}^{m} \exp \left[ \sum_{j=0}^{h}(\theta - \delta_{ij}) \right]}.$$  

- $\delta_{ij}$ ($j = 1, \ldots, m$): Item step difficulties, also known as
  - category boundaries;
  - category intersections.
- Notation: $\sum_{j=0}^{0}(\theta - \delta_{ij}) = 0.$
PCM (Masters, 1982)

- $\delta_{ij} = \theta$-value at which two consecutive ICRFs intersect:

$$P_{i(j-1)}(\delta_{ij}) = P_{ij}(\delta_{ij}).$$

- The higher the $\delta_{ij}$, the more difficult a particular step is.
- The $\delta_{ij}$'s aren’t necessarily ordered in the same sequence as the categories (reversals; such a case indicates that the item is probably not functioning as intended).

Special restriction of the PCM:
There must exist responses in every response category.
(Problematic for sparse data.)
PCM (Masters, 1982): Example

Item from a survey of morality \((N \approx 1,000)\).
Five-point Likert-type rating scale.

<table>
<thead>
<tr>
<th>Step Difficulties</th>
<th>(\delta_{i1})</th>
<th>(\delta_{i2})</th>
<th>(\delta_{i3})</th>
<th>(\delta_{i4})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-1.618)</td>
<td>(-0.291)</td>
<td>(0.414)</td>
<td>(2.044)</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\Pr_i(X = c) &= \Phi(\theta - \delta_{i1}) - \Phi(\theta - \delta_{i2}) \\
&= \Phi(\theta - \delta_{i3}) - \Phi(\theta - \delta_{i4})
\end{align*}
\]
PCM (Masters, 1982): Example

Interpretation:

- In this case the $\delta_{ij}$’s are ordered, so adjacent ICRFs intersect at locally optimal trait values.
- In particular, each answer option has the highest probability in some subinterval of the $\theta$-scale.
PCM (Masters, 1982): Example

Interpretation:

- In this case the $\delta_{ij}$’s are ordered, so adjacent ICRFs intersect at locally optimal trait values.
- In particular, each answer option has the highest probability in some subinterval of the $\theta$-scale.
Generalized partial credit model (GPCM)
The GPCM is a generalization of the PCM.

Idea: Add discrimination parameter (one per item).

So, in a way, PCM $\rightarrow$ GPCM just like 1PLM $\rightarrow$ 2PLM.

The GPCM is a “divide-by-total”, or “direct” model: The ICRFs are modeled directly.
GPCM (Muraki, 1992)

The ICRF for category $c$ ($c = 0, 1, \ldots, m$) is

$$P_{ic}(\theta) = \frac{\exp \left[ \sum_{j=0}^{c} \alpha_i(\theta - \delta_{ij}) \right]}{\sum_{h=0}^{m} \exp \left[ \sum_{j=0}^{h} \alpha_i(\theta - \delta_{ij}) \right]}.$$ 

- $\delta_{ij}$ ($j = 1, \ldots, m$): Item step difficulties (category intersections).
- $\alpha_i$: Item discrimination (slope parameters).
- Notation: $\sum_{j=0}^{0} \alpha_i(\theta - \delta_{ij}) = 0.$
GPCM (Muraki, 1992)

- $\delta_{ij} = \theta$-value at which two consecutive ICRFs intersect.
- $\alpha_i$ — Intuitive interpretation:
  - Small values (say, $\leq 1$) → ‘flatter’ ICRFs.
  - Large values (say, $\geq 1.5$) → more ‘peaked’ ICRFs.

In Muraki’s (1992, p. 162) words:

“[The $\alpha_i$’s] indicate the degree to which categorical responses vary among items as $\theta$ level changes.”
**GPCM (Muraki, 1992): Example**

- Items from the Neuroticism Extraversion Openness Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1992).
- Five-point Likert-type rating scale.  
  \((0 = \text{strongly disagree}; \ldots; 4 = \text{strongly agree}).\)
- \(N = 350.\)

Let’s see three items.

<table>
<thead>
<tr>
<th>Item</th>
<th>Content</th>
<th>Response category</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Feels tense and jittery</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>111</td>
</tr>
<tr>
<td></td>
<td></td>
<td>97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>101</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>6</td>
<td>Sometimes feels worthless</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>43</td>
</tr>
<tr>
<td>9</td>
<td>Feels discouraged, like giving up</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>128</td>
</tr>
<tr>
<td></td>
<td></td>
<td>66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>34</td>
</tr>
</tbody>
</table>
GPCM (Muraki, 1992): Example (slope $\simeq 1$)

Item 6 ‘Sometimes feels worthless’.
(0 = 72, 1 = 89, 2 = 52, 3 = 94, 4 = 43).

<table>
<thead>
<tr>
<th>Slope $\alpha_6$</th>
<th>Step Difficulties $\delta_{61}$</th>
<th>$\delta_{62}$</th>
<th>$\delta_{63}$</th>
<th>$\delta_{64}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.073</td>
<td>$-0.873$</td>
<td>0.358</td>
<td>$-0.226$</td>
<td>1.547</td>
</tr>
</tbody>
</table>

![Graph showing item response probabilities]
GPCM (Muraki, 1992): Example (slope < 1)

Item 5 ‘Feels tense and jittery’.
(0 = 17, 1 = 111, 2 = 97, 3 = 101, 4 = 24).

<table>
<thead>
<tr>
<th>Slope</th>
<th>Step Difficulties</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_5$</td>
<td>$\delta_{51}$</td>
</tr>
<tr>
<td>0.683</td>
<td>-3.513</td>
</tr>
</tbody>
</table>
GPCM (Muraki, 1992): Example (slope \(\approx 1.5\))

Item 9 ‘Feels discouraged, like giving up’.

(0 = 27, 1 = 128, 2 = 66, 3 = 95, 4 = 34).

<table>
<thead>
<tr>
<th>Slope</th>
<th>(\alpha_9)</th>
<th>Step Difficulties</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\delta_{91})</td>
<td>(\delta_{92})</td>
</tr>
<tr>
<td>1.499</td>
<td>-1.997</td>
<td>0.210</td>
</tr>
</tbody>
</table>
Rating scale model (RSM)
RSM (Andrich, 1978)

- Type of items: Polytomous with two or more ordinal categories.
- Requirement: All items of the measurement instrument have the same consistent structural response form. E.g.: When the set of responses is the same for all items.
- As a consequence, the response format is intended to function in the same way across all items.
- The RSM is an extension of the 1PLM. Moreover, the RSM can be seen as a special case of the PCM.

The RSM is a “divide-by-total”, or “direct” model: The ICRFs are modeled directly.
RSM (Andrich, 1978)

The ICRF for category \( c \) \((c = 0, 1, \ldots, m)\) is

\[
P_{ic}(\theta) = \frac{\exp \left \{ \sum_{j=0}^{c} [\theta - (\lambda_i + \delta_j)] \right \}}{\sum_{h=0}^{m} \exp \left \{ \sum_{j=0}^{h} [\theta - (\lambda_i + \delta_j)] \right \}}.
\]

- \( \lambda_i \): Item location parameter.
- \( \delta_j \) \((j = 1, \ldots, m)\): Category threshold parameters.
- Notation: \( \sum_{j=0}^{0} [\theta - (\lambda_i + \delta_j)] = 0. \)
RSM (Andrich, 1978)

- Two consecutive categories intersect at $\theta = (\lambda_i + \delta_j)$:
  \[
  P_{i(j-1)}(\lambda_i + \delta_j) = P_{ij}(\lambda_i + \delta_j).
  \]

- RSM is a special case of the PCM: Corresponding (across items) category intersections are equally spaced.
RSM (Andrich, 1978): Example (NEO-FFI)

Thresholds: \( \delta_1 = -1.600, \delta_2 = 0.224, \delta_3 = -0.184, \delta_4 = 1.560. \)
Graded response model (GRM)
GRM (Samejima, 1969)

- Type of items: Polytomous with two or more ordinal categories.
- Varying number of categories across items is possible.
- GRM = Applying the 2PLM at each category boundary (i.e., between two consecutive category responses).
- The GRM is an extension of the 2PLM.

The GRM is a “difference”, or “indirect” model: The ICRFs are modeled indirectly.
The ICRF for category $c$ ($c = 0, 1, \ldots, m$) is

$$P_{ic}(\theta) = P_{ic}^*(\theta) - P_{i(c+1)}^*(\theta),$$

where

$$P_{ic}^* = \frac{1}{1 + \exp[-\alpha_i(\theta - \beta_{ic})]}$$

(the 2PLM).

(And $P_{i0}^* \equiv 1$, $P_{im}^* \equiv 0$.)

For example, if $m = 4$ (i.e., $c = 0, 1, 2, 3$):

$$\begin{cases}
    P_{i0}(\theta) = 1 - P_{i1}^* \\
    P_{i1}(\theta) = P_{i1}^* - P_{i2}^* \\
    P_{i2}(\theta) = P_{i2}^* - P_{i3}^* \\
    P_{i3}(\theta) = P_{i3}^* - 0.
\end{cases}$$
GRM (Samejima, 1969)

- $\alpha_i$: Item slope parameter (one per item).
- $\beta_{ic}$: Category threshold parameters (one set $\{\beta_{i1}, \ldots, \beta_{im}\}$ per item). These are the $\theta$-values of transition between response categories.
- The $\beta_{ic}$'s are necessarily ordered.

\[
\begin{align*}
X_i &= c - 1 \\
\beta_{ic} \\
X_i &= c \\
\end{align*}
\]

\[
P(X_i \leq c) = .50 \quad P(X_i \geq c) = .50
\]
GRM (Samejima, 1969): Example (NEO-FFI)

Item 4 ‘Rarely feels lonely, blue’.

\(0 = 20, 1 = 90, 2 = 68, 3 = 125, 4 = 47\).

<table>
<thead>
<tr>
<th>Slope (\alpha_4)</th>
<th>Category thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta_{41})</td>
<td>(\beta_{42})</td>
</tr>
<tr>
<td>(\beta_{43})</td>
<td>(\beta_{44})</td>
</tr>
<tr>
<td>1.31</td>
<td>(-2.72)</td>
</tr>
<tr>
<td></td>
<td>(-0.81)</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>1.85</td>
</tr>
</tbody>
</table>

\(P_4(X = c)\) as a function of \(\theta\) (person ability).

- Red: Category 0
- Green: Category 1
- Blue: Category 2
- Brown: Category 3
- Purple: Category 4
Model selection
Model selection

- There are plenty of polytomous IRT models available (models + variants $> 10$).
- Choosing one model may be a hard enterprise.

Criteria to help choosing the ‘best’ model:

1. Data characteristics
2. Measurement philosophy
3. Mathematical approaches to check fit
Model selection

1 Data characteristics

- Dichotomous vs polytomous item scores.
- Nominal vs ordinal categories.
- Number of response categories.
  **E.g.:** The RSM requires the same number across items.

2 Measurement philosophy

- Does the model reflect the psychological reality that produced the data?
  **E.g.:** Can one conceptualize the answer to an item as being an ordered sequence of subtasks for which awarding partial credit to each is meaningful (i.e., PCM)?
Model selection

3 Mathematical approaches to check fit

- Check plots
  - Compare model-predicted vs empirical response functions.
  - Plot residuals.
Model selection

Mathematical approaches to check fit

- Statistical fit tests
  These may vary depending on their level of generality.
  (Assessing fit of all items, of a specific group of items, or of individual items.)

  - Residual-based measures.
    Based on differences between observed and expected item scores.

  - Multinomial distribution-based tests.
    Based on differences between observed and expected frequencies of response patterns.

  - Response function-based tests.
    Based on differences between observed and expected log-likelihood of response patterns.

  - Guttman error-based tests
    Nonparametric approach based on the number of Guttman errors.
Model selection

Mathematical approaches to check fit

- Goodness of fit
  Consider model fit \(\oplus\) number of estimated parameters.
  - Akaike’s information criterion (AIC; Akaike, 1977).
  - Procedures based on likelihood ratio of two comparing models.
Model selection

Some problems of statistical fit tests:

- The sampling distributions are often unknown.
- Some tests require very large sample sizes (on the hundreds), specially for $\chi^2$-based tests.
- Unknown influence of using estimated parameters or of mild model violations on the performance of the tests.
- Too large sample sizes invariably lead to rejections of the null hypothesis (effect size?).

A final reassurance:
Some comparative studies of polytomous IRT models suggest that results don’t vary much between models.
(E.g., Dodd, 1984; Maydeu-Olivares et al., 1994; Ostini, 2001; van Engelenburg, 1997; Verhelst et al., 1997.)
Software

- IRTPRO
- R: Several packages worth checking
  
  (see http://cran.r-project.org/web/views/Psychometrics.html)
  ltm, eRm, TAM, mcIRT, pcIRT,...