Workshop

Polytomous IRT models
(# 144, Remo Ostini and Michael L. Nering)

Jorge Tendeiro

16 April 2014

university of groningen

Literature

Presentation based on the book:

Ostini, R., & Nering, M. L. (2006). Polytomous item response theory models. Sage University Paper Series QASS. ("Little green book" # 144)

I also used a classic book:

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Chapter 5.

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:

$$\underbrace{P(X_i = c|\theta)}_{P_{ic}(\theta)} = f(\theta)$$

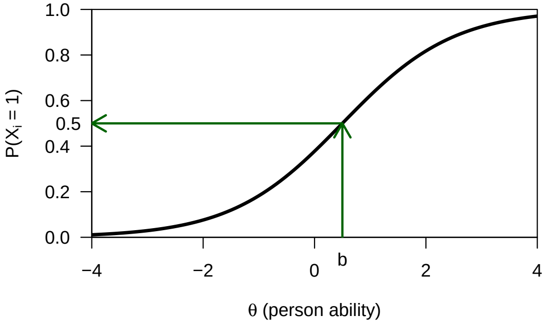
- ✓ X_i = Item i with discrete response categories.
- ✓ c = Coded response category:
  - If X is dichotomous, $c = 0, 1$;
  - If X is polytomous, $c = 0, 1, \ldots, m$ $(m > 1)$.
- ✓ θ = Person trait parameter.

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



- $\theta > b \longrightarrow$ person is more likely to answer $X_i = 1$.
- $\theta < b \longrightarrow$ 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 θ 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:

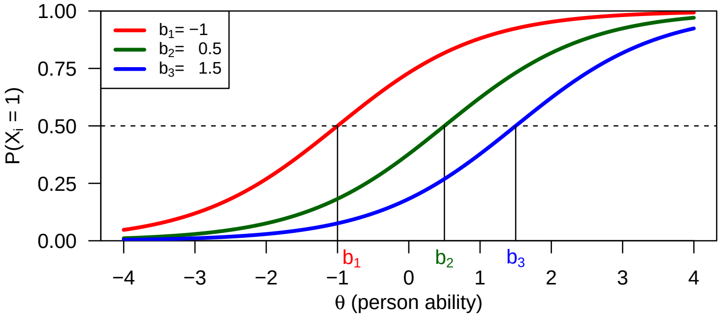
$$P_i(\theta) = P_{i1}(\theta).$$

IRT: Dichotomous models recap.

1PLM

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

- b_i = difficulty param.

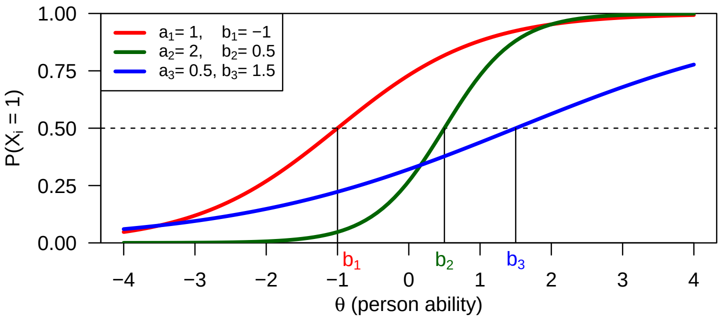


IRT: Dichotomous models recap.

2PLM

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

- b_i = difficulty param., a_i = discrimination param.

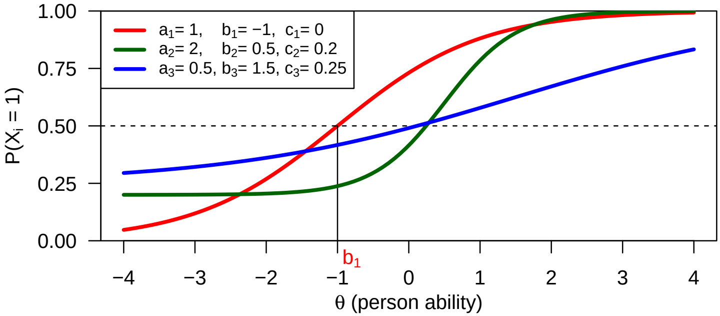


IRT: Dichotomous models recap.

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 $P_{ic}(\theta)$, $c = 0, 1, \ldots, m$:

$$\begin{cases} P_{i0}(\theta) = f_1(\theta) \\ \cdots \\ P_{im}(\theta) = f_m(\theta) \end{cases}.$$

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 θ.
- are related to an increase of statistical information when compared to dichotomous items.
- (in some settings) may help reducing test length (time$\searrow$, costs$\searrow$, respondents' motivation$\nearrow$).

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

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

- λ_{ih} = slope associated to category h of item i.
- ζ_{ih} = intercept associated to category h of item i.

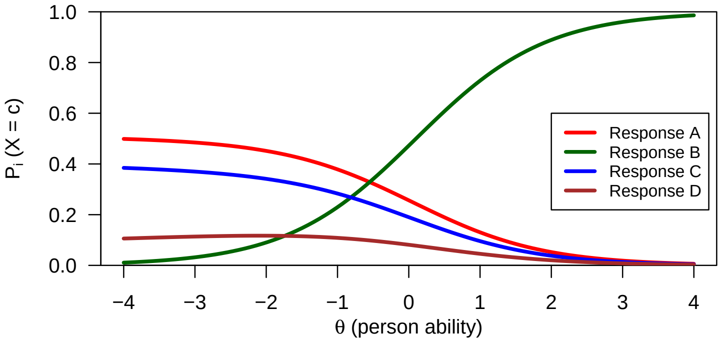
To identify the model (i.e., to estimate parameters), one of two constraints is typically imposed:

- $\sum_{h=0}^{m} \lambda_{ih} = \sum_{h=0}^{m} \zeta_{ih} = 0$, or
- $\lambda_{i0} = \zeta_{i0} = 0$.

NRM (Bock, 1972): Example

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

	Response options				
	A	B	C	D	$\sum$
λ_i	$-.30$	$.81$	$-.31$	$-.20$	$.000$
ζ_i	$.21$	$.82$	$-.09$	$-.94$	$.000$



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 ζ_{ic}.

For the example, in increasing order of popularity:

$$\text{Response D} < \text{Response C} < \text{Response A} < \text{Response B}.$$

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.

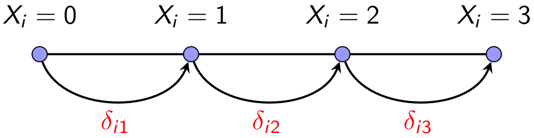
The PCM is a "divide-by-total", or "direct" model:
The ICRFs are modeled directly.

PCM (Masters, 1982)

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

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

- δ_{ij} ($j = 1, \dots, m$): 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:

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

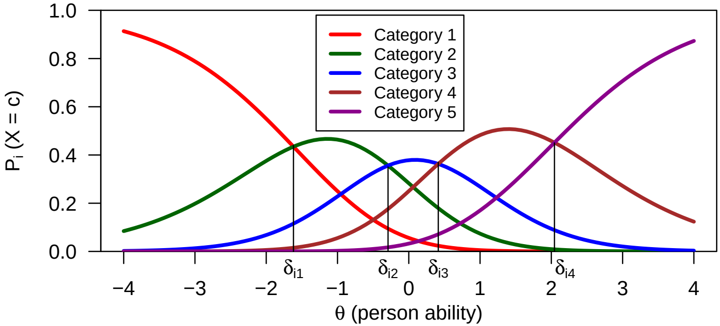
- The higher the δ_{ij}, the more difficult a particular step is.
- The δ_{ij}'s aren't necessarily ordered in the same sequence as the categories (reversals; such a case indicates that the item is probably not functioning as intended).

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 ($N \simeq 1,000$).
Five-point Likert-type rating scale.

Step Difficulties			
δ_{i1}	δ_{i2}	δ_{i3}	δ_{i4}
-1.618	-0.291	0.414	2.044



PCM (Masters, 1982): Example

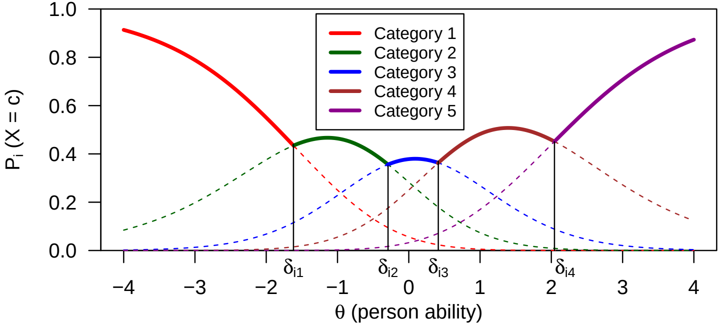
Interpretation:

- In this case the δ_{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 θ-scale.

PCM (Masters, 1982): Example

Interpretation:

- In this case the δ_{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 θ-scale.



Generalized partial credit model (GPCM)

GPCM (Muraki, 1992)

- 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.

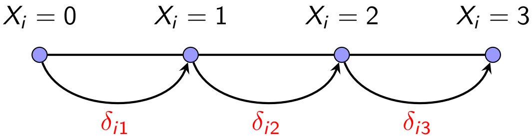
The GPCM is a "divide-by-total", or "direct" model:
The ICRFs are modeled directly.

GPCM (Muraki, 1992)

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

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

- δ_{ij} ($j = 1, \ldots, m$): Item step difficulties (category intersections).
- α_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.
- α_i — Intuitive interpretation:
  - Small values (say, ≤ 1) $\rightarrow$ 'flatter' ICRFs.
  - Large values (say, ≥ 1.5) $\rightarrow$ more 'peaked' ICRFs.

In Muraki's (1992, p. 162) words:

> *"[The α_i's] indicate the degree to which categorical responses vary among items as θ 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 = strongly disagree;. . . ; 4 = strongly agree.)
- $N = 350$.

Let's see three items.

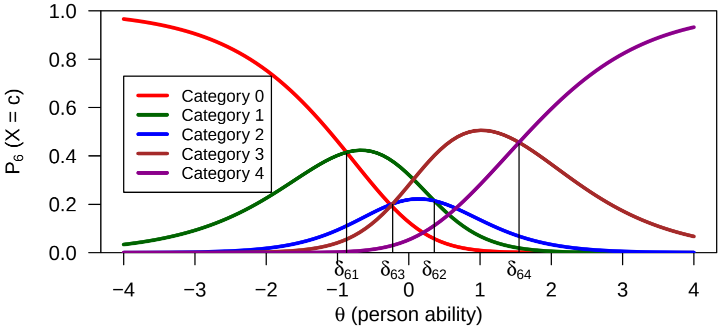
		Response category				
Item	Content	0	1	2	3	4
5	Feels tense and jittery	17	111	97	101	24
6	Sometimes feels worthless	72	89	52	94	43
9	Feels discouraged, like giving up	27	128	66	95	34

GPCM (Muraki, 1992): Example (slope $\simeq$ 1)

Item 6 '*Sometimes feels worthless*'.

$(0 = 72,\ 1 = 89,\ 2 = 52,\ 3 = 94,\ 4 = 43)$.

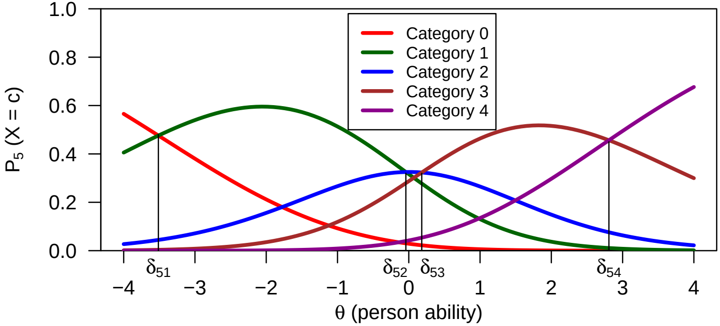
Slope	Step Difficulties			
α_6	δ_{61}	δ_{62}	δ_{63}	δ_{64}
1.073	−0.873	0.358	−0.226	1.547



GPCM (Muraki, 1992): Example (slope < 1)

Item 5 '*Feels tense and jittery*'.
$(0=17,\ 1=111,\ 2=97,\ 3=101,\ 4=24)$.

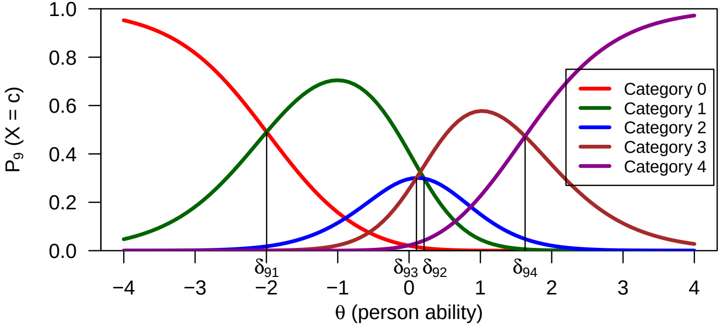
Slope	Step Difficulties			
α_5	δ_{51}	δ_{52}	δ_{53}	δ_{54}
0.683	−3.513	−0.041	0.182	2.808



GPCM (Muraki, 1992): Example (slope $\simeq$ 1.5)

Item 9 '*Feels discouraged, like giving up*'.
$(0 = 27,\ 1 = 128,\ 2 = 66,\ 3 = 95,\ 4 = 34)$.

Slope	Step Difficulties			
α_9	δ_{91}	δ_{92}	δ_{93}	δ_{94}
1.499	−1.997	0.210	0.103	1.627



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.

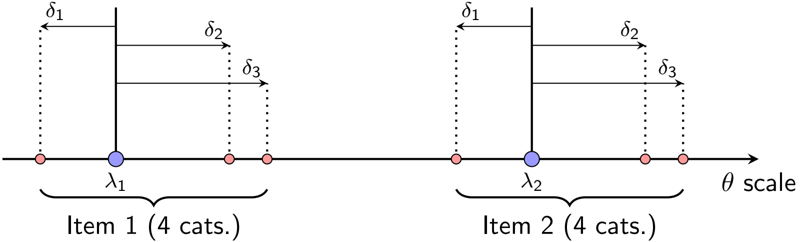
The RSM is a "divide-by-total", or "direct" model:
The ICRFs are modeled directly.

RSM (Andrich, 1978)

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

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

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



RSM (Andrich, 1978)

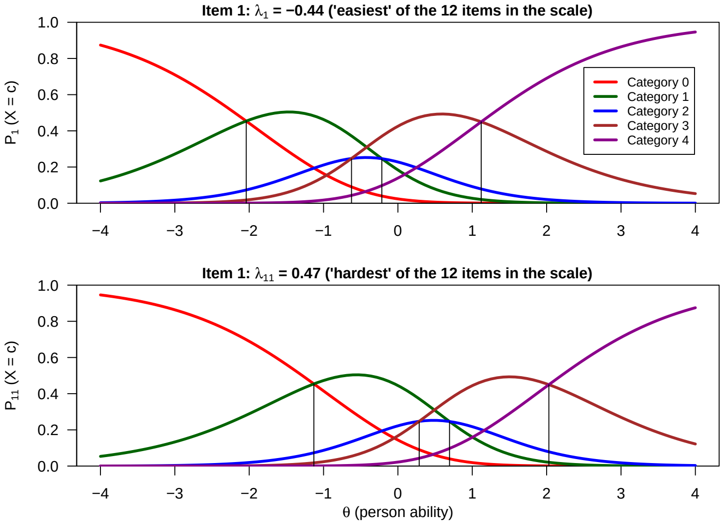
- Two consecutive categories intersect at $\theta = (\lambda_i + \delta_j)$:

$$P_{i(j-1)}(\lambda_i + \delta_j) = P_{ij}(\lambda_i + \delta_j).$$

- 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.

GRM (Samejima, 1969)

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

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

where

$$\underbrace{P^*_{ic}}_{P(X_i \geq c|\theta)} = \frac{1}{1 + \exp[-\alpha_i(\theta - \beta_{ic})]} \quad \text{(the 2PLM).}$$

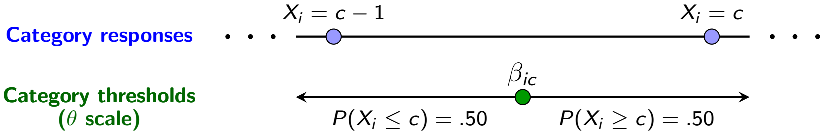
(And $P^*_{i0} \equiv 1$, $P^*_{im} \equiv 0$.)

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)

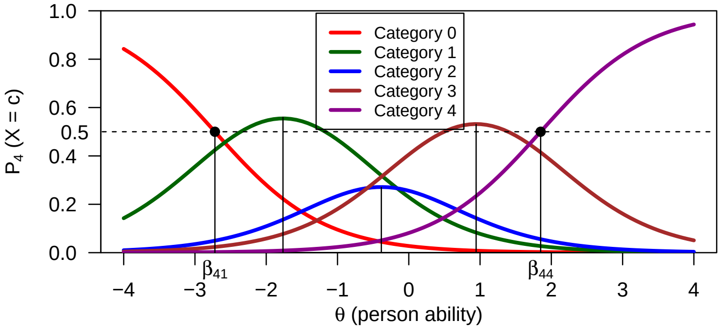
- α_i: Item slope parameter (one per item).
- β_{ic}: Category threshold parameters
  (one set $\{\beta_{i1}, \ldots, \beta_{im}\}$ per item).
  These are the θ-values of transition between response categories.
- The β_{ic}'s are necessarily ordered.



GRM (Samejima, 1969): Example (NEO-FFI)

Item 4 '*Rarely feels lonely, blue*'.
$(0 = 20,\ 1 = 90,\ 2 = 68,\ 3 = 125,\ 4 = 47)$.

Slope	Category thresholds			
α_4	β_{41}	β_{42}	β_{43}	β_{44}
1.31	−2.72	−0.81	0.04	1.85



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 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
     - $\hookrightarrow$ Compare model-predicted *vs* empirical response functions.
     - $\hookrightarrow$ Plot residuals.

Model selection

3. 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

3. Mathematical approaches to check fit
    - Goodness of fit
      Consider model fit $\oplus$ number of estimated parameters.
        - $\hookrightarrow$ Akaike's information criterion (AIC; Akaike, 1977).
        - $\hookrightarrow$ 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 χ^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 reassurence:

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,...