IRT Models for Polytomous Response Data

Lecture #4

ICPSR Item Response Theory Workshop

Lecture Overview

- Big Picture Overview
  - Framing “Item Response Theory” as a generalized latent variable modeling technique
  - Differentiating “RESPONSE Theory” from “Item RESPONSES”

- Nominal Response (but Categorical) Data
  - Ordered Category Models :: Graded Response Model
  - Partially Ordered Category Models :: Partial Credit Model
  - Unordered Category Models :: Nominal Response

- Brief introduction to even more types of data

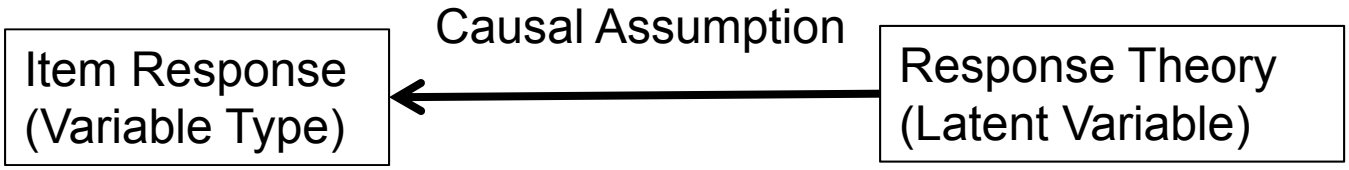
DIFFERENTIATING “RESPONSE THEORY” FROM “ITEM RESPONSES”

Fundamentals of IRT

- IRT is a type of *measurement model* in which transformed item responses are predicted using properties of **persons** (Theta) and properties of **items** (difficulty, discrimination)
- "**Rasch models**" are a subset of IRT models with more restrictive slope assumptions
- Items and persons are on the same latent metric: "**conjoint scaling**"
  - Anchor (identify) scale with either persons (z-scored theta) or items
- After controlling for a person's latent trait score (Theta), the item responses should be uncorrelated: "**local independence**"
- Item response models are re-parameterized versions of item factor models (for binary outcomes)
  - Thus, we can now extend IRT to "**polytomous responses**" (3+ options)

The Big Picture

- The key to working through the varying types of IRT models is understanding that IRT is all about the type of data you have that you intend to model
  - Once the data type is know, the nuances of a model family become evident (but mainly are due to data types)



- In latent variable modeling, we assume that variability in unobserved traits cause variability in item responses

IRT from the Big Picture Point-of-View

$$P(Y_{si} = 1|\theta_s) = \frac{\exp\left(1.7a(\theta_s - b_i)\right)}{1 + \exp\left(1.7a(\theta_s - b_i)\right)}$$

- Or…more conveniently re-organized:

$$ln\left(\frac{P(Y_{si} = 1|\theta_s)}{1 - P(Y_{si} = 1|\theta_s)}\right) = 1.7a(\theta_s - b_i) = b_i^* + a_i^*\theta_s$$

- The model has two parts:

Item Response (Variable Type)

Response Theory (Latent Variable)

Polytomous Items

- Polytomous items end up changing the left hand side of the equation
  - The *Item Response* portion
- Subsequently, minor changes are made to the right hand side
  - The *Response Theory* portion
- These changes frequently are related to the item more than to the theory
  - Think of the *c* parameter in the 3-PL (for guessing)
    - It cannot be present in an item that is scored continuously
- More commonly, nuances in IRT software reflect the changes in how models are constructed
  - But general theory remains the same

Polytomous Items

- Polytomous items mean more than 2 options (categorical)

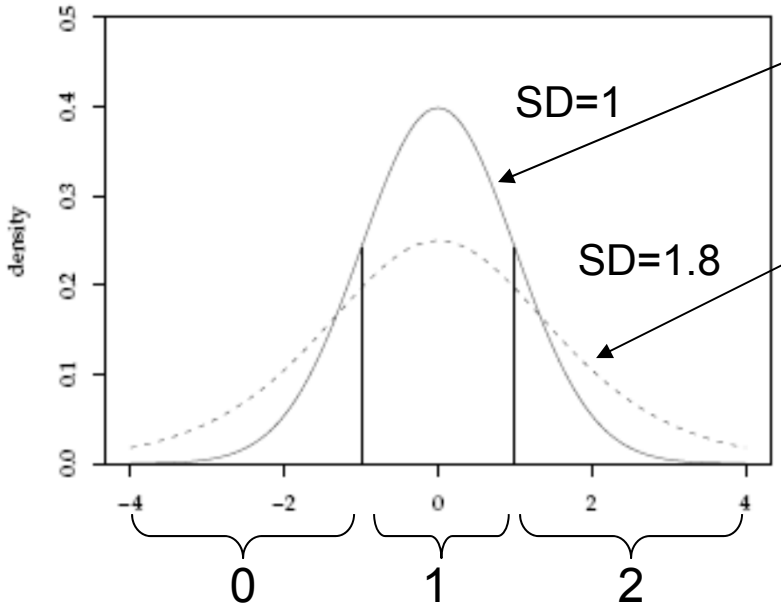
- Polytomous models are not named with numbers like binary models, but instead get called different names
  - Most have a 1-PL vs. 2-PL version that go by different names
  - Different constraints on what to do with multiple categories

- Three main kinds* of polytomous models:
  - Outcome categories are ordered (scoring rubrics, “Likert” scales)
    - Graded Response or Modified Graded Response Model
  - Outcome categories could be ordered
    - (Generalized) Partial Credit Model or Rating Scale Model
  - Outcome categories are not ordered (distractors/multiple choice)
    - Nominal Response Model

* Lots and lots more – these are the major categories

Threshold Concept for Binary and Ordinal Variables

- Each ordinal variable is really the chopped-up version of a hypothetical underlying continuous variable (Y*) with a mean of 0



thresholds = # options - 1

Probit (ogive) model: Pretend variable has a normal distribution (variance = 1)

Logit model: Pretend variable has logistic distribution (variance = $\pi^2/3$)

Polytomous models will differ in how they make use of multiple (k-1) thresholds per item

GRADED RESPONSE MODEL

Example Graded Response Item

From the 2006 Illinois Standards Achievement Test (ISAT):

www.isbe.state.il.us/assessment/pdfs/Grade_5_**ISAT**_2006_Samples.pdf

Mathematics Short-Response Sample Item

Below is a short-response sample item, followed by the short-response scoring rubric and 3 samples of student responses.

This short-response sample item is classified to assessment objective 10.8.07, "Represent all possible outcomes (sample space) for simple or compound events (e.g., tables, grids, tree diagrams)."

16

Using each digit only once, list all possible 3-digit numbers that can be made using the digits 2, 4, and 7.

ISAT Scoring Rubric

MATHEMATICS SCORING RUBRIC: A GUIDE TO SCORING SHORT-RESPONSE ITEMS

Note: Item-specific rubrics are developed for each item before scoring.

Score Level	
2	• Completely correct response, including correct work shown and/or correct labels/units if called for in the item
1	• Partially correct response
0	• No response, or the response is incorrect

Additional Example Item

- Cognitive items are not the only ones where graded response data occurs

- Likert-type questionnaires are commonly scored using ordered categorical values
  - Typically, these ordered categories are treated as continuous data (as with Factor Analysis)

- Consider the following item from the Satisfaction With Life Scale (e.g. SWLS, Diener, Emmons, Larsen, & Griffin, 1985)…

SWLS Item #1

- I am satisfied with my life.
    1. Strongly disagree
    2. Disagree
    3. Slightly disagree
    4. Neither agree nor disagree
    5. Slightly agree
    6. Agree
    7. Strongly agree

Graded Response Model (GRM)

- Ideal for items with clear underlying response continuum
- # response options (k) don't have to be the same across items
- Is an "indirect" or "difference" model
  - Compute difference between models to get probability of each response
- Estimate 1 a_i per item and k-1 difficulties (4 options → 3 difficulties)

- Models the probability of any given response category or higher, so for any given difficulty submodel, it will look like the 2PL
  - Otherwise known as "cumulative logit model"
  - Like dividing 4-category items into a series of binary items…
  - $\underbrace{0 \text{ vs. } 1,2,3}_{b_{1i}}$ $\underbrace{0,1 \text{ vs. } 2,3}_{b_{2i}}$ $\underbrace{0,1,2 \text{ vs. } 3}_{b_{3i}}$

  - **…But each threshold uses all response data in estimation**

Example GRM for 4 Options (0-3): 3 Submodels with common a

- Prob of 0 vs 123 :: $P_{i1}(Y_{si} \geq 1) = \frac{\exp(1.7a_i(\theta_s - b_{i1}))}{1+\exp(1.7a_i(\theta_s - b_{i1}))}$

- Prob of 01 vs 23 :: $P_{i2}(Y_{si} \geq 1) = \frac{\exp(1.7a_i(\theta_s - b_{i2}))}{1+\exp(1.7a_i(\theta_s - b_{i2}))}$

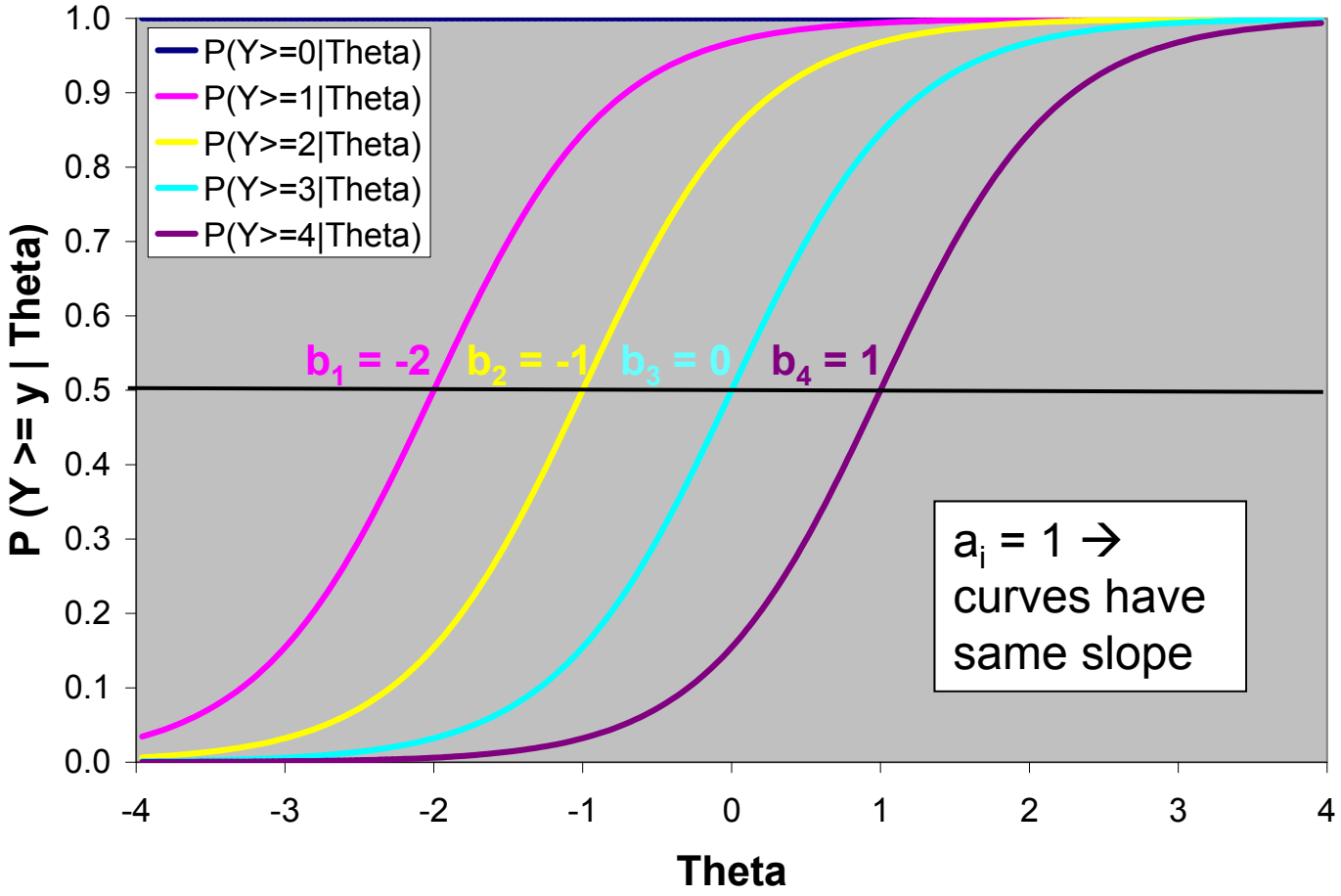
- Prob of 012 vs 3 :: $P_{i3}(Y_{si} \geq 1) = \frac{\exp(1.7a_i(\theta_s - b_{i3}))}{1+\exp(1.7a_i(\theta_s - b_{i3}))}$

- Prob of 0 → $1 - P_{i1}$
  Prob of 1 → $P_{i1} - P_{i2}$
  Prob of 2 → $P_{i2} - P_{i3}$
  Prob of 3 → $P_{i3} - 0$

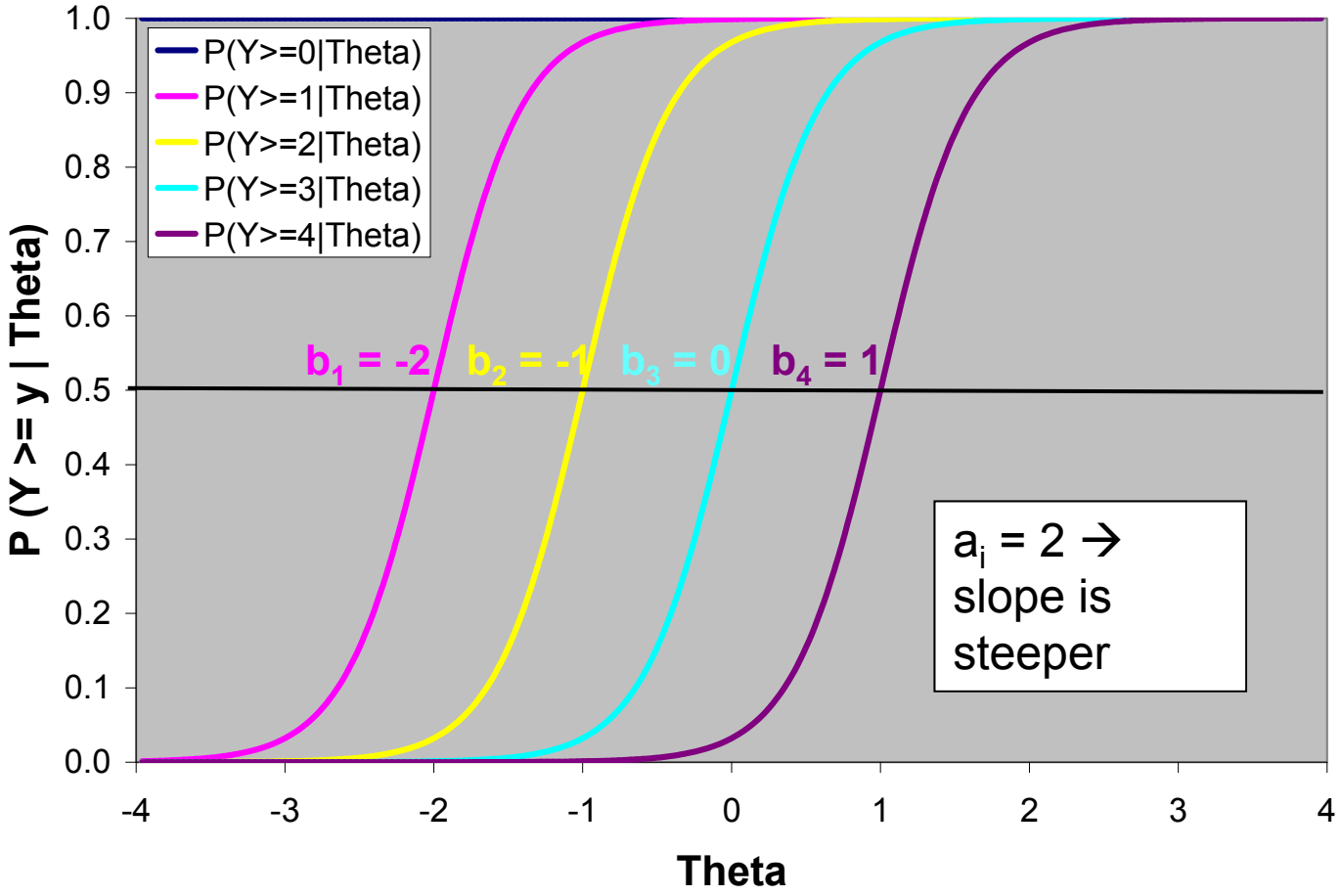
Note $\mathbf{a_i}$ is the same across thresholds :: only one slope per item

$\mathbf{b_{ik}}$ = trait level needed to have a 50% probability of responding in that category or higher

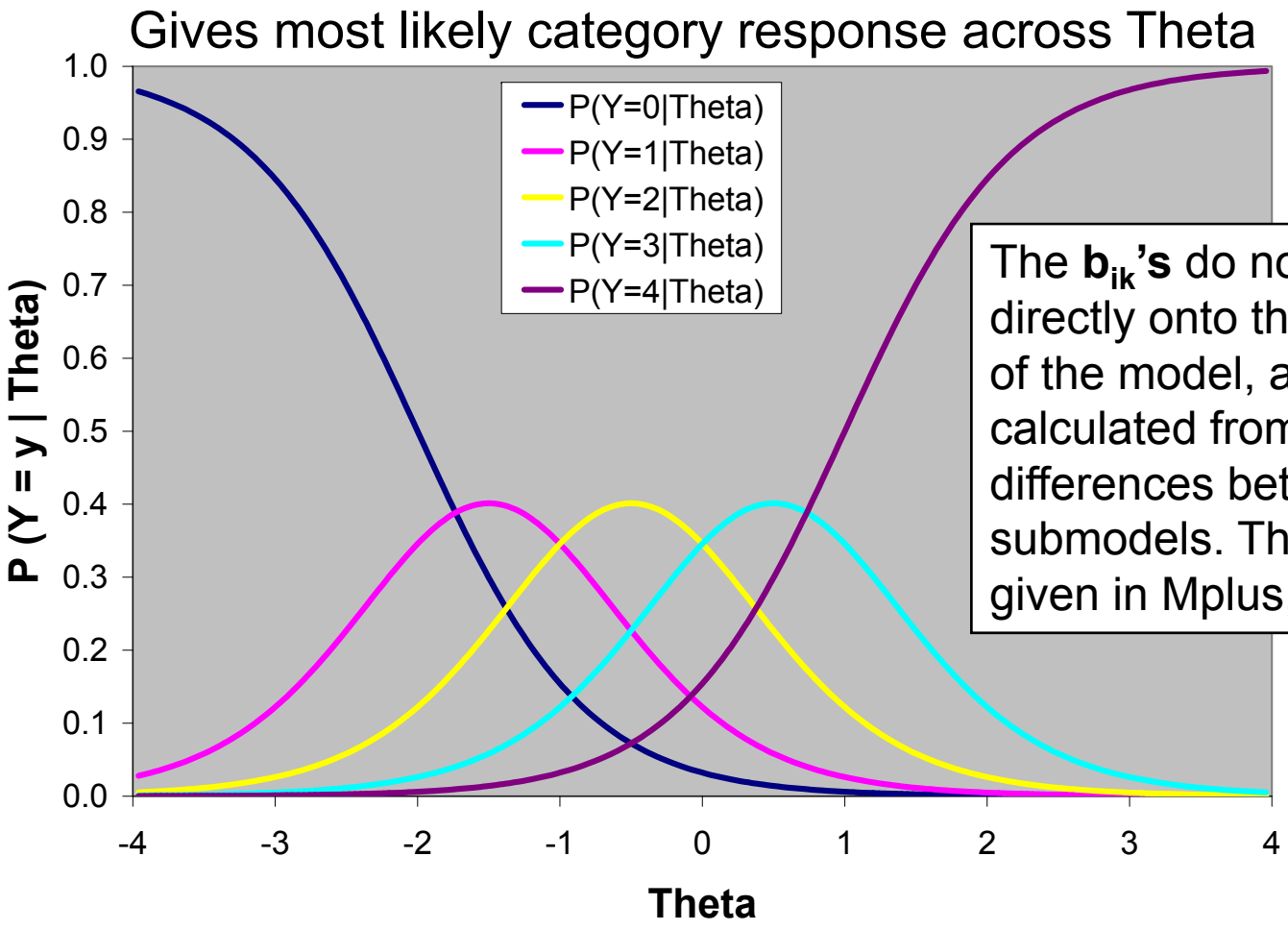
Cumulative Item Response Curves (GRM for 5-Category Item, $a_i = 1$)



Cumulative Item Response Curves (GRM for 5-Category Item, $a_i = 2$)

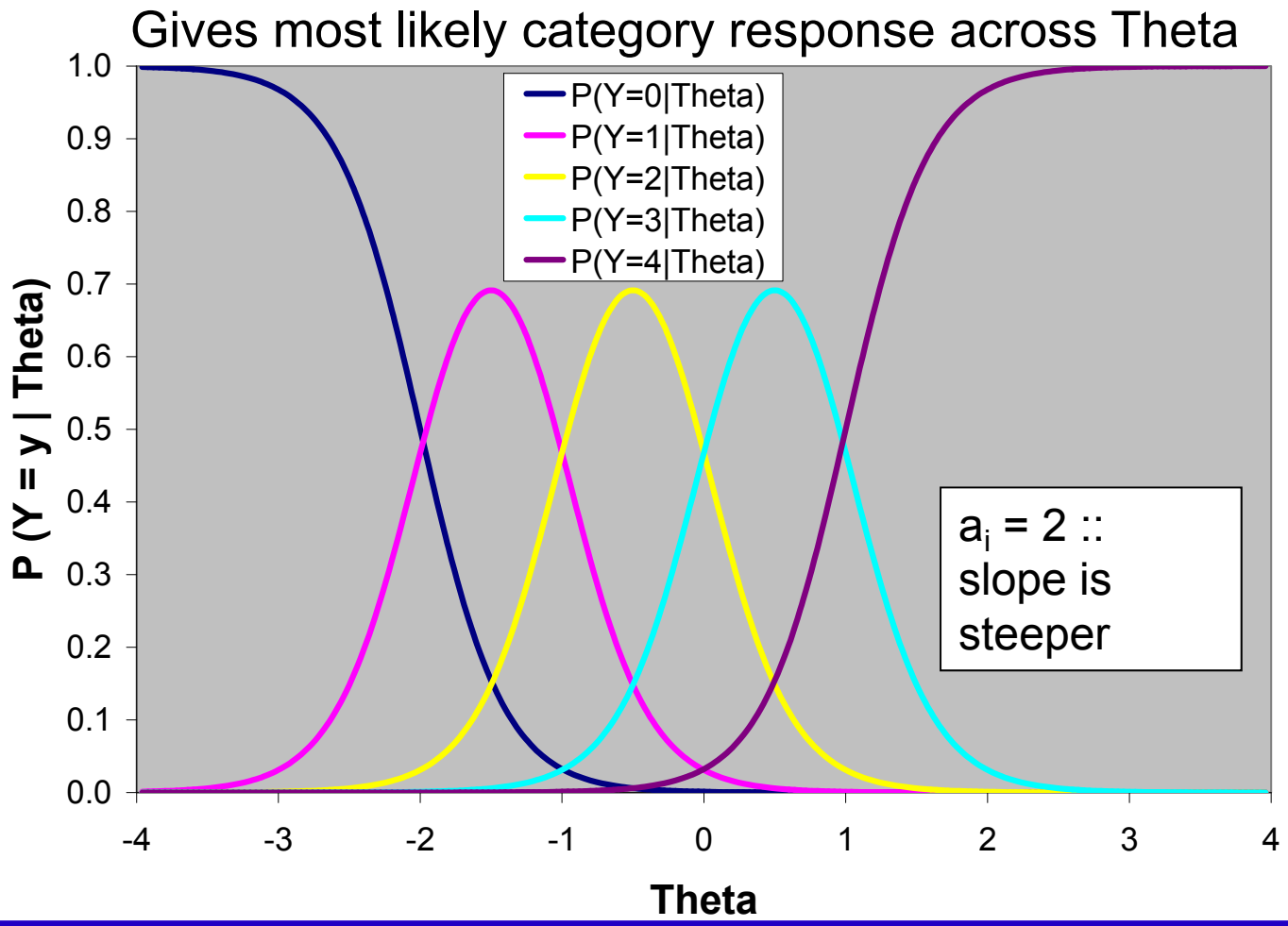


Category Response Curves (GRM for 5-Category Item, $a_i = 1$)

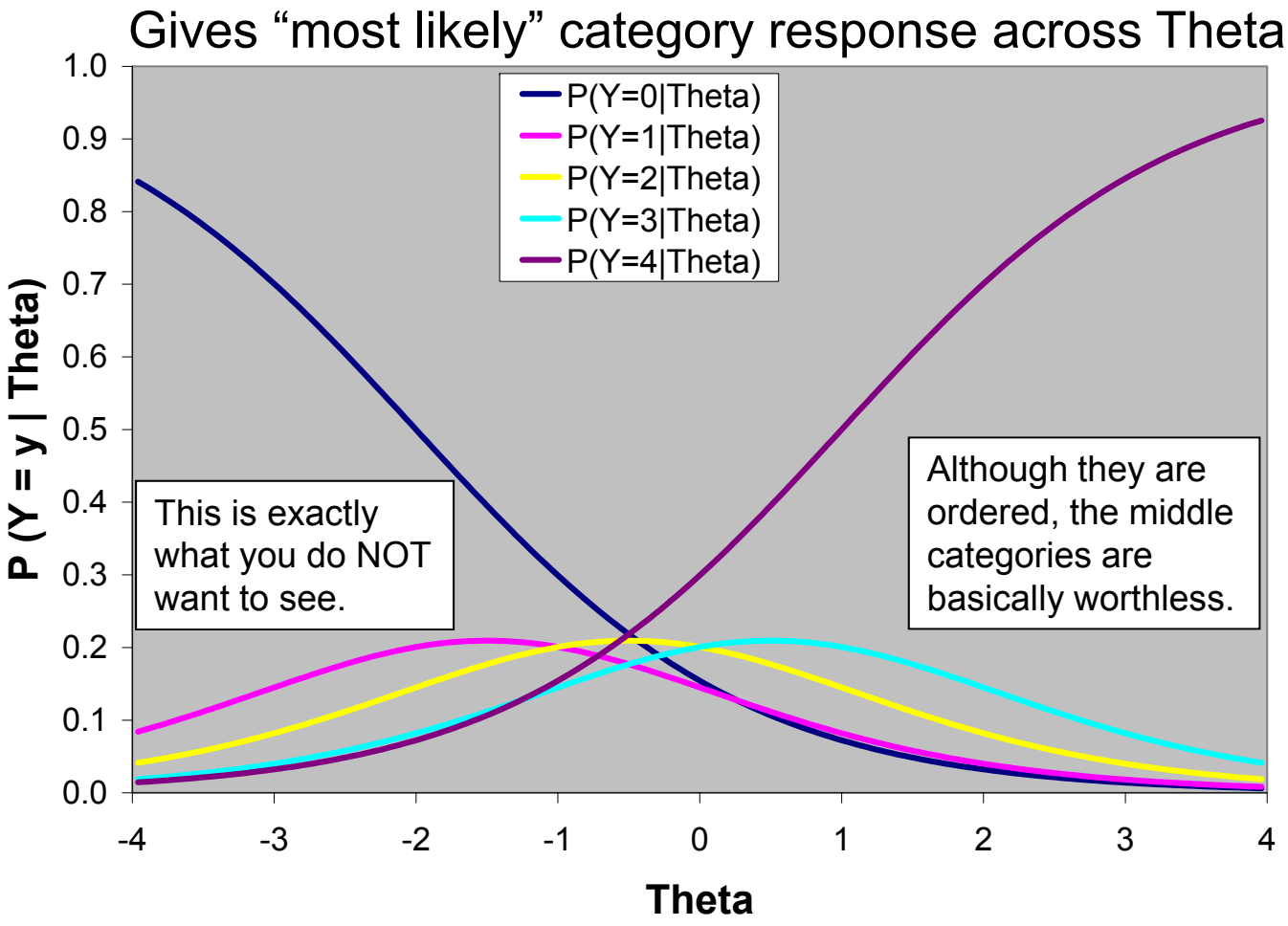


The **b_{ik}'s** do not map directly onto this illustration of the model, as these are calculated from the differences between the submodels. This is what is given in Mplus, however.

Category Response Curves (GRM for 5-Category Item, $a_i = 2$)



Category Response Curves (GRM 5-Category Item, a_i = .5)

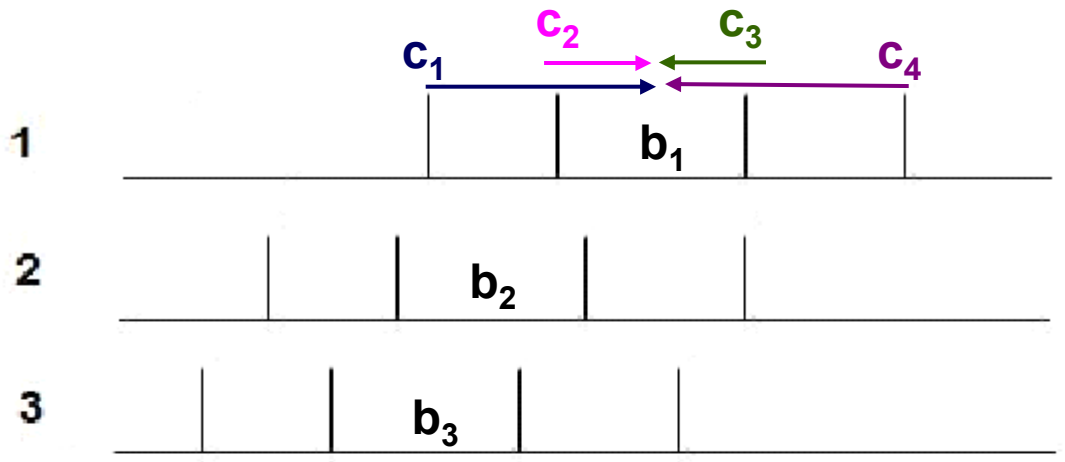


"Modified" ("Rating Scale") Graded Response Model

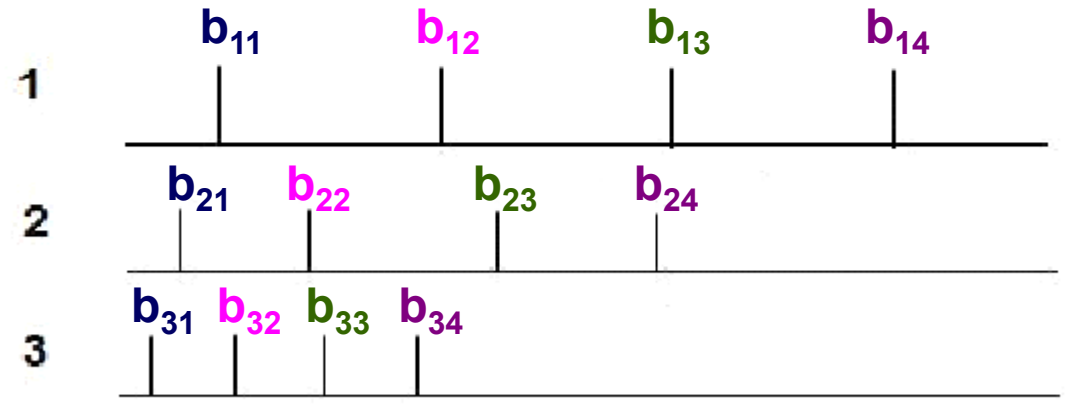
- Is more parsimonious version of graded response model
- Designed for items with same response format
- In GRM, there are (#options-1)*(#items) thresholds estimated + one slope per item
- In MGRM, each item gets own slope and own 'location' parameter, but the differences between categories around that location are constrained equal across items (get a "c" shift for each threshold)
  - Items differ in overall location, but spread of categories within is equal
  - So, different ai and bi per item, but same c1, c2, and c3 across items
- Prob of 0 vs 123 :: $P_{i1}(Y_{si} \geq 1) = \frac{\exp\left(1.7a_i(\theta_s-(b_i+c_i))\right)}{1+\exp\left(1.7a_i(\theta_s-(b_i+c_i))\right)}$

(and so forth for c2 and c3)

  - Not same 'c' as guessing parameter – sorry, they reuse letters…
  - Not directly available within Mplus, but pry could be using constraints



Modified GRM ::
1 Location, k-1 c's
All category distances are same across items



Original GRM ::
k-1 locations
All category distances are allowed to differ across items

Item Difficulty / Latent Ability

Summary of Models for Ordered Categorical Responses

Available in Mplus with "CATEGORICAL ARE" option	**Difficulty Per Item Only (category distances equal)**	**Difficulty Per Category Per Item**
Equal discrimination across items (1-PLish)?	(possible, but no special name)	(possible, but no special name)
Unequal discriminations (2-PLish)?	"Modified GRM" or "Rating Scale GRM" (same response options)	"Graded Response Model" "Cumulative Logit"

- GRM and Modified GRM are reliable models for ordered categorical data
  - Commonly used in real-world testing; most stable to use in practice
  - Least data demand because all data get used in estimating each b_{ik}
  - Only major deviations from the model will end up causing problems

PARTIAL CREDIT MODEL

Partial Credit Model (PCM)

- Ideal for items for which you want to test an assumption of an ordered underlying continuum
  - # response options doesn't have to be same across items
- Is a "direct, divide-by-total" model (probability of response given directly)

- Estimate k-1 thresholds (so 4 options :: 3 thresholds)

- Models the probability of adjacent response categories:
  - Otherwise known as "adjacent category logit model"
  - Divide item into a series of binary items, but without order constraints beyond adjacent categories because it only uses those 2 categories:
  - $\underbrace{0 \text{ vs. } 1}_{\delta_{1i}} \quad \underbrace{1 \text{ vs. } 2}_{\delta_{2i}} \quad \underbrace{2 \text{ vs. } 3}_{\delta_{3i}}$
  - No guarantee that any category will be most likely at some point

Partial Credit Model

- With different slopes (a_i) per item, then it's "generalized partial credit model"; otherwise 1-PLish version is Partial Credit Model
- **Still 3 submodels for 4 options, but set up differently:**
  - Given 0 or 1, prob of 1 :: $P_{i1}(\theta_s) = \frac{\exp(1.7a_i(\theta_s - \delta_{i1}))}{1+\exp(1.7a_i(\theta_s - \delta_{i1}))}$
  - Given 1 or 2, prob of 2 :: $P_{i2}(\theta_s) = \frac{\exp(1.7a_i(\theta_s - \delta_{i2}))}{1+\exp(1.7a_i(\theta_s - \delta_{i2}))}$
  - Given 2 or 3, prob of 3 :: $P_{i3}(\theta_s) = \frac{\exp(1.7a_i(\theta_s - \delta_{i3}))}{1+\exp(1.7a_i(\theta_s - \delta_{i3}))}$
- **δ** is the 'step parameter' :: latent trait where the next category becomes *more likely* – not necessarily 50%

- Other parameterizations also used – check the program manuals
- Currently not directly available in Mplus

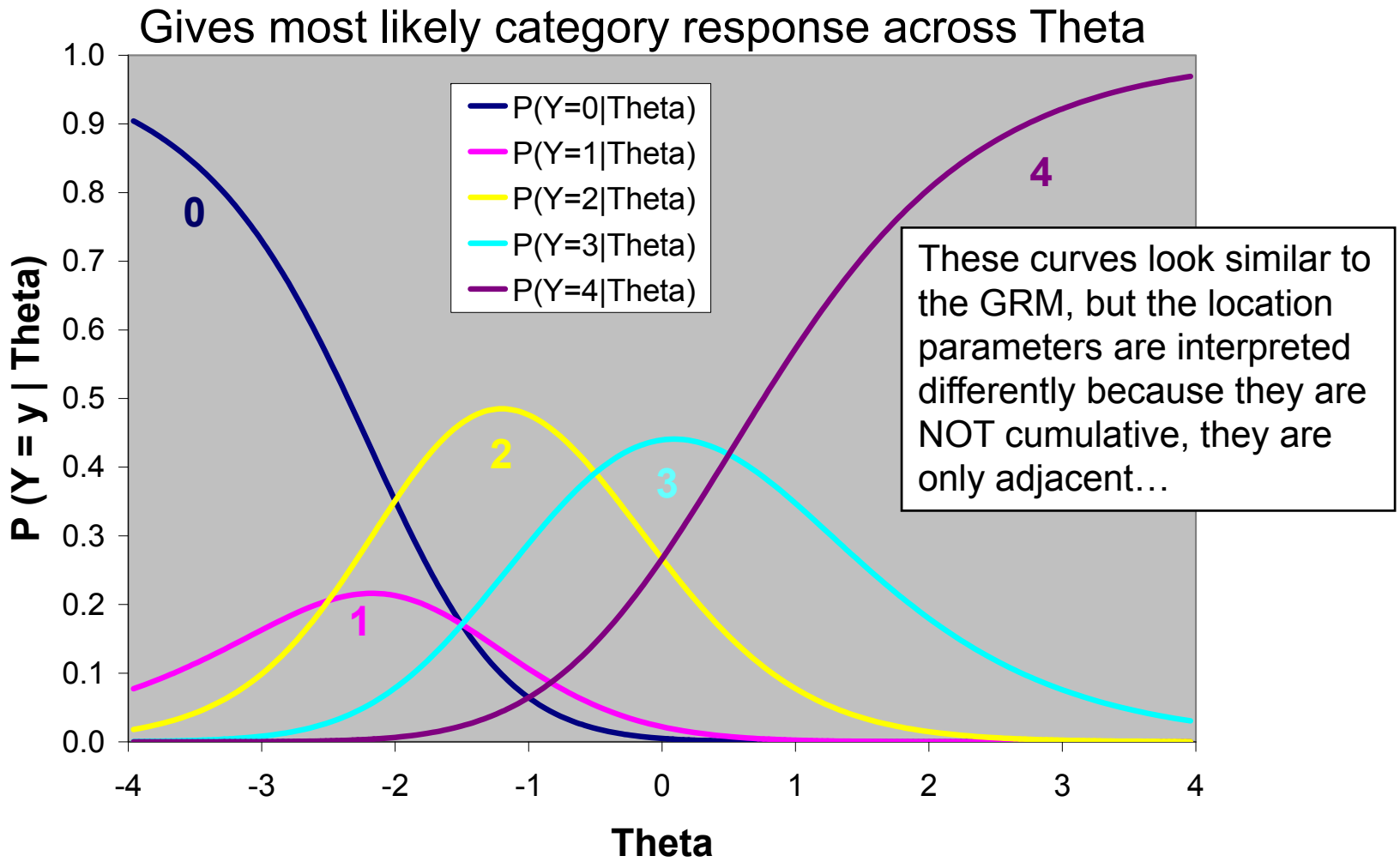
Generalized Partial Credit Model

- The item score category function

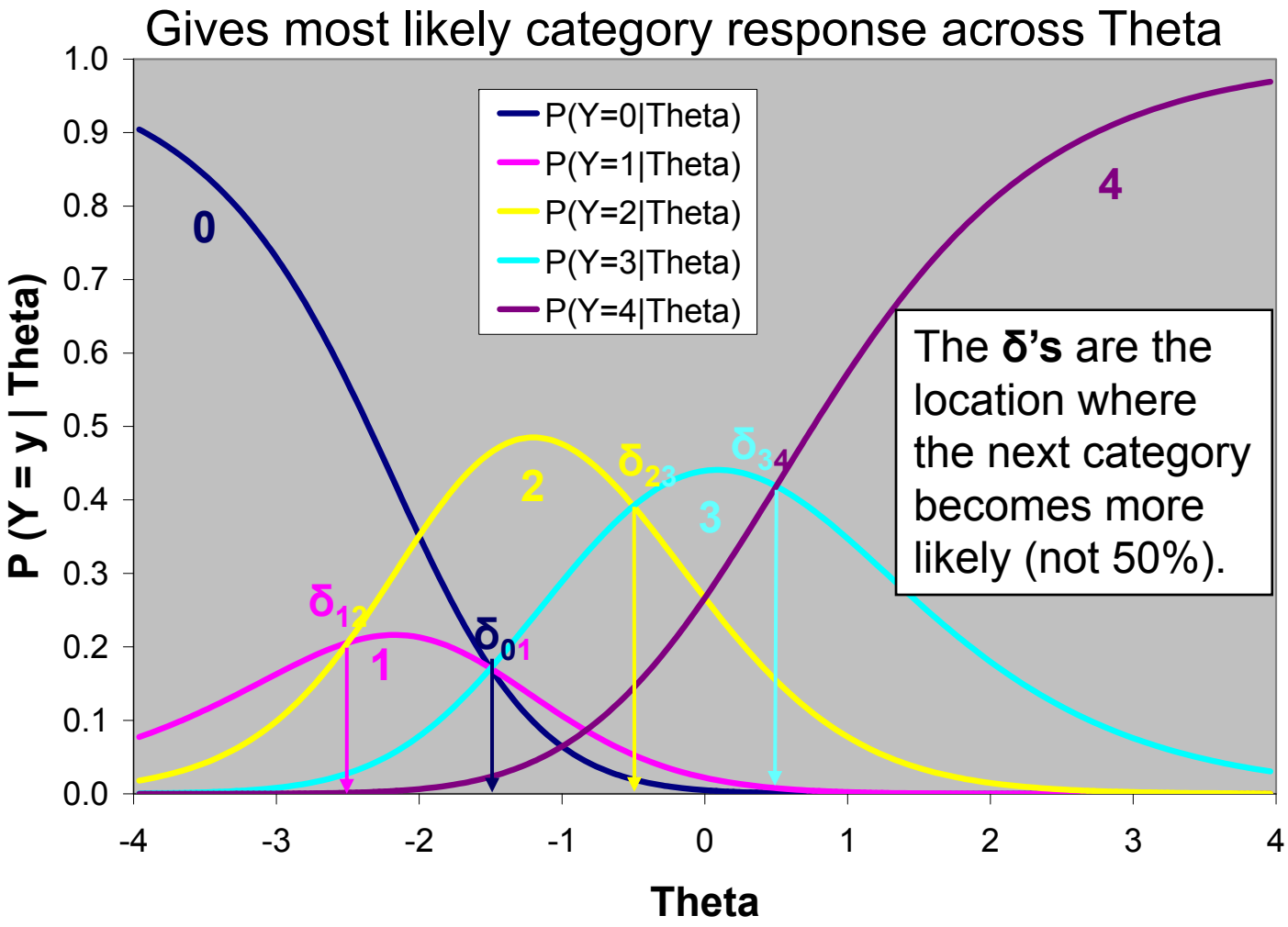
$$P_{iy}(\theta_s) = \frac{\exp\left[\sum_{k=0}^{y} 1.7a_i(\theta_s - \delta_{ik})\right]}{\sum_{h=0}^{m_i} \exp\left[\sum_{k=0h} 1.7a_i(\theta_s - \delta_{ik})\right]}$$

$$= \frac{e^{\text{Sum of terms for each category up to y}}}{\text{sum of numerator terms for all possible categories}}$$

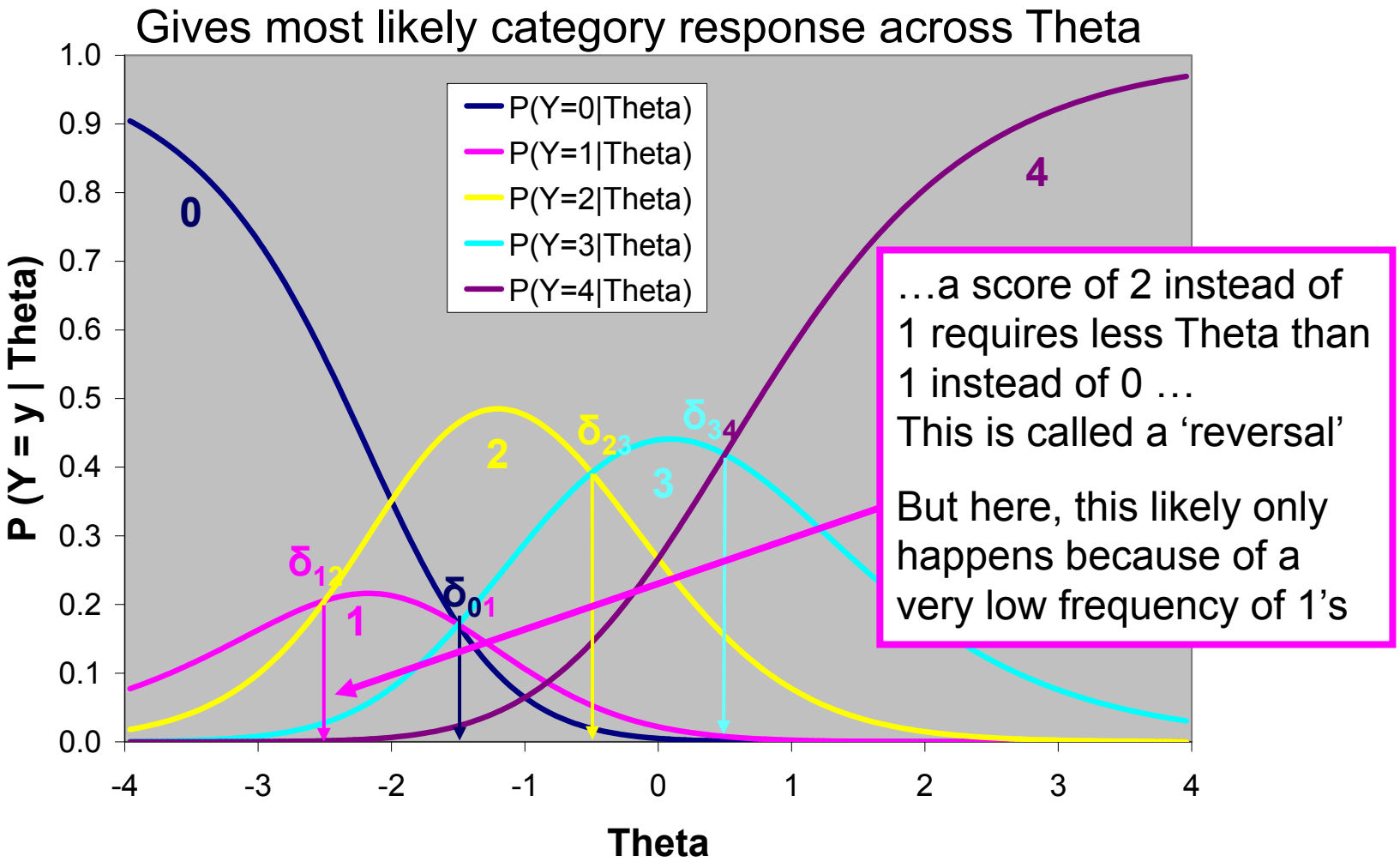
Category Response Curves (PCM for 5-Category Item, ai = 1)



Category Response Curves (PCM for 5-Category Item, $a_i = 1$)



Category Response Curves (PCM for 5-Category Item, $a_i = 1$)



Partial Credit Model vs. Graded Response Model

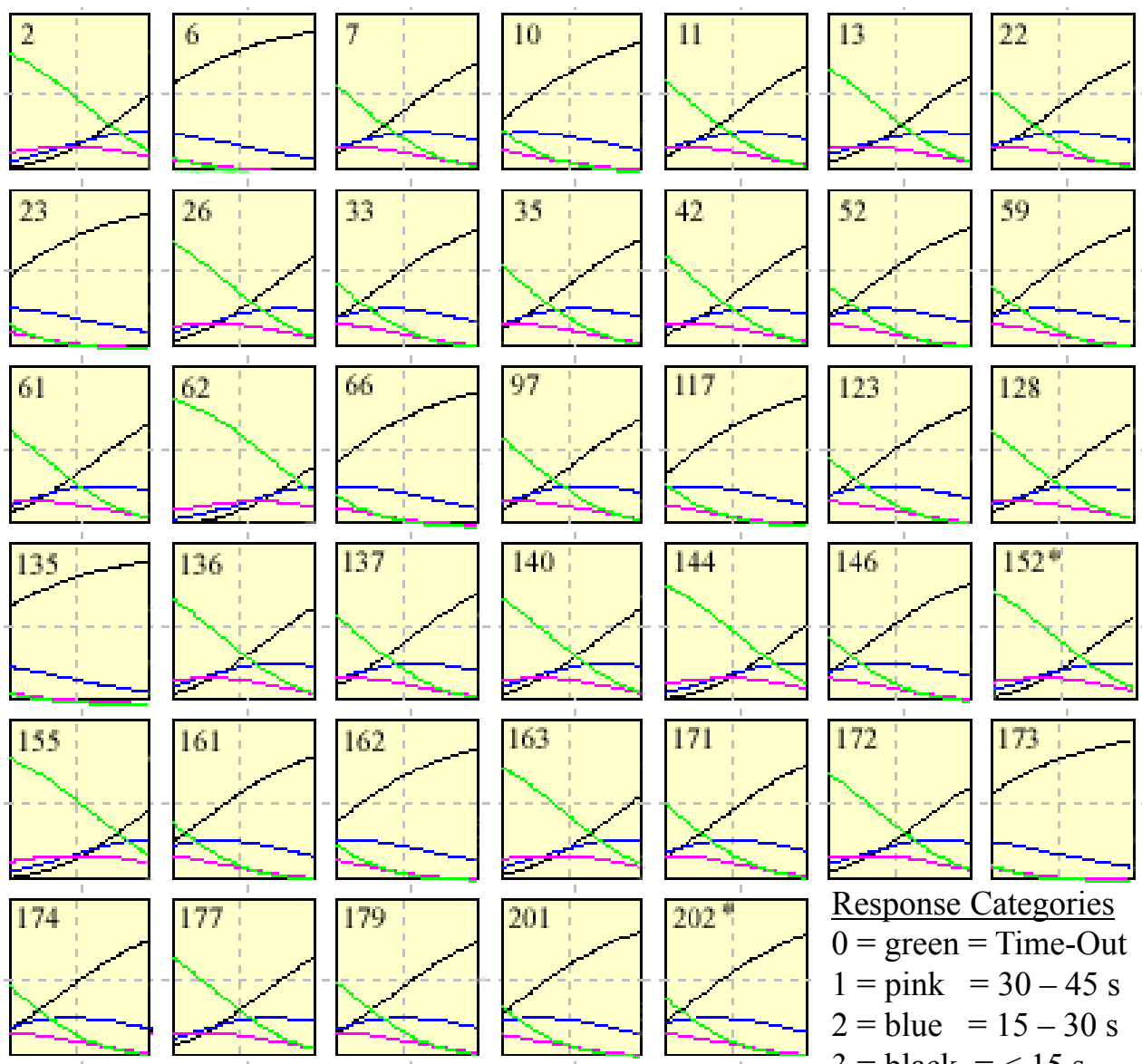
- The PCM is very similar to GRM
  - Except these models allow for the fact that one or more of the score categories may never have a point where the probability of x is greatest for a given q level

- Because of local estimation, there is no guarantee that category b-values will be ordered

- This is a flaw or a strength, depending on how you look at it…

PCM and GPCM vs. GRM

- GPCM and GRM will generally agree very closely, unless one or more of the score categories is underused
- GRM will force the categories boundary parameters to be ordered, GPCM and PCM do not
- For this reason, comparing results with the same data across models can point out interesting phenomena in your data

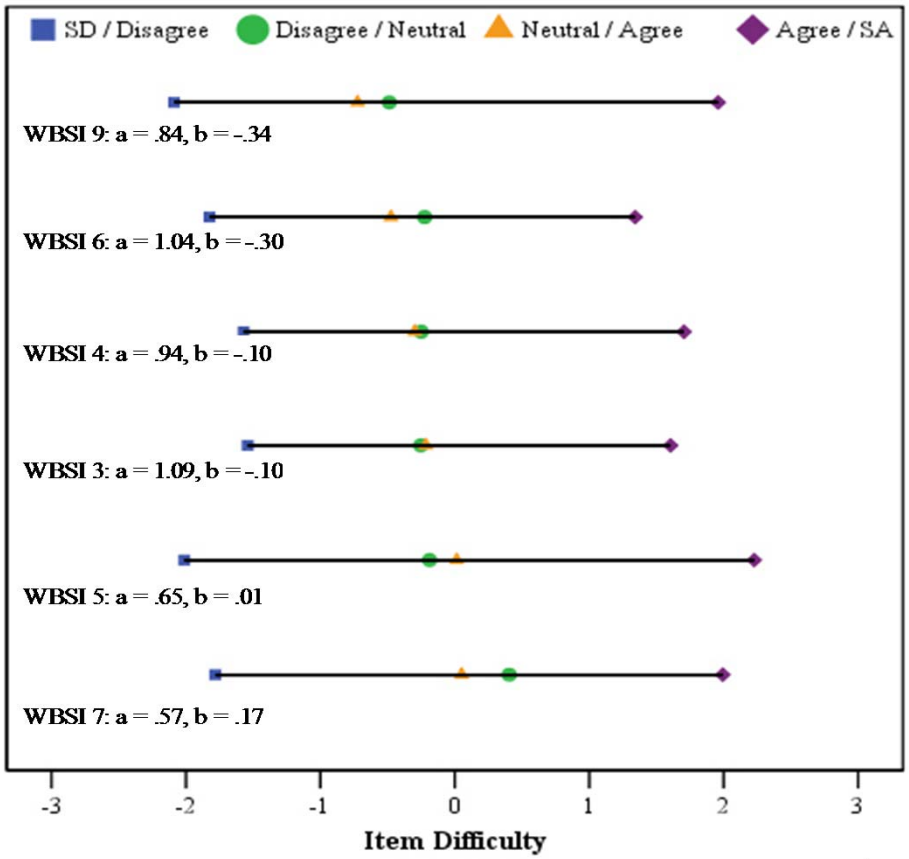


Response Categories
0 = green = Time-Out
1 = pink = 30 – 45 s
2 = blue = 15 – 30 s
3 = black = < 15 s
*Misfit ($p < .05$)

More of what you don't want to see… category response curves from a PCM where reversals are a plenty…

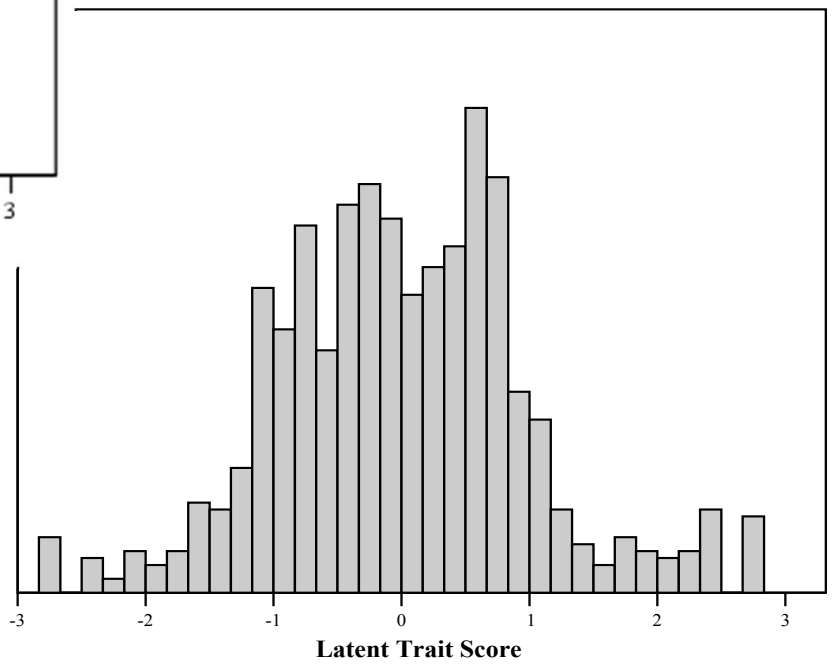
…and the middle categories are fairly useless.

PCM Example: General Intrusive Thoughts (5 options)

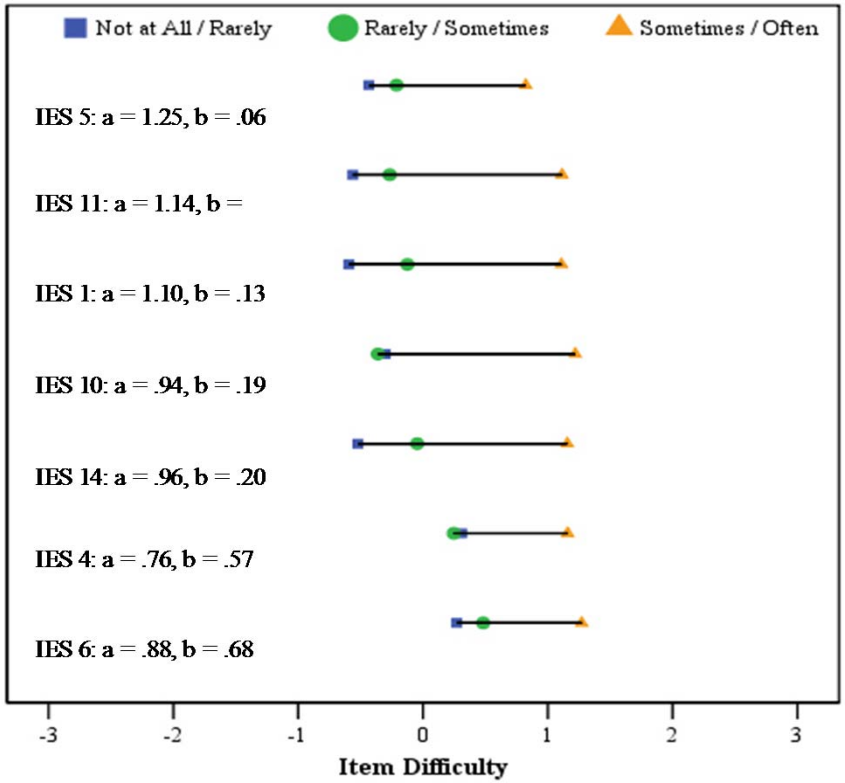


Note that the 4 thresholds **cover a wide range** of the latent trait, and what the distribution of Theta looks like as a result...

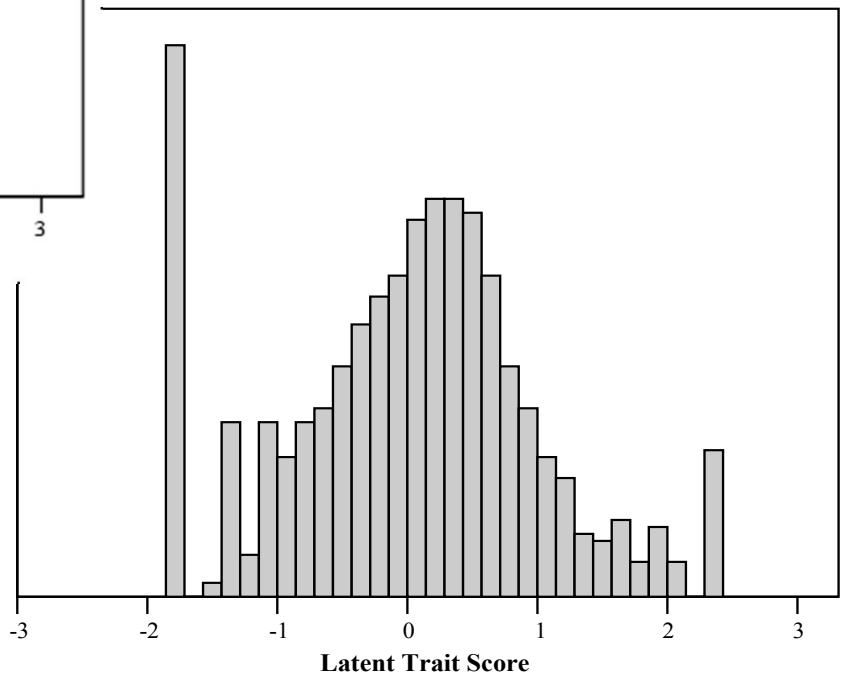
But the middle 3 categories are used infrequently &/or are not differentiable



Partial Credit Model Example: Event-Specific Intrusive Thoughts (4 options)

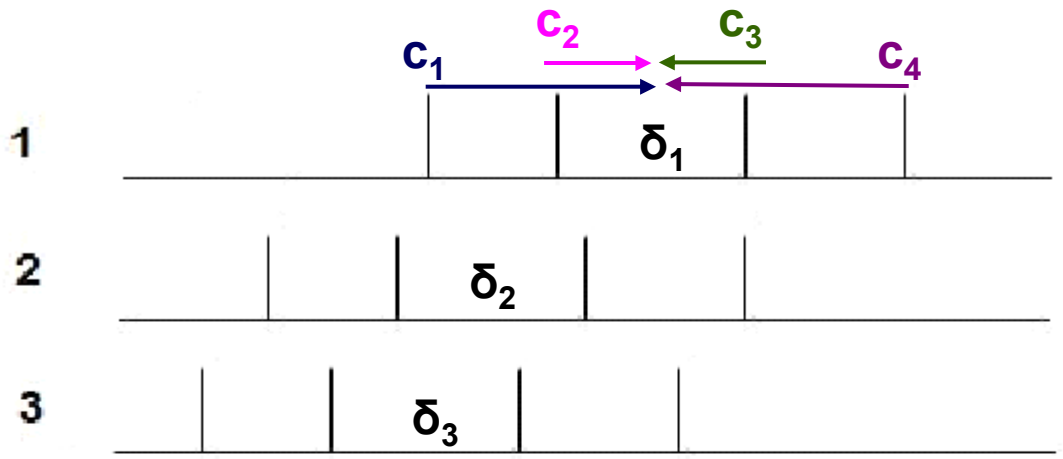


Note that the 3 thresholds **do not cover a wide range** of the latent trait, and what the distribution of theta looks like as a result…

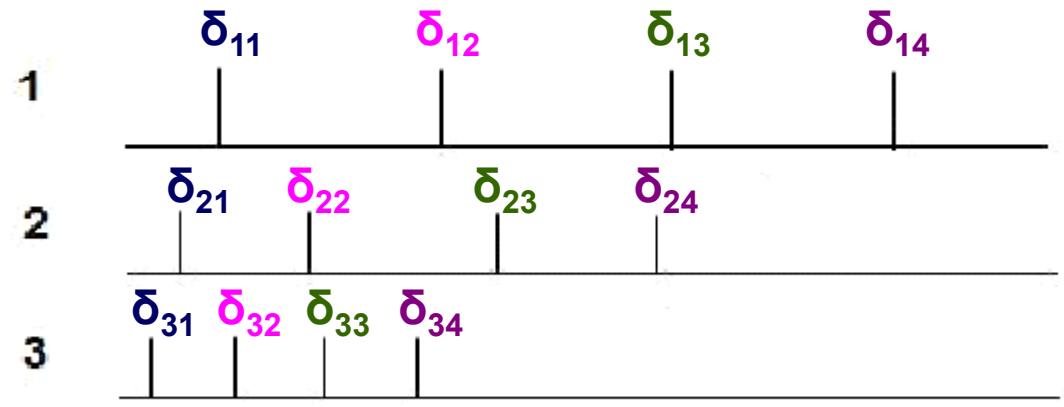


Rating Scale Model

- "Rating Scale" is to PCM what "Modified GRM" is to GRM
- Is more parsimonious version of partial credit model
- Designed for items with same response format
- In PCM, there are (#options-1)*(#items) step parameters estimated (+ one slope per item in generalized PCM version)
- In RSM, each item gets own slope and own 'location' parameter, but the differences between categories around that location are constrained equal across items
  - Items differ in overall location, but spread of categories within is equal
  - So, different δi per item, but same c1, c2, and c3 across items
- If 0 or 1, prob of 1 :: $P_{i1}(\theta_s) = \frac{\exp\left(1.7a_i\left(\theta_s-(\delta_i+c_1)\right)\right)}{1+\exp\left(1.7a_i\left(\theta_s-(\delta_i+c_1)\right)\right)}$
- (and so forth for δ2 and δ3)
  - δi is a 'location' parameter, and c is the step parameter as before
  - Constrains curves to look same across items, just shifted by δi



Rating Scale →
1 Location, k-1 c's
All category distances are same across items



Original PCM →
k-1 locations
All category distances are allowed to differ across items

Item Difficulty / Latent Ability

Summary of Models for Partially Ordered Categorical Responses

- Partial Credit Models test the assumption of ordered categories
  - This can be useful for item screening, but perhaps not for actual analysis
- These models have additional data demands relative to GRM
  - Only data from that threshold get used (i.e., for 1 vs. 2, 0 and 3 don't contribute)
  - So larger sample sizes are needed to identify all model parameters
  - Sometimes categories have to be consolidated to get the model to not blow up

Not directly available in Mplus	**Difficulty Per Item Only (category distances equal)**	**Difficulty Per Category Per Item**
Equal discrimination across items (1-PLish)?	"Rating Scale PCM"	"Partial Credit Model"
Unequal discriminations (2-PLish)?	"Generalized Rating Scale PCM"?? (same response options)	"Generalized PCM" "Adjacent Category Logit"

ADDITIONAL FEATURES OF ORDERED CATEGORICAL MODELS

Expected Scores

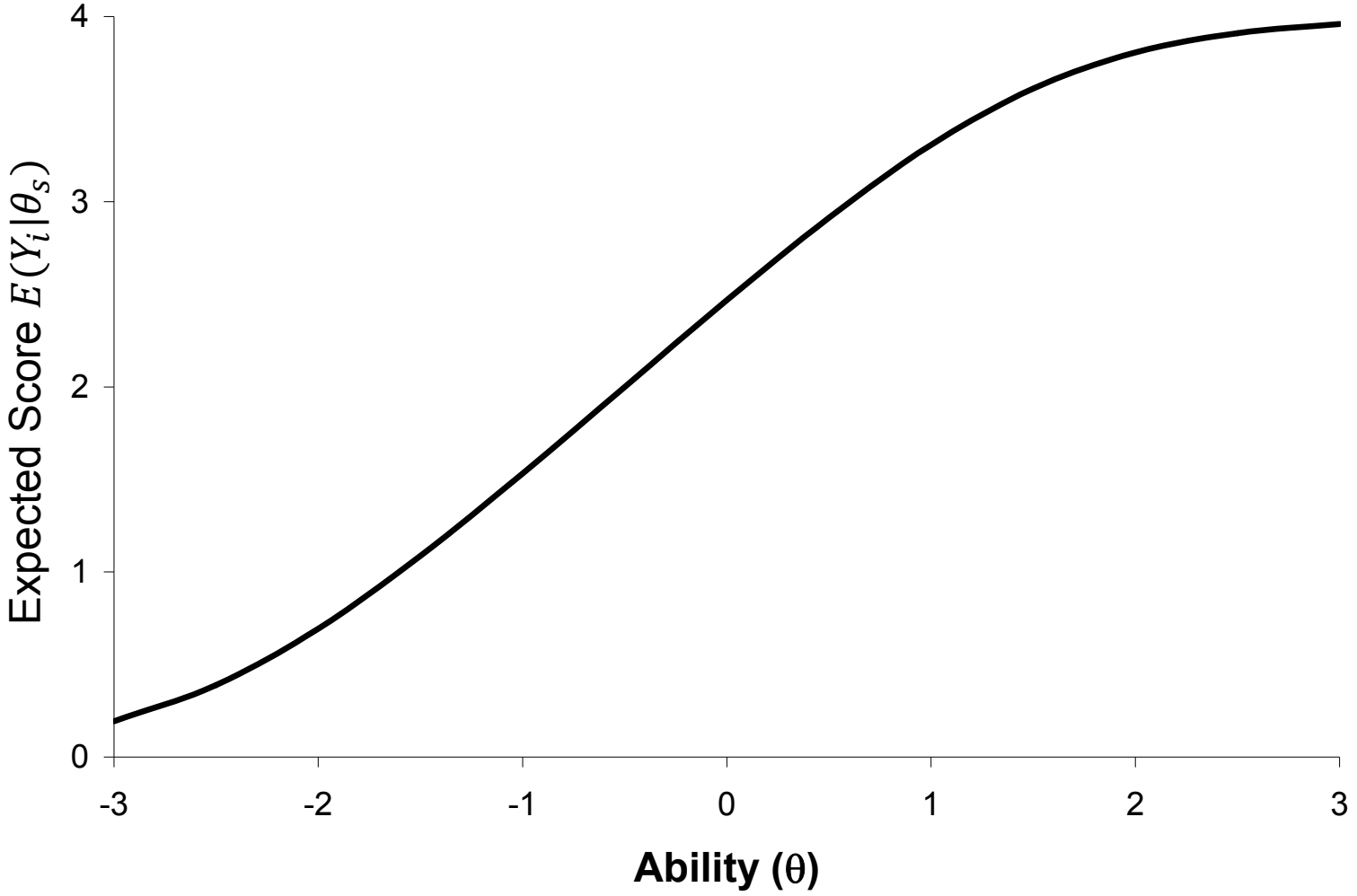
- It is useful to combine the probability information from categories into one function for an expected score:

$$E(Y_i|\theta_s) = \sum_{y=0}^{m_j} yP_{iy}(\theta_s)$$

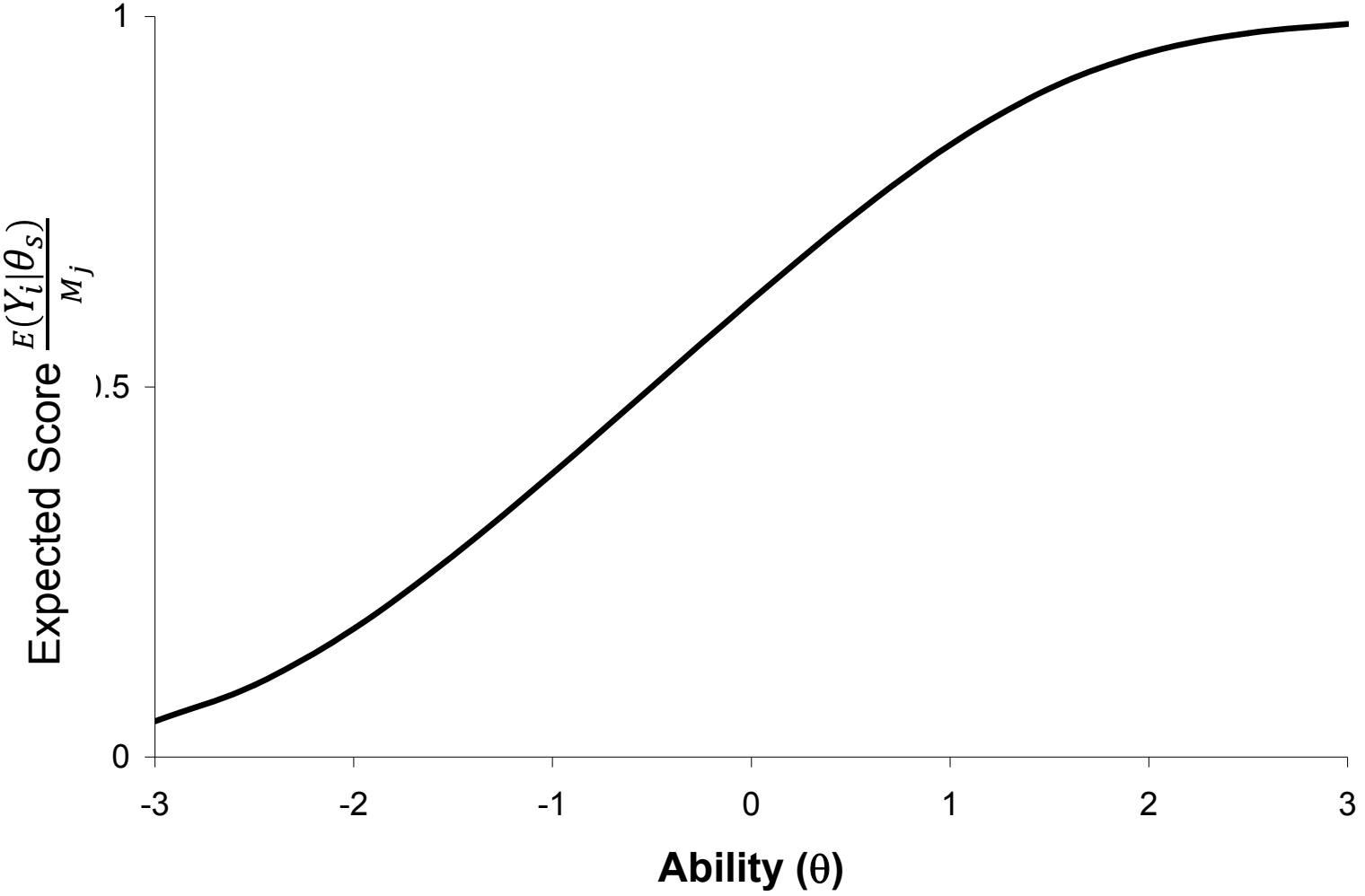
- Multiply each score by its P, add up over categories for any theta level

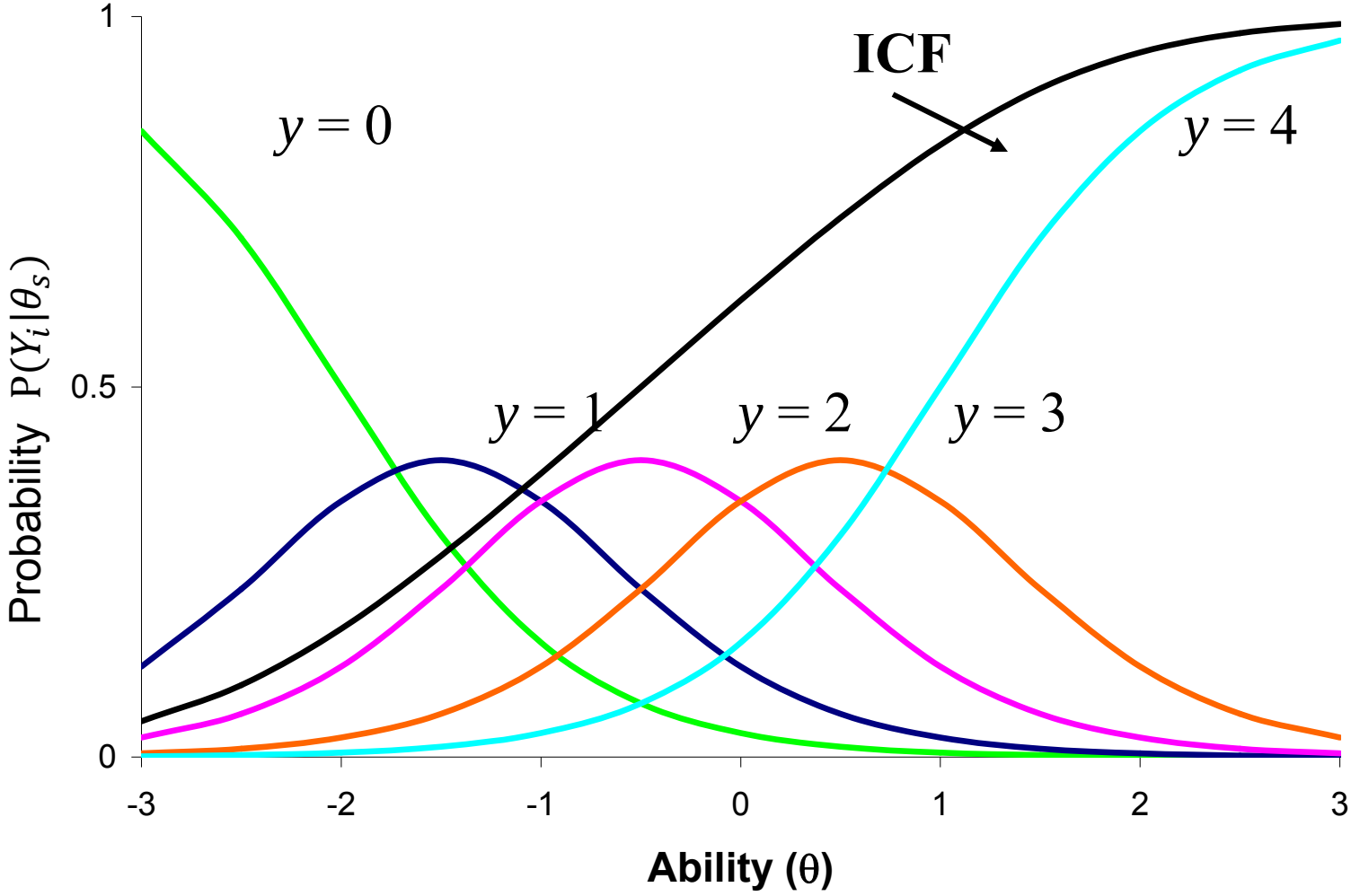
- This expected score function acts as a single Item Characteristic Function (analogous to the ICC for dichotomous/binary items)

Item Characteristic Function



Expected Proportion Correct




ICF
$y = 0$
$y = 1$
$y = 2$
$y = 3$
$y = 4$
Probability $P(Y_i|\theta_s)$
1
0.5
0
-3
-2
-1
0
1
2
3
Ability (θ)

Item/Test Characteristic Function

- ICF is a good summary of an item and is used in test development, DIF studies, model-data fit evaluations

- As before, the TCF is equal to the sum of expected scores over items

$$E(Y_s|\theta_s) = \sum_{j=1}^{n} \sum_{y=0}^{m_j} y P_{iy}(\theta_s)$$

- This could include dichotomous, polytomous, or mixed-format tests

NOMINAL RESPONSE MODELS

Nominal Response Model

- Ideal for items with no ordering of any kind (e.g., dog, cat, bird)
- # response options don't have to be same across items
- Is a "direct" model (probability of response given directly)
- Models the probability of one response category against all others
  - Still like dividing item into a series of binary items, but now each option is really considered as a separate item ("Baseline category logit")
  - 0 vs. 1,2,3 (c_{1i}) 1 vs. 0,2,3 (c_{2i}) 2 vs. 0,1,3 (c_{3i})

$$P(y=1) = \frac{\exp(1.7a_{i1}(\theta_s + c_{i1}))}{\sum_{y=0}^{3} \exp(1.7a_{iy}(\theta_s + c_{iy}))}$$

Estimate one slope (a_i) and one "intercept" (c_i) parameter *per item, per threshold,* such that sum(a's)=0, sum(c's)=0 (so a and c are only relatively meaningful within a single item)

  - Available in Mplus with NOMINAL ARE option
  - Can be useful to examine 'distractors' in multiple choice tests

Example Nominal Response Item

4

A company packs its coffee into cylindrical containers. The height of each container is 6 inches, and the radius of the container is 3 inches.

Which is closest to the volume of one of these cylindrical containers? (Use 3.14 for π.)

A 36 cubic inches
B 54 cubic inches
C 113 cubic inches
D 170 cubic inches

Additional Item Types

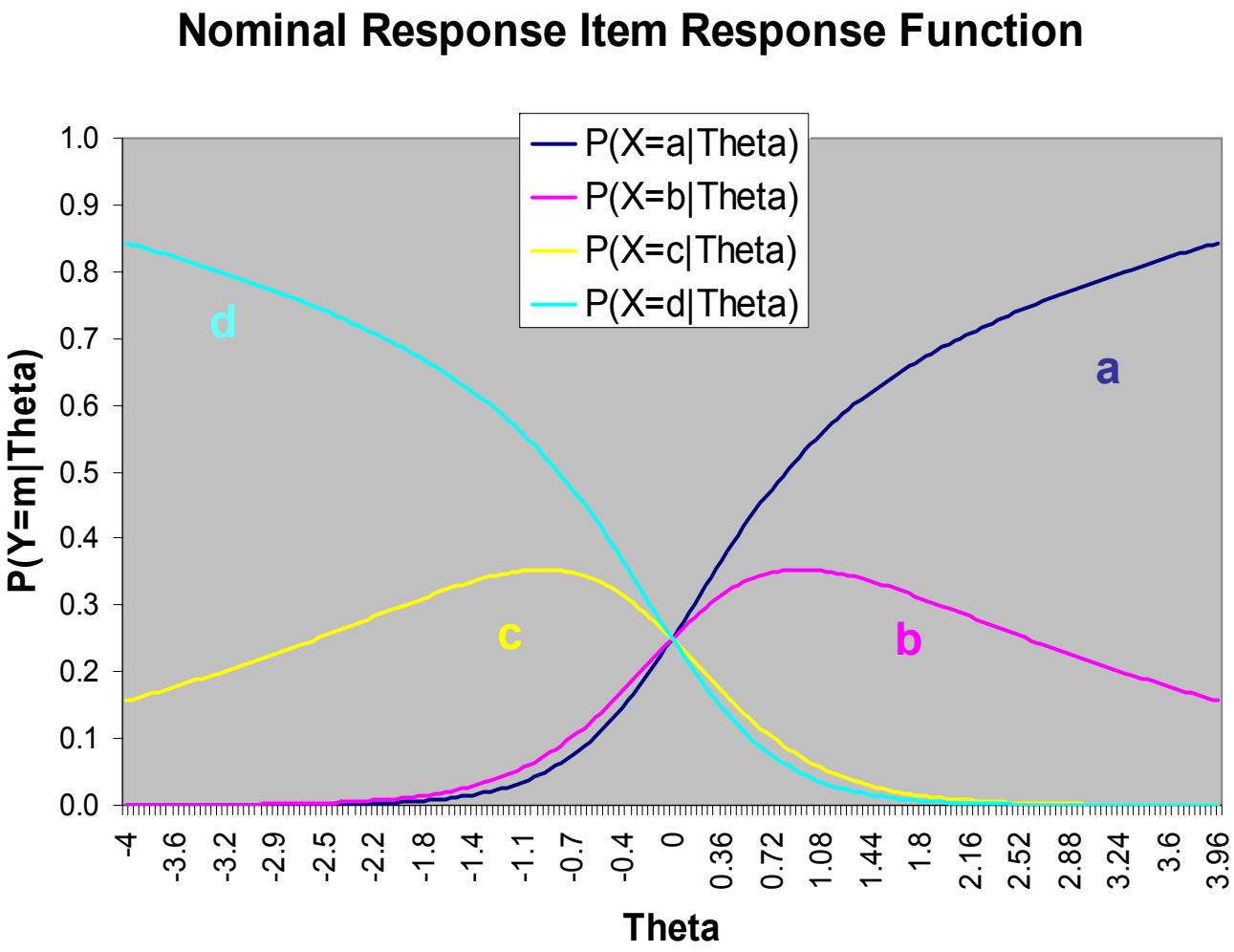
- Non-cognitive tests can also contain differing item types that could be modeled using a Nominal Response Model

- For example, consider an item from a questionnaire about political attitudes…

Which political party would you identify yourself with?

A. Democrat
B. Republican
C. Independent
D. Green
E. Unaffiliated

Category Response Curves (NRM for 5-Category Item)



Example Distractor Analysis:

People low in Theta are most likely to pick **d**, but **c** is their second choice

People high in Theta are most likely to pick **a**, but **b** is their second choice

CONCLUDING REMARKS

Summary: Polytomous Models

- Many kinds of polytomous IRT models…
- Some assume order of response options… (done in Mplus)
  - Graded Response Model Family :: "cumulative logit model"
    - Model cumulative change in categories using all data for each
- Some allow you to test order of response options… (no Mplus)
  - Partial Credit Model Family :: "adjacent category logit model"
    - Model adjacent category thresholds only, so they allow you to see reversals (empirical mis-ordering of your response options with respect to Theta)
    - PCM useful for identifying separability and adequacy of categories
    - Can be done using SAS NLMIXED (although very slowly… see example)
- Some assume no order of response options… (done in Mplus)
  - Nominal Model :: "baseline category logit model"
    - Useful to examine probability of each response option
    - Is very unparsimonious and thus can be hard to estimate

Up Next…

- Estimation of Parameters for IRT Models
  - Estimate person parameters when item parameters are known
  - Joint estimation of person and item parameters