Cumulative Scaling: Nonparametric Item Response Theory

Measurement, Scaling, and Dimensional Analysis
2019 ICPSR Summer Program
Prof. Adam M. Enders

Housekeeping

- How was lab on Friday?
- SRM and Reliability homework due **tomorrow**
- Blalock Lecture: "Teaching Statistics: What Every New (and Not-So-New) Instructor Needs to Know"
  - Tonight @ 7:30 in Angell Hall Auditorium D
  - Tim McDaniel, Regression II
- Blalock Lecture: "Issues of Rigor and the Emergence of Open Science Methods to Deal with Issues Emerging from the Replication Crisis"
  - Tomorrow @ 7:30 in Angell Hall Auditorium D
  - Pamela Davis-Kean, U-M (psychology)

Where Have We Been? Where Are We Going?

- With the summated rating model, we are able to array row objects (e.g., people, countries) along a single, latent continuum
  - Referred to, in Coombs' original data theory, as *subject-centered* scaling
- The summated rating model is a powerful one
  - Reduces dimensionality of dataset, increases level of measurement
  - Results in more reliable measure of the construct we're interested in
  - Can be used to test substantive theories about nature of variables
- Might also want to take into account variable characteristics
  - We might not want to assume parallel measures all the time

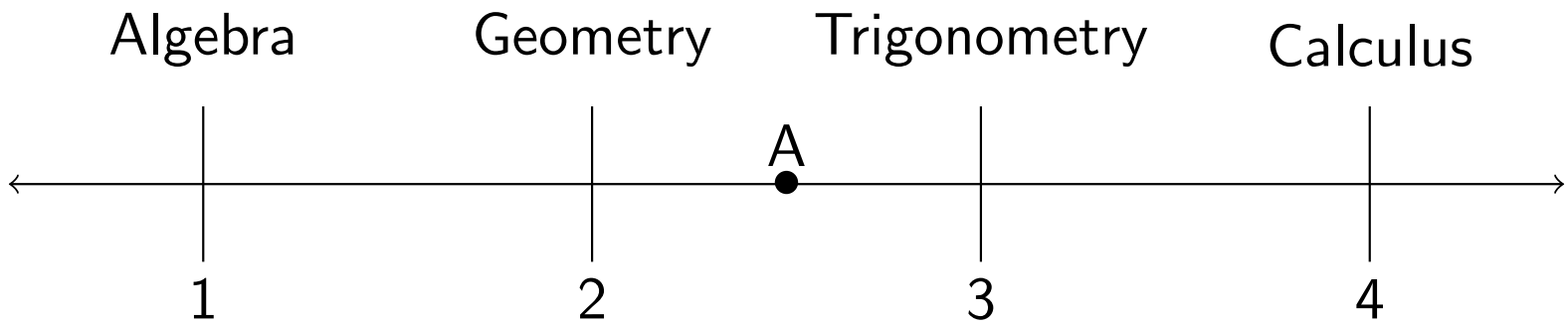
Some Examples

- Height questions, option 1:
  1. Are you *approximately* 5'11" (1.80 m) tall?
  2. Are you *approximately* 5'7" (1.70 m) tall?
- Height questions, option 2:
  1. Are you *at least* 5'11" (1.80 m) tall?
  2. Are you *at least* 5'7" (1.70 m) tall?
- In both cases, we can think of height as being a continuum that we can orient individuals along
- The questions are not, however, parallel measures – they require a different height for affirmative responses
- The former is a **proximity** relationship, the latter a **dominance** relationship (our focus, for now)

What is Cumulative Scaling?

- The cumulative family of scaling techniques makes similar assumptions about item response functions as the SRM
- However, cumulative scales, originally formulated by Guttman, also array column objects along the latent continuum
  - Unlike the SRM, column objects are not assumed to be parallel measures
  - We are taking into account item characteristics
- In some sense, cumulative scales provide more information about the data, but *only if the data conform to the assumptions of the model*
- Caution: this "family" of models is called different things
  - General: cumulative model, dominance model, item response theory (IRT)
  - Specific: nonparametric IRT, ordinal IRT, Mokken scaling, probabilistic Guttman scale

Example: Mathematical Ability



Items are arrayed along the latent dimension, "mathematical ability," according to their difficulty

Subjects are arrayed according to ability – placed in between items they successfully complete, or "dominate," and those they don't

Scale scores used to determine placement of subjects on continuum AND response patterns

Example: Mathematical Ability, cont'd

Consider the following cross tabulation of potential responses to an algebra and trigonometry question, where possible responses are {correct, incorrect}

		Trigonometry	
		Incorrect	Correct
Algebra	Incorrect	a	b
	Correct	c	d

Cells a, d, and c all make substantive sense – one could answer both questions correct, both incorrect, or algebra correct but not trigonometry

Cell b does not make substantive sense – why would someone get a trigonometry question correct, but not a simpler algebra question?

Objectives of Cumulative Scaling

- Exploit nominal level to gain an ordinal level estimate of the latent dimension
- Place row AND column objects along the same latent continuum in such a way that meaningful comparisons can be made between all pairs of row and column objects
- Leverage characteristics of the DGP to use row scale scores to be able to predict specific cell values
- NOTE: like with the SRM, cumulative scaling models are not good for testing dimensionality
  - We use these models when we can safely assume, or have empirical evidence that, a single substantive source of variance exists within the data
  - If you ask for a single dimension, you'll get one – doesn't mean you've captured anything useful or that accurately characterizes the attributes of the data

Guttman (Deterministic) Scaling

- For each subject, the response pattern to a set of k items can be represented as a row vector: $x = \{x_{i1}, x_{i2}..., x_{ik}\}$, where $i = 1, 0$
- 2^k response patterns
- Each subject falls somewhere along the latent variable, θ
- We want to use the response pattern constructed from each subject's responses to the k items to infer about their true position on the latent variable, θ
- We can conceptualize the position of a given item, x_j, along the latent dimension, θ, as the "item difficulty," δ_j

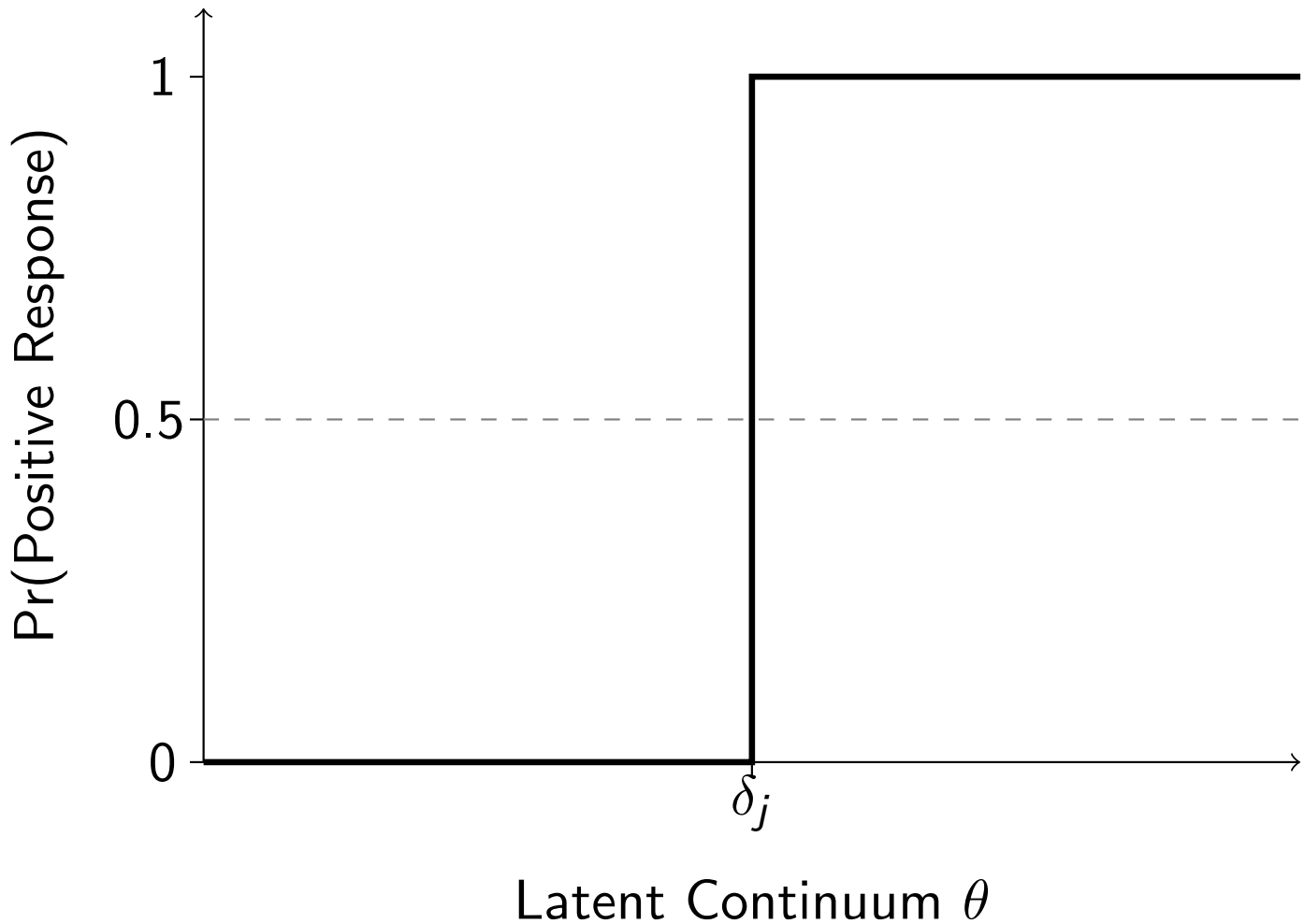
Guttman (Deterministic) Scaling, cont'd

- In the deterministic model, a given subject's response to a given item x_{ij} is completely determined (that is with a probability of 1 or 0) by the position of that item along the latent dimension, δ_j, and their own position, θ_i
- If the subject is "below" δ_j – that is, posseses less of the latent trait than item x_{ij} measures – they will provide a "negative" response to that item (or, receive a 0)
- If they are above δ_j, they will provide a positive response (or, receive a 1)
- In mathematical notation:

$$\begin{aligned} P\{x_{ij} = 1|\theta_i, \delta_j\} &= 0 \text{ if } \theta_i < \delta_j \\ &= 1 \text{ if } \theta_i \geq \delta_j \end{aligned}$$

- Only $k + 1$ response patterns exist with a perfect scale of all items, where 2^k response patterns are actually possible.

Deterministic Item Response Function



Error-Free Response Pattern

Initial matrix of data that forms perfect (deterministic) cumulative scale

	Stimuli					
Subjects	1	2	3	4	5	Scale Score
A	0	1	1	1	1	4
B	0	0	1	0	1	2
C	0	0	0	0	1	1
D	1	1	1	1	1	5
E	0	0	0	0	0	0
F	0	1	1	0	1	3
	0.17	0.50	0.67	0.33	0.83	

Error-Free Response Pattern, cont'd

Rearrange columns by probability of correct response, or "easiness"

	Stimuli					
Subjects	5	3	2	4	1	Scale Score
A	1	1	1	1	0	4
B	1	1	0	0	0	2
C	1	0	0	0	0	1
D	1	1	1	1	1	5
E	0	0	0	0	0	0
F	1	1	1	0	0	3
	0.83	0.67	0.50	0.33	0.17	

Error-Free Response Pattern, cont'd

Further rearrange data by rows in descending order of scale score

	Stimuli					
Subjects	5	3	2	4	1	Scale Score
D	1	1	1	1	1	5
A	1	1	1	1	0	4
F	1	1	1	0	0	3
B	1	1	0	0	0	2
C	1	0	0	0	0	1
E	0	0	0	0	0	0
	0.83	0.67	0.50	0.33	0.17	

Scale scores perfectly predict which stimuli the subject provided a positive/correct response to

A scale score of 2 means the subject responded positively/correctly to stimuli 5 and 3

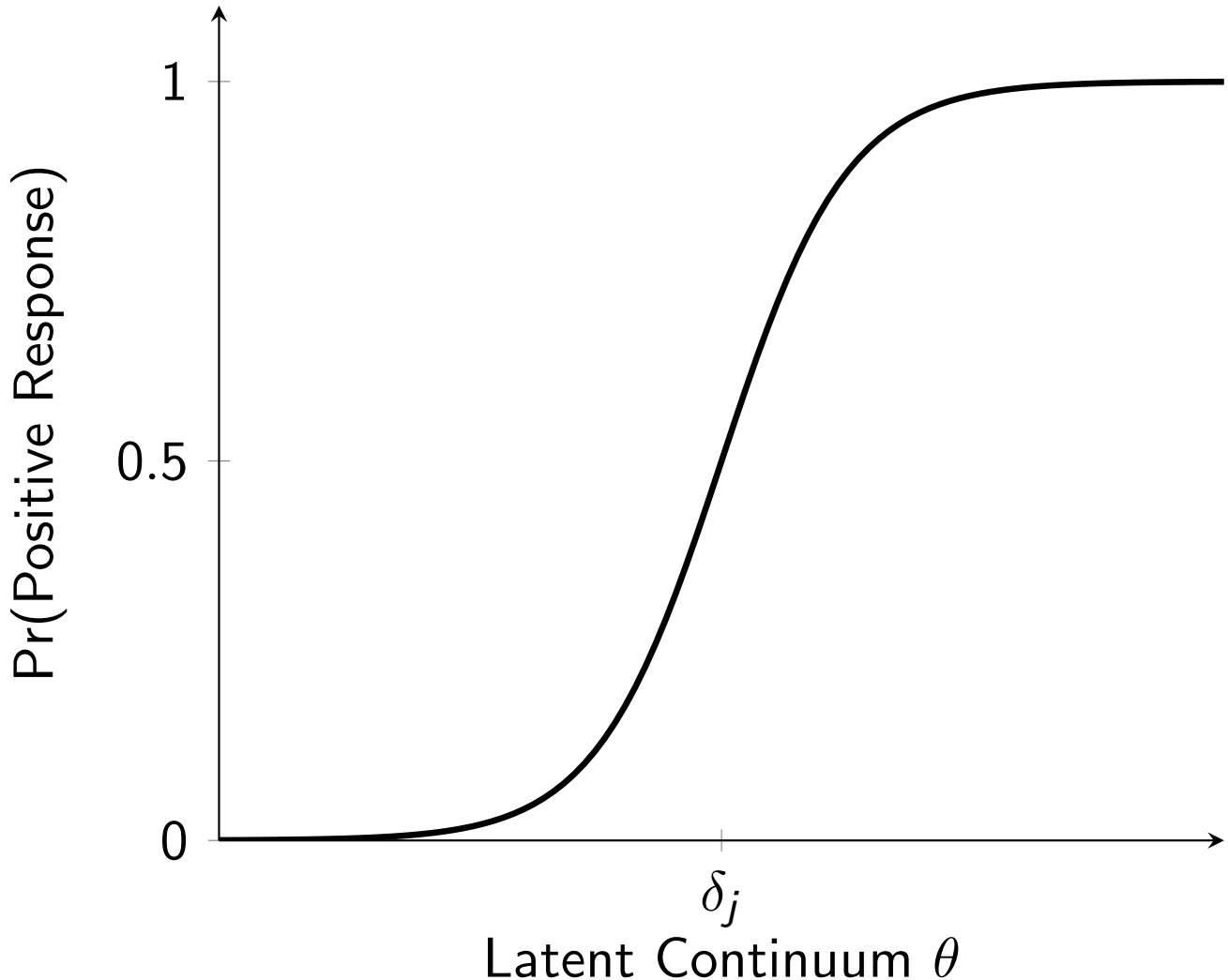
Limitations

- Deterministic cumulative scale is obviously extremely powerful
  - Like SRM, distills multivariate data to a single dimension, and moves up from nominal or ordinal data to interval level
  - Also provides information about stimuli, or column objects, and fuller information about contents of data matrix
- Problem: highly unlikely to find data that meet the strict assumptions of the deterministic model
  - 2^k possible response patterns, but only $k + 1$ form a perfect scale
  - For example, if $k = 8$: 2^8=256 possible response patterns, $8 + 1 = 9$ of which are acceptable in forming a perfect scale
- What do we do with errors – response patterns that don't make sense?
  - Discarding them will reduce the dataset and, therefore, the force of our substantive conclusion
  - We could try to score them, but their scores wouldn't mean the same thing as subjects with perfect response patterns

The Answer: Probabilistic Models

- The deterministic model assumes that if $\theta_i \geq \delta_j$ then $x_{ij} = 1$, and that if $\theta_i < \delta_j$ then $x_{ij} = 0$
- A probabilistic model would be less restrictive and allow for some error
  - On the one hand, if there are a lot of errors, then we might just conclude that the cumulative model is not the right one for the data
  - In other words, the DGP doesn't follow a cumulative pattern – subjects aren't conceptualizing the stimuli in the same way
  - On the other hand, if errors are fairly minimal (to be defined below!), we might simply conclude that errors are random or confined to a very small subset of subjects

Probabilistic Item Response Function



Nonparametric IRT

- The first model we will consider is nonparametric because we assume only monotonicity of item response functions
  - Less restrictive than parametric models (i.e., the Rasch model), but provides almost identical information
  - Particularly useful for survey data, and other types of "messy" (i.e., error-laden) data for which parametric assumptions are unlikely to hold
- Model originally developed by R.J. Mokken (1971) and frequently called "Mokken scaling"
- Simply the probabilisitic version of Guttman's deterministic cumulative scaling model
- Provides a set of tests to determine:
  1. If a cumulative pattern generally characterizes the dataset
  2. If items can be accurately arrayed along with subjects

Model Assumptions

1. Unidimensionality
   - Row and column objects can be arrayed along a single latent continuum, θ
   - This is testable – something we will learn next week
   - For now, let's just make the assumption where it seems "reasonable"
2. Monotonically nondecreasing IRFs
   - The probability of an affirmative response should never decrease for row objects with increasing values along the latent dimension, θ
3. Local stochastic independence
   - The probability of an affirmative response is contingent on θ alone, no other systematic influence
   - Difficult to test – usually just hope this is met...

Mokken Scaling

- There are two "submodels" of the nonparametric IRT model:
  1. The "Monotone Homogeneity" model
  2. The "Double Monotonicity" model
- Before considering the assumptions of these models, and tests of those assumptions, Mokken recommends checking for basic (cumulative) scalability
- Employs H "coefficient of homogeneity," first proposed by Loevinger (1948)
- $H = 1 - \frac{\#\text{observed errors}}{\#\text{expected errors}}$
- Observed and expected errors refer to the observed (response) patterns in the data
- H is bound between (0, 1), such that larger values represent better model fit

Example: Beliefs About Heaven and Hell

		Heaven		
		Yes	No	Total
Hell	Yes	24	6	30
	No	36	34	70
	Total	60	40	100

Steps for finding number of observed errors:

1. Make crosstab where more difficult item (lower proportion of "positive" responses) is the row variable
2. Number of observed errors = cell value in upper right top cell (hell yes, heaven no)

In a perfect scale, this value would be 0

Example: Beliefs About Heaven and Hell

		Heaven		
		Yes	No	Total
Hell	Yes	24	6	30
	No	36	34	70
	Total	60	40	100

The number of errors expected under statistical independence is the product of the probability of a positive response to the more difficult item and the negative response to an easier item, multiplied by n

Using data in the above table: $0.30 \times 0.40 = 0.12$; since $n = 100$, $0.12 \times 100 = 12$

We expect 12 errors under statistical independence (i.e., if responses to the items were truly random)

Calculating the H Coefficient

		Heaven		
		Yes	No	Total
Hell	Yes	24	6	30
	No	36	34	70
	Total	60	40	100

Observed errors $= 6$

Expected errors $= 12$

Therefore, $H = 1 - \frac{6}{12} = 0.50$

H Coefficients for Larger Scales and Items

- For scales with more than two items, we first construct 2×2 cross tabulations like above for each item pair
  - If there are 4 items, there are 6 pairs (tables)
- Then we sum the observed errors and expected errors across all pairs/tables to calculate H
- Can also calculate H_i for individual items
  - Sum the observed and expected errors over all pairs (tables) that include the item in question
  - Items with low H_i's are more likely to violate the assumptions we'll discuss below, even though they are weakly scalable by the scale H criterion at this stage

Interpreting the H Coefficient

- In scales with more than two items, we construct 2×2
- Again, it is bound between (0, 1)
- But, Is $H = 0.50$ good or bad?
- No simple answer
- Rules of thumb that most live by:
  - $H < 30$ is unacceptable – data not homogenous to form a cumulative scale
  - $H \geq 0.50$ is excellent
  - Everything in between: proceed to checking model assumptions and see what you've got
- Note that software also conducts a hypothesis test that $H = 0$, since H could be ≥ 0.30 in a small sample

Monotone Homogeneity

- Three fundamental assumptions of Monotone Homogeneity model:
  1. There is a unidimensional continuum/trait, θ, on which subjects and column objects (items) can be located where relationships between pairs of subjects/items are "meaningful"
     - For example: if the scale values of a subject and item are identical, the probability of a positive response $p(x = 1|\theta = \delta_i) = 0.50$, and so on
  2. IRFs are monotonically nondecreasing
     - The probability of a positive response to an item i increases (or, at least does not decrease) with increasing subject value θ_i
  3. Responses by the same subject are locally stochastically independent
     - Responses to two or more items by the same subject are influenced only by the latent trait, θ

Tests of Monotone Homogeneity

- Visual inspection via an item analysis, just like with the SRM
  - Plot item responses against rest scores and look for "serious" violations of monotonicity
  - Less formal than a statistical test, but gives pretty much the same info
- Examination of proportions of positive responses by item step (rest score group) to check
  - If there is a decrease in proportion from a "lower" rest score group to a "higher" one:
    - Ignore if difference in proportions is 0.03 or less than
    - Test of statistical independence of cross tab of rest score groups in question (rows) and item responses (columns)

Testing Monotone Homogeneity

Group #	*Rest score value(s)*	*N*	*Frequencies per item value*	*Mean*	*Proportions of positive responses per item*
1	0	240	222	18	0.07
2	1	166	137	29	0.17
3	2	115	78	37	0.32
4	3–4	181	94	87	0.48
5	5	140	46	94	0.67
6	6	112	21	91	0.81
7	7	178	3	175	0.98
8	8	107	7	100	0.93

Note: usually combine adjacent rest score groups if subjects within group account for more than 2% of group

Testing Monotone Homogeneity

Examine crosstab of rest score groups 7 and 8, since the probability of a positive response dropped from group 7 to 8

Table 4 Rest score groups 7 and 8 by item scores

	Item: score 0	*Item: score 1*	*Total*
Rest score 7	3	175	178
Rest score 8	7	100	107
Total	10	275	285

$$z = 2 \times \frac{\sqrt{(f_{11}+1)\times(f_{00}+1)} - \sqrt{(f_{01}\times f_{10})}}{\sqrt{(n+1)}}$$

$$z = 2 \times \frac{\sqrt{(100+1)\times(3+1)} - \sqrt{(175\times 7)}}{\sqrt{(285+1)}}$$

$z = 1.762 \longrightarrow$ significant one-sided test at $\alpha = 0.05$

Example: Political Participation

- Data are dummy-scored (0=no, 1=yes) questions about political participation
- Asked after presidential elections: "During the past year, did you participate in any of the following activities?"
  - `Talk`: Discuss the election with friends, family, or coworkers
  - `Sign`: Display a yard sign, bumper sticker, or wear apparel supporting a particular candidate/party
  - `Rally`: Attend a political rally
  - `Donate`: Donate to a political party or candidates
  - `Work`: Volunteer for a political party or candidate

Example: Political Participation, cont'd

Proportion of individuals who gave the positive response to the deleted item (columns) in each restscore group:

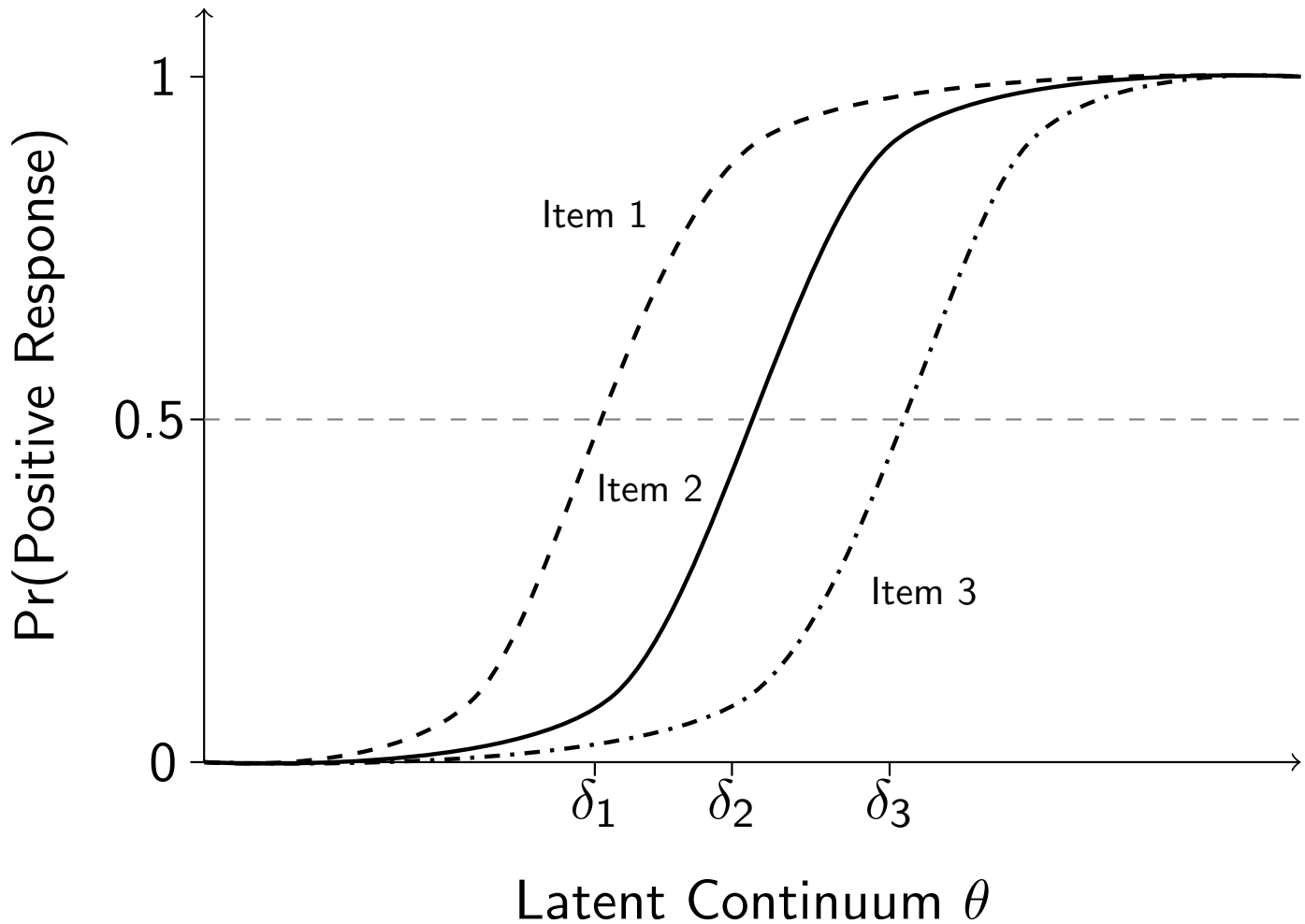
Group	Restscore	Work	Donate	Rally	Sign	Talk
1	0	0.00	0.00	0.00	0.39	0.43
2	1	0.05	0.26	0.19	0.74	0.76
3	2	0.08	0.51	0.42	0.88	0.88
4	3	0.59	0.76	0.82	0.97	0.98
5	4	0.86	0.85	1.00	1.00	1.00

We observe no violations of Monotone Homogeneity

Double Monotonicity

- Satisfying the assumptions of the Monotone Homogeneity model are **not** sufficient for establishing a unique rank ordering of items along the latent continuum
- Patterns in data must satisfy one more assumption:
  1. IRFs for individual items must not (significantly) cross each other
- If IRFs for items i and j cross, we can't tell if $\delta_i > \delta_j$, or vice versa
- This is the assumption that most sets the cumulative model apart from the summated rating model
- BUT, it's also somewhat more difficult to meet

Double Monotonicity Assumption



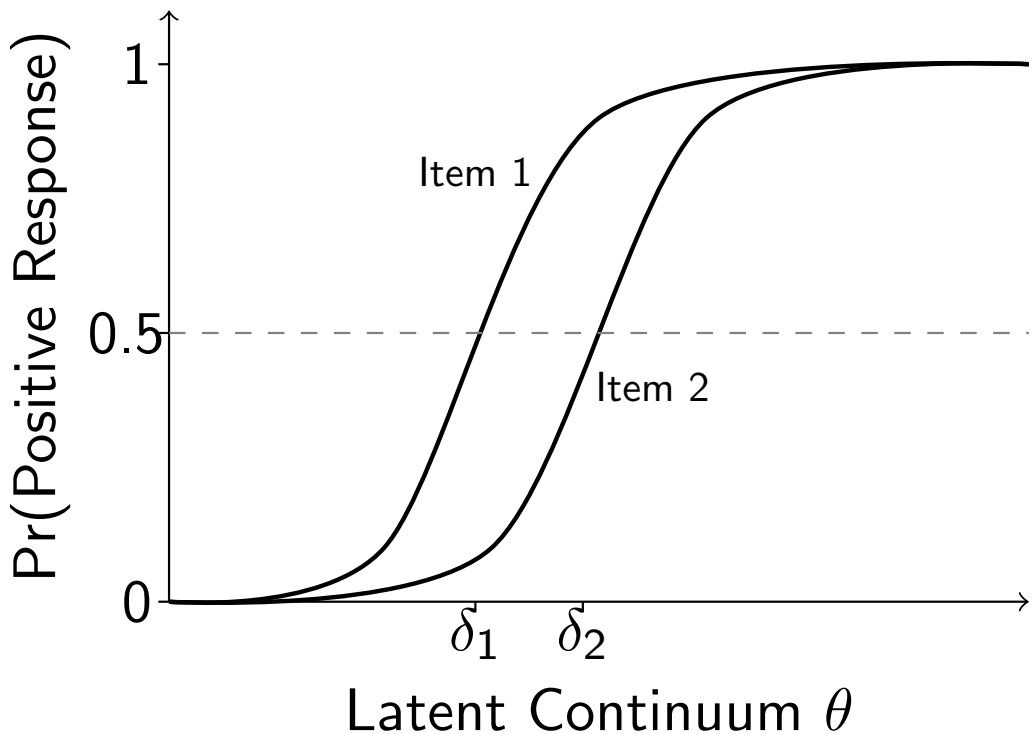
Tests of Double Monotonicity

- At this point, there are four major tests that have been developed to statistically test whether item IRFs significantly cross:
  1. Examination of the "P(+,+)" and "P(-,-)" matrices
  2. The "restscore" method
  3. The "restsplit" method
  4. The "manifest invariant item ordering (MIIO)" method
- Unfortunately, they test the invariance of item ordering in slightly different ways, and can turn up slightly (but not usually radically) different results
- We'll focus on the restscore and P matrix methods, though we'll look at results from other methods in an example
- Much like the basic nondecreasing monotonicity assumption of the MH model, we can also use visual inspections of empirical IRFs to guide us

Testing Double Monotonicity, cont'd

All tests are designed to determine whether the IRFs cross, and whether the overlap/crossing can be chalked up to sampling variability

Software even shows this graphically, plotting pairs of estimated IRFs with confidence bands



Example: Political Participation, cont'd

Number of respondents in restscore groups if two items (columns) are deleted:

Restscore	Work Donate	Work Rally	Work Sign	Work Talk	Donate Rally	Donate Sign	Donate Talk	Rally Sign	Rally Talk	Sign Talk
0	28	28	46	51	28	58	51	47	57	113
1	58	53	85	77	70	123	142	125	110	111
2	120	239	232	237	256	171	158	309	313	254
3	994	880	837	835	846	848	849	719	720	722

Since we're now looking at pairs, we only have 5 rest score groups, instead of 4 ($k - 1$)

Example: Political Participation, cont'd

Look at crosstabs for restscore groups against potential response patterns (00, 01, 10, 11) for pairs of items

Table for Donate (first value) by Rally (second value) pictured below:

Group	Score	n	00	01	10	11	p(Donate)	p(Rally)	Diff	z
1	0	28	28	0	0	0	0.00	0.00		
2	1	70	39	9	14	8	0.31	0.24	0.07	0.83
3	2	256	48	37	52	119	0.67	0.61	0.06	1.49
4	3	846	1	129	0	716	0.85	1.00		
Total		1200	116	175	66	843	0.76	0.85	0.13	

Since Rally is easier than Donate, we shouldn't observe higher numbers of individuals in cells with pattern (10) than in cells for (01)

Here, we observe two violations, neither of which is statistically significant (column z – test statistic for difference in column "Diff")

Example: Political Participation, cont'd

Information for the P(+,+) and P(-,-) matrices

Lower triangle: frequencies in (1,1) cell; Upper triangle: frequencies in (0,0) cell

	Work	Donate	Rally	Sign	Talk
Work		157	177	64	61
Donate	724		116	60	48
Rally	853	843		54	59
Sign	851	898	1,001		29
Talk	852	890	1,010	1,091	

Since total $n = 1,200$, we divide each entry in the lower triangle by 1,200 to construct the P(+,+) matrix, and do the same for the upper triangle to form the P(-,-) matrix

Example: Political Participation, cont'd

P(+,+) matrix

Cell values should increase as we move down columns

Why? Because the joint probability of providing the affirmative response to two easier item pairs should increase!

		Work	Donate	Rally	Sign	Talk
	p	0.71	0.76	0.85	0.94	0.94
Work	0.71					
Donate	0.76	0.60				
Rally	0.85	0.71	0.70			
Sign	0.94	0.71	0.75	0.83		
Talk	0.94	0.71	0.74	0.84	0.91	

We observed one (seemingly small) violation moving from cell (6,4) to (7,4)

Example: Political Participation, cont'd

P(-,-) matrix

Cell values should decrease as we move down columns

Why? Because the joint probability of providing the negative response to two easier item pairs should decrease!

		Work	Donate	Rally	Sign	Talk
	$1-p$	0.28	0.24	0.15	0.06	0.06
Work	0.28					
Donate	0.24	0.13				
Rally	0.15	0.15	0.10			
Sign	0.06	0.05	0.05	0.04		
Talk	0.06	0.05	0.04	0.05	0.02	

We observed two (seemingly small) violations in moving from cell (4,2) to (5,3), and moving from (6,5) to (7,5)

Sources of Error

- We tend to blame items for error
- That is, when a model assumption is violated or simply doesn't fit, we most frequently begin discarding items in a backward fashion
- In cumulative scaling, where items and subjects are located along a latent dimension, either could be the culprit
- Some items might not fit with the others, or some subjects just might be "deviant" in some way (think about the items in different ways than others)
- Possible to transpose the data matrix and calculate H values for individuals
- Not really recommended, though
  - Dropping problematic subjects will likely be interpreted as cherry-picking cases
  - Also poses serious problem for generalizability/statistical inference
  - Can't calculate H when columns have 0 variance

Dealing with Error

- When is error "sufficient" enough to discard items?
- Tradeoff: strict adherence to model assumptions vs. discarding valuable data
- Never an easy answer...
  - Consider theoretical expectations
  - Consider potential substantive explanations for error; that is, errors can sometimes be useful and interesting!
  - Consider how difficult it might be to convey technical aspects to reviewers (e.g., will Reviewer 2 accept that there are slight violations?)
  - Consider magnitude of violation, especially with respect to other model violations or variable characteristics
- Ultimately, we want to be good Popperian scholars
  - Accept that we're only investigating temporary, contingent "truths"
  - Falsifiability is key – keep trying to prove yourself wrong, and learn about the data along the way

Dealing with Error: Row Objects

- May be useful to examine Guttman errors
  - These are just the number of violations of the deterministic Guttman "scalogram" pattern implied by the item ordering
- Generally speaking, good-fitting scales have lots of row objects with errors of 0 (maybe 1, 2, 3)
- Might look into the particular observations that have large errors – could be substantively interesting information
  - Maybe large error is evidence for an "outlier," like in regression analysis?
  - Maybe large errors correspond to individuals who have lots of missing data and/or DK responses (i.e., they weren't paying attention, didn't care)?
  - Maybe the row objects that have large error form their own subscale?

Reliability

- Also useful to check reliability of the resultant scale (especially since reviewers will want this)
- Can use Cronbach's alpha, just as most people do
- BUT, α is even worse for scales of dichotomously-scored variables
  - Assumes linearity, using correlations, rather than weaker assumption of monotonicity
  - Pearson product-moment correlation biased downward for dichotomous items
- Molenaar and Sijtsma (1987) developed more appropriate estimator of reliability
  - Uses probabilities, rather than correlations, like Cronbach's alpha
  - Interpretation is the same as α, reliability still increases with more scale items

Steps in Conducting Mokken Scale Analysis

1. Most analyses start with "bottom up" (or, semi-exploratory) automated item selection procedure based on H coefficients (more in later slides)
2. Once a subset of scalable items has been identified, move on to checking model assumptions:
   - First, check monotonicity assumptions of Monotone Homogeneity model
   - Then, check stricter assumptions of Double Monotonicity model
3. If some items do not meet these assumptions, consider removal *depending on scope of problem and what you're trying to do*
4. Estimate reliability of scale of remaining items
5. Interpret scale and use it to test your theory!

Software

- `mokken` package in R
  - Completes statistical and graphical tests of both models
  - Also employs "automated item selection procedure" to help determine which items belong in the scale
- "msp" user-written command for Stata
  - Will compute scale, item, and item pair H coefficients, and perform automated item selection procedure
  - Does not conduct statistical or graphical tests of assumptions of Monotone Homogeneity or Double Monotonicity model

Automated Item Selection Procedure

- Allows for an informed exploratory analysis of data that doesn't require strict theory about which items belong and their difficulty ordering along the latent dimension
  1. Starts with scale comprised of item pair with largest H coefficient
     - If no item pair forms a scale with an $H \geq 0.30$, the scale analysis is terminated and items are deemed unscalable
  2. Then, searches for next "best" item to add to the scale (i.e., item that reduces scale's H coefficient the least)
  3. Proceeds until next "best" item reduces scale H coefficient to pre-specified lower boundary (default in software, and norm in community, is 0.30), when procedure is terminated
- Even though the procedure is "bottom up," items could be later excluded due to model violations

What to Do with Model Violations

- What should you do when the MH or DM model are violated?
- What should you do when model checking procedures turn up different results (i.e., violations under one test, but not another)?
- My advice:
  1. First, assess the extent of the violation. This can be done with "crit" (40>) or visually.
  2. If the largest or only significant model violations doesn't look too problematic (e.g., barely intersecting IRFs, one out of hundred(s) possible violations), don't worry about it
  3. If there are several items that see problematic, start by taking out the worst offender and re-checking everything with new scale
  4. Proceed like this until no (large, significant) violations are present, or you're otherwise satisfied
- Use theory to guide expectations, be transparent in your written assessment, utilize the web appendix to iron out details of decision-making

Isn't This Just a "Count"?

- If I add up the number of questions answered correctly, number of campaign activities participated in, etc., why bother with the cumulative model?
- Two problems with conceptualizing (at least some constructs) as counting processes
  1. Counts are inherently interval-level, and we may not feel comfortable making that assumption about social data
  2. Conceptualizing social variables as a count also requires us to assume that the column objects are equally difficult – they are perfect parallel measures that can simply be stacked
     - Under the cumulative scaling model, items of equal difficulty would be located at the same point on the latent dimension
     - We know what a score of δ_j would correspond to, but everything above and below would be a guess (almost treat it like error)
- Even though we sum across columns to estimate the latent dimension, we're doing much more than merely counting...

Data: Political Knowledge

- 1/0, yes/no, political knowledge questions from 2012 ANES
  1. `preztimes`: "Do you happen to know how many times an individual can be elected President of the United States under current laws?"
  2. `deficit`: "Is the U.S. federal budget deficit – the amount by which the government's spending exceeds the amount of money it collects – now bigger, about the same, or smaller than it was during most of the 1990s?
  3. `senterm`: "For how many years is a United States Senator elected – that is, how many years are there in one full term of office for a U.S. Senator?"
  4. `medicare`: "What is Medicare?" (4 response options)
  5. `leastspend`: "On which of the following does the U.S. federal government currently spend the least?"
  6. `speaker`: "John Boehner. What job or political office does he NOW hold?"
  7. `viceprez`: "Joe Biden. What job or political office does he NOW hold?"
  8. `primemin`: "David Cameron. What job or political office does he NOW hold?"

Applications of Mokken Scaling

- Cingranelli and Richards. 1999. "Measuring the Level, Pattern, and Sequence of Government Respect for Physical Integrity Rights." *International Studies Quarterly*
  - Create measure of government respect for a subset of human rights known as physical integrity rights; patterns in types of human rights violations (columns) help classify/cluster governments/states
  - Don't really test DM assumptions (at least not reported)...
- Jacoby. 1995. "The Structure of Ideological Thinking in the American Electorate." *American Journal of Political Science*
  - Ideological thinking is a cumulative latent trait whereby some indicators of ideological sophistication are "easier" than others
- Mondak and Anderson. 2004. "The Knowledge Gap: A Reexamination of Gender-Based Differences in Political Knowledge." *Journal of Politics*
  - Examine differences in H coefficients between men and women, and for open-ended and multiple choice question types

Cingranelli and Richards (1999)

TABLE 3. Physical Integrity Scale Scores and Mokken Scale Predictions of Patterns of Government Respect for Particular Physical Integrity Rights: The Pattern of Respect

	Government Respect for Physical Integrity Rights			
Scale Score	*Disappearances*	*Killing*	*Imprisonment*	*Torture*
0	None	None	None	None
1	Partial	None	None	None
2	Partial	Partial	None	None
3	Full	Partial	None	None
4	Full	Partial	None	Partial
5	Full	Partial	Partial	Partial
6	Full	Full	Partial	Partial
7	Full	Full	Full	Partial
8	Full	Full	Full	Full

Generalization for Polytomous Items

- The nonparametric model easily generalizes from dichotomous variables to polytomous (multiple category) variables
- Instead of items having single item response functions, there are $m - 1$ item *step* response functions, where m corresponds to the number of distinct values a variable can take on
  - This is really only useful for ordinal variables, which presumably take on a fairly small number of categories
  - (Parametric) IRT models with interval and ratio level data are basically unidimensional factor analysis models...we'll see this later in the course
  - In other words, IRT models are special cases of the general factor analysis model where a certain dimensionality and link function are assumed (and, re-parameterization, usually)

Generalization for Polytomous Items, cont'd

Table 6 Hypothetical data set of six variables with four categories each [which, after recoding (0,1 = 0, 2,3 = 1) is identical to the data set of Table 1]

Response type	*V1*	*V2*	*V3*	*V4*	*V5*	*V6*	*Frequency of occurrence of response pattern*
1	0	0	0	0	0	0	25
2	0	0	0	0	0	1	25
3	0	0	0	0	0	2	1
4	0	0	0	0	1	2	2
5	0	0	0	1	2	2	1
6	0	0	0	1	2	3	1
7	0	0	1	2	2	3	17
8	0	0	1	2	3	3	18
9	0	1	2	2	3	3	2
10	0	1	2	3	3	3	3
11	0	2	2	3	3	3	1
12	1	2	3	3	3	3	2
13	2	2	3	3	3	3	1
14	3	3	3	3	3	3	1

Generalization for Polytomous Items, cont'd

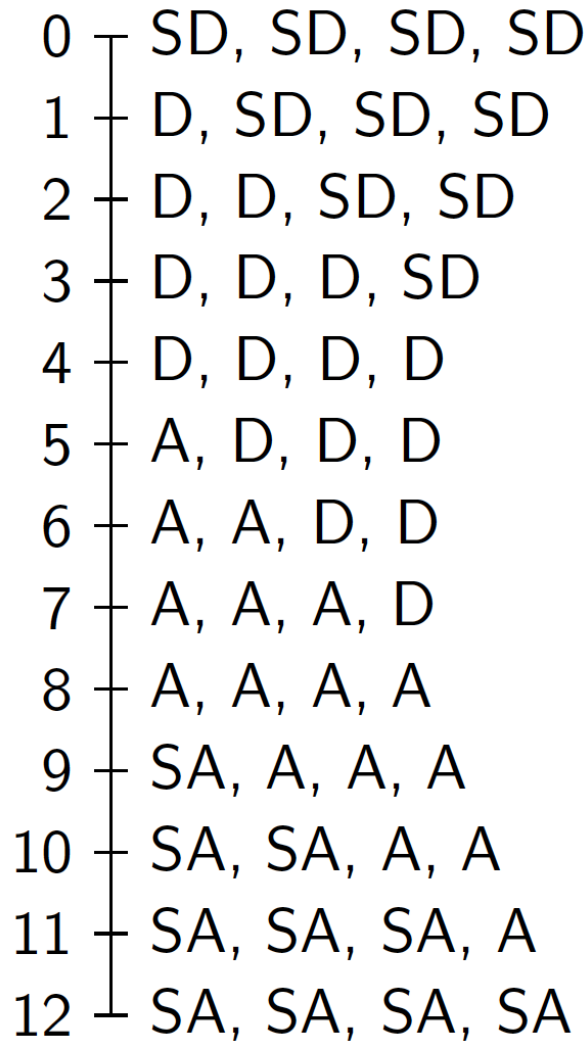
- General procedure:
  1. Generate $k(m-1)$ new "dichotomized" versions of the original k ordinal variables with m categories
  2. Conduct AISP as described above using newly dichotomized variables
  3. Check assumptions of MH and DM using pairs of item step responses

Example: Political Suspicion

- From 2014 Cooperative Congressional Election Study:
  1. "Politicians often lie, deflect blame, and find other ways to look innocent" (`lies`)
  2. "Government institutions are largely controlled by elite outside interests" (`outsideint`)
  3. "In national politics, nothing happens by accident" (`accident`)
  4. "You can see patterns, designs, and secret activities everywhere once you know where to look" (`secrets`)
- Response options: "strongly agree" (3), "agree" (2), "disagree" (1), "strongly disagree" (0)

Example: Political Suspicion

Order: Politicians Lie, Outside Interests, No Accidents, Secret Designs



Example: Government Spending Preferences (Jacoby 2000)

- "Should spending in the following areas, and on the following problems, be decreased (0), increased (2), or kept about the same (1)?"
  1. Solving the problem of the homeless
  2. Poor people
  3. Child care
  4. Assistance to the unemployed
  5. Assisting blacks
  6. Food stamps
  7. Welfare programs
- 7 items (k) and 3 response categories (m) means $k(m-1)+1$, or 15 distinct scale scores

Example: Government Spending Preferences (Jacoby 2000)

TABLE A1 Scaled Order of Program-Specific Spending Alternatives

Scale Position	Government Spending Alternative
0	Decrease spending, all programs
1	Maintain current spending, programs to help the homeless
2	Maintain current spending, programs to help the poor
3	Maintain current spending, child care
4	Maintain current spending, unemployment programs
5	Maintain current spending, programs to help blacks
6	Increase spending, programs to help the homeless
7	Maintain current spending, Food Stamps
8	Maintain current spending, Welfare
9	Increase spending, Programs to help the poor
10	Increase spending, child care
11	Increase spending, unemployment programs
12	Increase spending, programs to help blacks
13	Increase spending, Food stamps
14	Increase spending, Welfare
15	Increase spending, all programs

Concluding Thoughts

- Nonparametric IRT models (Mokken scales) are not as popular as their parametric counterparts
- However, relaxing the assumptions about the functional form of the IRF means the nonparametric models will fit a larger variety of data
  - On the one hand, this requires scaling at the ordinal level, at least when it comes to the stimuli
  - On the other hand, most researchers ignore difficulty or employ parametric models where stimuli cannot be neatly arrayed along the latent dimension
- The nonparametric approach is all about finding the subset of items that fit the model, rather than finding the right model to fit the data
  - Cumulative structure is an a priori hypothesis
  - Most practitioners of parametric IRT are concerned with finding the model that best fits the data
  - Both approaches are useful in different situations...