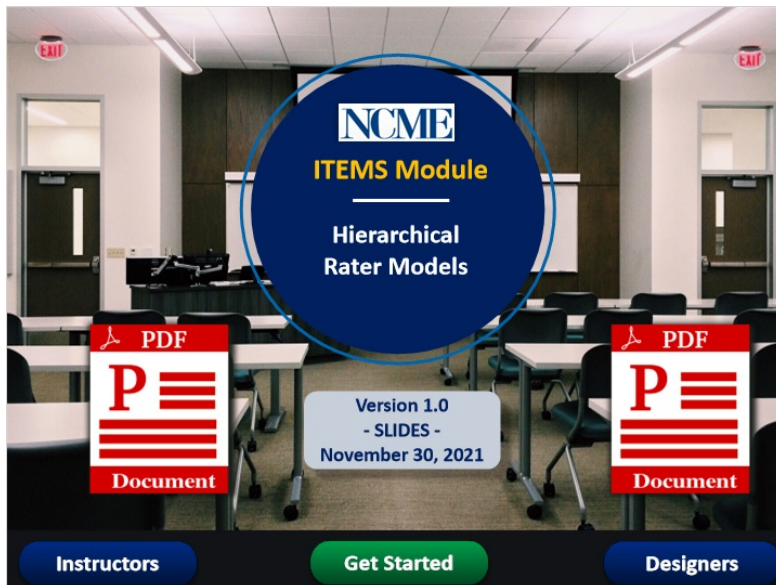


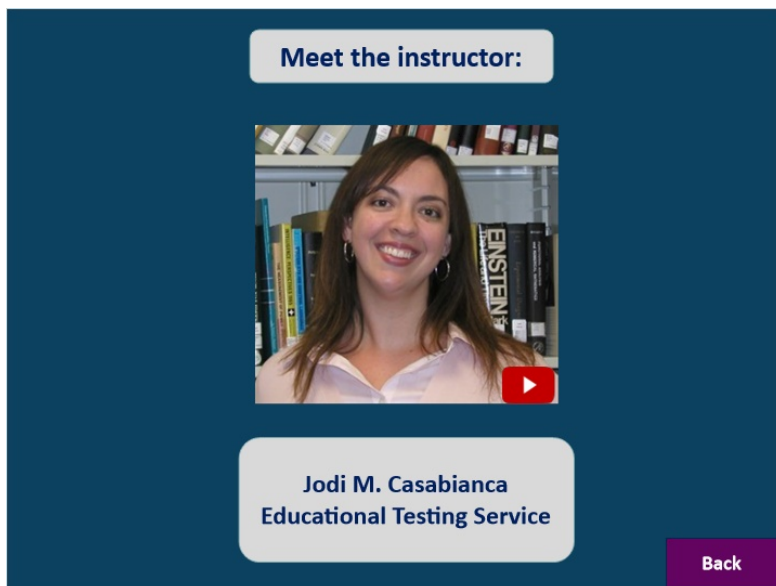
# DM27 SLIDES (Rater Models, Version 1.0)

## 1. Module Overview

### 1.1 Module Cover (START)




### 1.2 Instructor



### 1.3 Designer


Meet the designer:



André A. Rupp  
Mindful Measurement

Back

### 1.4 Welcome



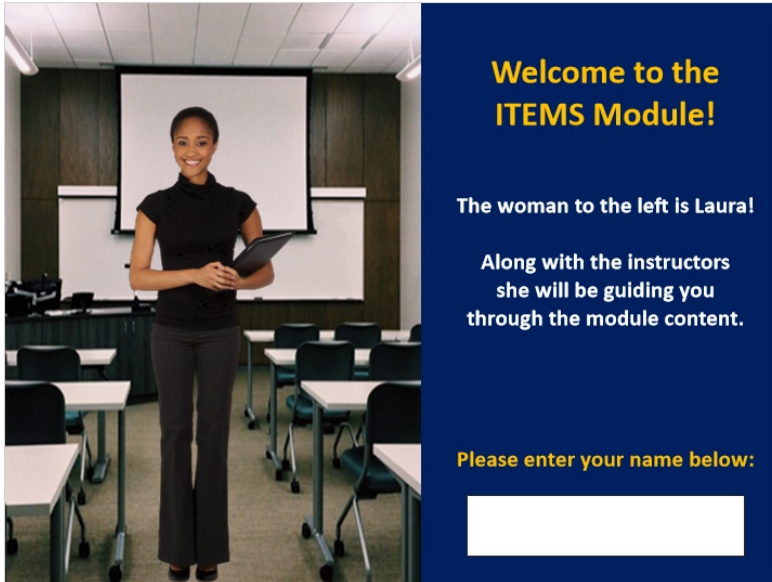
**Welcome to the  
ITEMS Module!**

The woman to the left is Laura!

Along with the instructors  
she will be guiding you  
through the module content.

Please enter your name below:

## Untitled Layer 1 (Slide Layer)



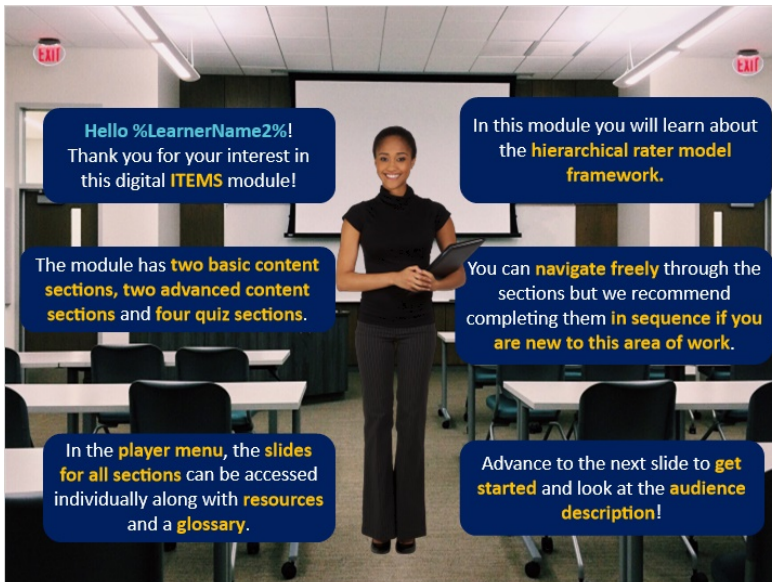
**Welcome to the  
ITEMS Module!**

The woman to the left is Laura!

Along with the instructors  
she will be guiding you  
through the module content.

Please enter your name below:

### 1.5 Overview



Hello %LearnerName2%!  
Thank you for your interest in  
this digital **ITEMS** module!

The module has **two basic content sections, two advanced content sections** and **four quiz sections**.

In the **player menu**, the **slides for all sections** can be accessed individually along with **resources** and a **glossary**.

In this module you will learn about the **hierarchical rater model framework**.

You can **navigate freely** through the sections but we recommend completing them **in sequence if you are new to this area of work**.

Advance to the next slide to **get started** and look at the **audience description!**

## 1.6 Target Audience

### Target Audience

Anyone who would like a gentle statistical introduction to this topic:

- graduate students and faculty in Master's, Ph.D., or certificate programs
- psychometricians and other measurement professionals
- data scientists / analysts
- research assistants or research scientists
- technical project directors
- assessment developers



However, we hope that you find the information in this module useful no matter what your official title or role in an organization is!

## 1.7 Expectations (I)

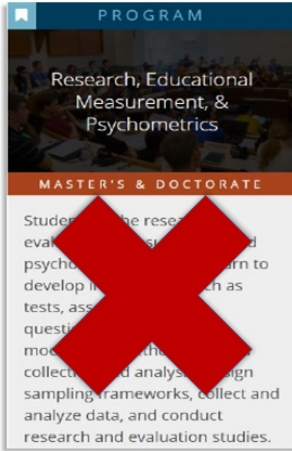
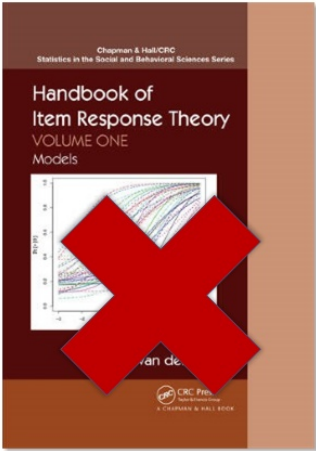


Let's discuss expectations....




## 1.8 Expectations (II)

### ITEMS Modules in Context



## 1.9 Learning Objectives

### Learning Objectives



1. Describe the common rater effects and their impact on test taker scores
2. Understand the main advantage of the HRM framework over other IRT rater models and when it is appropriate to use
3. Discuss the HRM-based rater parameters and describe how they capture different rater behaviors
4. Understand the HRM framework components and conceptualize special cases relevant to specific data sets

## 1.10 Prerequisites


**Prerequisites**

1. Familiarity with constructed response scoring, rater reliability, and rater agreement
2. Completed a two-semester graduate level applied statistics course series and a two-semester psychometric theory course series
3. Completed a graduate level item response theory course

## 1.11 Module Citation

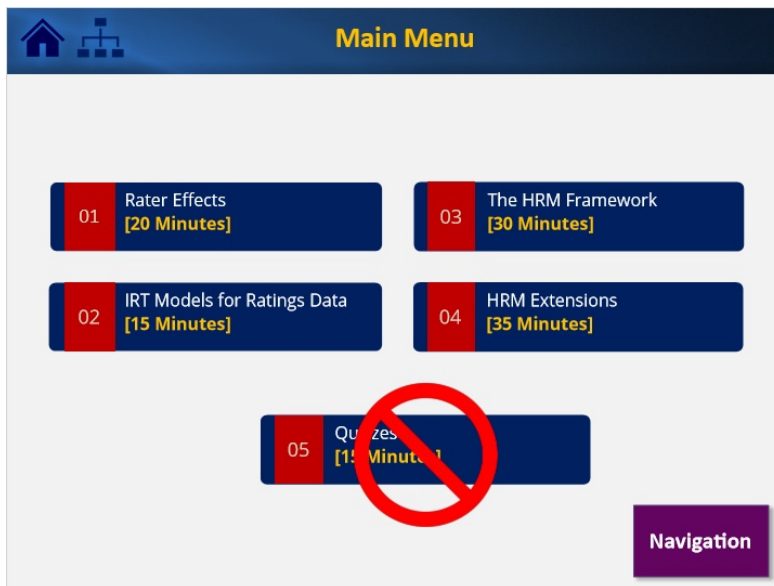
**Module Citation**

Casabianca, J. M. (2021). Hierarchical rater models (Digital ITEMS Module 27). *Educational Measurement: Issues and Practice*, 40(4).



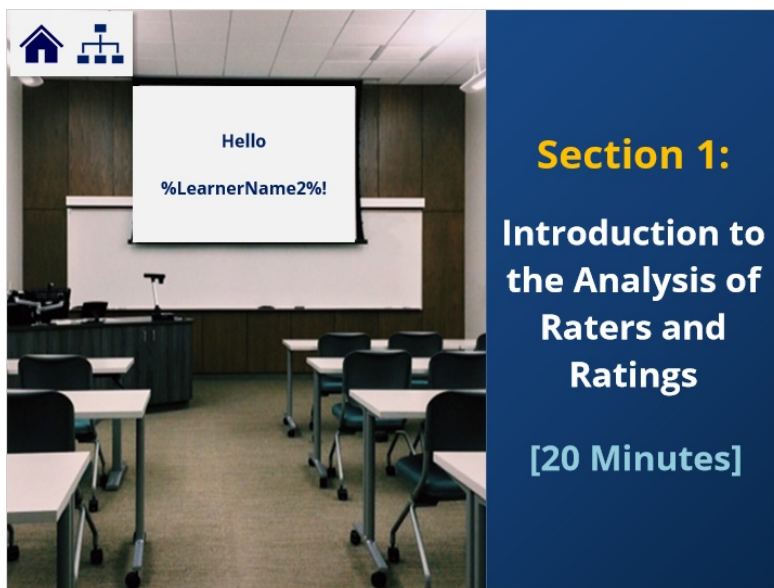
**FREE WEB RESOURCES**

## 1.12 Main Menu





## 2. Section 1: Rater Effects


### 2.1 Cover: Section 1



## 2.2 Objectives: Section 2





### Learning Objectives



1. Understand how interrater agreement and rater accuracy are measures of rating quality
2. Describe the common rater effects
3. Discuss the impacts of rater effects on scores

## 2.3 Tasks with Ratings



### Ratings of Constructed Responses

*Types of Constructed Response Items or Performance Tasks*

Multiple choice	Short Answer (Show work)	Essay (Explain, discuss, compare, etc.)	...	Performance, Activity, or Behavior
Right/wrong; Machine scored	Right/wrong with partial credit; different correct responses	Partial credit; many correct responses with different degrees of quality; different scoring designs	...	Holistic or rubric-based rating
(SAT, Praxis)	(NAEP, PISA)	(GRE Analytical Writing)		(CLASS-S)

Increasing scoring complexity

## 2.4 Agreement vs. Accuracy



### Measuring Rating Quality





**Rater Agreement**



**Rater Accuracy**

- **Agreement** concerns concordance between multiple raters' evaluation of the same work
- **Accuracy** concerns concordance between a rater and the true score assigned by an expert rater

### Accuracy (Slide Layer)



### More on Accuracy



- Cronbach (1955) decomposed rater accuracy into four parts:
  1. elevation, or overall accuracy
  2. differential elevation, or discrimination among test takers,
  3. stereotype accuracy, or discrimination among traits, and
  4. differential accuracy, or discrimination among test takers within traits.

Different measures of rater accuracy that are linked to specific components may not be correlated.
- Accuracy can also be criterion- or norm-referenced.

**Back**





## 2.5 Measuring Interrater Agreement



### Measuring Interrater Agreement

- Reliability scoring/sampling
- Interrater agreement/reliability measures
  - ✓ Correlation between rater  $r$  and rater  $r'$
  - ✓ Rates of exact agreement, adjacent agreement, non-adjacent agreement
  - ✓ Quadratically weighted Kappa (QWK)
  - ✓ Intraclass correlation coefficient (ICC)
- Using interrater agreement as a measure of rating quality is misleading because there could be high agreement and inaccurate ratings



## 2.6 Measuring Rater Accuracy



### Measuring Rater Accuracy

- Rater accuracy is a complex notion. Agreement with experts is not the only way to measure accuracy
- For now, let's consider rater accuracy the extent to which a rater agrees with expert scores (or true scores). We estimate this using:
  - ✓ Correlation between rater  $r$  and true score
  - ✓ Rates of exact agreement, adjacent agreement, non-adjacent agreement
  - ✓ Quadratically weighted Kappa (QWK)
  - ✓ Intraclass correlation coefficient (ICC)
- Operational setting -> performance on validity responses

## 2.7 Agreement Statistics Do Not Provide





### What do these measures provide?

- Low agreement statistics may indicate poor quality ratings, but may not help target specific rater issues.
- *Example:* Validity agreement rates for five raters.
  - Aside from inaccuracy, what type of errors are these raters making?

Validity Agreement Rates			
Rater ID	% Exact Agreement	% Adjacent Agreement	% Discrepant
0001	95%	2%	3%
0002	100%	0%	0%
0003	85%	10%	5%
0004	85%	10%	5%
0005	60%	10%	30%



## 2.8 Common Rater Effects



### Common Rater Effects

• Severity/Leniency (negative/positive bias)
• Centrality/Extremity
• Restriction of range
• Halo effects
• Accuracy/Inaccuracy
• and more...

## 2.9 Impact on Scores (I)





### Impact on Scores (I)

Generalizability studies reveal the proportions of variance attributable to rater error. For example, approximately 35% of variance in ratings of emotional support in the classroom  
(Casabianca, Lockwood, & McCaffrey, 2015)

- **Severity/Leniency** – shifts in average ratings
- **Centrality/Extremity** – decrease or increase in ratings SD
- **Inaccuracy/Accuracy**
  - low rate of exact agreement
  - low  $r(\text{observed}, \text{true})$

## 2.10 Impact on Scores (II)



### Impact on Scores (II)



**Table 2. Descriptive Statistics for Simulated Raw Scores by Rater Effect Group**

Group	N	Score Level					Mean	SD	$r$
		0	1	2	3	4			
Normal	84	11%	21%	37%	22%	9%	1.97	1.11	0.65
Lenient	2	3%	8%	29%	30%	30%	2.77	1.04	0.60
Central	2	7%	23%	43%	22%	5%	1.97	0.97	0.42
Inaccurate	2	11%	22%	34%	22%	12%	2.02	1.17	0.46

Note:  $r$  = the correlation between the average of the scores assigned to each examinee by raters in the Normal group and the scores assigned by a single randomly chosen rater the group represented by the row of the table.

Reference



## Reference (Slide Layer)



Reference

Back

## 2.11 Implications



Implications

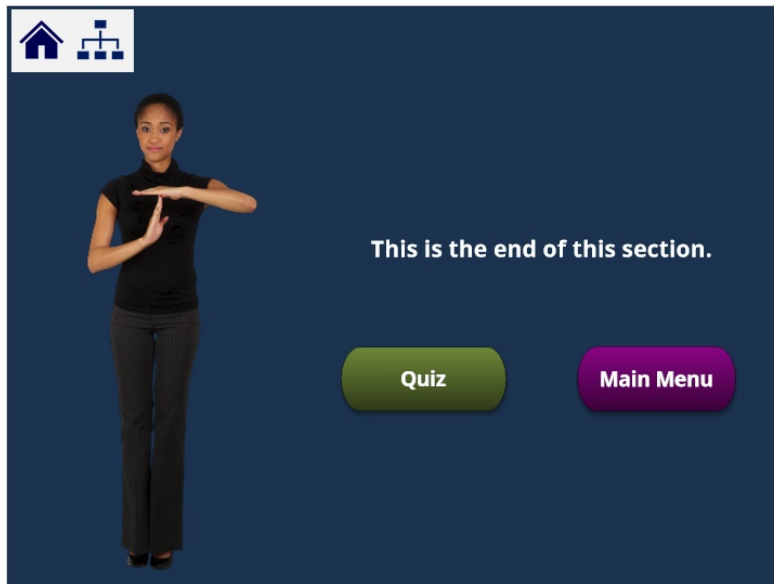
**Importance of Rater Effects**

- Ignoring rater effects impacts the scores resulting from ratings
- Identifying rater effects provides guidance for rater diagnosis and remediation

**Mitigation of Rater Effects**

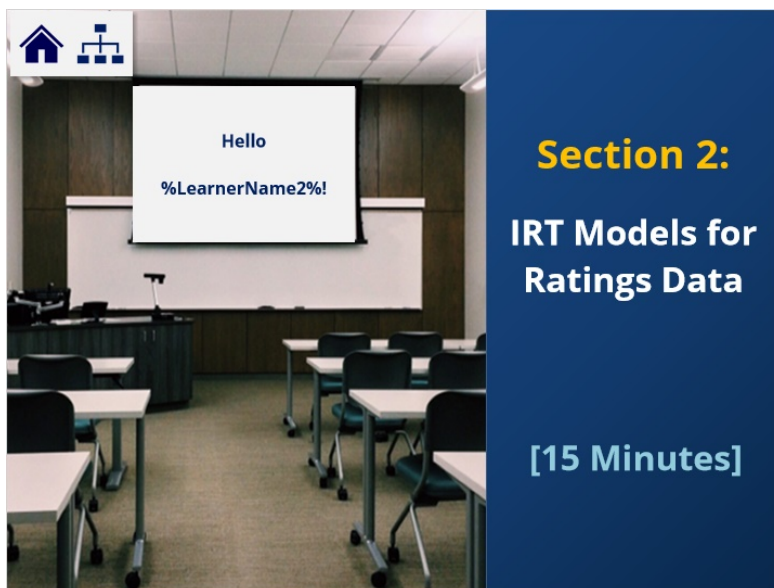
- Reduce errors with rater training, monitoring, calibration, and certification qualifications, and remediation
- Refine scores with item response theory (IRT) rater models

## 2.12 Bookend: Section 1





## 3. Section 2: IRT Models for Ratings Data

### 3.1 Cover: Section 2






### 3.2 Objectives: Section 2





## Learning Objectives



1. Discuss the various types of IRT models for ratings at a high level.
2. Understand the differences between item response and rater response modeling approaches.
3. Describe how different rater effects manifest in the parameter estimates of IRT rater models.
4. Describe the differences in the local independence assumptions and the information accumulation problem.

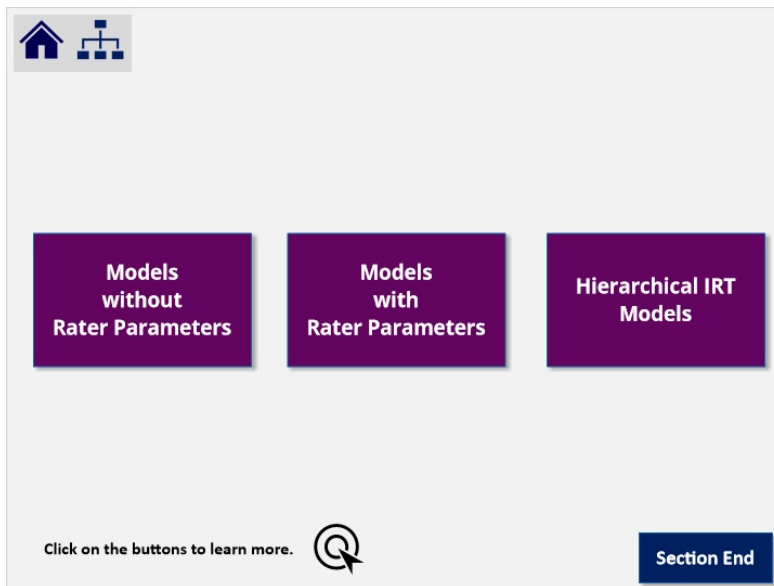
### 3.3 IRT Models for Ratings Data



## IRT Models for Ratings Data


- Polytomous IRT models without rater parameters
- Polytomous IRT models with rater parameters
  - ✓ Multi-faceted Rasch Model (Linacre, 1989)
  - ✓ Muraki's Raters'-effect model (Muraki, 1993)
- Hierarchical IRT rater models

### 3.4 Topic Selection



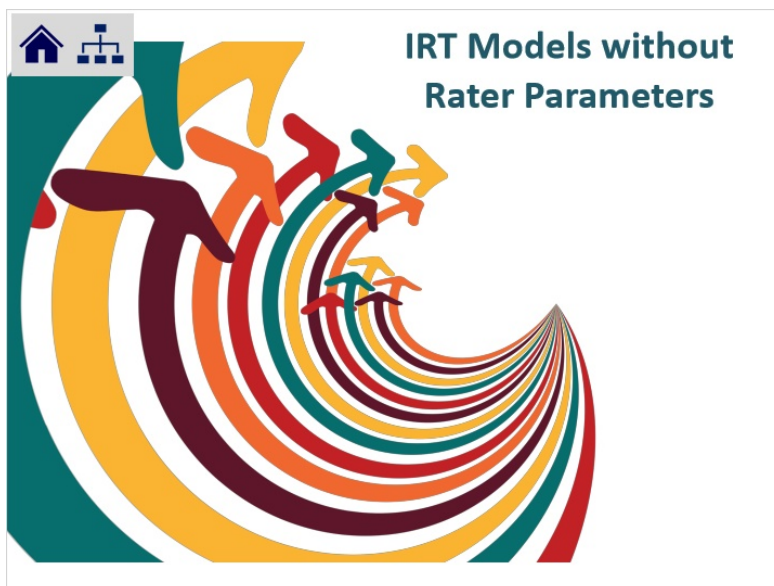
Home icon | Tree icon

Models without Rater Parameters    Models with Rater Parameters    Hierarchical IRT Models

Click on the buttons to learn more. 

Section End

### 3.5 Bookmark: Polytomous Models w/o Rater Parameters





Home icon | Tree icon

IRT Models without Rater Parameters

Decorative graphic of colorful, curved arrows pointing right.

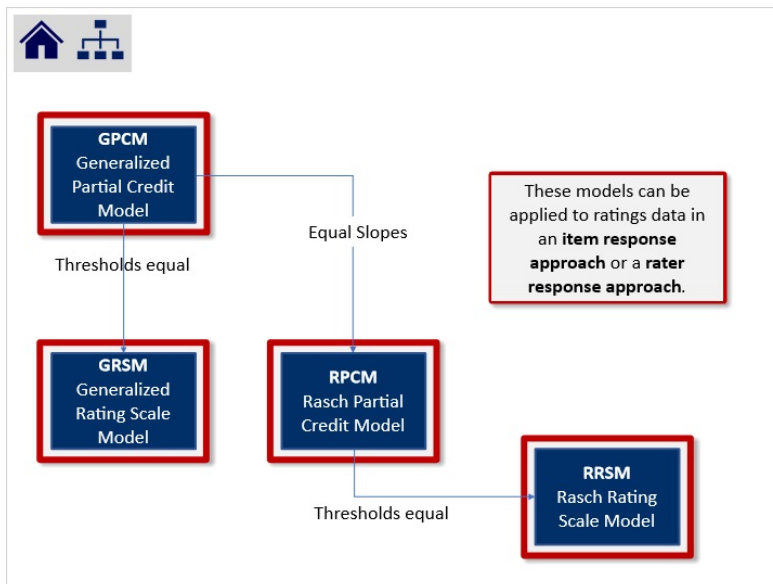
### 3.6 General Principles




## General Principles

- **Item response models** treats each rating as an item response, even if there are multiple ratings per item  
  
Example: N=1,000 test takers, 15 CR items, 50 raters, 2 ratings per item per test taker -> IRT model considers this a test administration with N=1,000 test takers and 30 items.
- **Rater response models** can be applied to a lone CR item to analyze raters' responses instead of analyzing items  
  
Example: N=1,000 test takers, 1 CR item, 50 raters, 2 ratings per item -> IRT model considers this a test administration with N=1,000 test takers and 50 items. Item parameter estimates describe characteristics of raters instead of items.

### 3.7 Model Comparison



### 3.8 Rater Response Modeling Approach



# Rater Response Modeling Approach

Model	Index	Severity/ Leniency	Centrality/ Extremity	Accuracy/ Inaccuracy
RRSM	$\rho_r$	+ / -		
	$PPMC_{r,B}$		- <sup>a</sup> / N	+ / -
	$PPMC_{res,exp}$		- / +	+ <sup>a</sup> / - <sup>a</sup>
	MSU		+ / +	- / +
RPCM	$\rho_r$	+ / -		
	$PPMC_{r,B}$		- <sup>a</sup> / N	+ / -
	$PPMC_{res,exp}$		+ <sup>a</sup> / N	+ / -
	$SDT_{rk}$		+ / -	+ / N
GRSM	MSU		- <sup>a</sup> / + <sup>a</sup>	- / +
	$\rho_r$	+ / -		
	$PPMC_{r,B}$		- <sup>a</sup> / + <sup>a</sup>	+ / -
	$\alpha_r$		- <sup>a</sup> / + <sup>a</sup>	+ / -
GPCM	$\rho_r$	+ / -		
	$PPMC_{r,B}$		- <sup>a</sup> / N	+ / -
	$SDT_{rk}$		+ / -	- <sup>a</sup> / + <sup>a</sup>
	$\alpha_r$		+ <sup>a</sup> / - <sup>a</sup>	+ / -

Reference

Notation

#### Reference (Slide Layer)

Reference	
-----------	--

Wolfe, E. W. (2014). Methods for monitoring rating quality: Current practices and suggested changes. (White Paper). Iowa City, IA: Pearson Education.

Back

### 3.9 Using IRT Model Parameter Estimates for Rater Diagnosis



**RPCM-based Rater Effects Estimates**

Table 3. Validity and Agreement Percentages and Correlations for the Example Data

Index	True Score	Normal	Lenient	Severe	Central	Extreme	Inaccurate	Accurate
$\rho_p$	-0.03	0.00	-3.28	3.49	0.09	-0.05	-0.03	-0.07
$PPMC_{\theta}$	0.95	0.87	0.82	0.85	0.79	0.82	0.76	0.94
$SD\tau_w$	4.16	4.02	4.09	4.27	8.65	0.95	4.30	4.09
MSU	0.37	0.91	1.08	0.89	0.99	1.59	1.93	0.40

Note: Blue shading = reference values. Dark orange shading = outliers. Light orange shading = slightly inflated/deflated values. Normal = average index value for the five Normal raters.

Reference

#### Reference (Slide Layer)

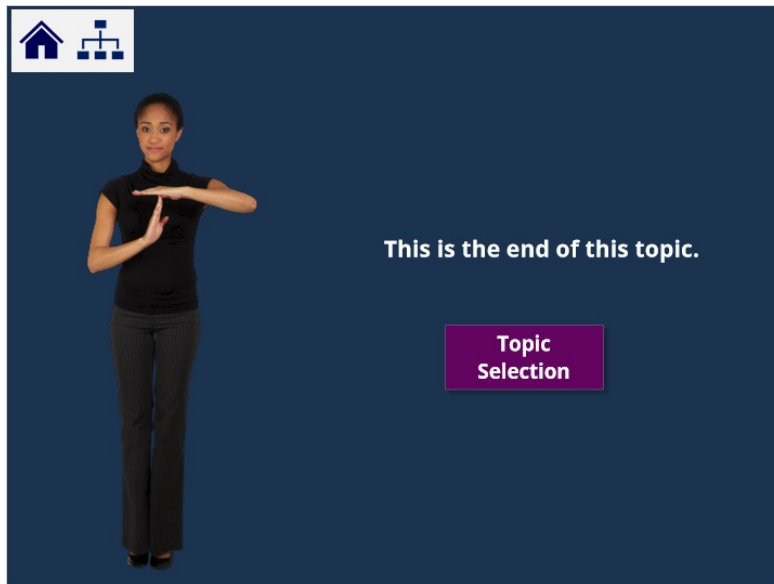


Wolfe, E. W. (2014). Methods for monitoring rating quality: Current practices and suggested changes. (White Paper). Iowa City, IA: Pearson Education.

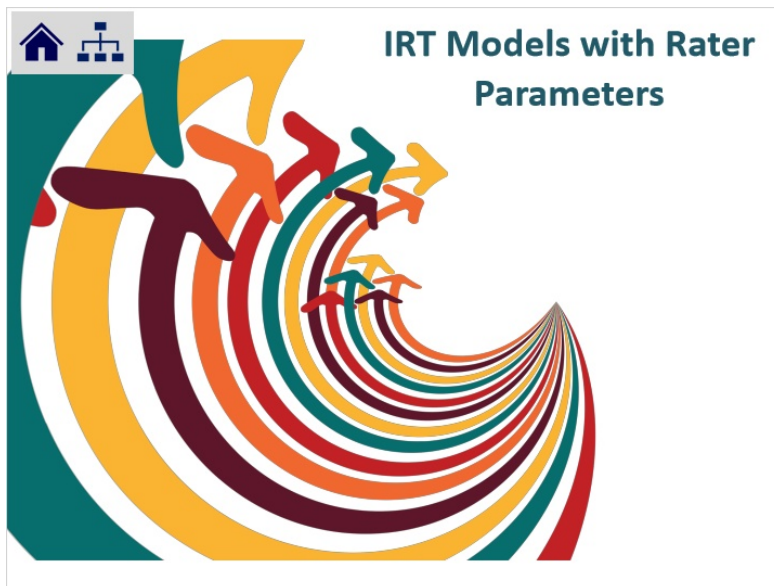
Back





### ***3.10 Bookend: Polytomous Models w/o Rater Parameters***



### ***3.11 Bookmark: Polytomous Models with Rater Parameters***



### 3.12 Multi-faceted Rasch Model (I)



#### Multi-faceted Rasch Model (I)



- **Multifaceted Rasch model** (MFRM; Linacre, 1989)

$$\log \left[ \frac{P(X_{ijr} = k | \theta_i)}{P(X_{ijr} = k - 1 | \theta_i)} \right] = \theta_i - \beta_j - \gamma_{jk} - \phi_r$$

where  $\phi_r$  is a **rater severity parameter**.

- **Muraki's rater effect model** (Muraki, 1993) generalizes the facets model by adding a discrimination parameter (a-parameter) for the item.



### 3.13 Multi-faceted Rasch Model (II)



#### Multi-faceted Rasch Model (II)

- Expanding the MFRM is relatively straightforward and there are numerous variants of the MFRM (see Myford & Wolfe, 2003)
- The MFRM can be used to detect leniency/severity, central tendency, accuracy, randomness, halo, and differential leniency/severity (Linacre, 1989; Myford & Wolfe, 2003, 2004)
  - ✓ In a variant of the MFRM that includes rater-item-specific thresholds,  $\gamma_{rkr}$  rater centrality can be captured via the SD of these thresholds
  - ✓ Rater accuracy can be approximated using rater fit statistics which tend to correspond to expert ratings (Wind & Engelhard, 2012), and post-model estimation approximations (Linacre, 2004; Wind, Engelhard, & Wesolowski, 2016)



### 3.14 Multi-faceted Rasch Model (III)




#### Multi-faceted Rasch Model (III)

- The underlying IRT assumption of local independence given the latent trait technically applies in these MFRM
- However, it is unreasonable to make that assumption since it is unlikely that ratings assigned to the same response will be independent
  - ✓ In some scenarios, we want to consider each individual rater as contributing a different expert opinion
  - ✓ In many scenarios, we ask raters to use a scoring rubric, and expect consistency in their use of the scoring rubric.

### 3.15 Bookend: Polytomous Models with Rater Parameters







This is the end of this topic.

Topic Selection

### 3.16 Bookmark: Hierarchical IRT Models



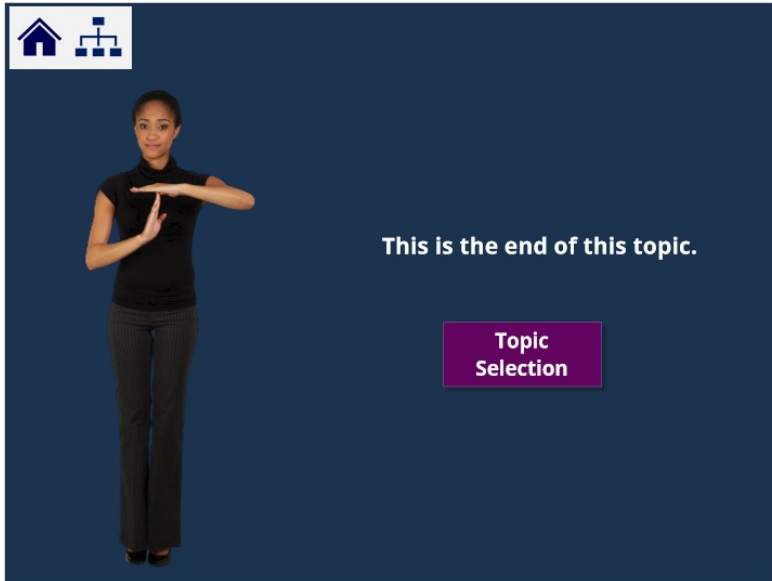
### 3.17 General Principles



## General Principles

**Hierarchical models** account for the **nesting structure** in data sets where multiple raters are rating the same work. **This corrects the information accumulation problem** (Mariano, 2002).

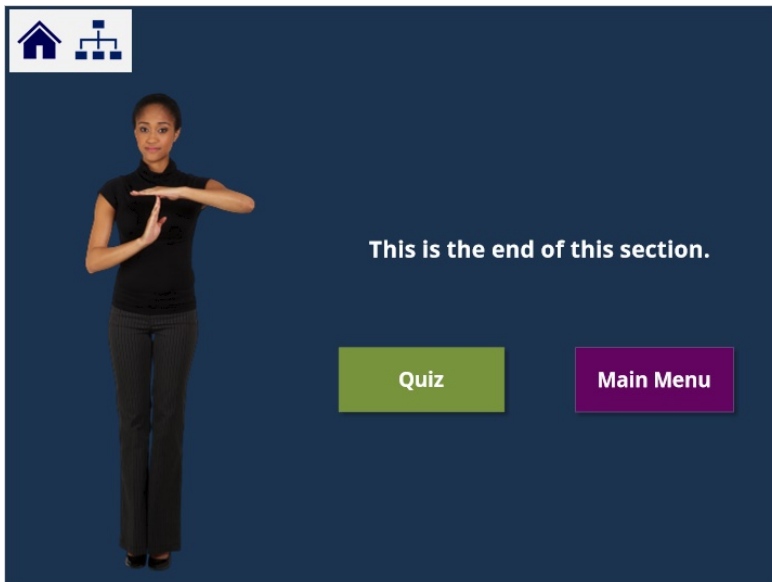
### 3.18 Bookend: Hierarchical IRT Models



This is the end of this topic.

Topic Selection

### 3.19 Bookend: Section 2



This is the end of this section.

Quiz Main Menu



## 4. Section 3: The HRM Framework

### 4.1 Cover: Section 3

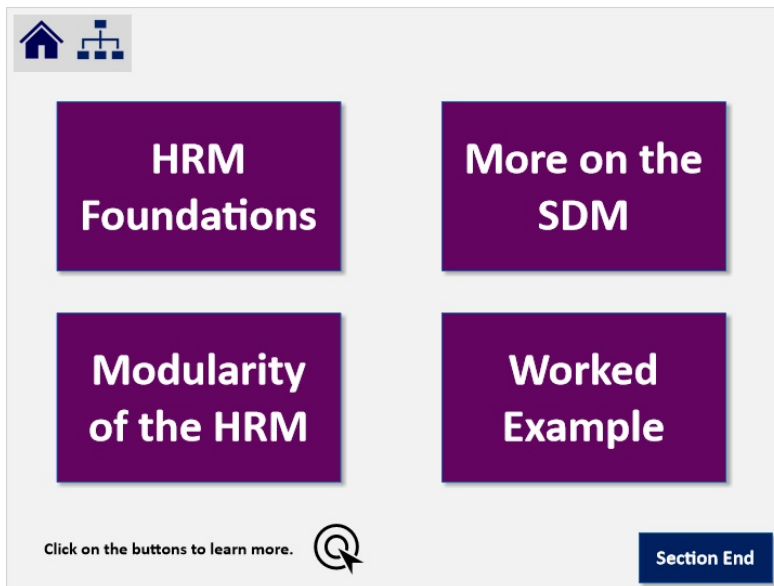


### 4.2 Objectives: Section 3

### Learning Objectives

1. Discuss the repeated ratings problem and the nesting structure of common rater datasets.
2. Describe the different levels of the HRM and the two stage measurement process.
3. Compare and contrast the different signal detection models.
4. Understand how the modularity of the HRM framework permits various special cases.

### 4.3 Topic Selection




This interface features a navigation bar at the top with a home icon and a hierarchical tree icon. Below this, four purple rectangular buttons are arranged in a 2x2 grid. The buttons are labeled: 'HRM Foundations', 'More on the SDM', 'Modularity of the HRM', and 'Worked Example'. At the bottom left, there is a text prompt 'Click on the buttons to learn more.' followed by a magnifying glass icon. At the bottom right, there is a dark blue button labeled 'Section End'.

HRM Foundations

More on the SDM

Modularity of the HRM

Worked Example

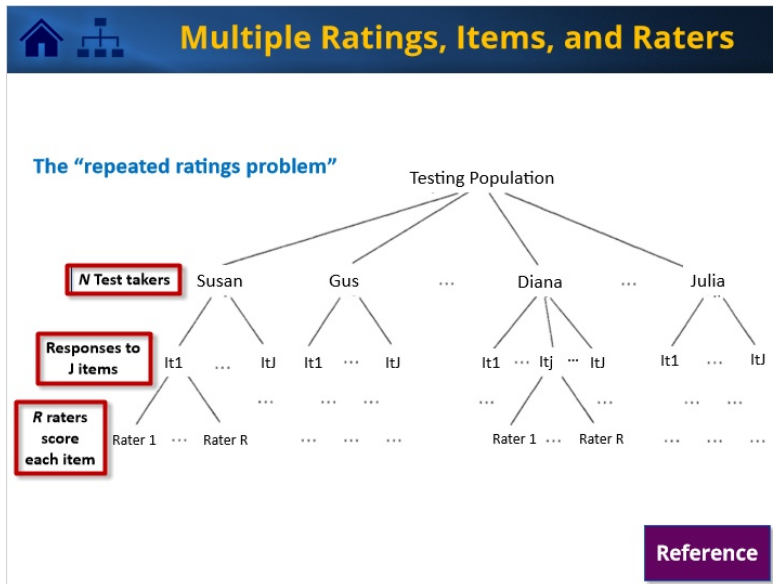
Click on the buttons to learn more. 

Section End

### 4.4 Bookmark: HRM Foundations



## 4.5 The Completely Crossed HRM



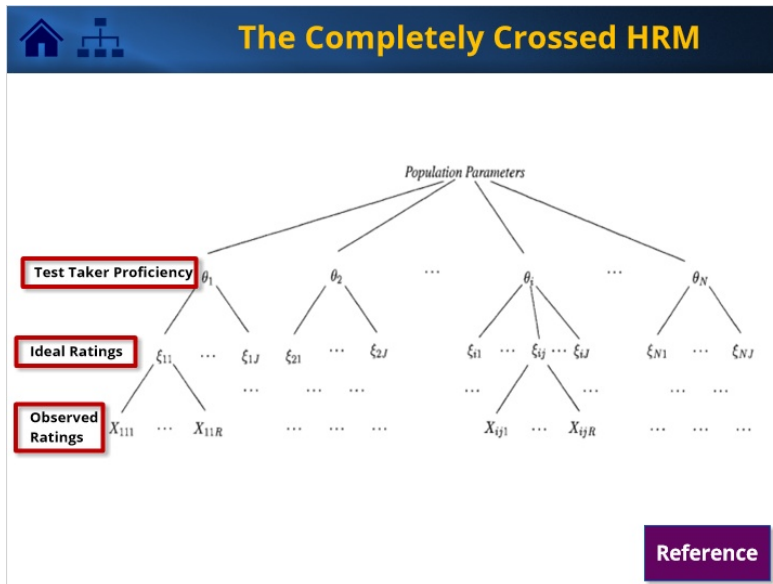
### Reference (Slide Layer)

**Reference**

Source: Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to largescale educational assessment data. *Journal of Educational and Behavioral Statistics*, 27, 341-384.

Back

## 4.6 The Completely Crossed HRM





### Reference (Slide Layer)

Source: Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to largescale educational assessment data. *Journal of Educational and Behavioral Statistics*, 27, 341-384.

Back



## 4.7 Ideal Ratings



### Ideal Ratings

- **Ideal ratings** are per-item (per-test taker) **latent variables** for estimating an test taker's **true response** to an item
  - Think: “true score”
  - The test taker's response scored without rater bias or variability
- We collect **observed ratings** to estimate **ideal ratings**

## 4.8 Two-stage Model I



### Two-stage Process: Stage 1

**Stage 1:** An IRT model defines the relationship between ideal ratings and latent traits



Example:  $K$ -category Generalized Partial Credit Model (GPCM; Muraki, 1992) relates test taker proficiency to ideal ratings

$$P[\xi_{ij} = \xi | \theta_i, \beta_j, \gamma_{jk}] = \frac{\exp\left\{\sum_{k=1}^{\xi} \alpha_i (\theta_i - \beta_j) - \gamma_{jk}\right\}}{\sum_{h=0}^{K-1} \exp\left\{\sum_{k=1}^h \alpha_i (\theta_i - \beta_j) - \gamma_{jk}\right\}}$$

$i = 1, \dots, N; j = 1, \dots, J; k = 1, \dots, K$

- $\xi_{ij}$  is ideal rating for test taker  $i$  on item  $j$
- $\alpha_i$  is item discrimination
- $\beta_j$  is item difficulty
- $\gamma_{jk}$  step parameter at category  $k$

## 4.9 Two-stage Model II



### Two-stage Process: Stage 2

**Stage 2:** A signal detection model (SDM) defines the relationship between ideal ratings and observed ratings.



For example: A discrete SDM defines the probability that rater  $r$  rates a score of  $k$ , given the ideal rating is  $\xi$

$$p_{\xi kr} = P[X_{ijr} = k | \xi_{ij} = \xi] \propto \exp \left\{ -\frac{1}{2\psi_r^2} [k - (\xi + \phi_r)]^2 \right\}$$

$i = 1, \dots, N; j = 1, \dots, J; r = 1, \dots, R$

- $\psi_r$  is rater variability
- $\phi_r$  is rater bias

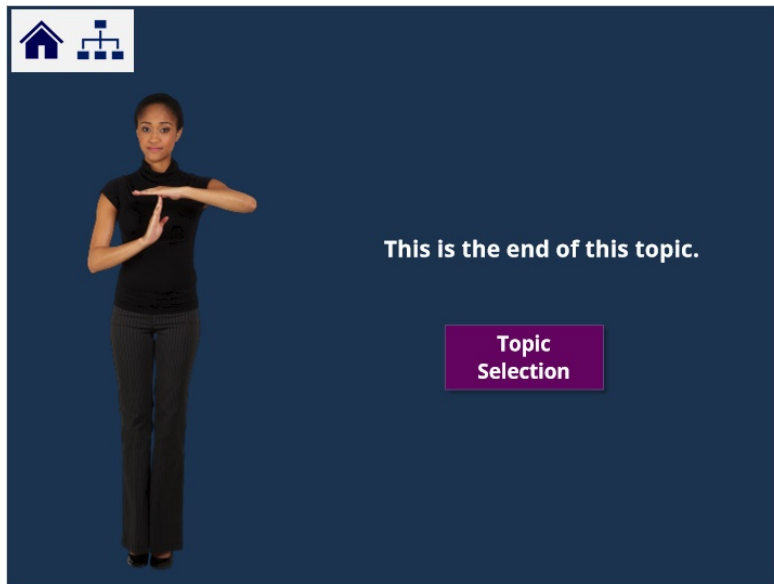
## 4.10 Two-stage Model II



### Model Estimation

- The HRM is estimated as a Bayesian model using Markov chain Monte Carlo (MCMC) estimation.
- Some good references:
  - Kim, J. S., & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, 26(4), 38-51.
  - Junker, B. W., Patz, R. J., & VanHoudnos, N. M. (2016). Markov chain Monte Carlo for item response models. In W. van der Linden (Ed.), *Handbook of item response theory* vol 2 (pp. 271-312).

#### **4.11 Bookend: HRM Foundations**



#### **4.12 Bookmark: SDM Component**





#### 4.13 Modeling Rater Behavior in the

Modeling Rater Behavior I: Patz et al. (2002)

**Matrix of Rating Probabilities**

$p_{\xi kr} = P[\text{Rater } r \text{ rates } k \mid \text{ideal rating } \xi]$  in each row of this matrix

Ideal Rating ( $\xi$ )	Observed Rating (k)				
	0	1	2	3	4
0	$p_{00r}$	$p_{01r}$	$p_{02r}$	$p_{03r}$	$p_{04r}$
1	$p_{10r}$	$p_{11r}$	$p_{12r}$	$p_{13r}$	$p_{14r}$
2	$p_{20r}$	$p_{21r}$	$p_{22r}$	$p_{23r}$	$p_{24r}$
3	$p_{30r}$	$p_{31r}$	$p_{32r}$	$p_{33r}$	$p_{34r}$
4	$p_{40r}$	$p_{41r}$	$p_{42r}$	$p_{43r}$	$p_{44r}$

#### 4.14 Modeling Rater Behavior in the

Modeling Rater Behavior II: Patz et al. (2002)

**Probabilities** in each row of the matrix can be made proportional to a **Normal density**:



**The normal density:**

$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[x - \mu]^2\right\}$$

**The discrete SDM in the Patz et al. (2002) HRM:**

$$p_{\xi kr} = P[X_{jr} = k | \xi_{ij} = \xi] \propto \exp\left\{-\frac{1}{2\psi_r^2}[k - (\xi + \phi_r)]^2\right\}$$

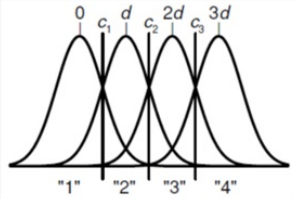
#### 4.15 Modeling IV: De Carlo et al. (2011)



**Modeling IV: De Carlo et al. (2011)**

DeCarlo et al. (2011) noted that the Patz et al. (2002) HRM only captures rater bias and variability and proposed a different model for Level 1



**A Latent Class SDT model**




$d$  = distances between perceptual distributions (here, equidistant)  
 $c$  = response criteria locations; divide the scoring decision space

Reference

#### Reference (Slide Layer)



**Reference**



Back

#### 4.16 Modeling V: De Carlo et al. (2011)



##### Modeling V: De Carlo et al. (2011)

###### Latent Class SDT model

$$p(X_{rj} \leq k | \xi_j = \xi) = F(c_{rkj} - d_{rj}\xi_j)$$

- $X_{rj}$  = Rater  $r$ 's score for item  $j$
- $F$  = a cumulative distribution function (logistic or normal)
- $c_{rkj}$  =  $K-1$  ordered response criteria locations,  $c_{r1j} < c_{r2j} < \dots < c_{r(K-1)j}$
- $d_{rj}$  = rater- and item-specific "detection" parameter (for item  $j$  and rater  $r$ ); considered rater precision / discrimination, or rater's ability to detect the latent categories

#### 4.17 Model Selection



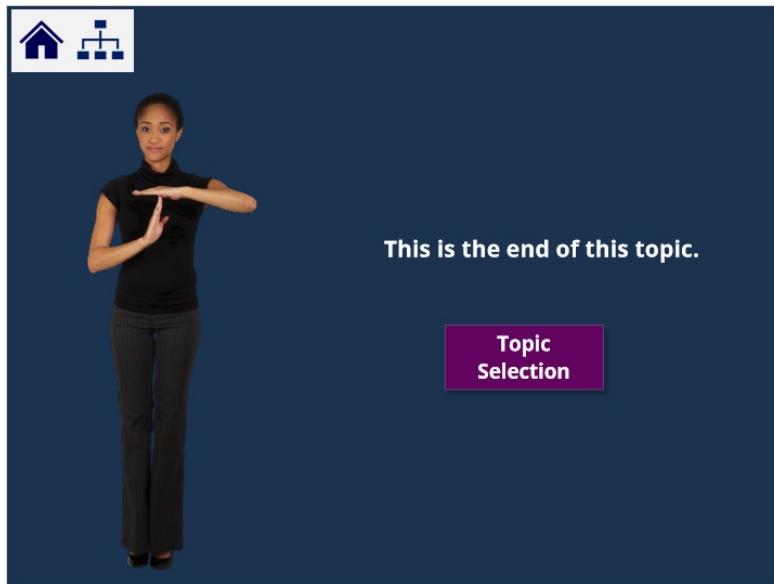
##### Model Selection

"If your purpose is to **fully understand rating behavior**, then the signal detection model in the Patz et al. HRM may not be enough; **DeCarlo's SDT** may be better! If your goal is to **account for major sources of variation in raters** when modeling or estimating  $\theta_i$ , you may not need a **fully realistic rater model**."

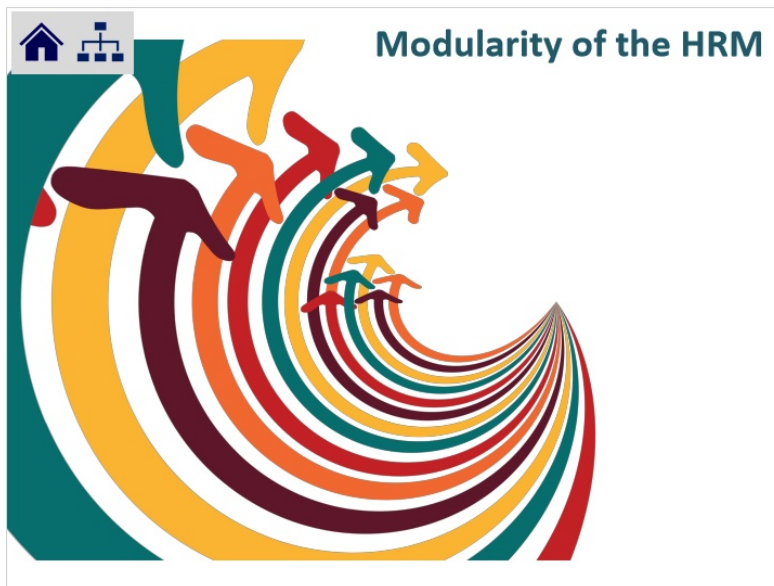
(Junker, 2016)

**Must consider parsimony**



#### 4.18 Bookend: SDM Component



#### 4.19 Bookmark: Modularity





## 4.20 Modularity



### Modularity

- Junker (2016) highlighted that the HRM is flexible because we can insert different models at different levels
- Any IRT model can be specified to model ideal ratings
  - ✓ 2PL, GPCM, PCM
  - ✓ MGPCM
- The SDM has traditionally been modeled using a discrete SDM (~normal density as per Patz et al., 2002); alternatives do exist (DeCarlo, Kim, & Johnson's HRM-SDT (2011))

## 4.21 HRM Framework I: Formulation





### HRM Framework I: Formulation

$\theta_{itm}$  : a longitudinal model  
 $\xi_{ijt}$  : a multidimensional polytomous IRT model  
 $X_{ijrt}$  : a multidimensional polytomous signal detection model

**Notation**

- $i = 1, \dots, N$  test takers
- $j = 1, \dots, J$  items
- $r = 1, \dots, R$  raters
- $t = 1, \dots, T$  time points
- $m = 1, \dots, M$  dimensions
- $k = 1, \dots, K$  score levels
- Covariates and other features are not depicted here

## 4.22 HRM Framework II: Formulation



### HRM Framework II: Formulation



$\theta_{itm}$  : a longitudinal model  
 $\xi_{ijt}$  : a multidimensional polytomous IRT model  
 $X_{ijrt}$  : a multidimensional polytomous signal detection model


**Special cases**

- T=1 and M=1 A cross-sectional, unidimensional test/rubric
- T=5 and M=1 A longitudinal design using a unidimensional test
- T=1 and M=3 A cross-sectional, multidimensional test/rubric
- T=5 and M=2 A longitudinal design using a multidimensional test

*Note: Indices t and m refer to models for the trait. Indexing may be more complicated if modeling changes in raters (or changes in raters and traits).*

## 4.23 Bookend: Modularity

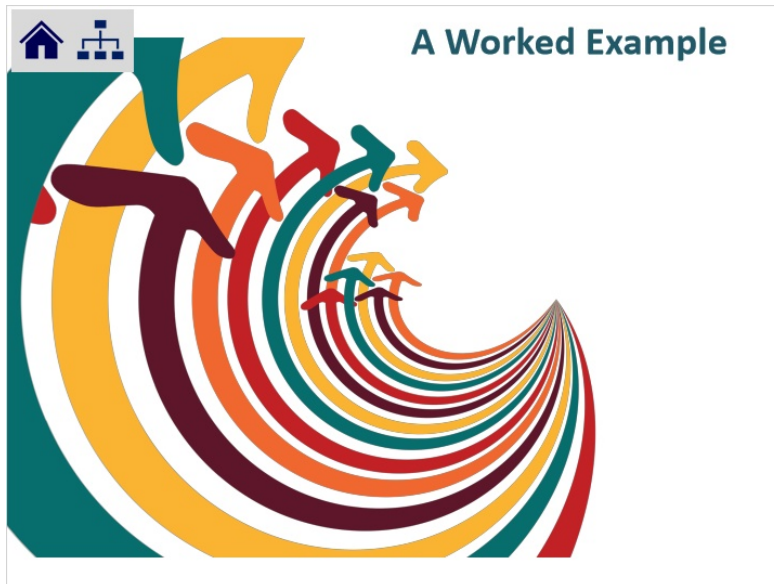





This is the end of this topic.

Topic Selection

#### 4.24 Bookmark: Worked Example





#### 4.25 Rating Design and Data Structure

 **Rating Design and Data Structure**

- Consider a large-scale assessment, which involves the rating of  $N = 500$  test takers by  $R = 10$  raters
- The test contains  $J = 8$  CR items, each scored on a  $K = 4$  point scale
- Test takers were scored by 3 different raters on all items with each rater assigned to multiple individuals (partially crossed design)
- 12,000 total ratings

## 4.26 Analysis Goals and Plan





### Analysis Goals and Plan

1. Evaluate the performance of this group of raters in terms of bias and variability
2. Measure individuals' underlying latent trait while accounting for these rater effects
3. Describe the characteristics of the items

**Let's fit the basic HRM with:**

- Conventional SD model from Patz et al. (2002)
- GPCM

## 4.27 Distribution of Ratings by Item



### Distribution of Ratings by Item

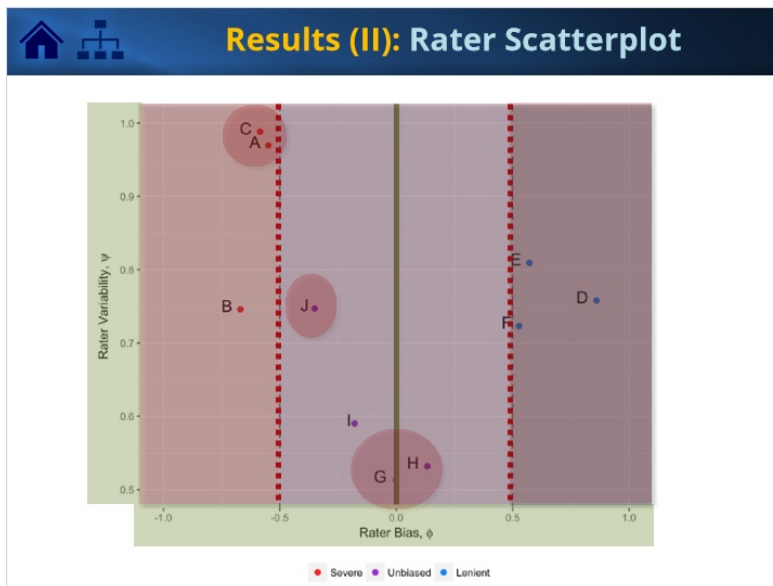
Item Score	Item								Percent
	1	2	3	4	5	6	7	8	
1	440	470	433	522	146	298	340	367	25.13
2	589	417	501	516	314	426	363	587	30.94
3	359	343	344	314	473	455	371	367	25.22
4	112	270	222	148	567	321	426	179	18.71



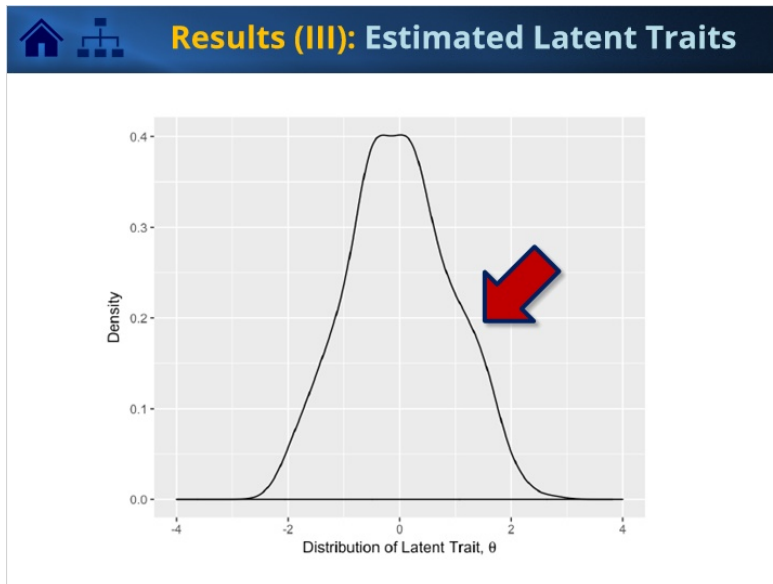
## 4.28 Results (I): Summary Statistics

	mean	sd	X2.5.	X25.	X50.	X75.	X97.5.	Rhat	n.eff
alpha[1]	1.5062	0.2198	1.1070	1.3536	1.4879	1.6498	1.9822	1.0013	2400
alpha[2]	1.2515	0.1699	0.9550	1.1293	1.2401	1.3599	1.6103	1.0007	3000
alpha[3]	2.0403	0.2991	1.5159	1.8323	2.0153	2.2326	2.6814	1.0010	3000
gamma[1,2]	-1.0588	0.1439	-1.3620	-1.1508	-1.0541	-0.9603	-0.7912	1.0007	3000
gamma[2,2]	-0.0598	0.1698	-0.3739	-0.1722	-0.0681	0.0456	0.3009	1.0006	3000
gamma[3,2]	-0.7923	0.1130	-1.0210	-0.8679	-0.7885	-0.7148	-0.5811	1.0013	2500
gamma[4,2]	-0.2752	0.1187	-0.5120	-0.3533	-0.2770	-0.1969	-0.0400	1.0018	1500
phi.r[1]	-0.5496	0.0478	-0.6440	-0.5805	-0.5481	-0.5179	-0.4572	1.0008	3000
phi.r[2]	-0.6695	0.0394	-0.7484	-0.6956	-0.6692	-0.6429	-0.5933	1.0014	3000
phi.r[3]	-0.5850	0.0465	-0.6762	-0.6166	-0.5851	-0.5525	-0.4939	1.0010	3000
psi.r[1]	0.9694	0.0342	0.9064	0.9456	0.9679	0.9915	1.0383	1.0009	3000
psi.r[2]	0.7459	0.0272	0.6926	0.7278	0.7457	0.7632	0.7989	1.0011	3000
psi.r[3]	0.9881	0.0332	0.9244	0.9652	0.9878	1.0102	1.0542	1.0028	860
theta[1]	-0.4755	0.3997	-1.2750	-0.7326	-0.4754	-0.1957	0.2753	1.0007	3000
theta[2]	0.8032	0.3391	0.1405	0.5700	0.8066	1.0307	1.4824	1.0007	3000
theta[3]	0.4746	0.3878	-0.2961	0.2212	0.4870	0.7343	1.2238	1.0035	670

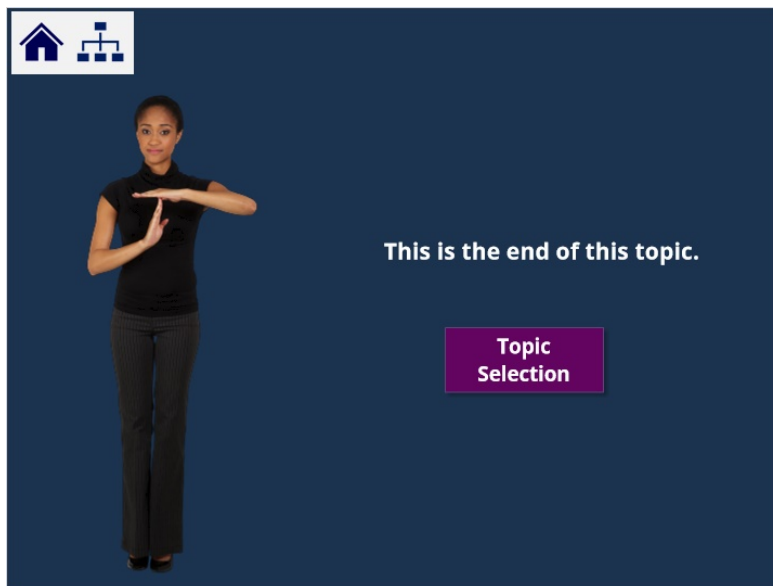
## 4.29 Results (II): Rater Scatterplot



### 4.30 Results (III): Estimated Latent Traits

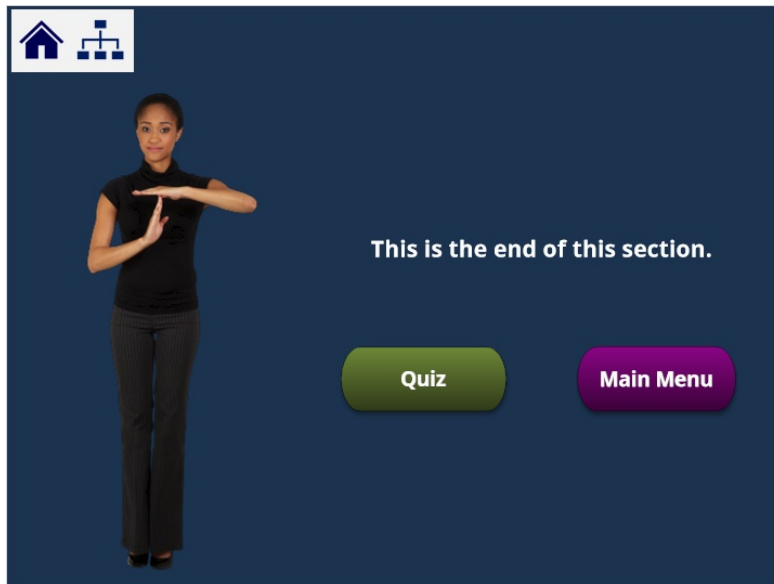


### 4.31 Bookend: Worked Example



This slide features a woman in a black top and dark pants standing on the left side, gesturing with her hands. In the top left corner, there is a small icon of a house with a tree. On the right side, the text "This is the end of this topic." is displayed. Below this text is a purple button with the text "Topic Selection" in white.

### 4.32 Bookend: Section 3





## 5. Section 4: HRM Extensions


### 5.1 Cover: Section 4



## 5.2 Objectives: Section 4





### Learning Objectives



1. Understand the general approach to modeling rater covariates.
2. Discuss the components of the longitudinal HRM that model changes in traits.
3. Describe how the HRM may be adapted to accommodate multidimensionality in traits and rater behavior.
4. Compare use of the M-HRM to other modeling approaches that ignore the multidimensionality.


## 5.3 Topic Selection



Covariates

Longitudinal HRM

Multidimensional HRM



Click on the buttons to learn more. 

Section End

## 5.4 Bookmark: Covariates





## 5.5 SDM Adaptation I



### SDM Adaptation I

- To incorporate **rater covariates** and estimate their effects, we adapt the SDM:
$$p_{\xi kv} = P[X_{ijv} = k | \xi_{ij} = \xi] \propto \exp \left\{ -\frac{1}{2\omega_v^2} [k - (\xi + \rho_v)]^2 \right\}$$
- Here, **bias** is  $\rho_v$  and rater SD is  $\omega_v$  for **pseudorater**  $v$  ( $v = 1, \dots, V$ ). There is a pseudorater for every unique combination of rater and rater covariate.
- Specify with a  $V \times (R+S)$  design matrix  $\mathbf{Y}_v$ 
  - ✓  $R$  columns contain binary indicators per rater
  - ✓  $S$  columns contain rater covariates (or covariate factors)



## 5.6 SDM Adaptation II



### SDM Adaptation II

- Linear model for rating bias  $\rho_v$  depending on covariates as defined by design matrix  $\mathbf{Y}_v$ 
$$\rho_v = \mathbf{Y}_v \boldsymbol{\eta}$$
where  $\boldsymbol{\eta} = (\phi_1, \dots, \phi_R, \eta_1, \dots, \eta_S)^T$
- Linear model for rating variability  $\omega_v^2$  depending on covariates as defined by design matrix  $\mathbf{Y}_v$ 
$$\ln \omega_v^2 = \mathbf{Y}_v (\ln \boldsymbol{\tau}^2)$$
where  $\ln \boldsymbol{\tau}^2 = (\ln \psi_1^2, \dots, \ln \psi_R^2, \ln \tau_1^2, \dots, \ln \tau_S^2)^T$

## 5.7 SDM Adaptation III





### SDM Adaptation III

**Random** or **fixed** rating effects modeling approach

- Random*: pseudorater bias and variability are treated as random draws
  - ✓ Bias is drawn from a normal distribution centered at a linear function of the covariates
  - ✓ Variability is drawn from a similarly centered log-normal distribution (or normal, if transformed)
- Fixed*: the linear structures for rating bias and variability replace the original rater parameters



## 5.8 Examples




### Examples

- Mariano, L. T., & Junker, B. W. (2007). Covariates of the rating process in hierarchical models for multiple ratings of test items. *Journal of Educational and Behavioral Statistics*, 32(3), 287-314.
- Casabianca, J. M., Junker, B. W., & Patz, R. J. (2016). Hierarchical rater models. In W. van der Linden (Ed.), *Handbook of item response theory Vol 2* (pp. 449-465). Chapman and Hall/CRC.
- Check back in future versions of this module!

## 5.9 Bookend: Covariates







This is the end of this topic.

Topic Selection

## 5.10 Bookmark: Longitudinal HRM





## 5.11 Longitudinal Assessment Scenarios

 **Longitudinal Assessment Scenarios**

<ul style="list-style-type: none"><li>• <b>Large-scale testing</b> Example: Tracking learning progressions (Winsight, 2017)</li></ul>
<ul style="list-style-type: none"><li>• <b>Educational interventions</b> Example: Social and Character Development (IES, 2010)<ul style="list-style-type: none"><li>✓ 5 time points (over 3 years)</li><li>✓ Teachers and parents are raters of children's behavior</li></ul></li></ul>
<ul style="list-style-type: none"><li>• <b>Studies of teaching quality</b> Example: Measures of Effective Teaching (BMGF, 2012)<ul style="list-style-type: none"><li>✓ No intervention</li><li>✓ 4 time points (over 2 years)</li><li>✓ Trained raters (certain % of classrooms double-scored) evaluate teacher and student interactions</li></ul></li></ul>



## 5.12 Longitudinal Modeling in the IRT





### Longitudinal Modeling in the IRT

IRT must make assumptions about the stability of the latent trait (Andersen, 1985; Embretson, 1991, 1997; McArdle, Petway & Hishinuma, 2015; Millsap, 2010; Robert & Ma, 2006):

- ✓ Changes we observe in individuals are due to real changes or growth
- ✓ The construct definition does not change over time
- ✓ Counter-example: items on a depression scale may perform differently for pre-adolescents versus adolescents

## 5.13 The L-HRM



### The L-HRM

- **The L-HRM presented here assumes:**
  - ✓ Scalar invariance (Horn & McArdle, 1992; Little, 2013; Meredith, 1993)
  - ✓ one common set of item parameters for all time points
  - ✓ permits changes in the latent traits to be attributed to real changes, and not changes in the relationship between the items and the construct (Meredith, 1993).
  - ✓ Dual change score modeling approach which incorporates growth and autoregressive components (McArdle et al., 2015)
- **Estimates overall growth** (not individual growth/trends)
- **Permits the explicit modeling of different types of growth** even though we discuss linear growth only

## 5.14 Model Formulation I

Model Formulation I

$$\begin{aligned} \theta_{it} | \rho, g &\sim \text{longitudinal model, } t = 1, \dots, T, \text{ for each } i \\ \xi_{ijt} | \theta_{it}, \alpha_j, \beta_j, \gamma_{jk} &\sim \text{polytomous IRT model, } j = 1, \dots, J, \text{ for each } i, t \\ X_{ijrt} | \xi_{ijt}, \psi_r^2, \phi_r &\sim \text{polytomous signal detection model, } r = 1, \dots, R, \text{ for each } i, j, t \end{aligned}$$

## 5.15 Model Formulation II

Model Formulation II

$$\begin{aligned} \theta_{it} | \rho, g &\sim \text{longitudinal model, } t = 1, \dots, T, \text{ for each } i \\ \xi_{ijt} | \theta_{it}, \alpha_j, \beta_j, \gamma_{jk} &\sim \text{polytomous IRT model, } j = 1, \dots, J, \text{ for each } i, t \\ X_{ijrt} | \xi_{ijt}, \psi_r^2, \phi_r &\sim \text{polytomous signal detection model, } r = 1, \dots, R, \text{ for each } i, j, t \end{aligned}$$



$$\theta_{it} = \delta_t + Z_{it}$$

**Time Series Component**  
 (Box, Jenkins, & Reinsel, 2013)

**Growth**

The longitudinal component here is in the level of the traits, not the rater parameters.

### 5.16 Model Formulation III



#### Model Formulation III



$$\theta_{it} = \delta_t + Z_{it}$$

$\delta_t = g * ([t-1]/[T-1])$  for linear growth  
Note:  $g$  is overall growth, which is what we estimate

$$Z_{it} = U_{it} + \varepsilon_{it} + \eta \varepsilon_{i(t-1)}$$

$U_{it}$  is an autoregressive term,  $U_{it} \sim N(\rho * U_{i(t-1)}, \tau_\theta)$   
 $\rho \sim \text{Unif}(-1, 1)$  is the autocorrelation  
 $\varepsilon_{it}$  is a random error,  $\varepsilon_{it} \sim N(0, \omega_\varepsilon)$   
 $\eta \sim \text{Unif}(-1, 1)$  is a moving average parameter

### 5.17 Model Formulation II



#### Model Formulation II



The model can be restated using two steps implemented at each time point  $t$

Step 1: AR(1) process (with no trend),

- When  $t = 1$  and there is no lagged value,  $Z_{it} \sim N(0, \tau_\theta)$
- When  $t > 1$ , we place a normal prior with different (hyper)parameters on this quantity, namely,  
$$Z_{it} \sim N(\rho \times Z_{i(t-1)}, \tau_\theta / (1 - \rho^2))$$

Step 2: The latent trait at time  $t$  is computed as an additive function of the estimated parameters:  $\theta_{it} = \delta_t + Z_{it}$



### 5.18 What about Rater Drift?



#### Rater Drift

- Most data designs **do not permit the decoupling** of changes in test takers and changes in raters; the times of the response and the measurement are completed conflated
- The longitudinal component of the L-HRM can be **applied to traits only, rater parameters only, or both** (if the design permits); see Casabianca, Lockwood, & McCaffrey (2015)



### 5.19 Other Parameterizations



#### Other Parameterizations

- Different **levels of invariance** in item parameters
- Different **types of growth structures**
  - ✓ Individual growth trends
  - ✓ Different types of trends
  - ✓ Unequally-spaced time points



## 5.20 Examples




### Examples

- Casabianca, J. M., Junker, B. W., Nieto, R., & Bond, M. A. (2017). A hierarchical rater model for longitudinal data. *Multivariate Behavioral Research*, 52(5), 576-592.
- Check back in future versions of this module!

## 5.21 Bookend: Multidimensional HRM







This is the end of this topic.

Topic Selection

## 5.22 Bookmark: Multidimensional HRM

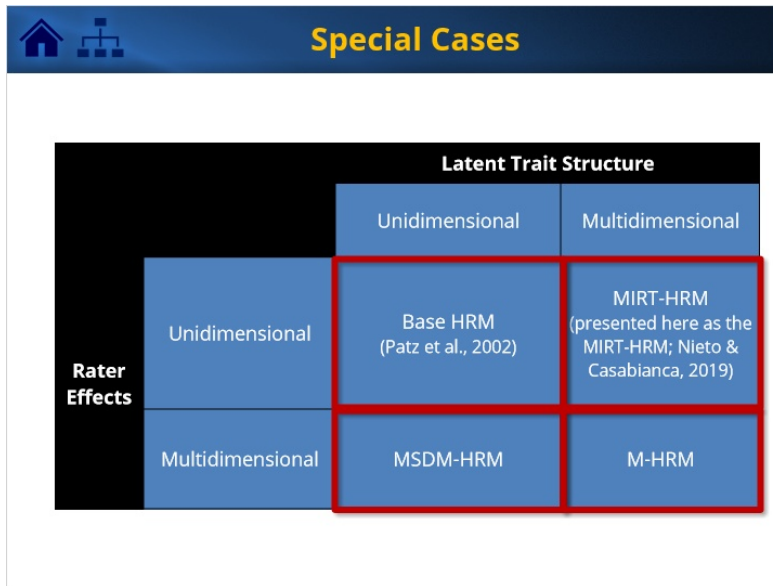


## 5.23 General Principles





  **General Principles**

- ▶ A multidimensional HRM (M-HRM) offers a framework for modeling these two processes
  1. Multidimensional IRT (MIRT) used to locate individuals in this measurement space
  2. A signal-detection process sensitive to dimension-specific rater effects
- ▶ Analysis should capture the complexity with which rating instruments are constructed, and address the measurement goals



## 5.24 Possible Special Cases Under this Multidimensional Framework



## 5.25 Model Variants


	<b>Base HRM</b> $\begin{cases} \theta_i \sim \mathcal{N}(\mu, \sigma^2) \\ \xi_{ij} \sim \text{unidimensional IRT model} \\ X_{ijr} \sim \text{unidimensional signal detection model} \end{cases}$
	<b>MIRT HRM</b> $\begin{cases} \theta_j \sim \mathcal{N}(\mu, \Sigma) \\ \xi_{ij} \sim \text{MIRT model} \\ X_{ijr} \sim \text{unidimensional signal detection model} \end{cases}$
	<b>MSDM HRM</b> $\begin{cases} \theta_i \sim \mathcal{N}(\mu, \sigma^2) \\ \xi_{ij} \sim \text{unidimensional IRT model} \\ X_{ijr} \sim \text{multidimensional signal detection model} \end{cases}$
	<b>Multidimensional HRM</b> $\begin{cases} \theta_i \sim \mathcal{N}(\mu, \Sigma) \\ \xi_{ij} \sim \text{MIRT model} \\ X_{ijr} \sim \text{multidimensional signal detection model} \end{cases}$

## 5.26 Subtopic Selection





**Primer on  
Multidimensional IRT**

**Parameterization of the  
Multidimensional HRM**

Click on the buttons to learn more. 

## 5.27 General Principles

**General Principles**

- ▶ MIRT models describe the interaction between multiple latent traits, characteristics of the items, and item responses
- ▶ These models relate the probability of responses to a location in the parameter space
- ▶ Goals
  1. Describe each individual using multiple latent trait scores, one per dimension
  2. Describe associations between items and individual dimensions using a set of item parameters



## 5.28 Multidimensional 2PL Model

Multidimensional 2PL Model

$$\text{logit}\{\Pr(y_{ij} = 1|\theta_i, \alpha_j, \delta_j)\} = \alpha_j \theta'_i + \delta_j$$

- ▶ Let the number of dimensions be denoted by  $m = (1, \dots, M)$
- ▶  $\alpha_j = \alpha_{jm}$ : vector of dimension-specific item discrimination parameters
- ▶  $\theta_i = \theta_{im}$ : vector of dimension-specific latent traits
- ▶  $\delta_j$ : (scalar) intercept parameter (related to item difficulty)
- ▶  $\alpha_j \theta'_i = \alpha_{j1} \theta_{i1} + \dots + \alpha_{jM} \theta_{iM}$

## 5.29 Multidimensional Generalized Partial Credit Model (MGPCM; Yao & Schwarz, 2006)

Multidimensional GPCM

$$\Pr(y_{ij} = k|\theta_i, \alpha_j, \gamma_{jk}) = \frac{\exp\{(k-1)\alpha_j \theta'_i - \sum_{k=1}^k \gamma_{jk}\}}{\sum_{h=1}^{K_j} \exp\{(k-1)\alpha_j \theta'_i - \sum_{k=1}^h \gamma_{jk}\}}$$

- ▶ Let  $k = (1, \dots, K)$  denote the category of a polytomous item  $j$  with  $K$  total categories
- ▶  $\theta_i$  and  $\alpha_j$  as previously defined
- ▶  $\gamma_{jk}$ : threshold parameters, assumed constant across dimensions
- ▶ This parameterization of the MGPCM is a multidimensional variation of the nominal response model

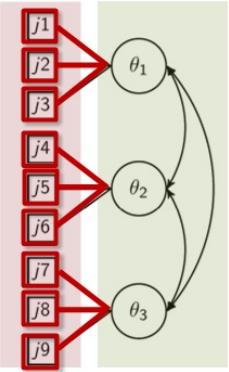
### 5.30 Multidimensional Generalized Partial Credit Model (MGPC; Yao & Schwarz, 2006)

Types of Multidimensionality

- ▶ This distinction refers to how dimensions influence item responses
- ▶ **Between-item dimensionality**—each item measure a *single* dimension
  - ▶ An item response is influenced by a single dimension
  - ▶  $\alpha_j$  contains at most one non-zero value
  - ▶ AKA: simple factorial structure
- ▶ **Within-item dimensionality**—each item measures *multiple* constructs
  - ▶ Responses influenced by a composite of dimensions
  - ▶  $\alpha_j$  may contain more than one non-zero value
  - ▶ AKA: complex factorial structure

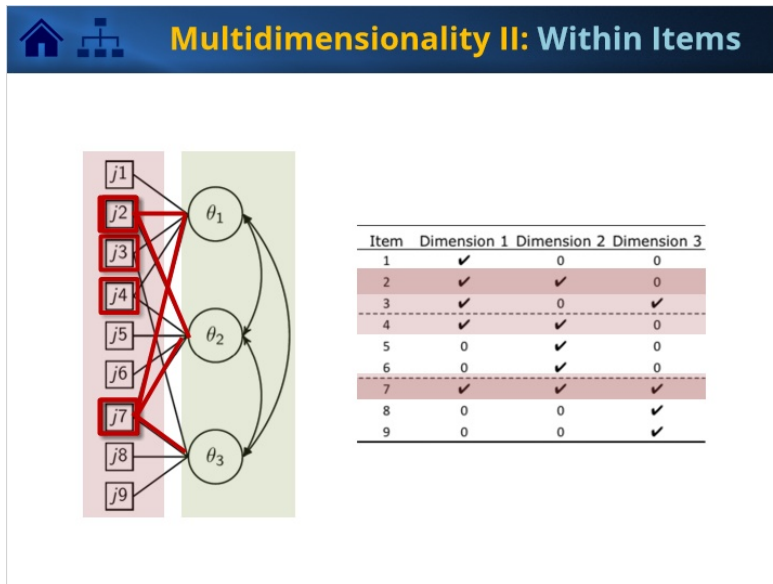
### 5.31 Between-Item Dimensionality

Multidimensionality I: Between Items




Item	Dimension 1	Dimension 2	Dimension 3
1	✓	0	0
2	✓	0	0
3	✓	0	0
4	0	✓	0
5	0	✓	0
6	0	✓	0
7	0	0	✓
8	0	0	✓
9	0	0	✓

### 5.32 Within-Item Dimensionality





### 5.33 Modeling Approaches: Confirmatory vs. Exploratory

 **Modeling Approaches**

- ▶ Describes the manner in which MIRT models are applied, and is analogous to the distinction in factor analysis/structural equation modeling
- ▶ **Confirmatory** approach assumes the number of dimensions, and their relationship with items, are known a priori
- ▶ **Exploratory** approach allows one to find the number of dimensions empirically

### 5.34 Modeling Approaches: Confirmatory vs. Exploratory






More on MIRT

- Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22(3), 37-51.

- Check back in future versions of this module!

### 5.35 Bookend: Primer on Multidimensional IRT







This is the end of this subtopic.

Subtopic Selection

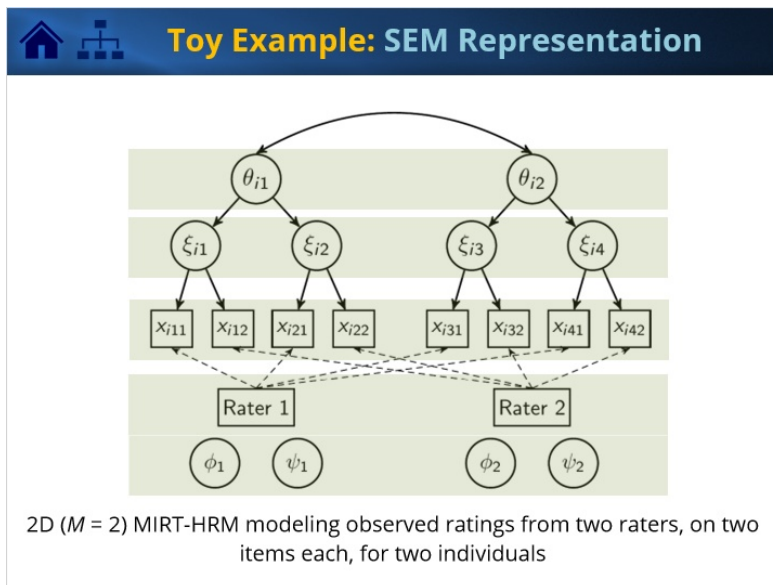
### 5.36 General Principles





## General Principles

- ▶ Within the MIRT-HRM, the MGPCM is used to model *ideal ratings*
- ▶ Accounts for the correlation among dimensions by jointly modeling the covariance structure associated with the distribution of the latent traits
  - ▶ I.e.,  $\theta_i \sim \mathcal{N}(\mu, \Sigma)$
- ▶ For now, assumes rater effects are constant across dimensions for each rater (i.e., we are focusing on the MIRT-HRM)

### 5.37 2D ( $M = 2$ ) MIRT-HRM modeling observed ratings from two raters, on two items each, for two individuals



### 5.38 Parameterizing the MIRT-HRM





## Parameterization

$$\theta_i \sim \mathcal{N}(\mu, \Sigma)$$

$$\Pr(\xi_{ij} = \xi | \theta_i, \alpha_j, \gamma_{j\xi}) = \frac{\exp\{(\xi - 1)\alpha_j\theta'_i - \sum_{k=1}^{\xi} \gamma_{jk}\}}{\sum_{h=1}^{K_j} \exp\{(\xi - 1)\alpha_j\theta'_i - \sum_{k=1}^h \gamma_{jk}\}}$$

$$\Pr(X_{ijr} = k | \xi_{ij} = \xi) \propto \exp\left\{-\frac{[k - (\xi + \phi_r)]^2}{2\psi_r^2}\right\}$$

### 5.39 Alternatives I





## Alternatives I

What are the Alternatives?

- ▶ Option 1: Assume unidimensionality and fit a base HRM
- ▶ Option 2: Take the “consecutive approach” and fit multiple base HRMs



## 5.40 Alternatives II



### Alternatives II: Fit Single Base HRM

- ▶ This implies ignoring the structure of the instrument and treating item responses as measuring a single skill or ability
- ▶ Limits the quality of diagnostic information available for individuals
- ▶ Potentially disregards the intended purpose of the instrument
- ▶ Ignores associations among dimensions



## 5.41 Alternatives II



### Alternatives II: Fit Multiple Base HRMs

- ▶ Fit multiple base HRMs, one for each dimension
- ▶ Leads to multiple latent trait scores, and thus reflects the structure of the instrument
- ▶ However, ignores the associations among dimensions
  - ▶ When only few items measure each dimension, the correlations among dimensions serve as collateral information and improve precision of estimates (de la Torre & Patz, 2005; Wu & Wang, 2016)
  - ▶ Note that this is a common scenario for rating assessments

### 5.42 The M-HRM Thus Far






## Examples

- Nieto, R., & Casabianca, J. M. (2019). Accounting for rater effects with the hierarchical rater model framework when scoring simple structured constructed response tests. *Journal of Educational Measurement*, 56(3), 547-581.

- Check back in future versions of this module!

### 5.43 Bookend: Parameterization of the Multidimensional HRM



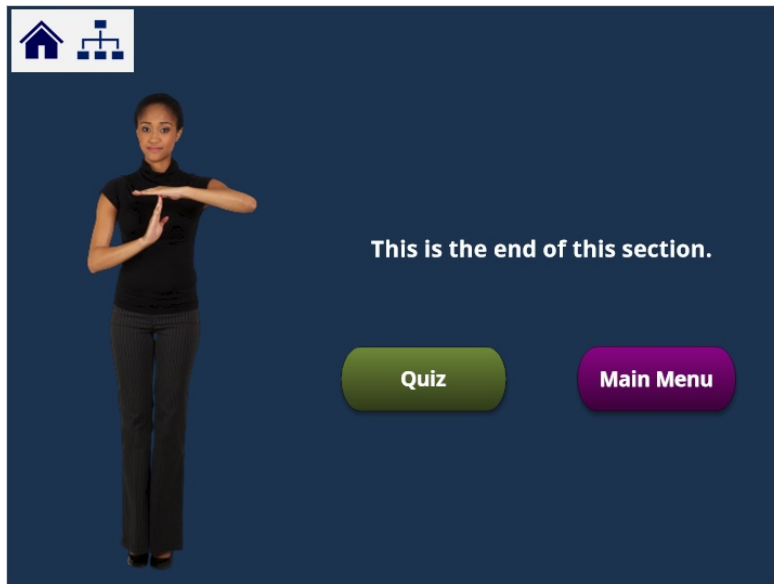


This is the end of this subtopic.

Subtopic Selection



#### 5.44 Bookend: Section 4



#### 5.45 Module Cover (END)

