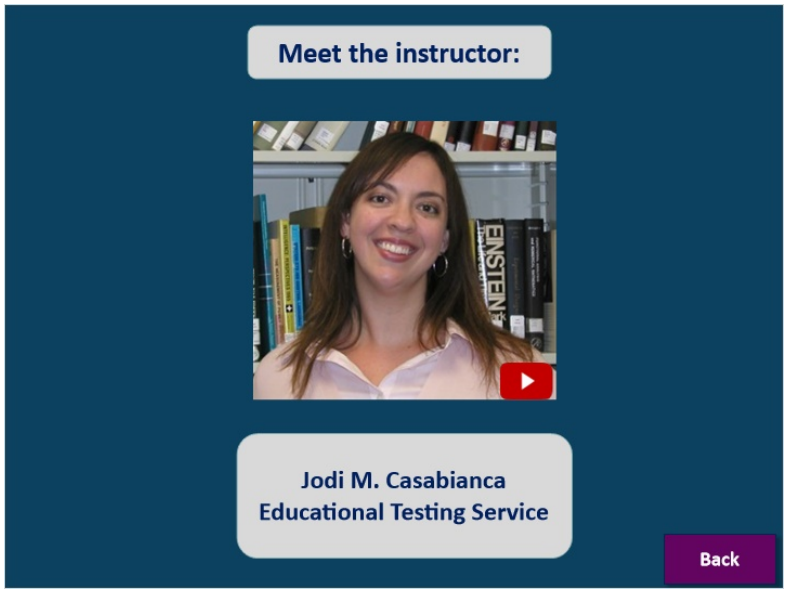
DM27 SLIDES (Rater Models, Version 1.0)

1. Module Overview

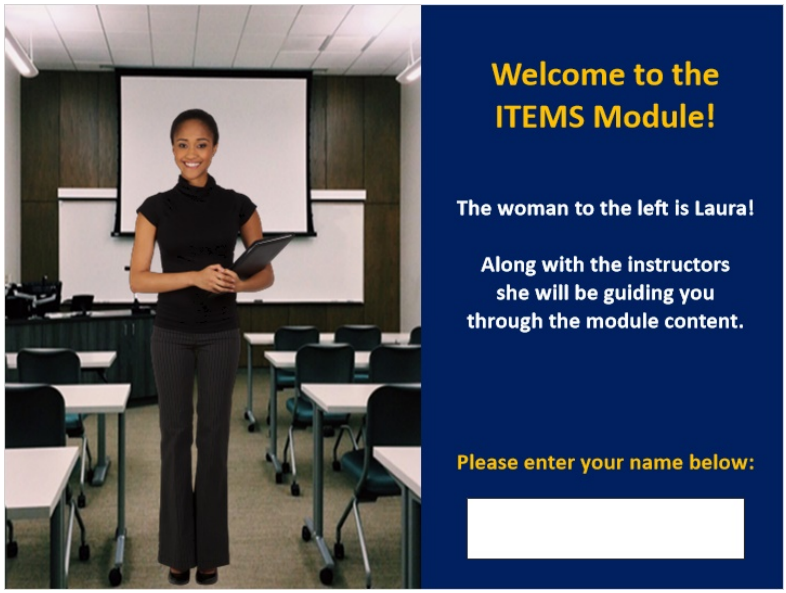
1.1 Module Cover (START)

1.2 Instructor

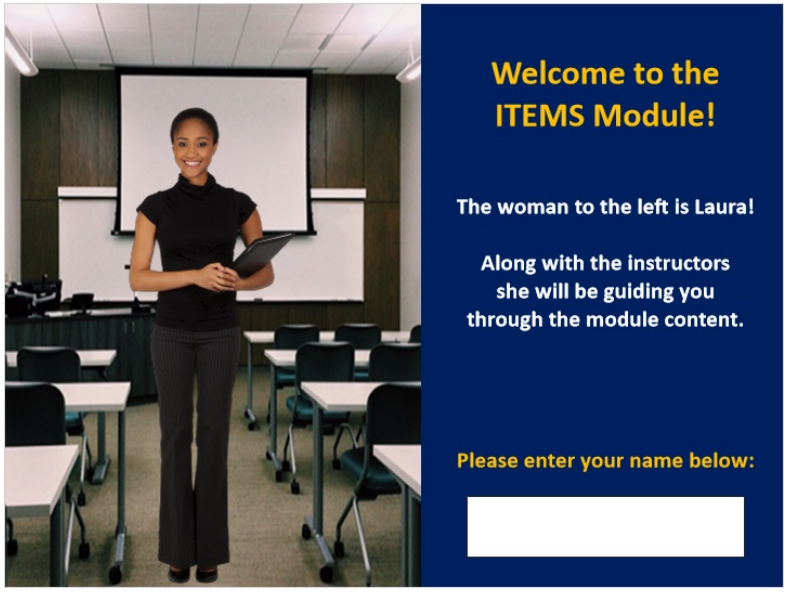


1.3 Designer

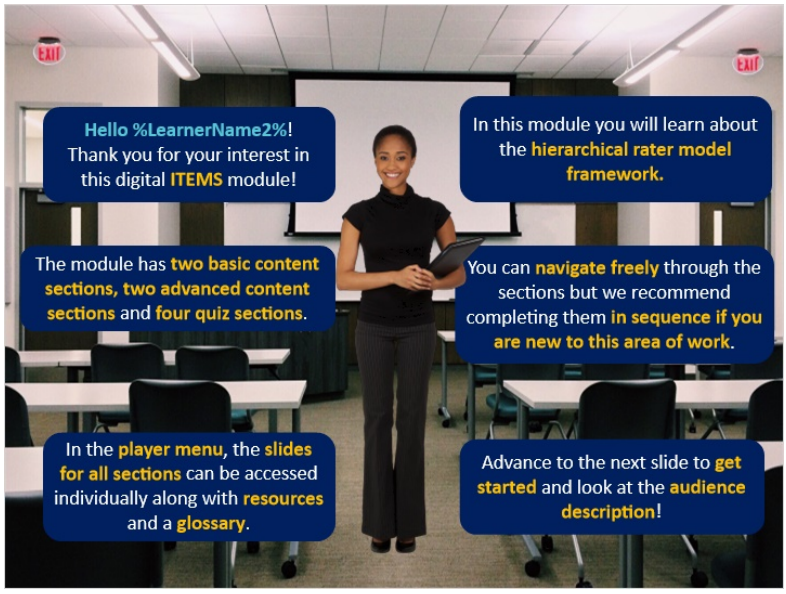
1.4 Welcome



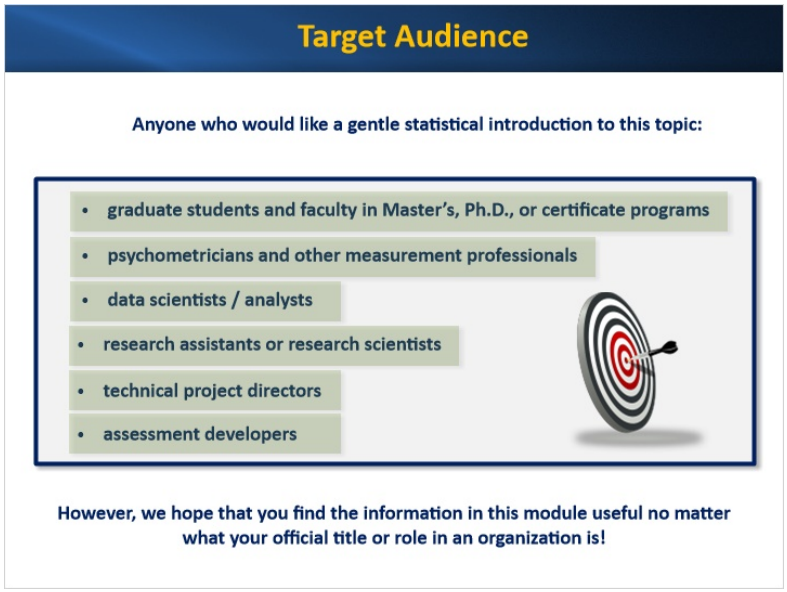
Untitled Layer 1 (Slide Layer)



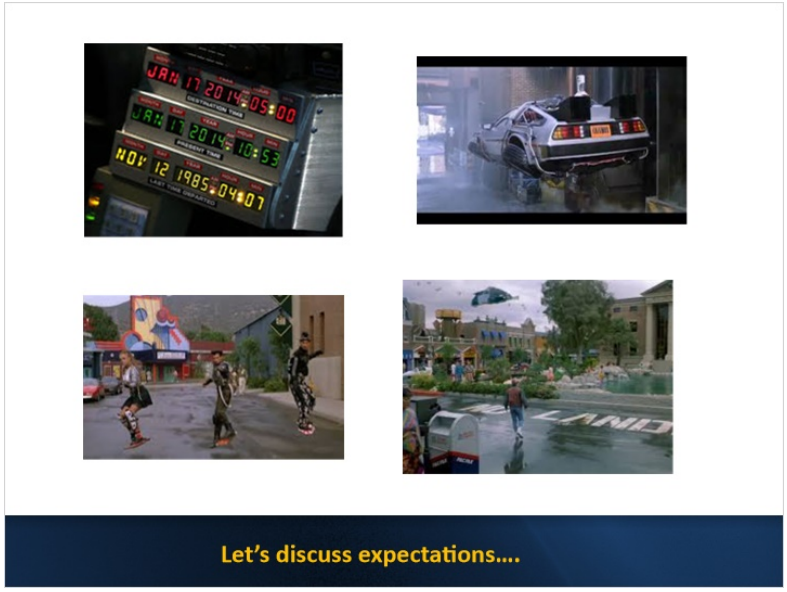
1.5 Overview



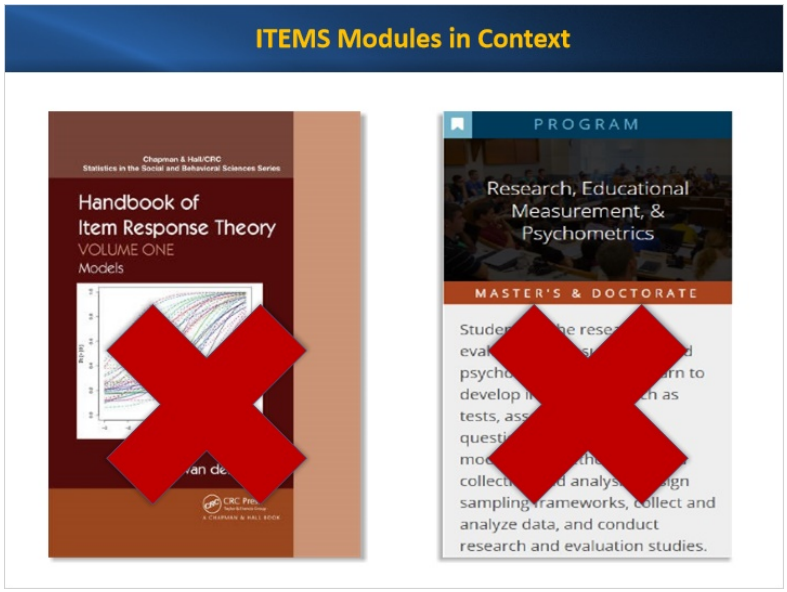
1.6 Target Audience



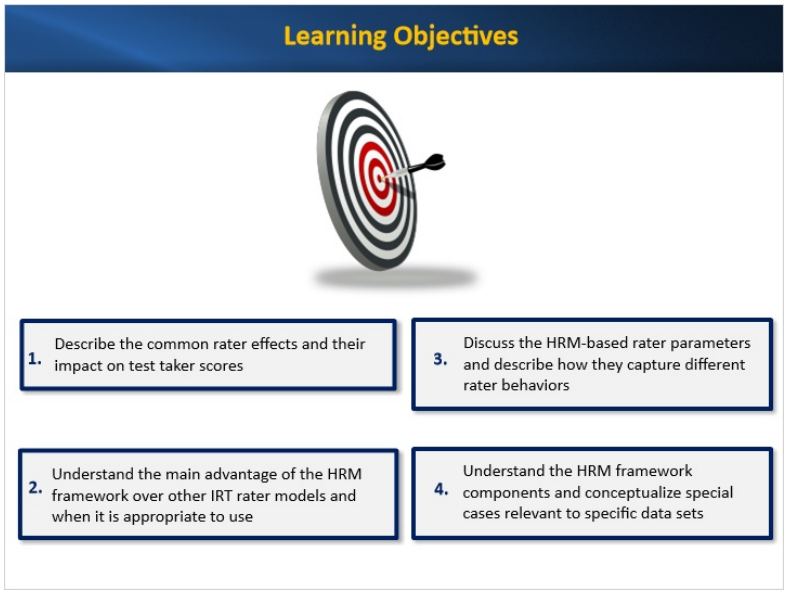
1.7 Expecations (I)



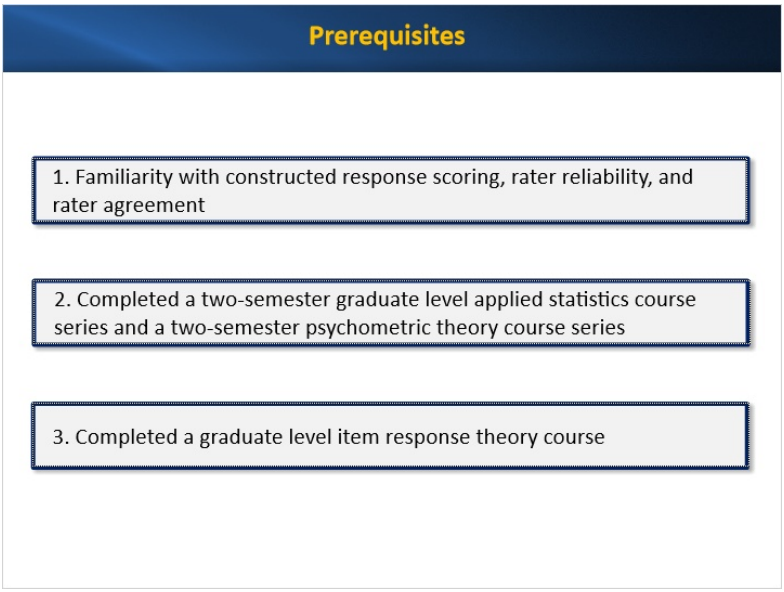
1.8 Expectations (II)



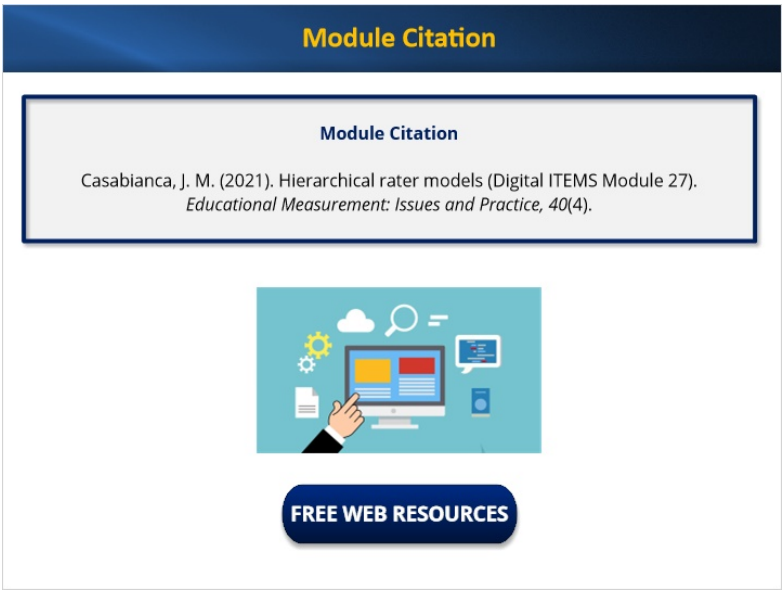
1.9 Learning Objectives



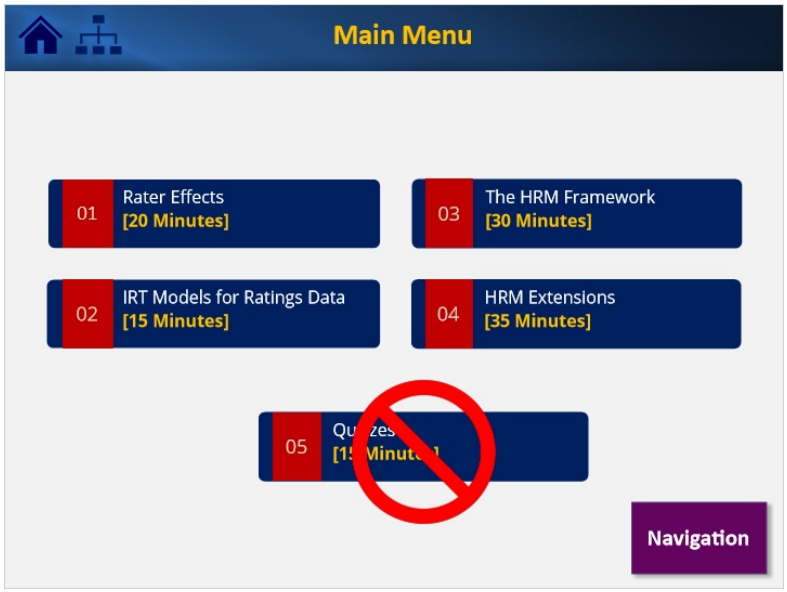
1.10 Prerequisites



1.11 Module Citation

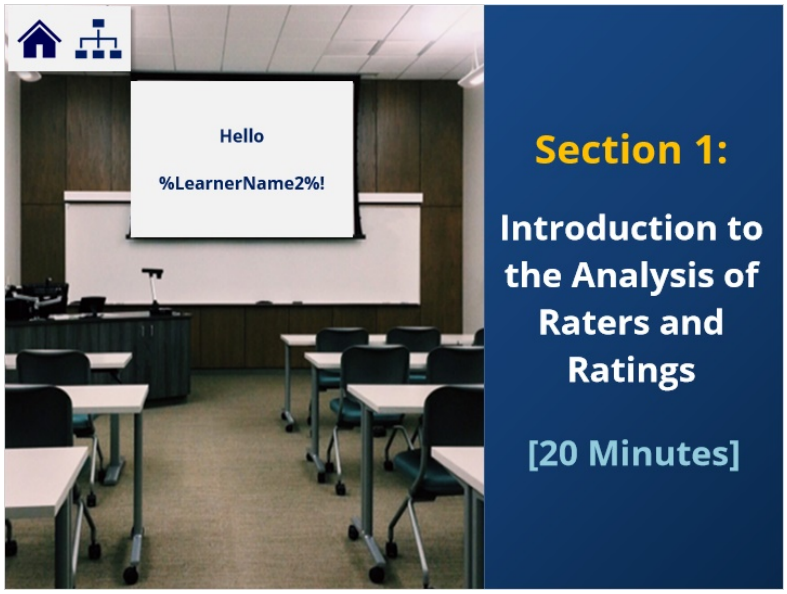


1.12 Main Menu

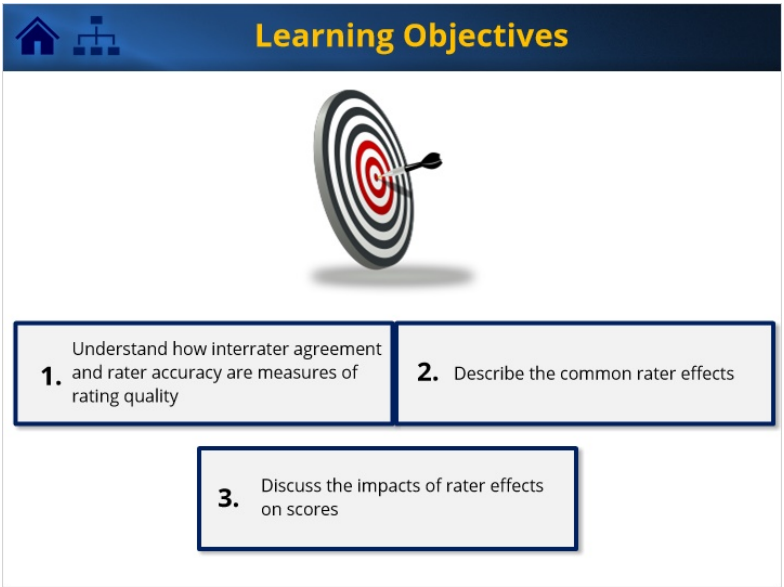


2. Section 1: Rater Effects

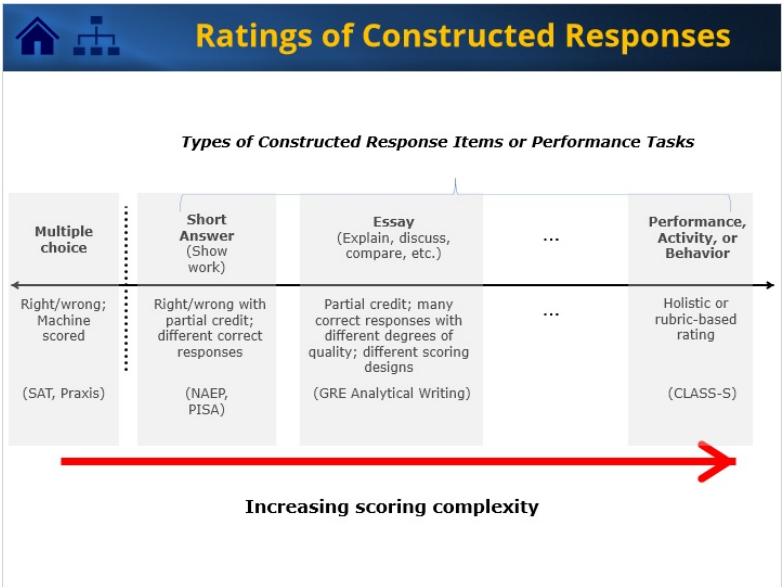
2.1 Cover: Section 1



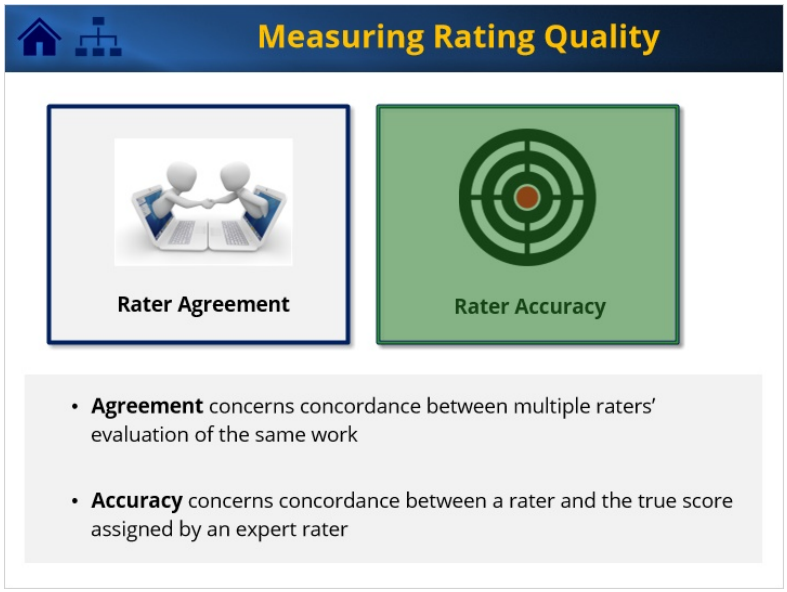
2.2 Objectives: Section 2



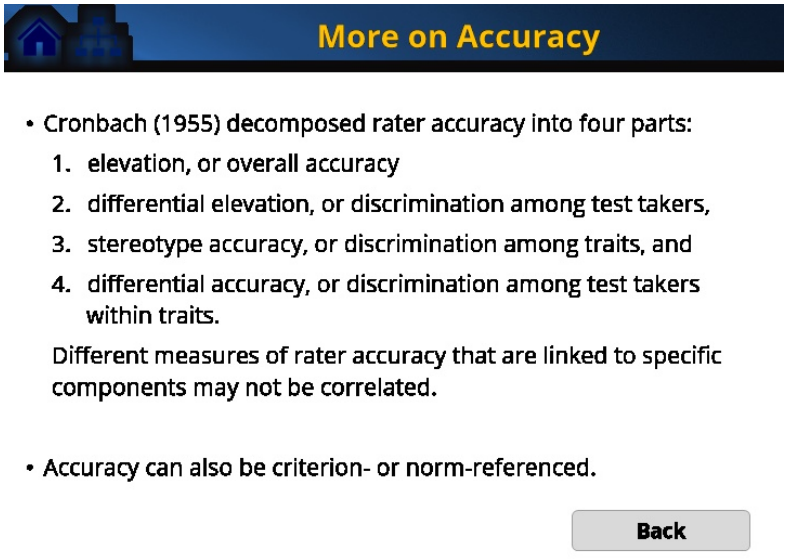
2.3 Tasks with Ratings



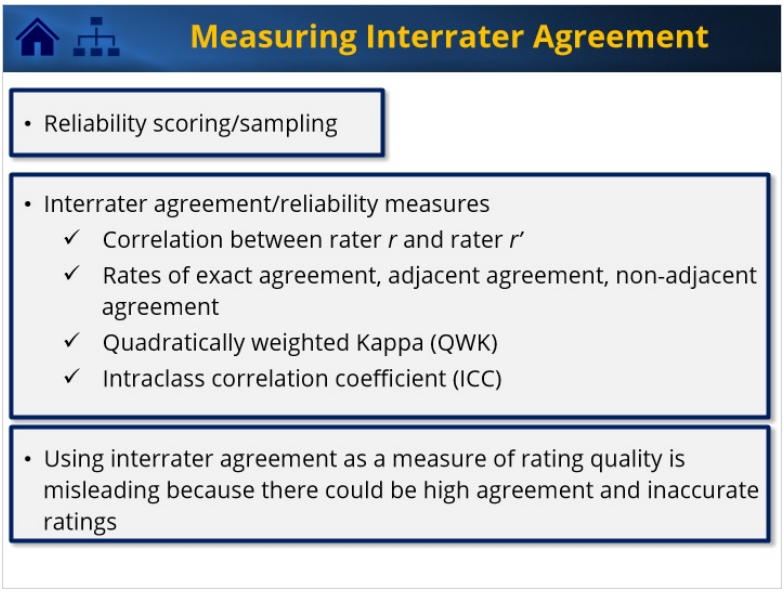
2.4 Agreement vs. Accuracy



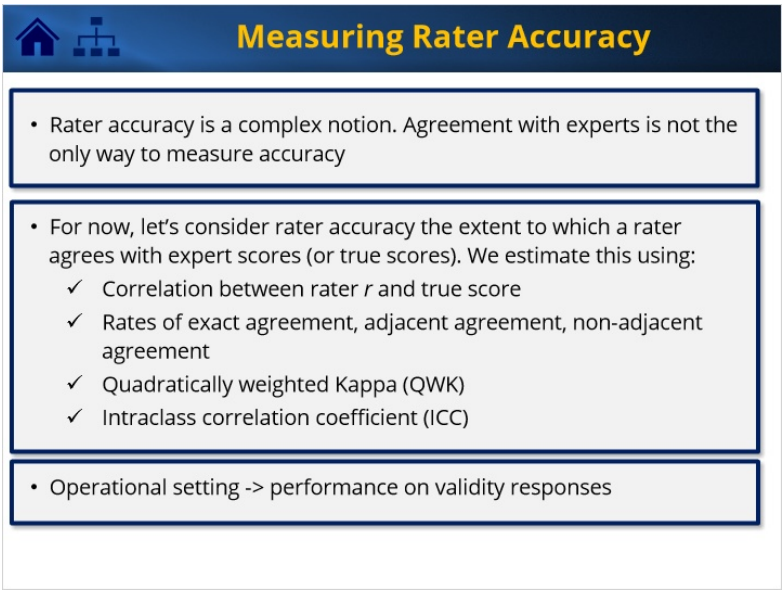
Accuracy (Slide Layer)



2.5 Measuring Interrater Agreement



2.6 Measuring Rater Accuracy



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*(observed,true)

2.10 Impact on Scores (II)

Impact on Scores (II)

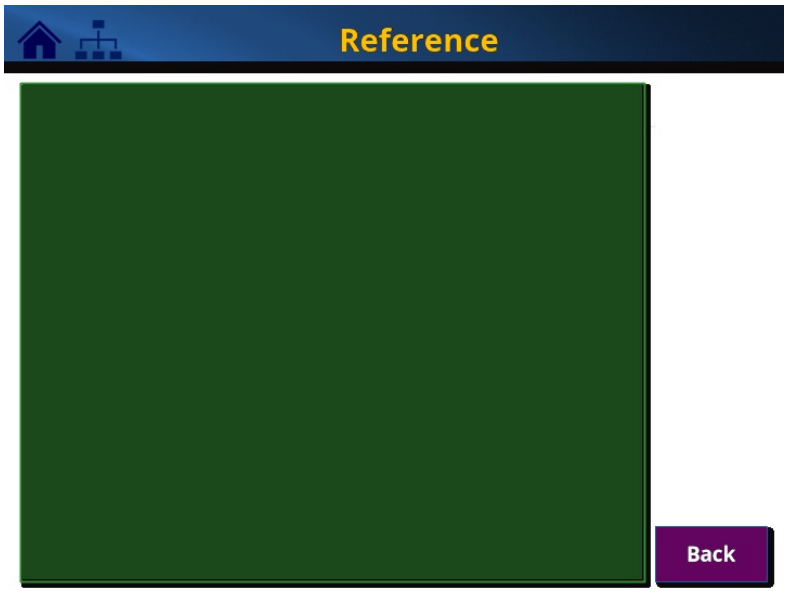
Table 2. Descriptive Statistics for Simulated Raw Scores by Rater Effect Group

		Score Level							
Group	*N*	0	1	2	3	4	Mean	*SD*	*r*
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

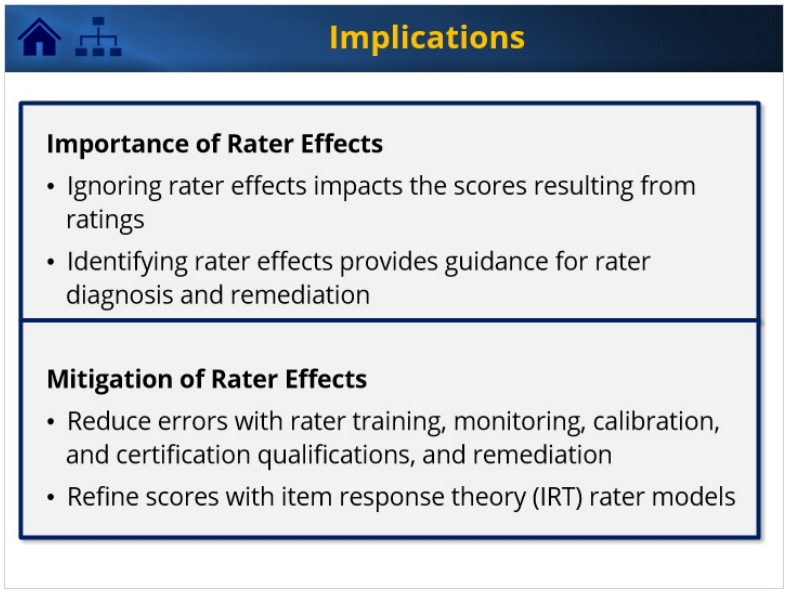
Note: *r* = the correlation between the average of the scores assigned to each examinee by raters in the Normal group and the scores assigned by a single randomly chosen rater the group represented by the row of the table.

Reference

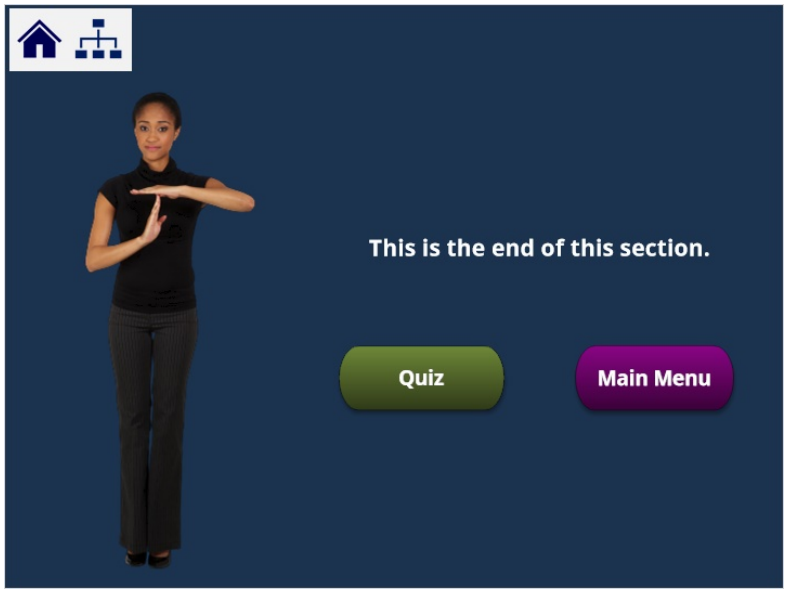
Reference (Slide Layer)



2.11 Implications

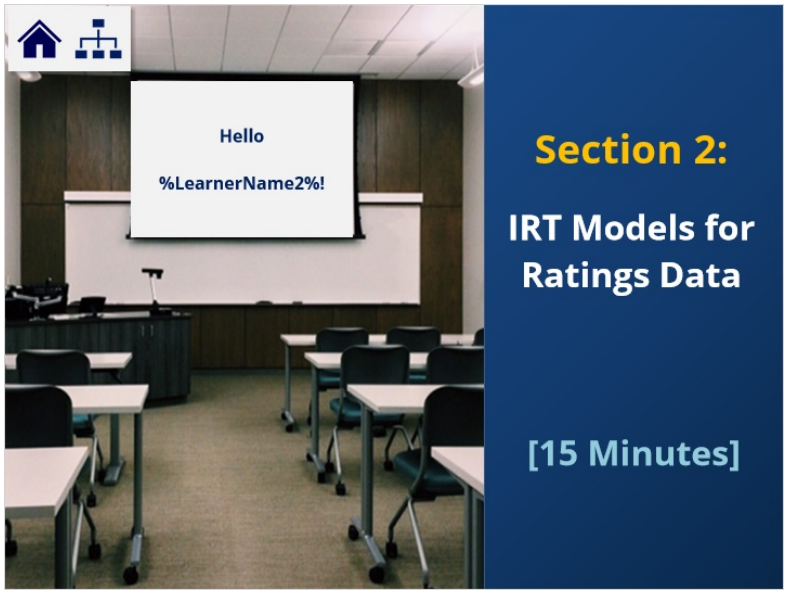


2.12 Bookend: Section 1

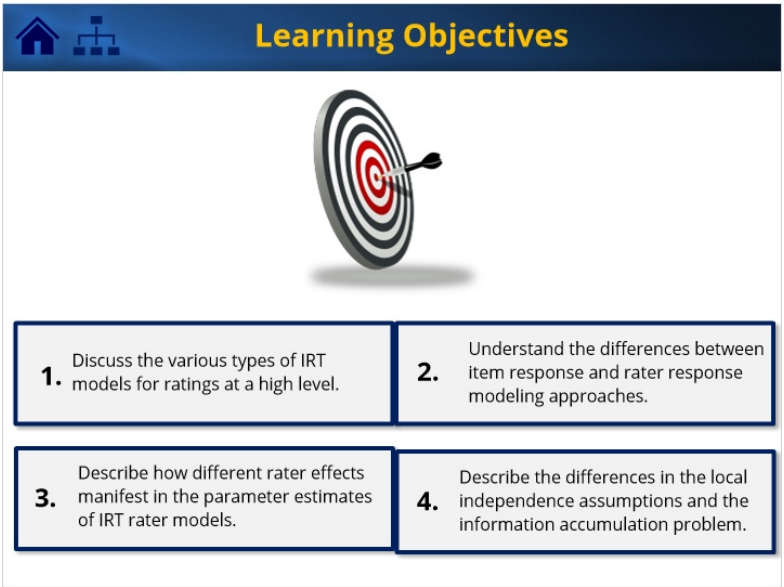


3. Section 2: IRT Models for Ratings Data

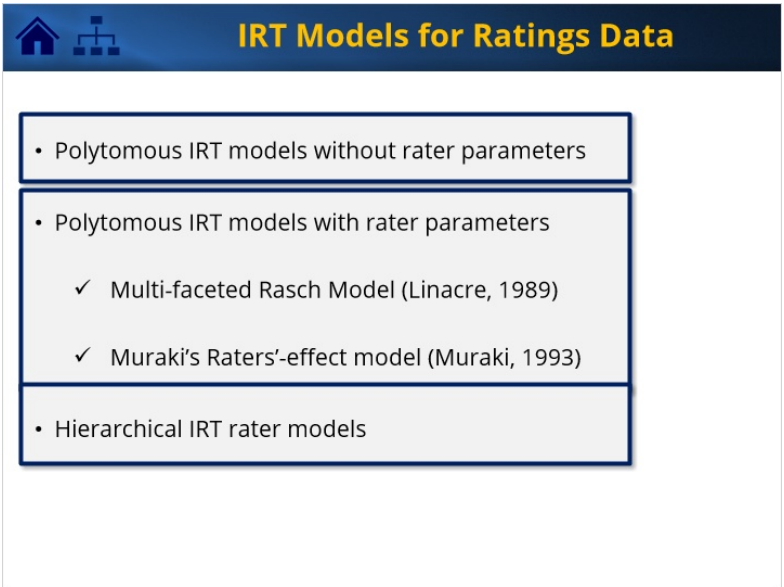
3.1 Cover: Section 2



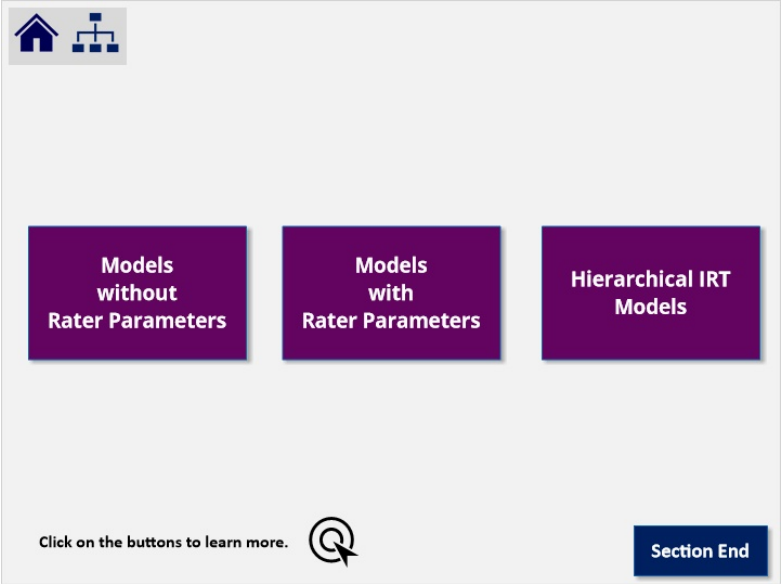
3.2 Objectives: Section 2



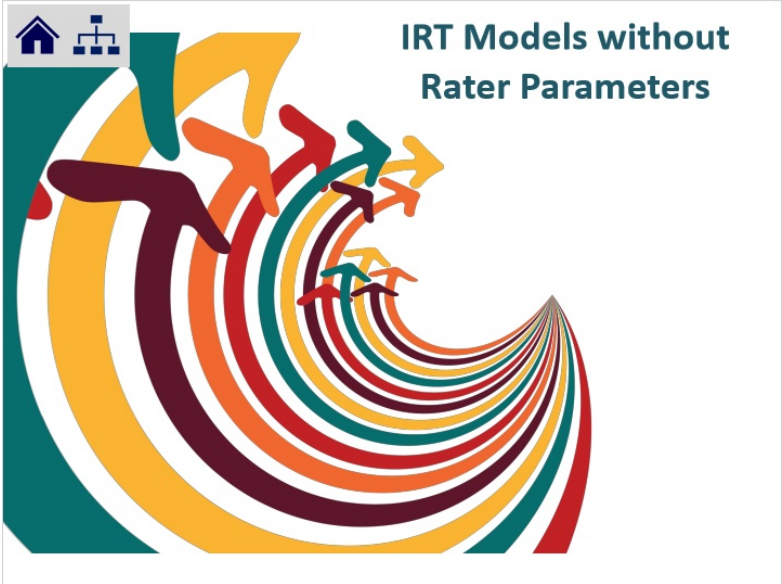
3.3 IRT Models for Ratings Data



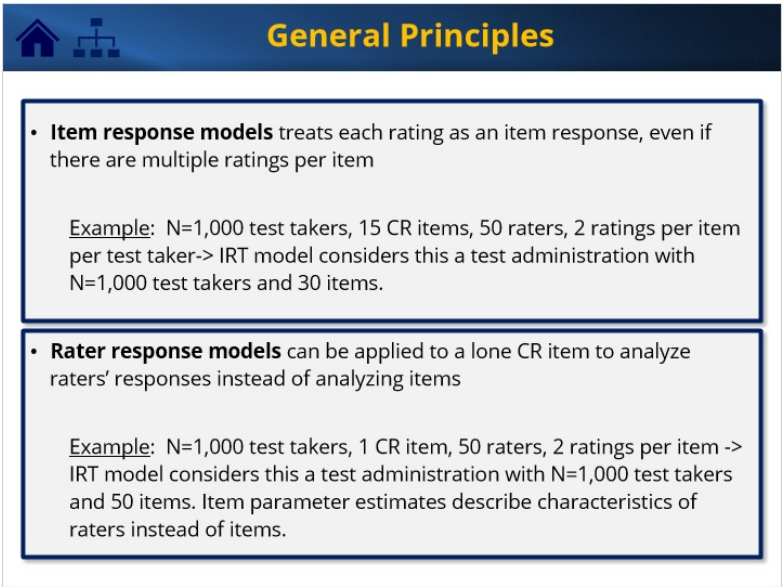
3.4 Topic Selection



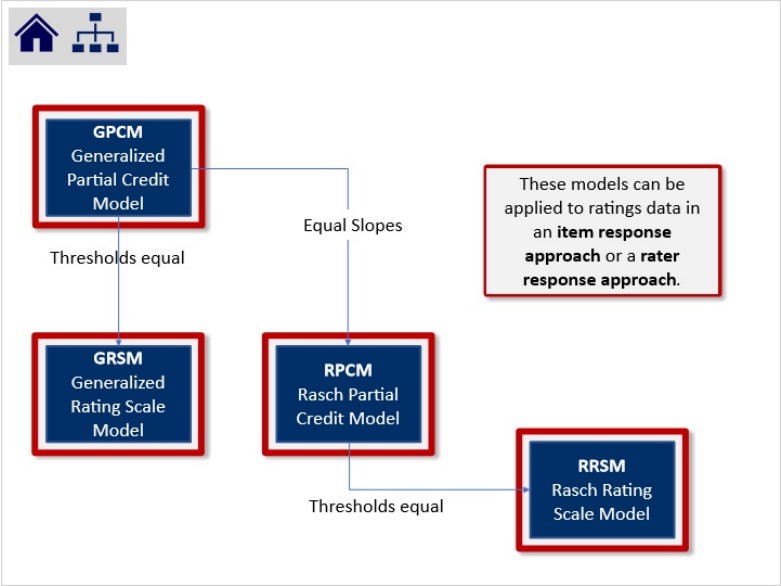
3.5 Bookmark: Polytomous Models w/o Rater Parameters



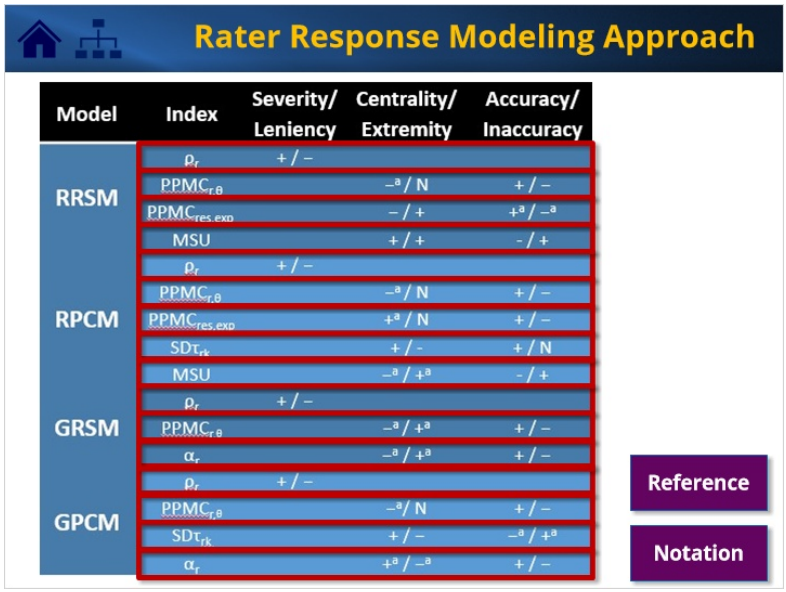
3.6 General Principles



3.7 Model Comparison



3.8 Rater Response Modeling Approach

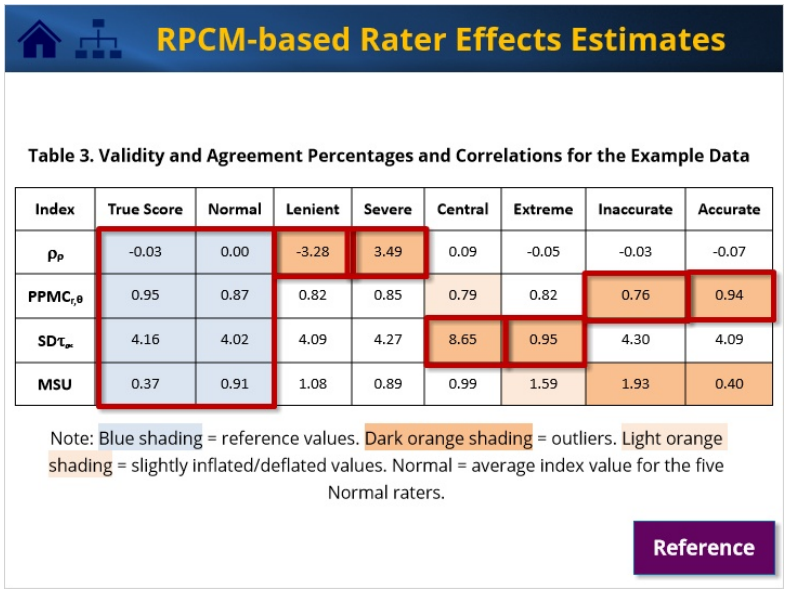


Reference (Slide Layer)

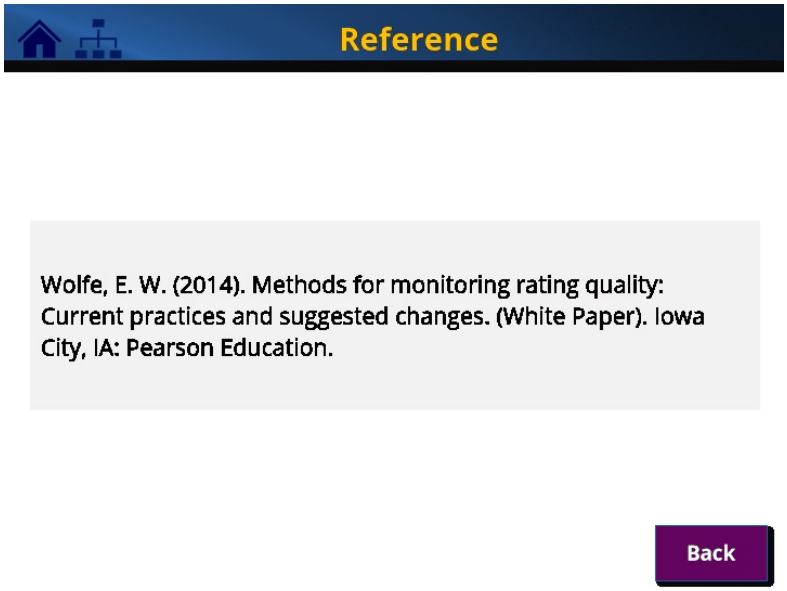
Wolfe, E. W. (2014). Methods for monitoring rating quality: Current practices and suggested changes. (White Paper). Iowa City, IA: Pearson Education.

Back

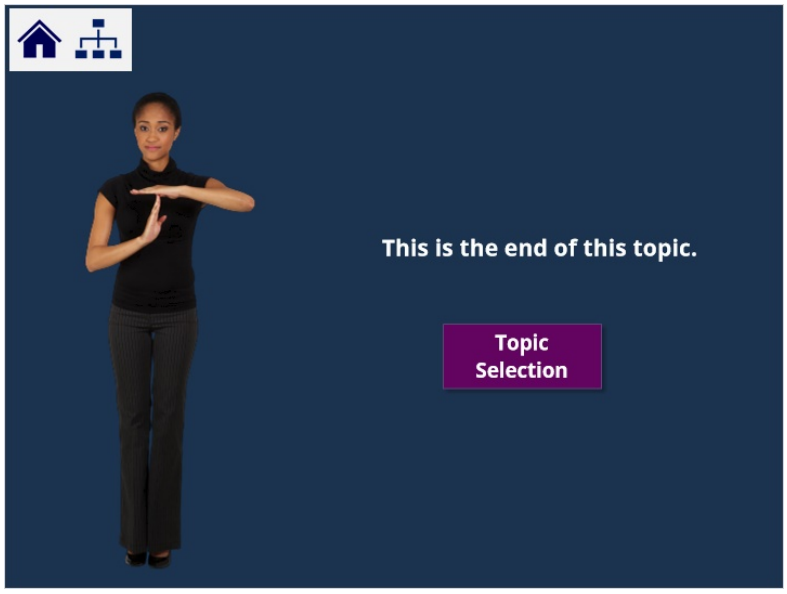
3.9 Using IRT Model Parameter Estimates for Rater Diagnosis



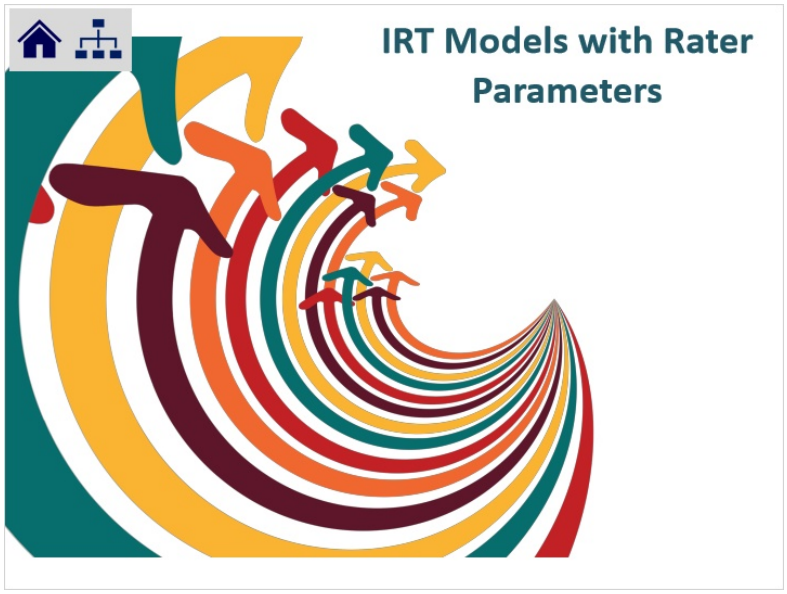
Reference (Slide Layer)



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)

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

where ϕ_r is a **rater severity parameter**.

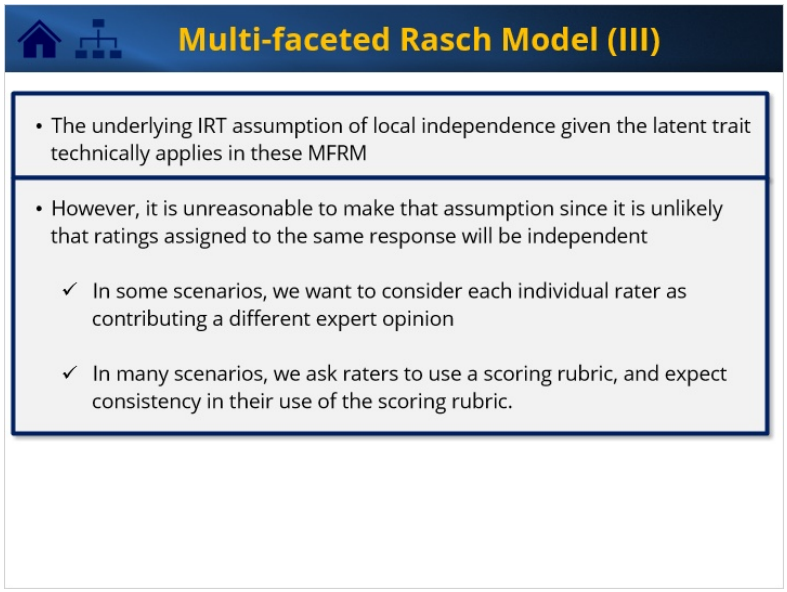
- **Muraki's rater effect model** (Muraki, 1993) generalizes the facets model by adding a discrimination parameter (a-parameter) for the item.

3.13 Multi-faceted Rasch Model (II)

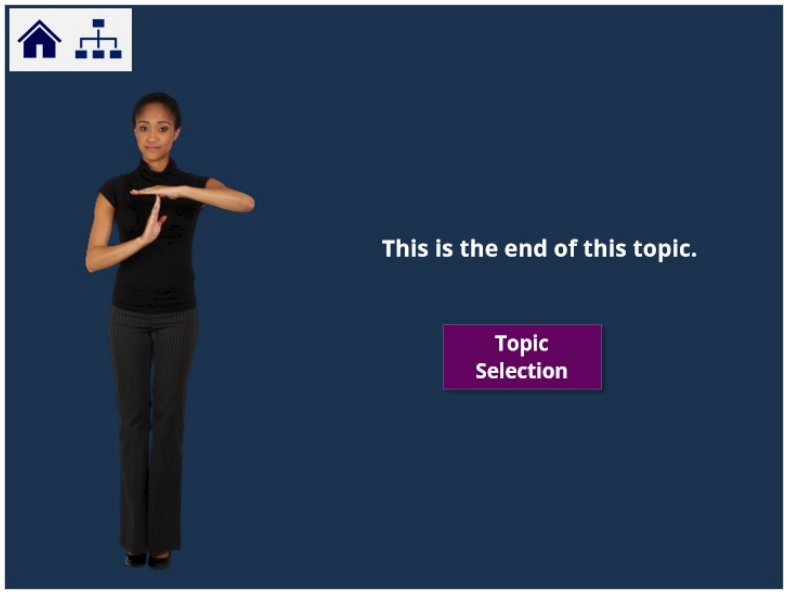
Multi-faceted Rasch Model (II)

- 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, γ_{jrk}, 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)

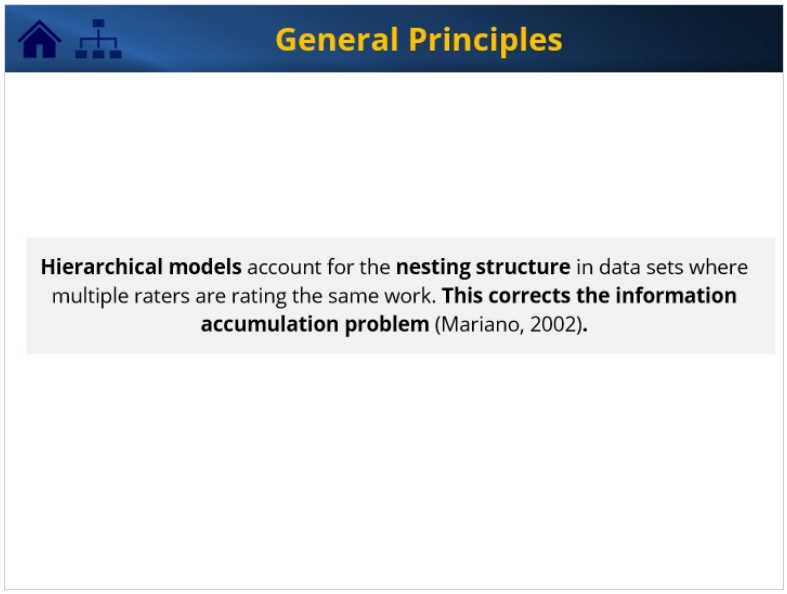


3.15 Bookend: Polytomous Models with Rater Parameters

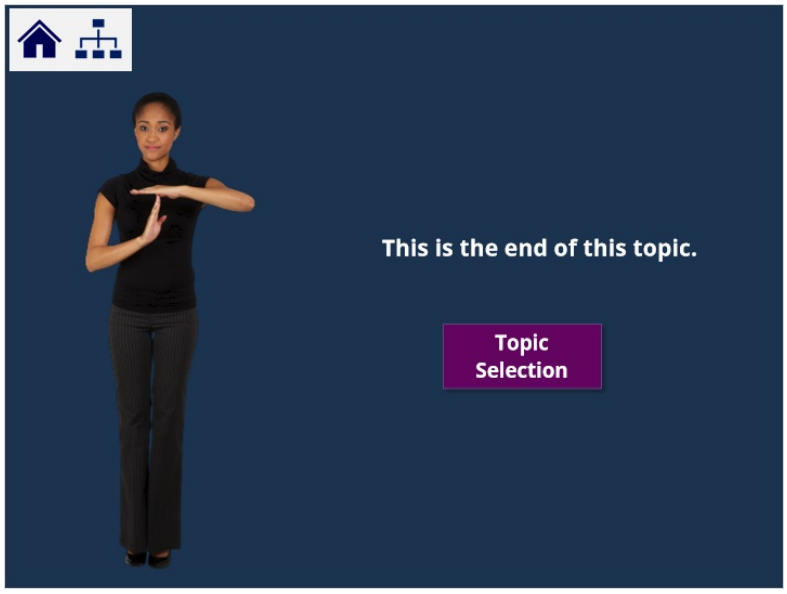


3.16 Bookmark: Hierarchical IRT Models

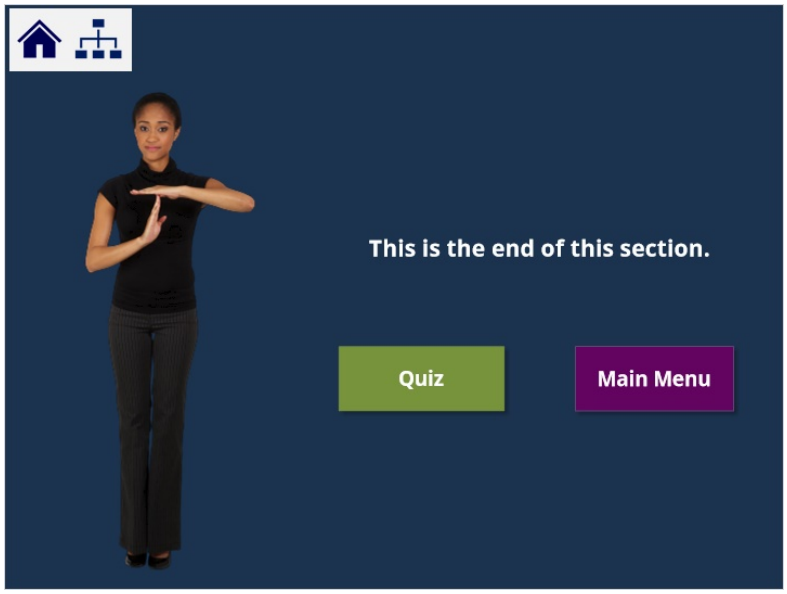
3.17 General Principles



3.18 Bookend: Hierarchical IRT Models



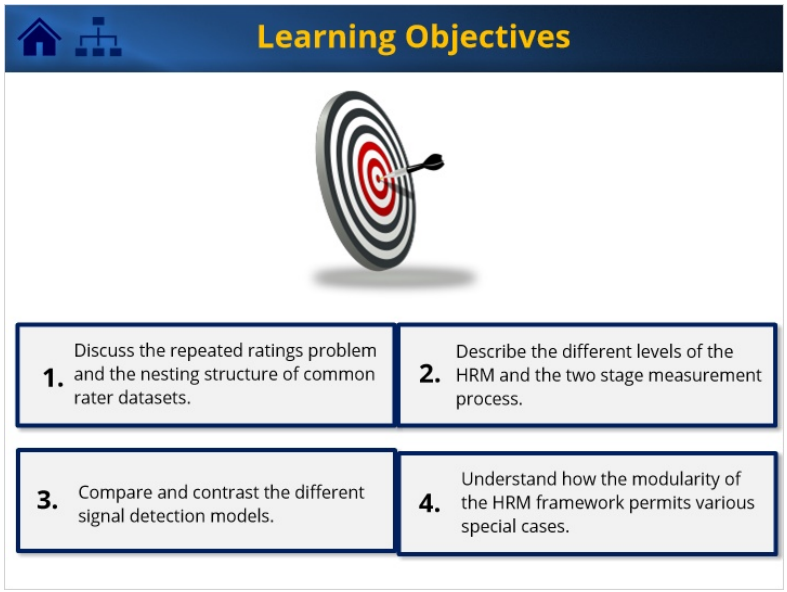
3.19 Bookend: Section 2



4. Section 3: The HRM Framework

4.1 Cover: Section 3

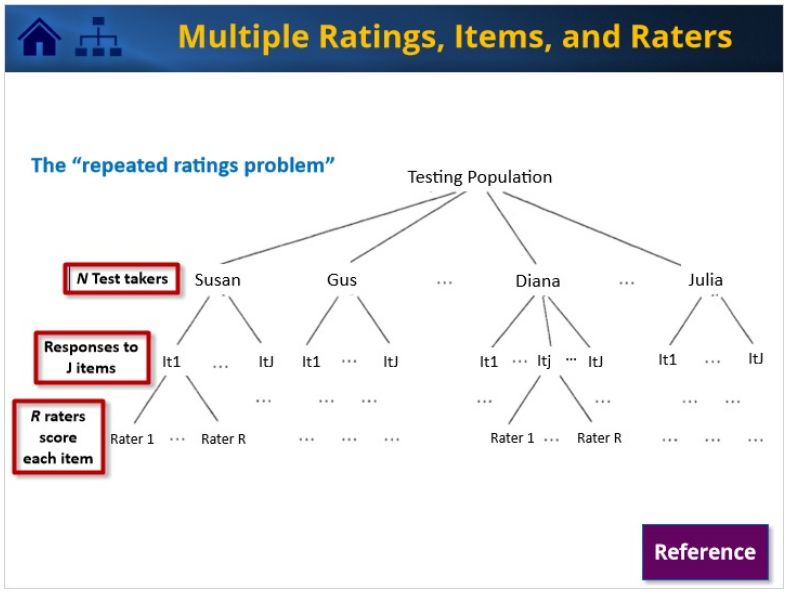
4.2 Objectives: Section 3



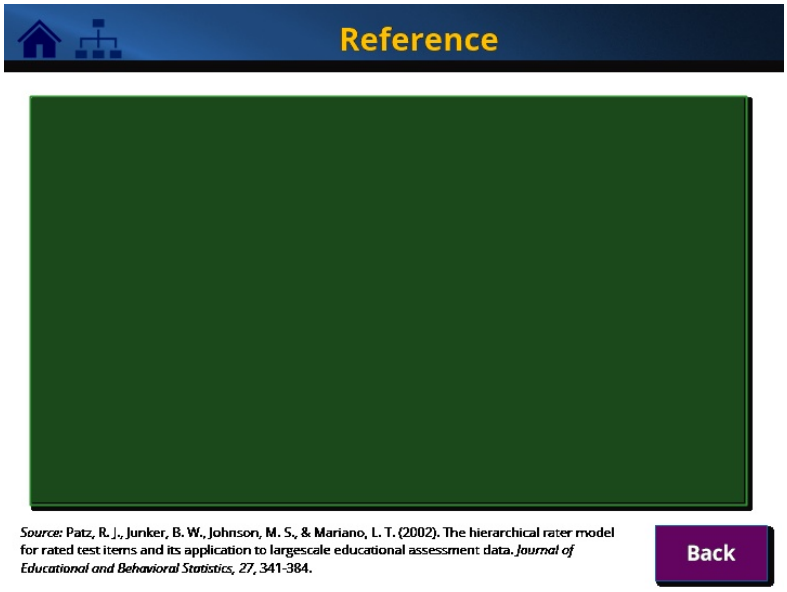
4.3 Topic Selection

4.4 Bookmark: HRM Foundations

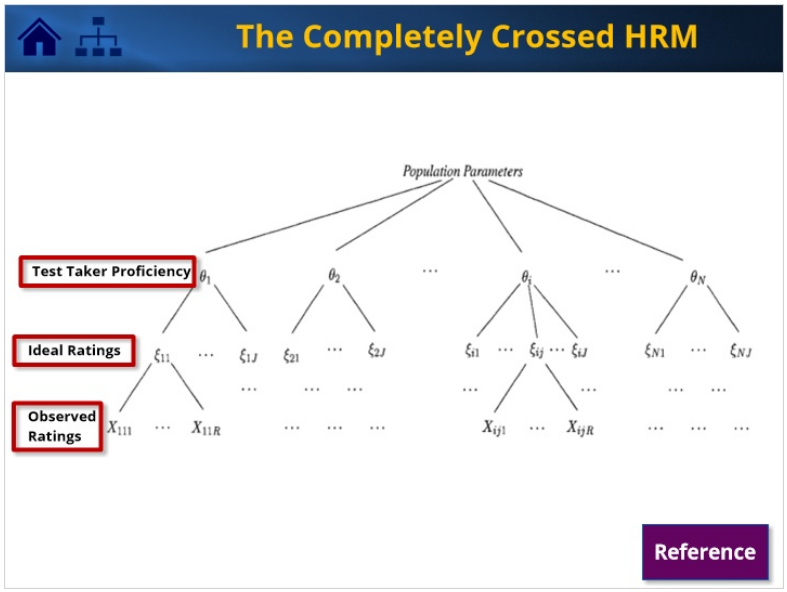
4.5 The Completely Crossed HRM



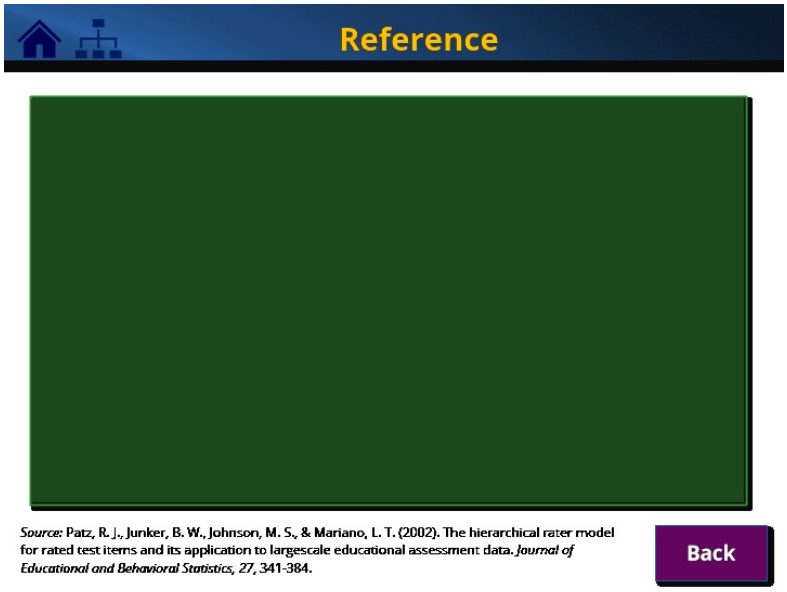
Reference (Slide Layer)



4.6 The Completely Crossed HRM



Reference (Slide Layer)



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

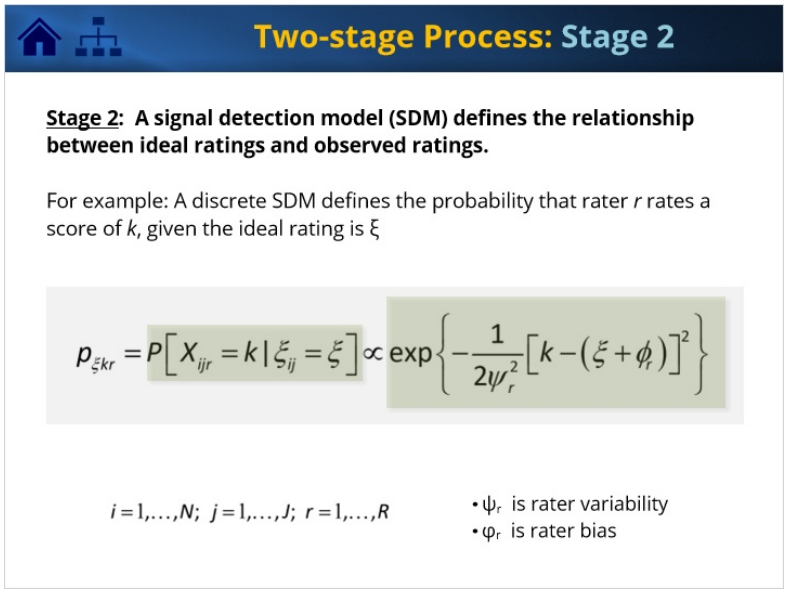
Example: *K*-category Generalized Partial Credit Model (GPCM; Muraki, 1992) relates test taker proficiency to ideal ratings

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

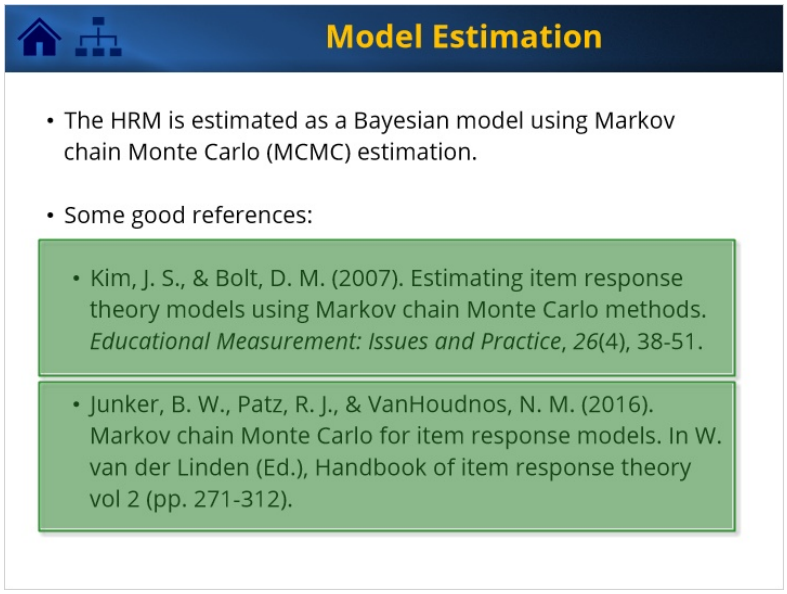
$i=1,\ldots,N;\ j=1,\ldots,J;\ k=1,\ldots,K$

- ξ_{ij} is ideal rating for test taker i on item j
- α_j is item discrimination
- β_j is item difficulty
- γ_{jk} step parameter at category k

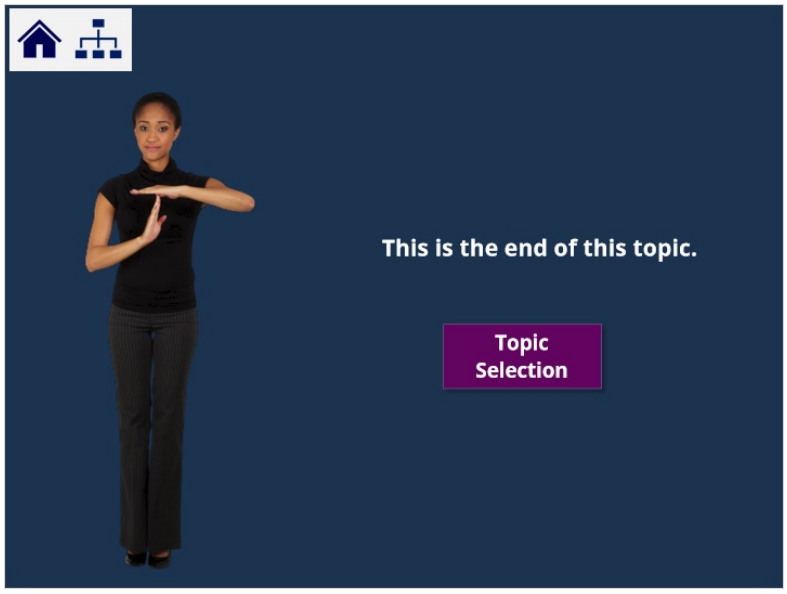
4.9 Two-stage Model II



4.10 Two-stage Model II



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[Rater *r* rates *k* | ideal rating ξ] in each row of this matrix

Ideal Rating (ξ)	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**:

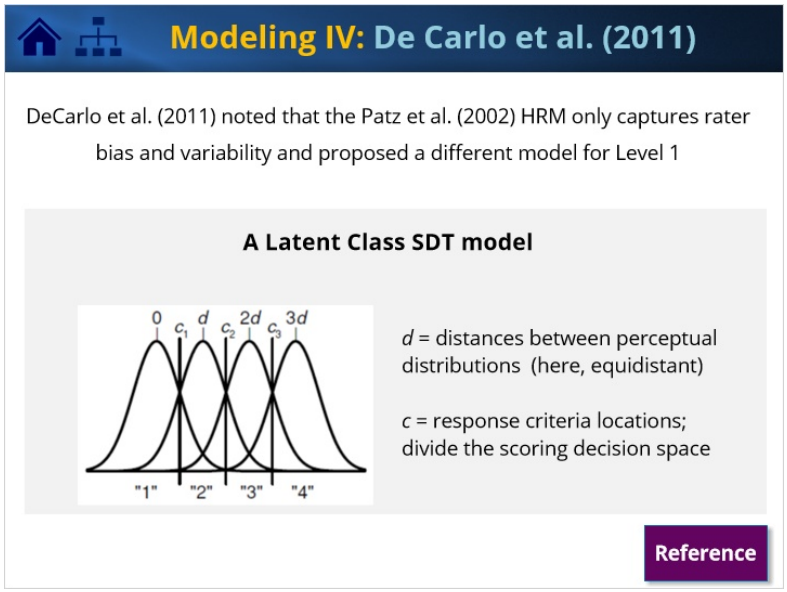
The normal density:

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

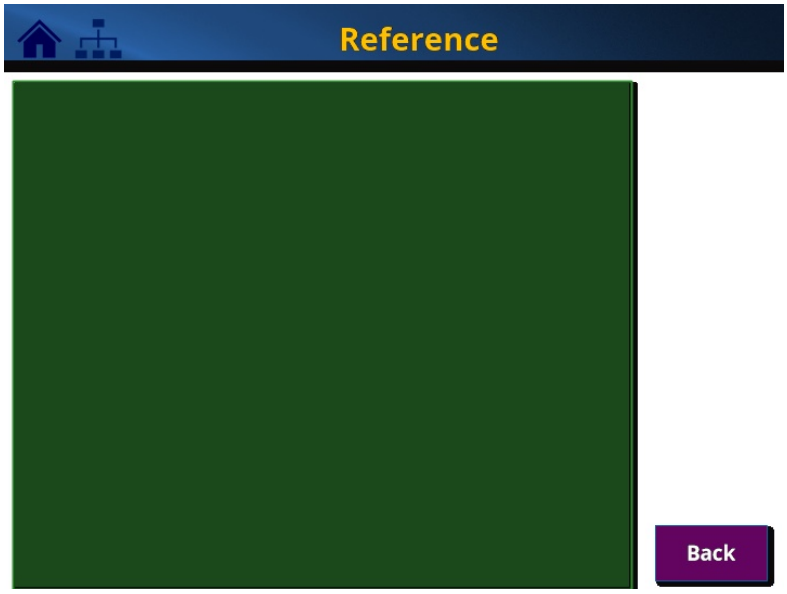
The discrete SDM in the Patz et al. (2002) HRM:

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

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



Reference (Slide Layer)



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

Modeling V: De Carlo et al. (2011)

Latent Class SDT model

$$p\left(X_{rj} \le k \mid \xi_j = \xi\right) = F\left(c_{rkj} - d_{rj}\xi_j\right)$$

- 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, $c_{r1j} < c_{r1j} < \ldots < c_{r,K-1,j}$
- 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

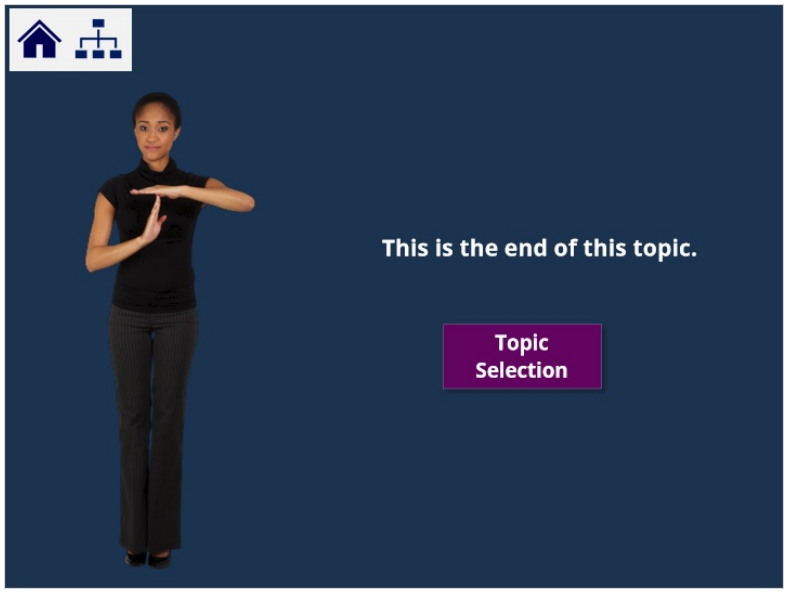
Model Selection

"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 θ_i, you may not need a **fully realistic rater model**."

(Junker, 2016)

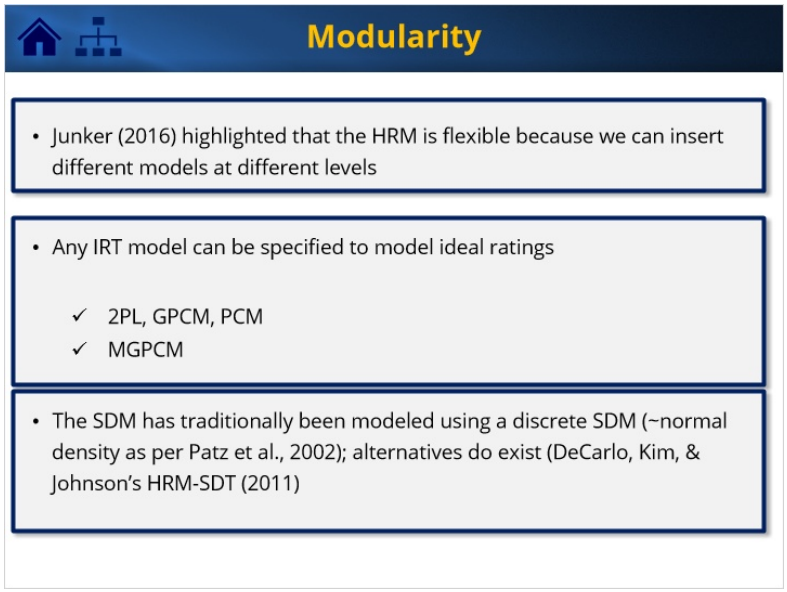
Must consider parsimony

4.18 Bookend: SDM Component

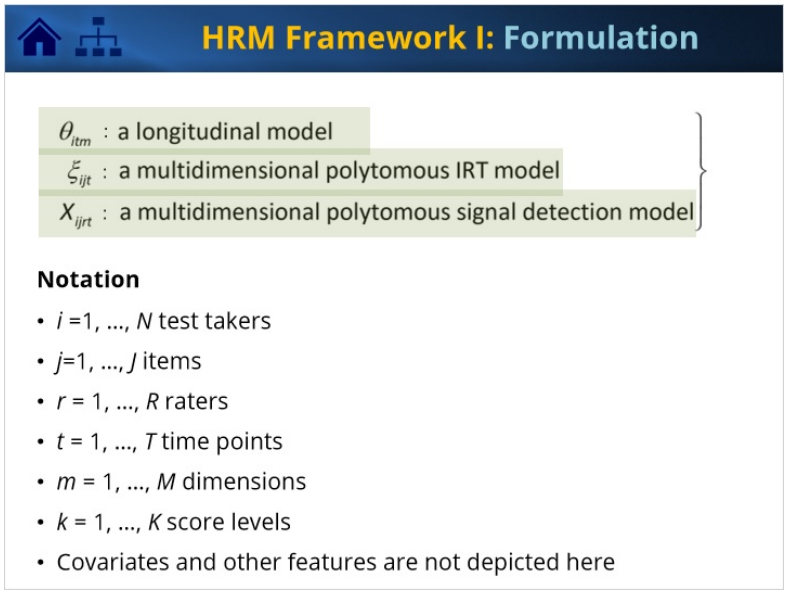


4.19 Bookmark: Modularity

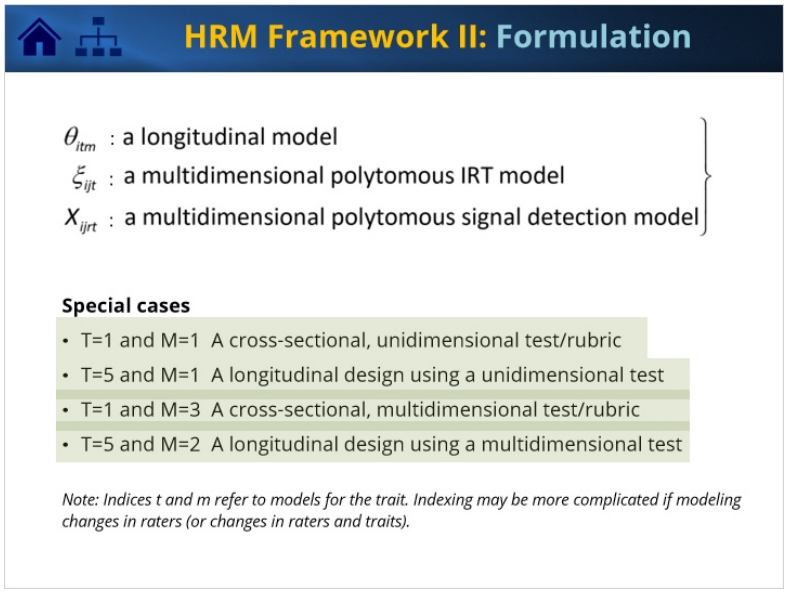
4.20 Modularity



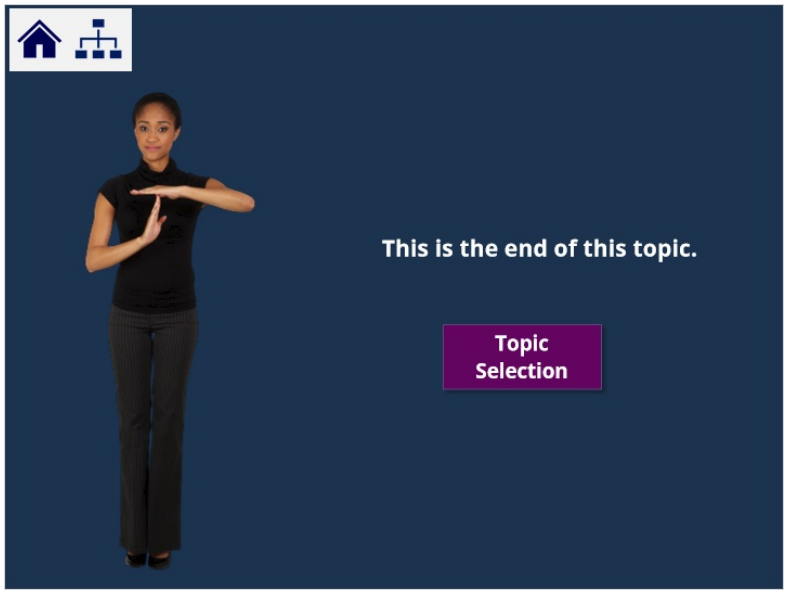
4.21 HRM Framework I: Formulation



4.22 HRM Framework II: Formulation

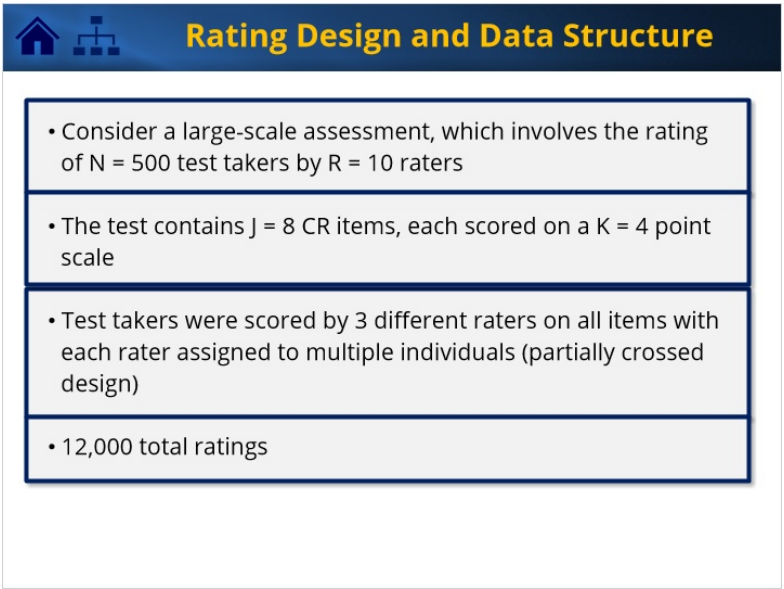


4.23 Bookend: Modularity

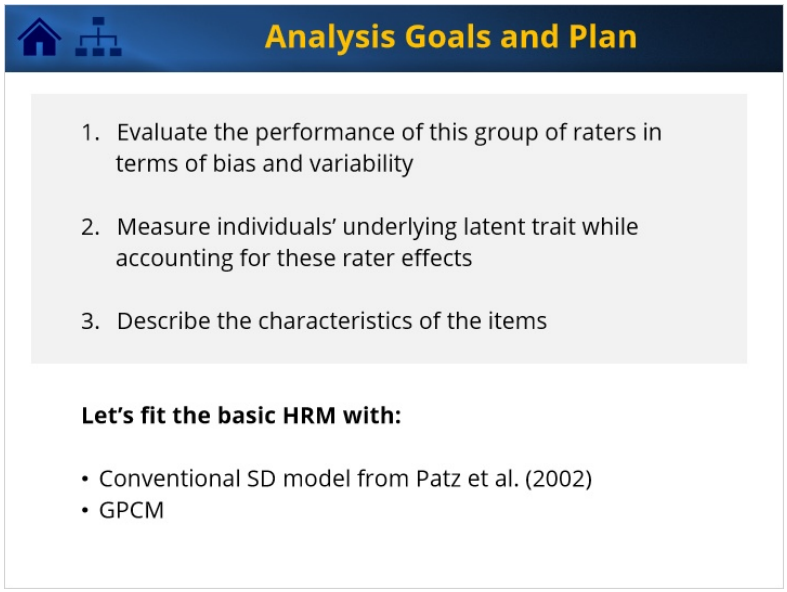


4.24 Bookmark: Worked Example

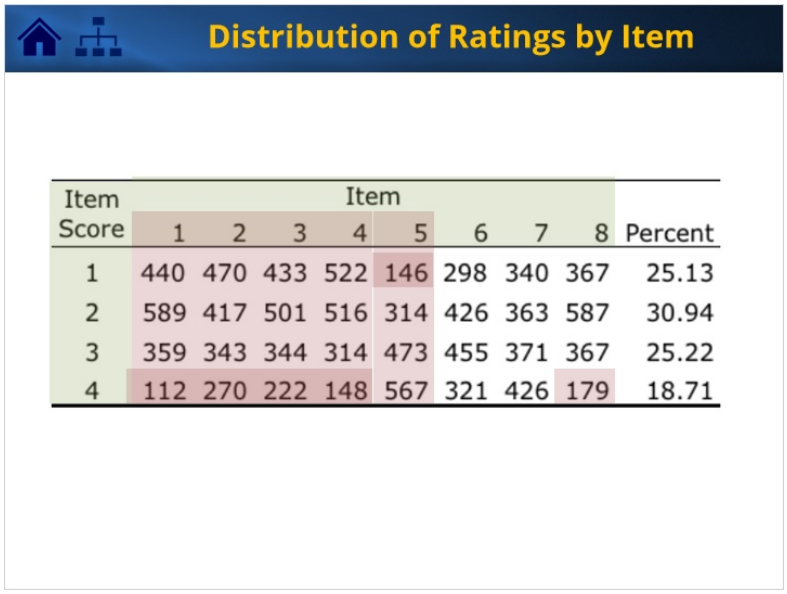
4.25 Rating Design and Data Structure



4.26 Analysis Goals and Plan



4.27 Distribution of Ratings by Item



Item Score	Item 1	2	3	4	5	6	7	8	Percent
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

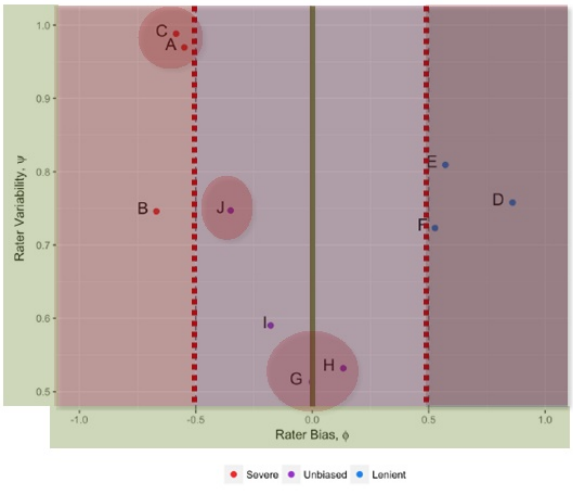
4.28 Results (I): Summary Statistics

Results (I): Summary Statistics

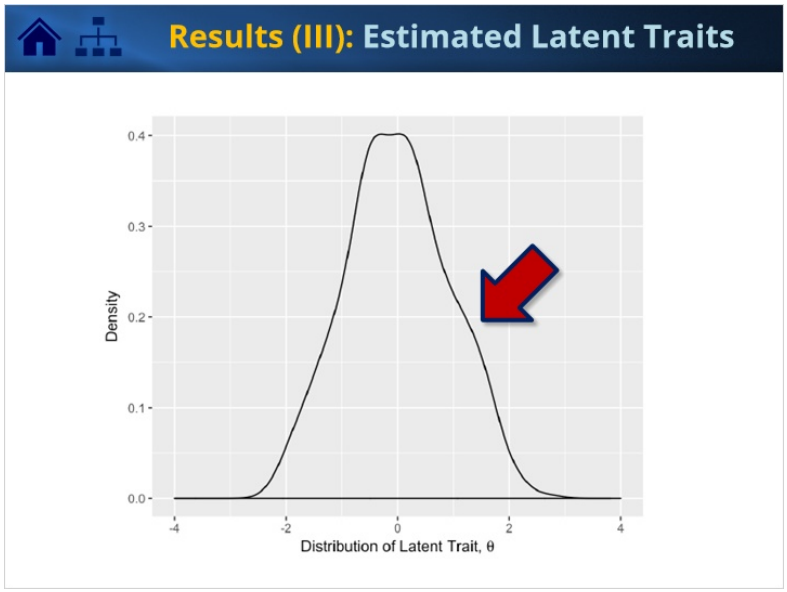
	mean	sd	X2.5.	X25.	X50.	X75.	X97.5.	Rhat	n.eff
alpha[1]	1.5062	0.2198	1.1070	1.3536	1.4879	1.6498	1.9822	1.0013	2400
alpha[2]	1.2515	0.1699	0.9550	1.1293	1.2401	1.3599	1.6103	1.0007	3000
alpha[3]	2.0403	0.2991	1.5159	1.8323	2.0153	2.2326	2.6814	1.0010	3000
gamma[1,2]	-1.0588	0.1439	-1.3620	-1.1508	-1.0541	-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

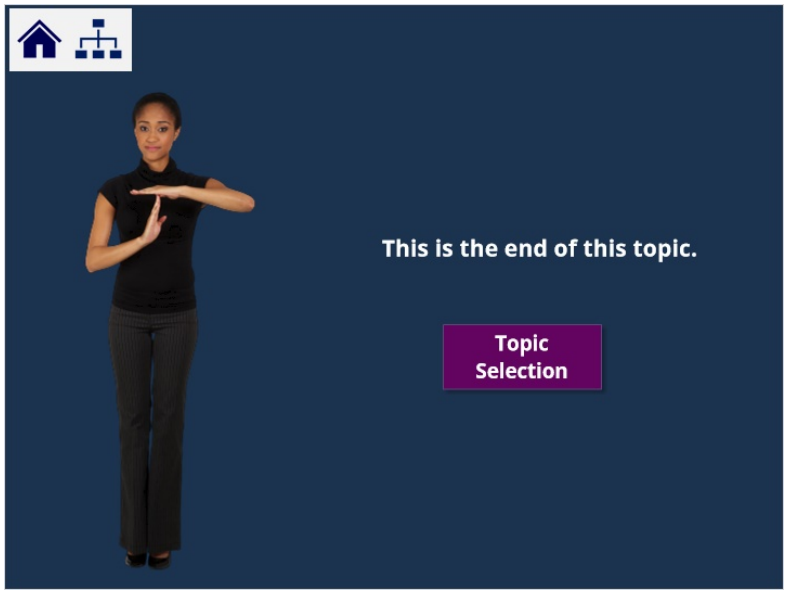
Results (II): Rater Scatterplot



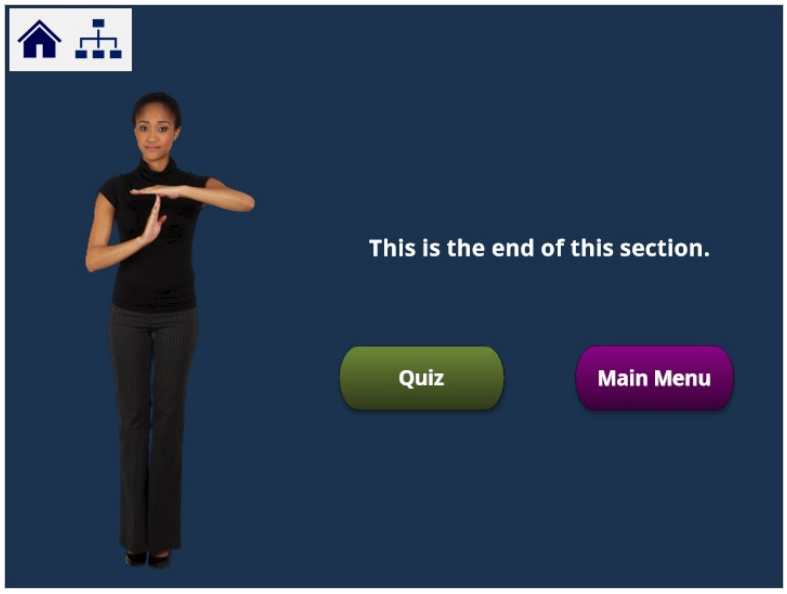
4.30 Results (III): Estimated Latent Traits



4.31 Bookend: Worked Example



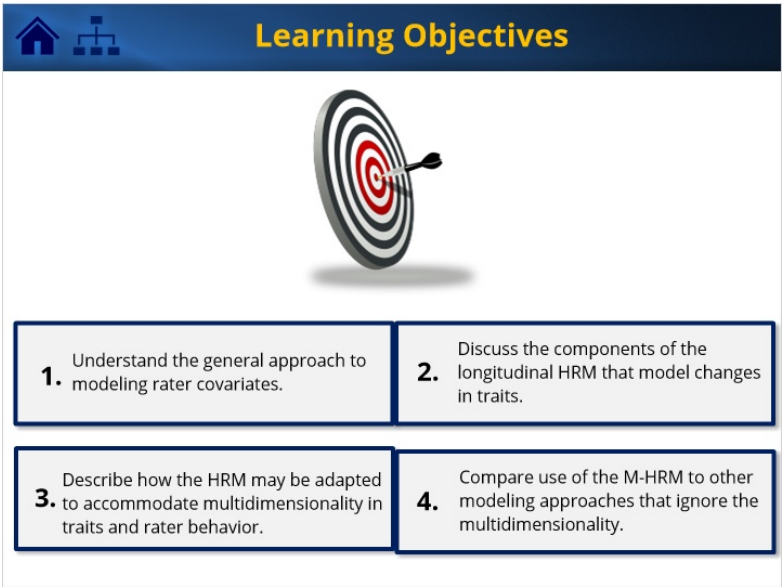
4.32 Bookend: Section 3



5. Section 4: HRM Extensions

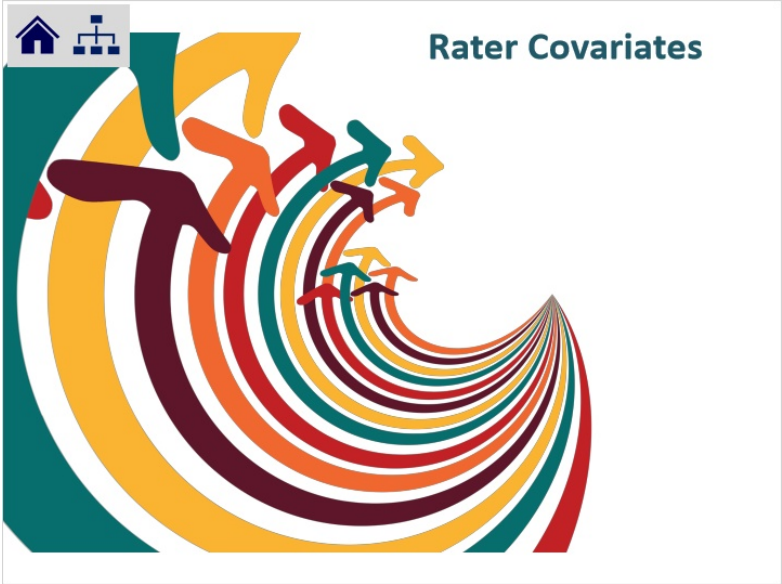
5.1 Cover: Section 4

5.2 Objectives: Section 4

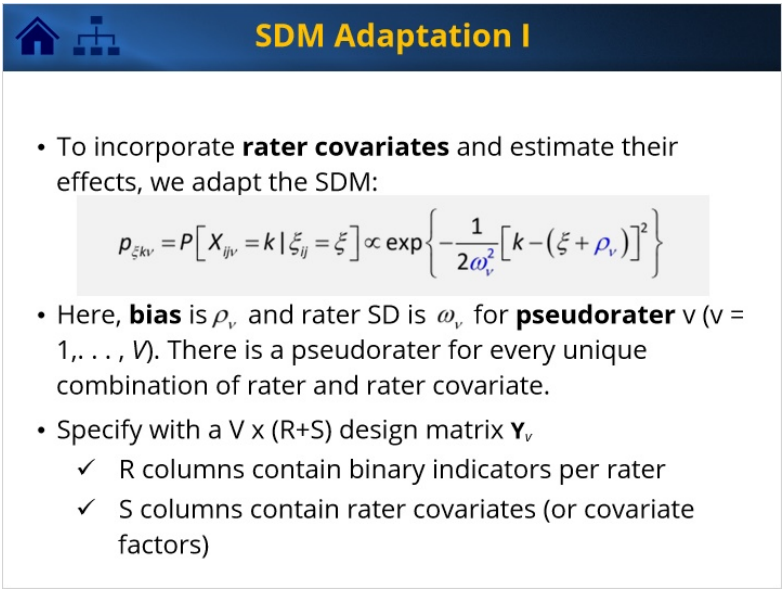


5.3 Topic Selection

5.4 Bookmark: Covariates



5.5 SDM Adaptation I



5.6 SDM Adaptation II

SDM Adaptation II

- Linear model for rating bias ρ_v depending on covariates as defined by design matrix $\mathbf{Y}_v$

$$\rho_v = \mathbf{Y}_v \eta$$

where $\eta = (\phi_1, \ldots, \phi_R, \eta_1, \ldots, \eta_S)^T$

- Linear model for rating variability ω_v^2 depending on covariates as defined by design matrix $\mathbf{Y}_v$

$$\ln \omega_v^2 = \mathbf{Y}_v \left(\ln \boldsymbol{\tau}^2 \right)$$

where $\ln \boldsymbol{\tau}^2 = \left(\ln \psi_1^2, \ldots, \ln \psi_R^2, \ln \tau_1^2, \ldots, \ln \tau_S^2 \right)^T$

5.7 SDM Adaptation III

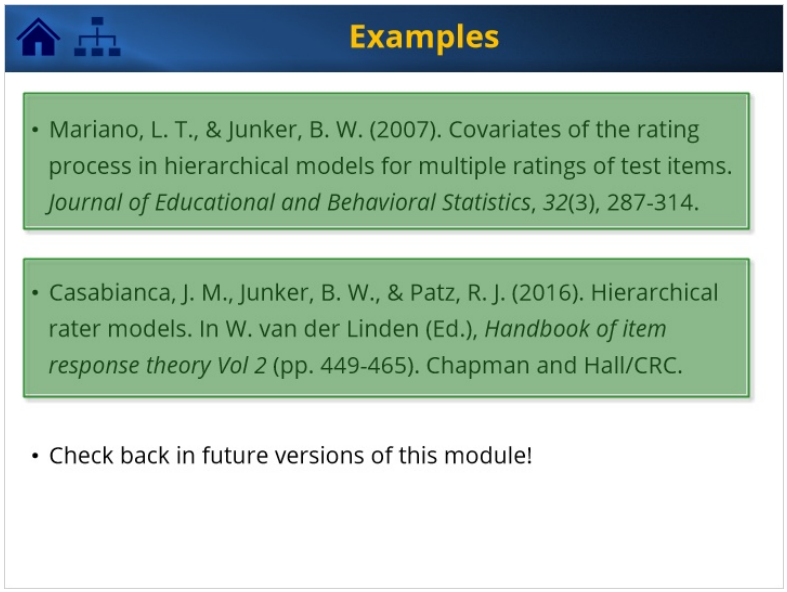
SDM Adaptation III

Random or **fixed** rating effects modeling approach

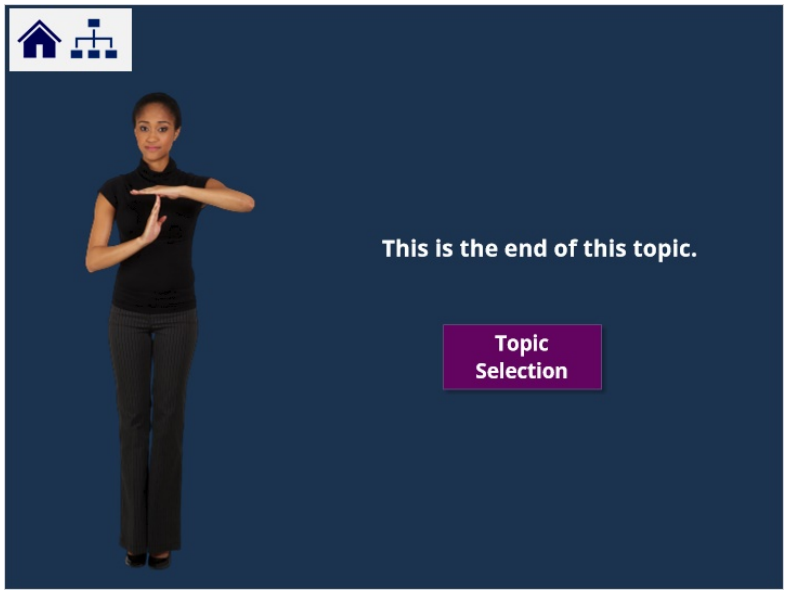
- *Random:* pseudorater bias and variability are treated as random draws
  - ✓ Bias is drawn from a normal distribution centered at a linear function of the covariates
  - ✓ Variability is drawn from a similarly centered log-normal distribution (or normal, if transformed)

- *Fixed:* the linear structures for rating bias and variability replace the original rater parameters

5.8 Examples

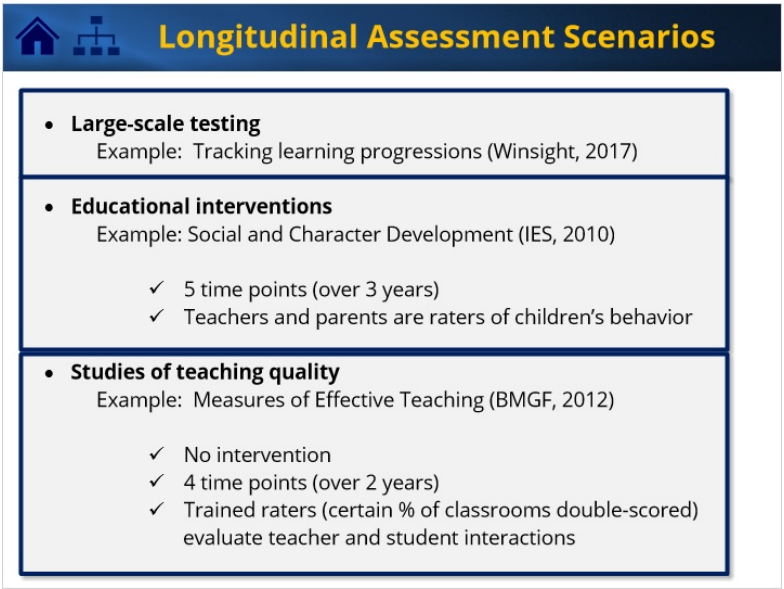


5.9 Bookend: Covariates

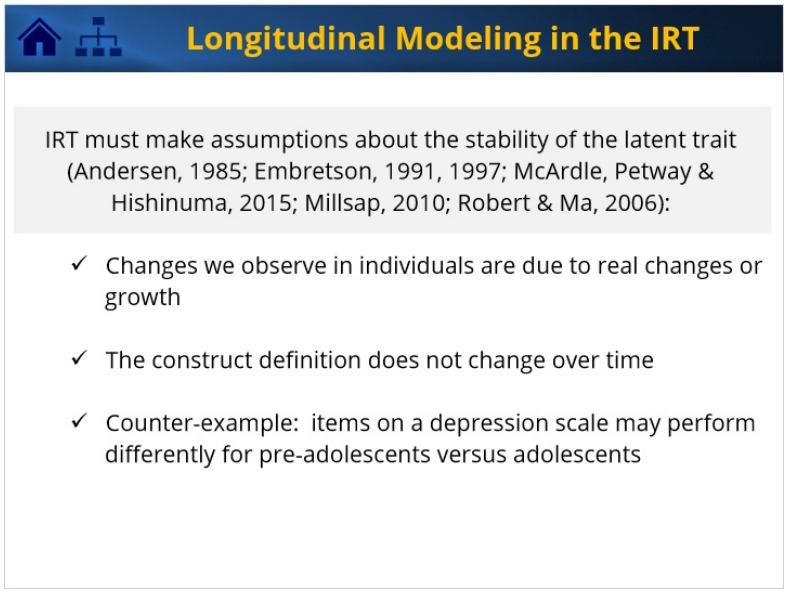


5.10 Bookmark: Longitudinal HRM

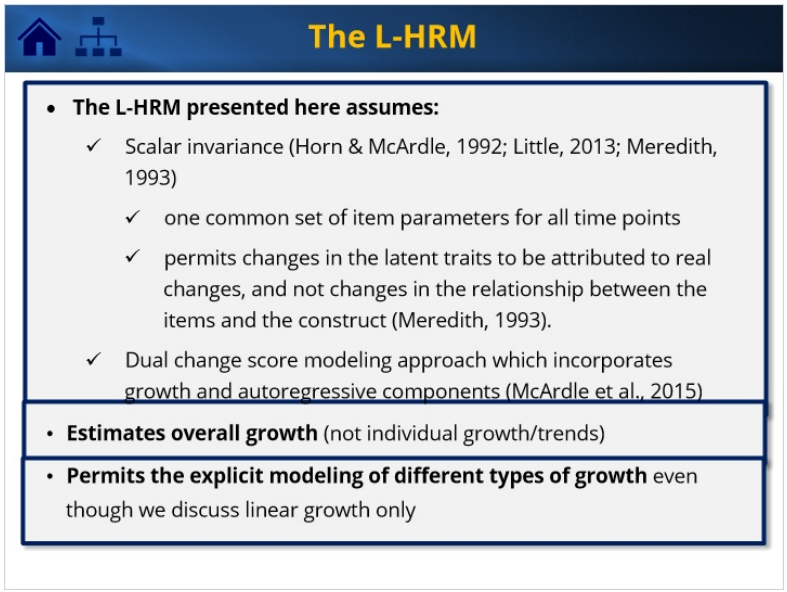
5.11 Longitudinal Assessment Scenarios



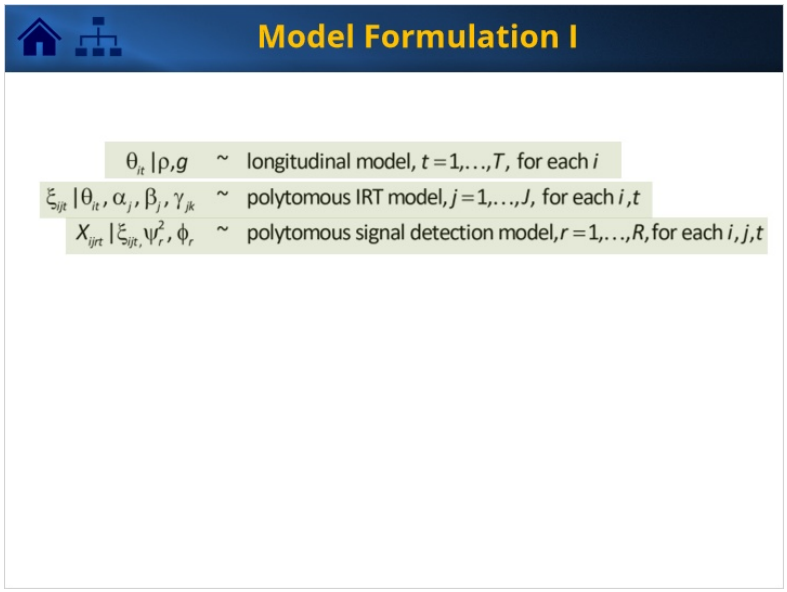
5.12 Longitudinal Modeling in the IRT



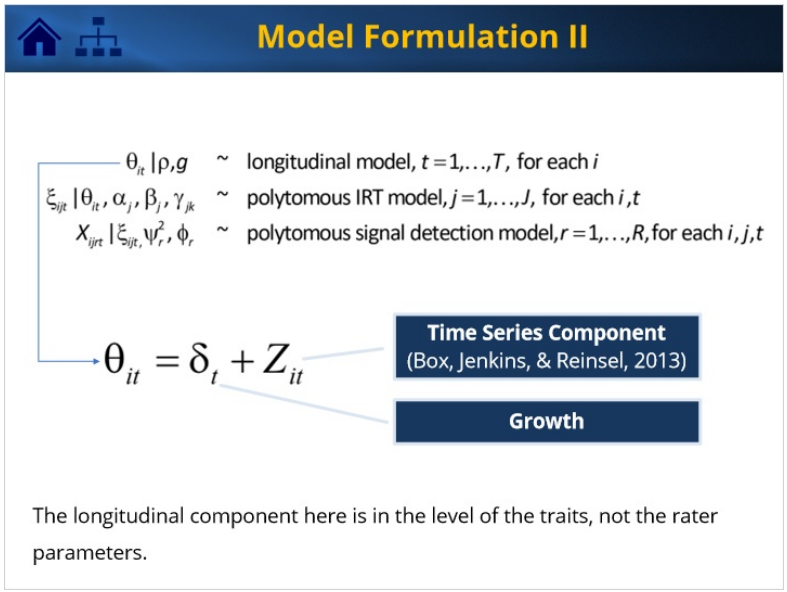
5.13 The L-HRM



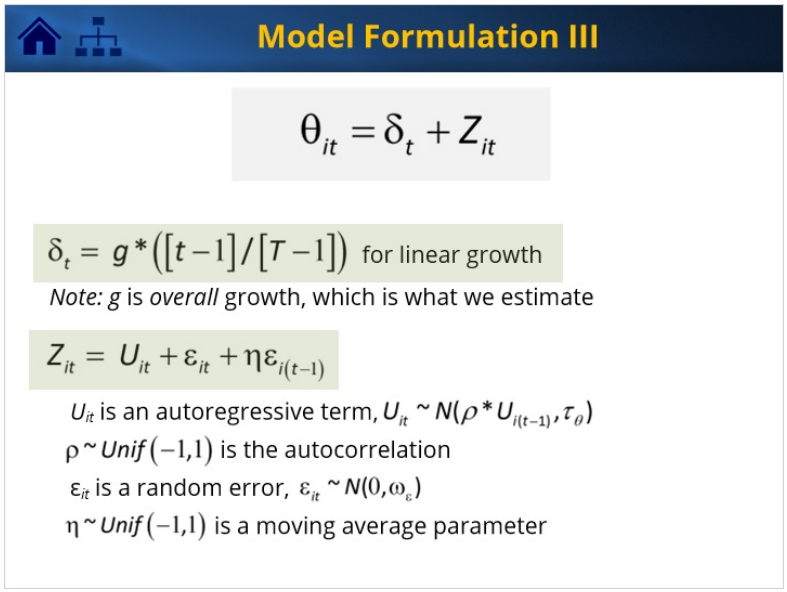
5.14 Model Formulation I



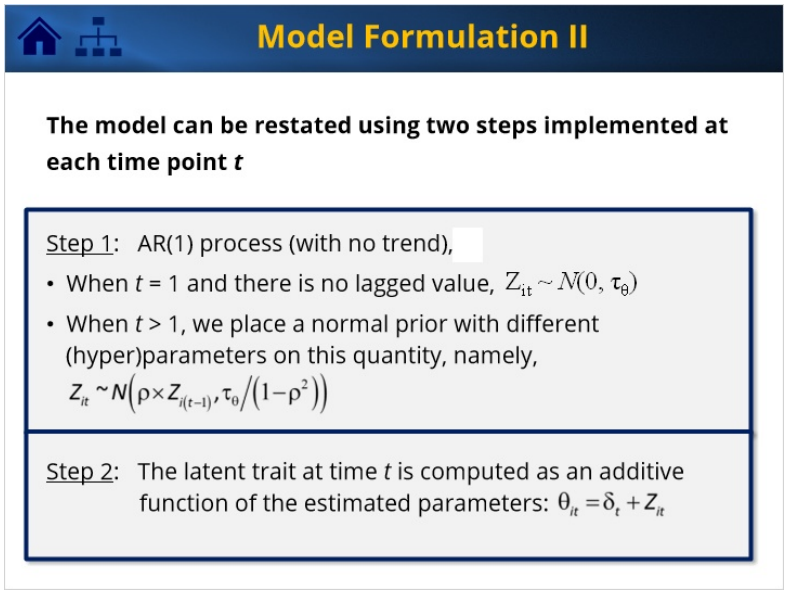
5.15 Model Formulation II



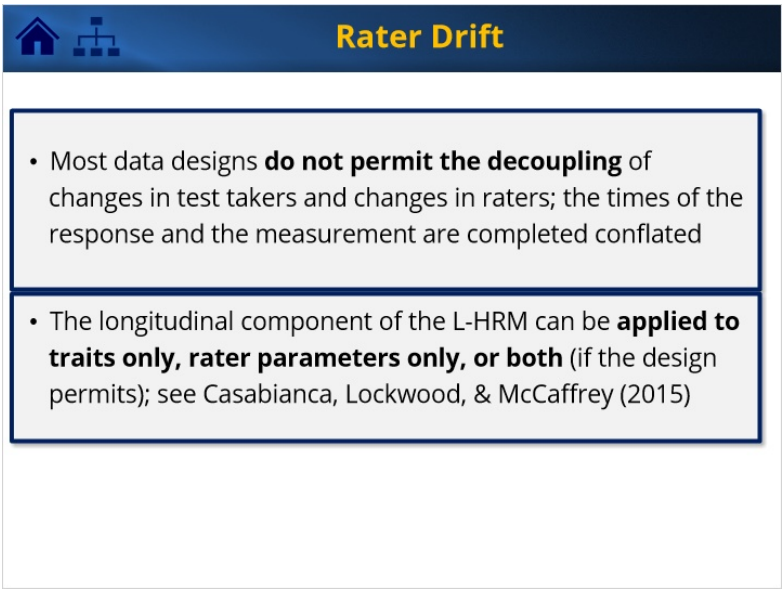
5.16 Model Formulation III



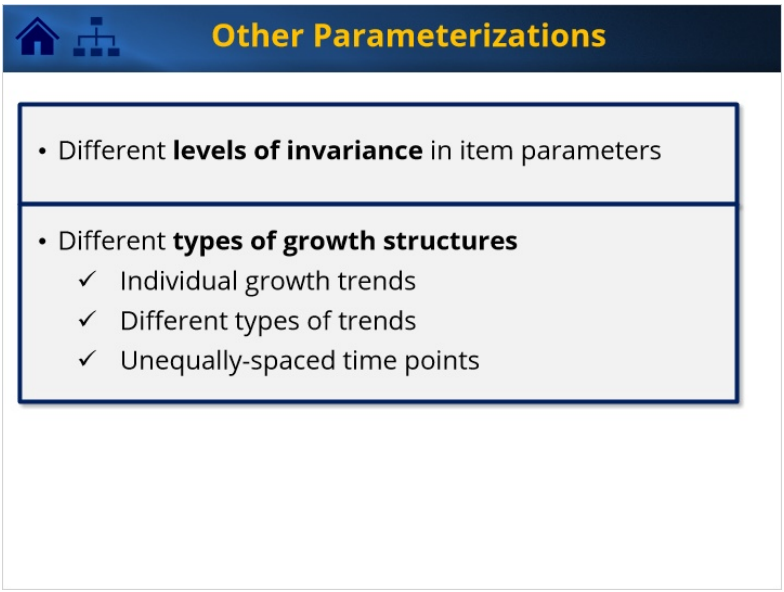
5.17 Model Formulation II



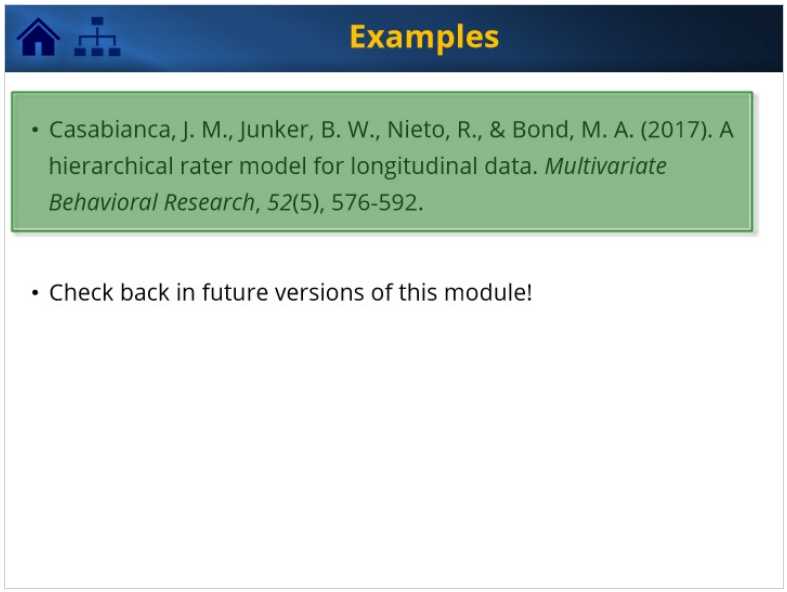
5.18 What about Rater Drift?



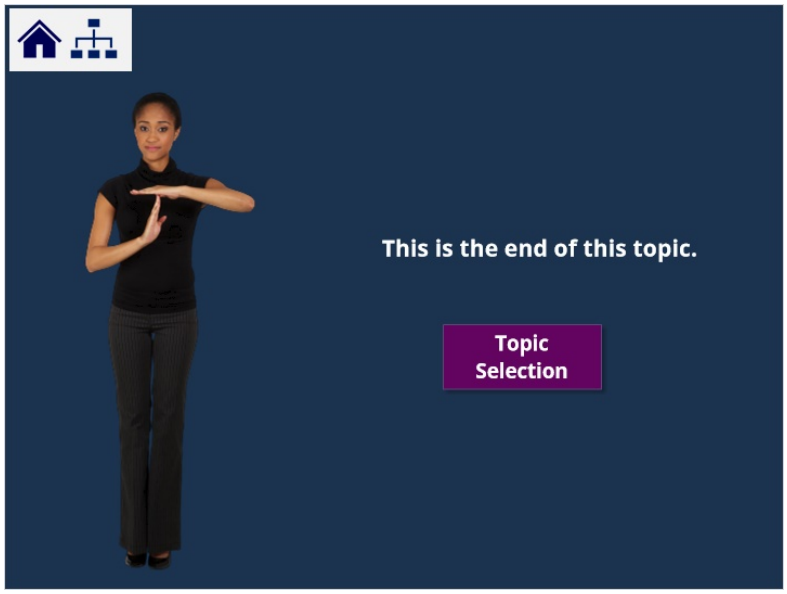
5.19 Other Parameterizations



5.20 Examples

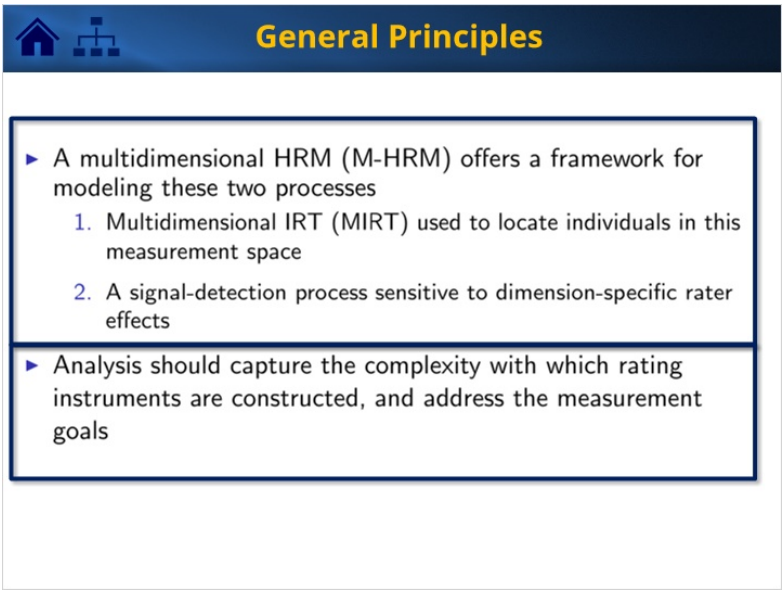


5.21 Bookend: Multidimensional HRM

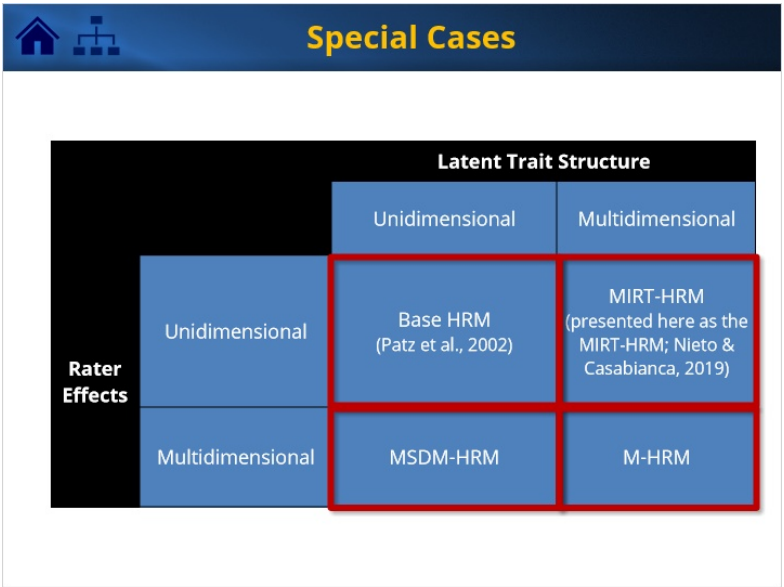


5.22 Bookmark: Multidimensional HRM

5.23 General Principles



5.24 Possible Special Cases Under this Multidimensional Framework



5.25 Model Variants

Base HRM

$$\begin{cases} \theta_i \sim \mathcal{N}(\mu, \sigma^2) \\ \xi_{ij} \sim \text{unidimensional IRT model} \\ X_{ijr} \sim \text{unidimensional signal detection model} \end{cases}$$

MIRT HRM

$$\begin{cases} \boldsymbol{\theta_j} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \xi_{ij} \sim \text{MIRT model} \\ X_{ijr} \sim \text{unidimensional signal detection model} \end{cases}$$

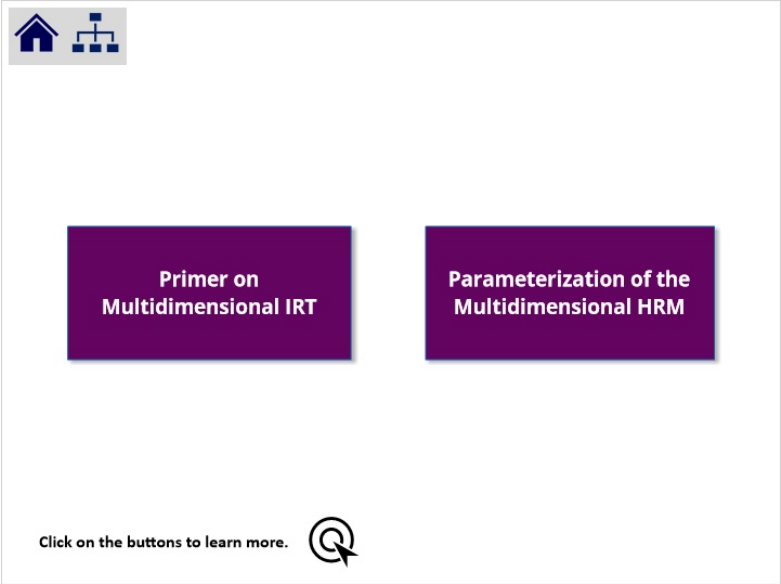
MSDM HRM

$$\begin{cases} \theta_i \sim \mathcal{N}(\mu, \sigma^2) \\ \xi_{ij} \sim \text{unidimensional IRT model} \\ X_{ijr} \sim \text{multidimensional signal detection model} \end{cases}$$

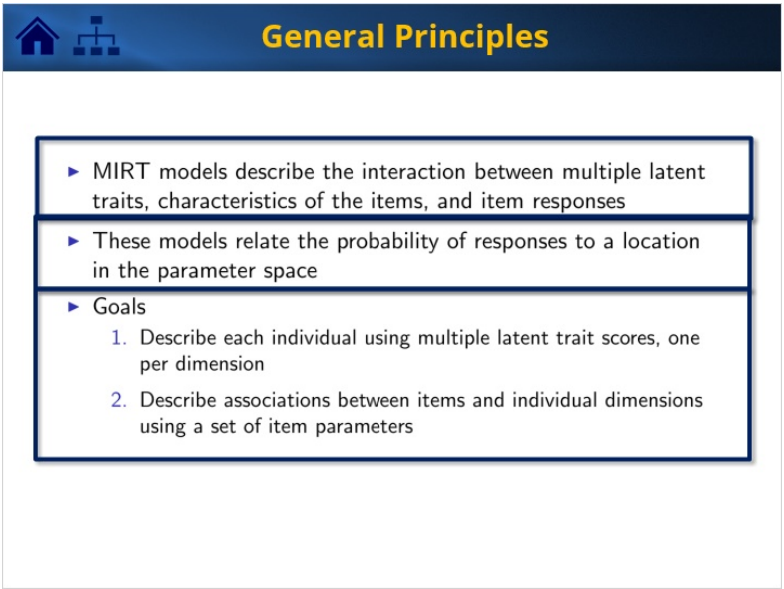
Multidimensional HRM

$$\begin{cases} \boldsymbol{\theta_i} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \xi_{ij} \sim \text{MIRT model} \\ X_{ijr} \sim \text{multidimensional signal detection model} \end{cases}$$

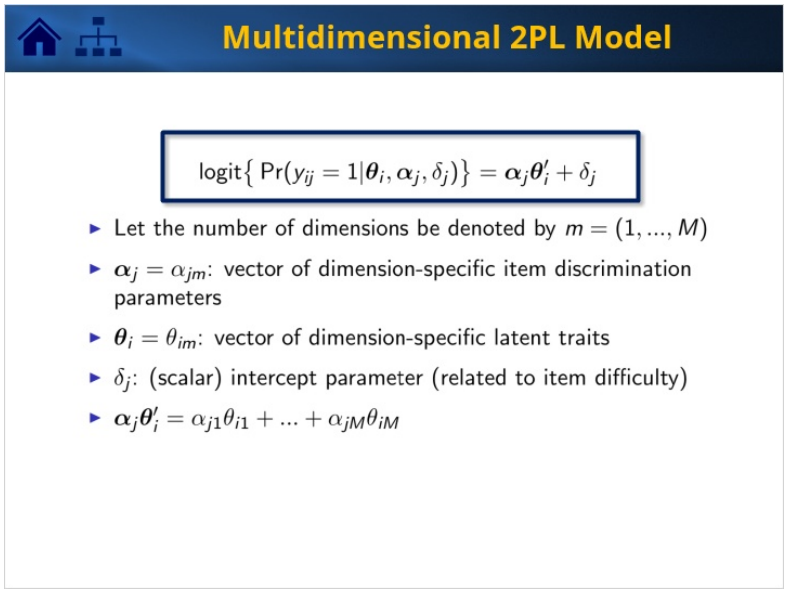
5.26 Subtopic Selection



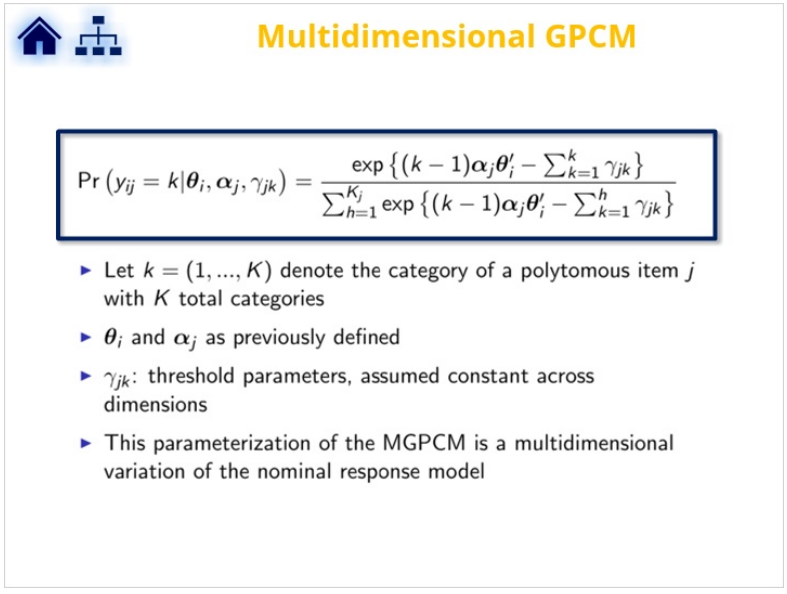
5.27 General Principles



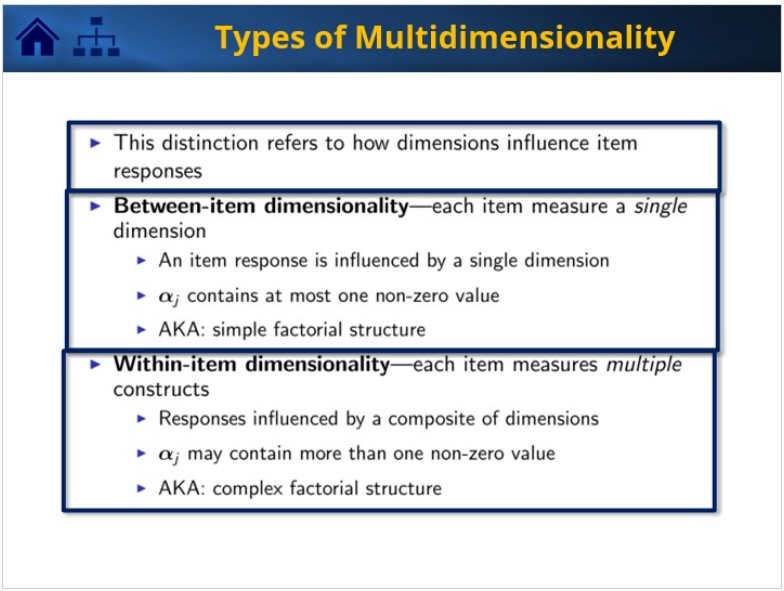
5.28 Multidimensional 2PL Model



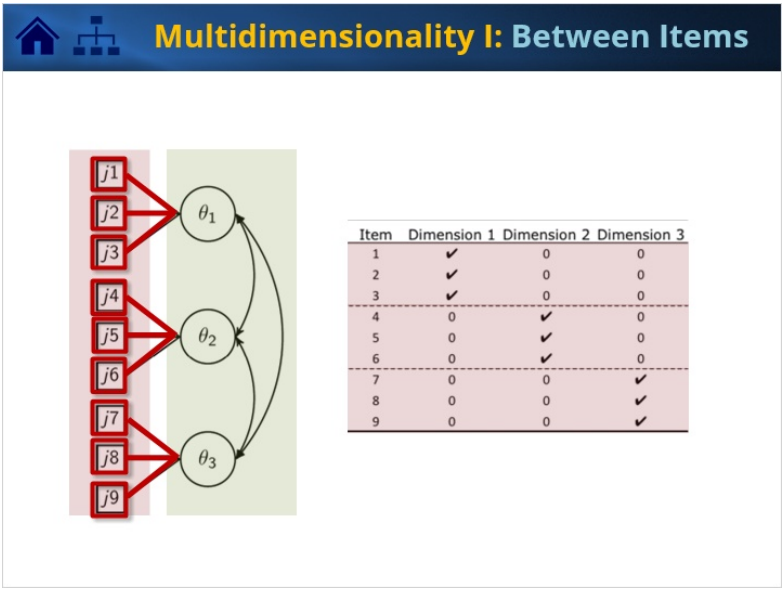
5.29 Multidimensional Generalized Partial Credit Model (MGPC; Yao & Schwarz, 2006)



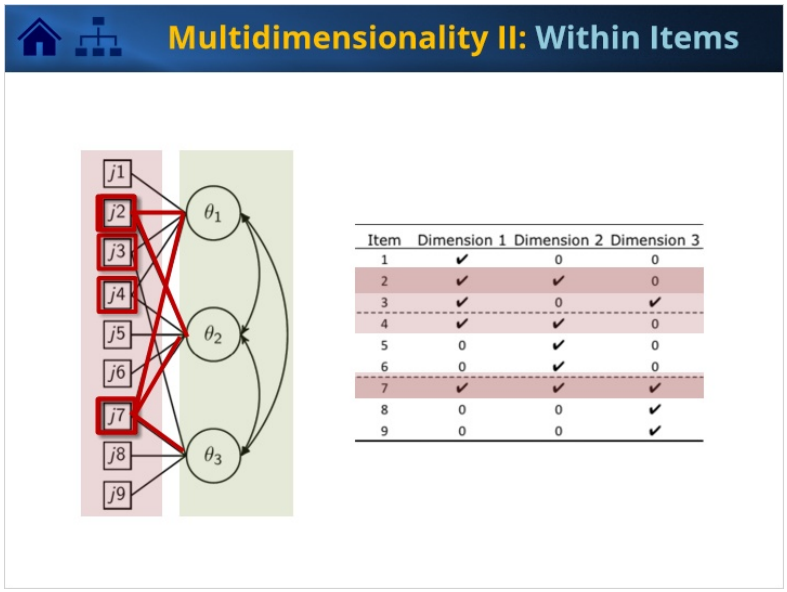
5.30 Multidimensional Generalized Partial Credit Model (MGPC; Yao & Schwarz, 2006)



5.31 Between-Item Dimensionality

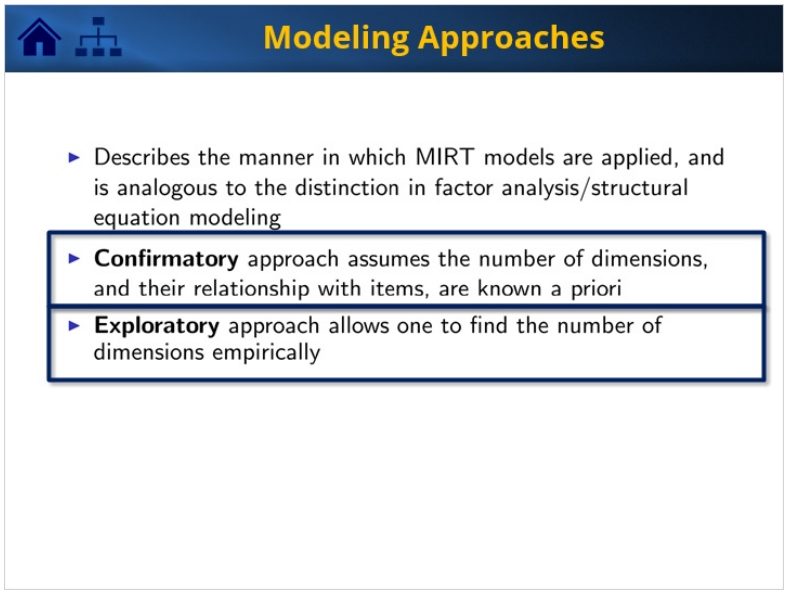


5.32 Within-Item Dimensionality

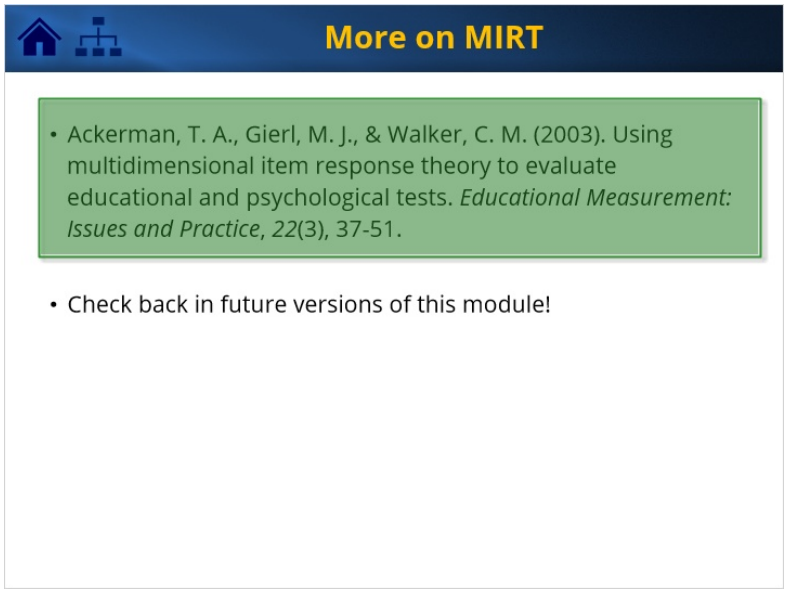


Item	Dimension 1	Dimension 2	Dimension 3
1	✔	0	0
2	✔	✔	0
3	✔	0	✔
4	✔	✔	0
5	0	✔	0
6	0	✔	0
7	✔	✔	✔
8	0	0	✔
9	0	0	✔

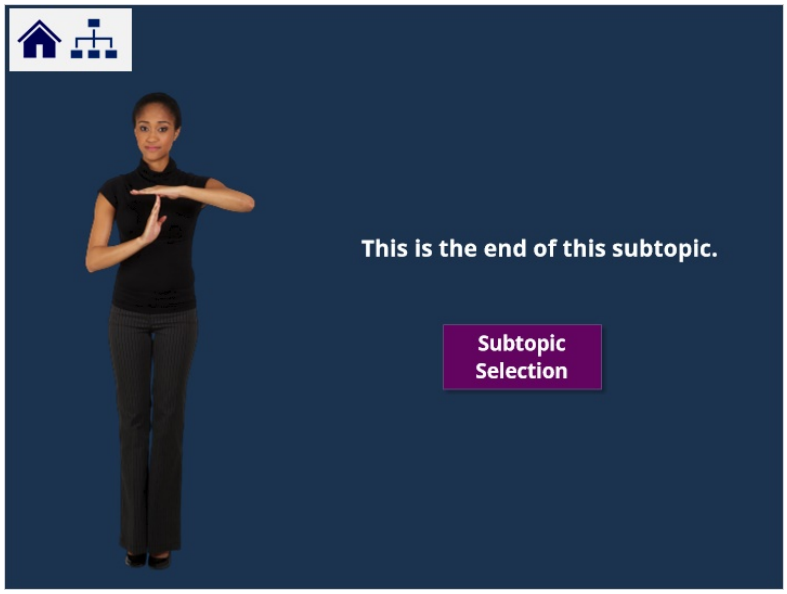
5.33 Modeling Approaches: Confirmatory vs. Exploratory



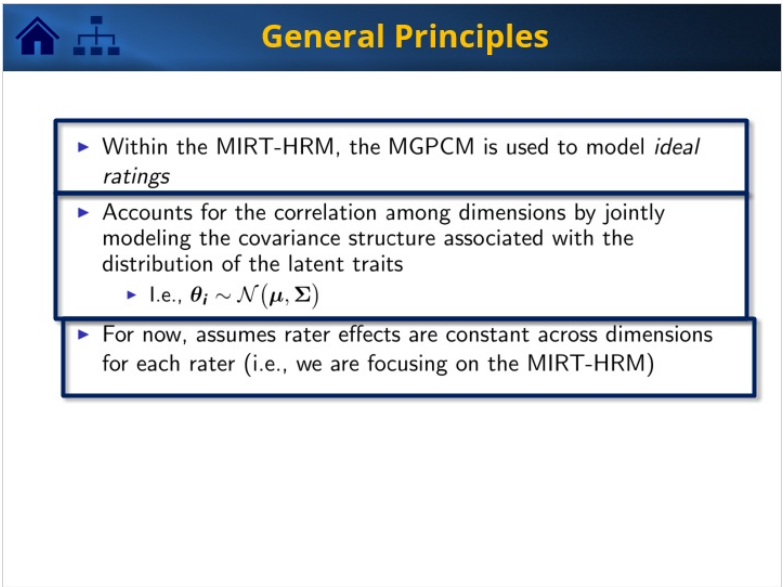
5.34 Modeling Approaches: Confirmatory vs. Exploratory



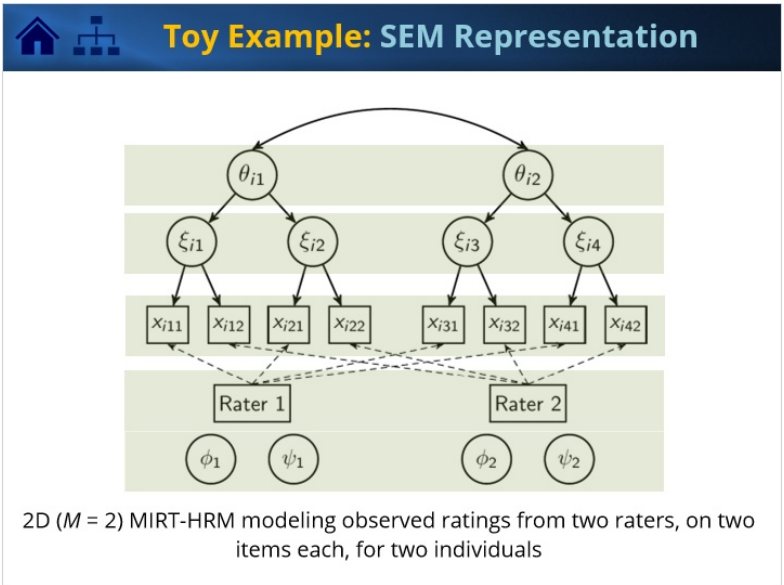
5.35 Bookend: Primer on Multidimensional IRT



5.36 General Principles

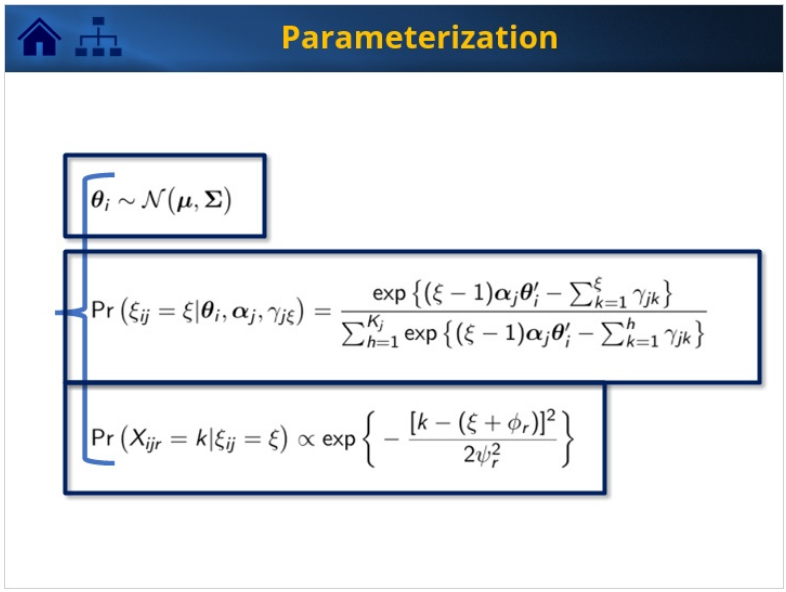


5.37 2D (M = 2) MIRT-HRM modeling observed ratings from two raters, on two items each, for two individuals

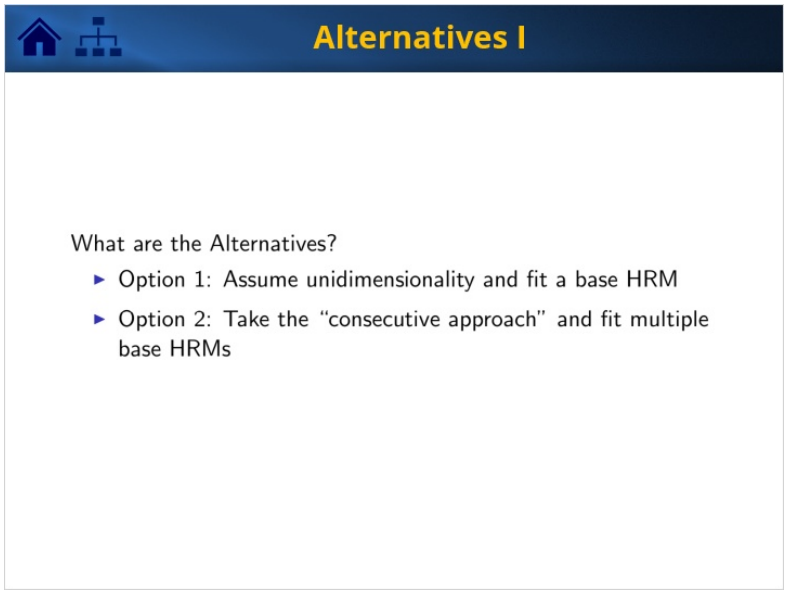


2D (M = 2) MIRT-HRM modeling observed ratings from two raters, on two items each, for two individuals

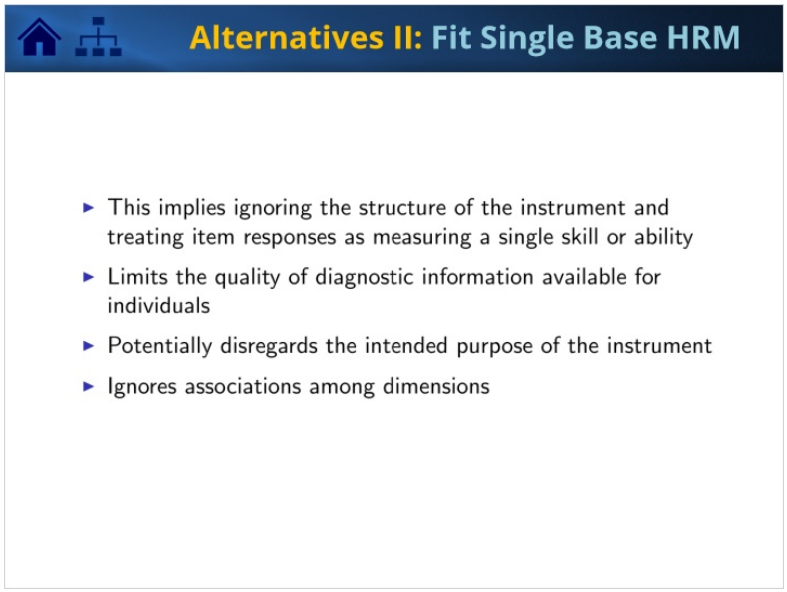
5.38 Parameterizing the MIRT-HRM



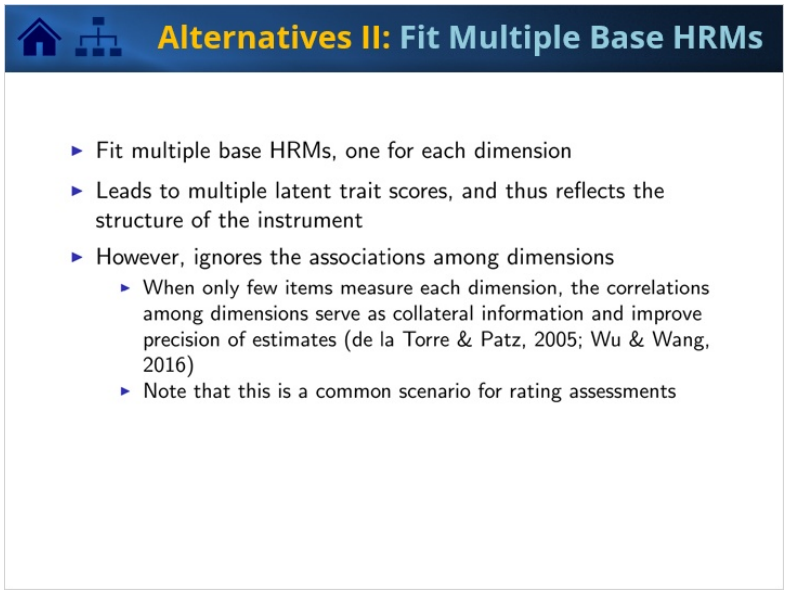
5.39 Alternatives I



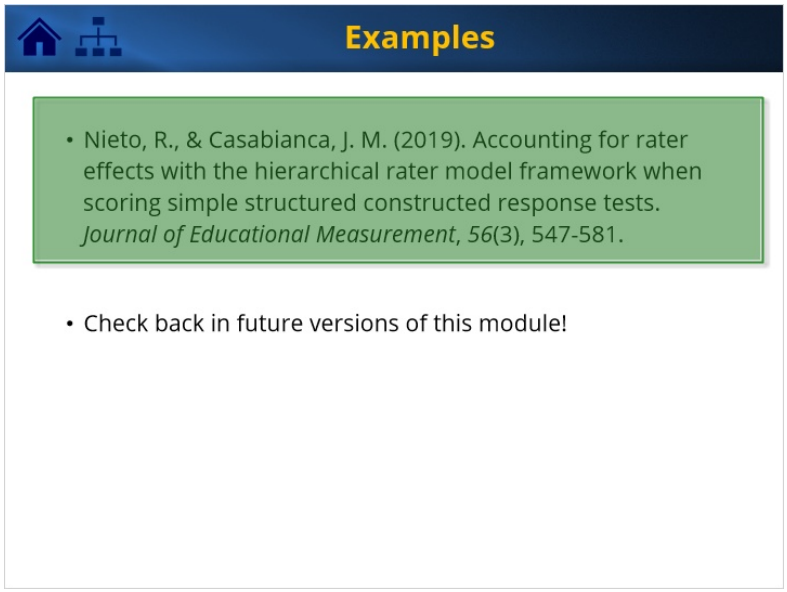
5.40 Alternatives II



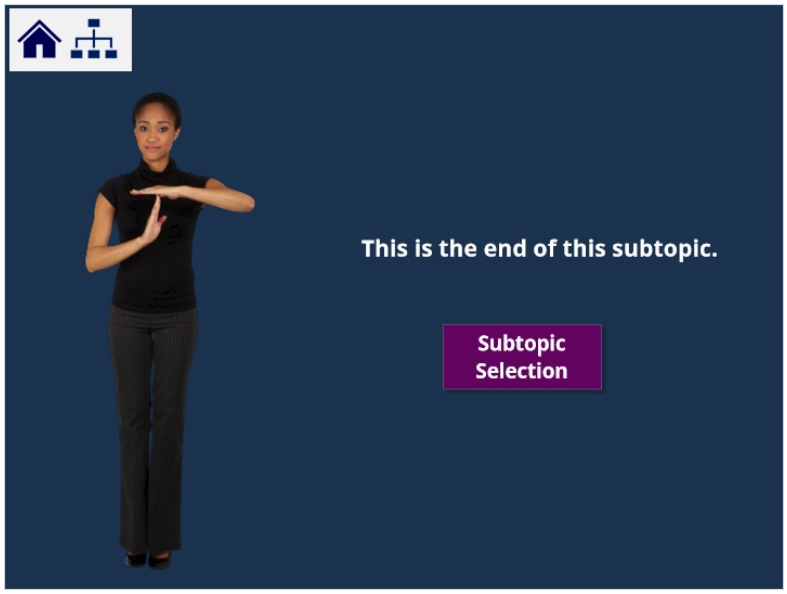
5.41 Alternatives II



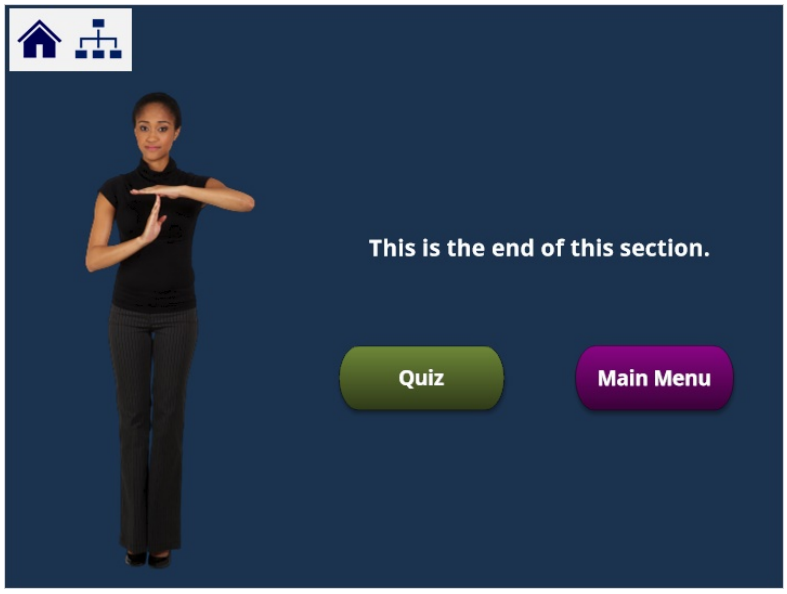
5.42 The M-HRM Thus Far



5.43 Bookend: Parameterization of the Multidimensional HRM



5.44 Bookend: Section 4



5.45 Module Cover (END)