

## ITEMS Digital Module 03: Nonparametric Item Response Theory

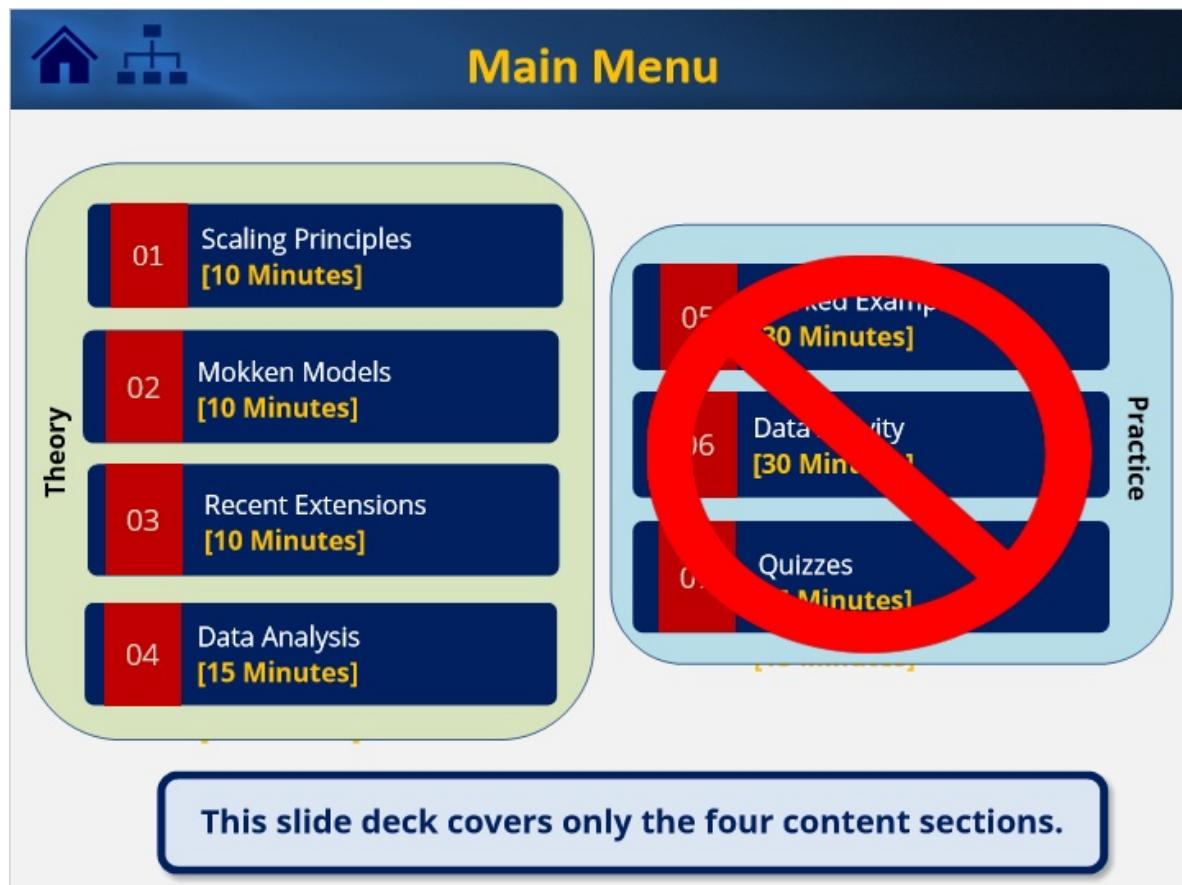
---

This document contains all core content slides from sections 1-4 with the exception of slides that show video screens. In the digital module all slides can be accessed individually.

---

### Module Organization

The module starts with an introductory section that leads to the main menu from which learners can select individual content and activity sections:



# DM03 SLIDES (Version 1.3)

## 1. Module Overview

### 1.1 Module Cover (START)



### 1.2 Instructor



### 1.3 Designers

Meet the instructional design team:



Special thanks:



André A. Rupp  
ETS

Xi Lu  
Florida State  
University

### 1.4 Welcome



Welcome to the  
ITEMS Module!

The woman to the left is Laura!

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

## Untitled Layer 1 (Slide Layer)

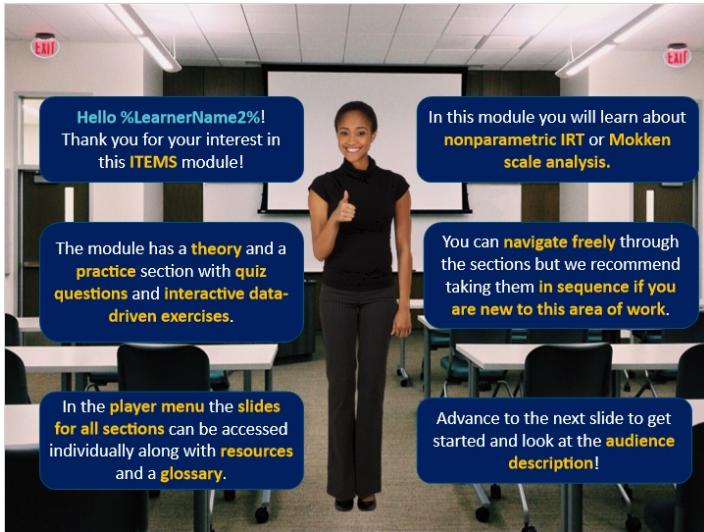


Welcome to the  
ITEMS Module!

The woman to the left is **Laura**!

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

### 1.5 Overview



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

In this module you will learn about  
**nonparametric IRT** or **Mokken**  
scale analysis.

The module has a **theory** and a  
**practice** section with **quiz**  
**questions** and **interactive data-**  
**driven exercises**.

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

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

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)



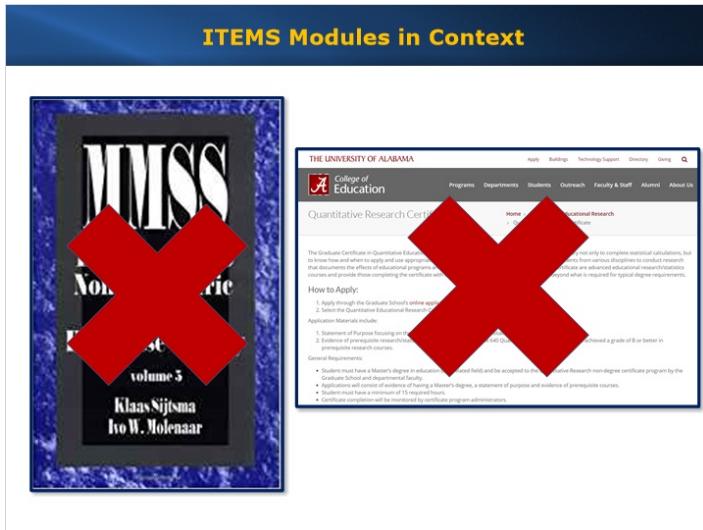




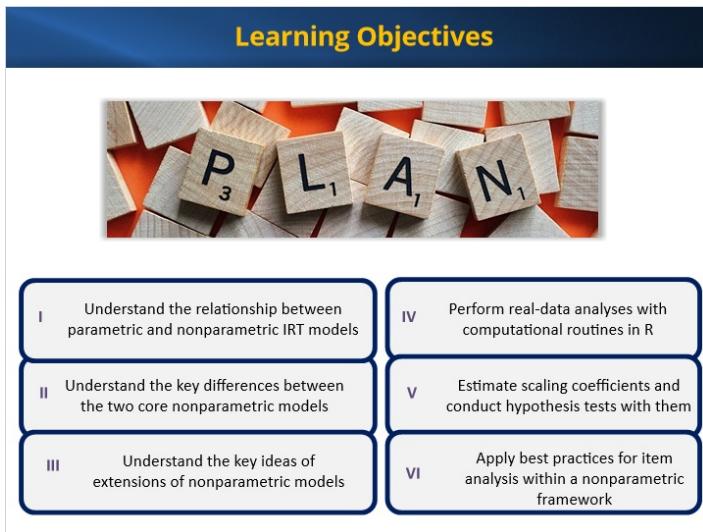


**Let's discuss expectations....**

## 1.8 Expectations (II)



## 1.9 Learning Objectives



I	Understand the relationship between parametric and nonparametric IRT models
II	Understand the key differences between the two core nonparametric models
III	Understand the key ideas of extensions of nonparametric models
IV	Perform real-data analyses with computational routines in R
V	Estimate scaling coefficients and conduct hypothesis tests with them
VI	Apply best practices for item analysis within a nonparametric framework

## 1.10 Prerequisites

**Prerequisites**

- Working knowledge of foundational statistical concepts:
  - Means, variances, and standard deviations
  - Standard errors
  - Statistical hypothesis testing, specifically  $t$ -tests
- Working knowledge of foundational measurement concepts:
  - Construct definitions / latent variables
  - Assessment formats
  - Item / task types
  - Scales and scale scores
  - Basic aspects of assessment development
- Optional: Basic experience with R for the practice exercises

*Note:* Not technically required as introductory videos are provided

## 1.11 Main Menu

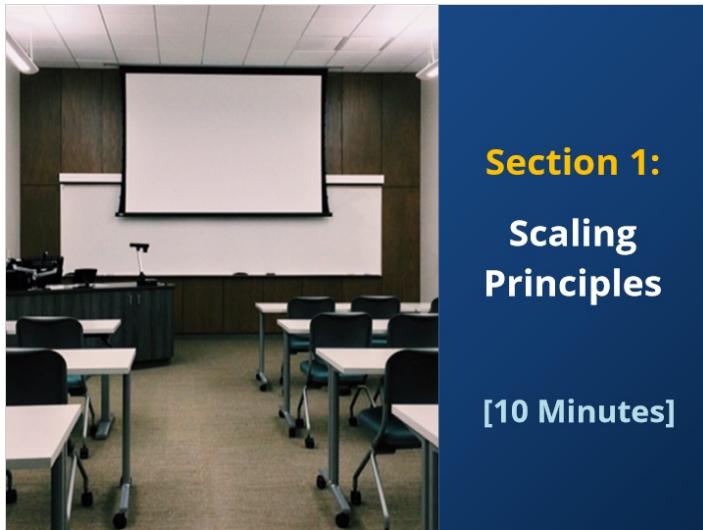
**Main Menu**

The slide displays a main menu with two main sections: Theory and Practice. The Theory section on the left contains four items: 01 Scaling Principles [10 Minutes], 02 Mokken Models [10 Minutes], 03 Recent Extensions [10 Minutes], and 04 Data Analysis [15 Minutes]. The Practice section on the right contains three items: 05 Simulated Exam [30 Minutes], 06 Data Quality [30 Minutes], and 07 Quizzes [15 Minutes]. A large red 'no' symbol is overlaid on the Practice section, indicating that it is not covered by the slide deck.

This slide deck covers only the four content sections.

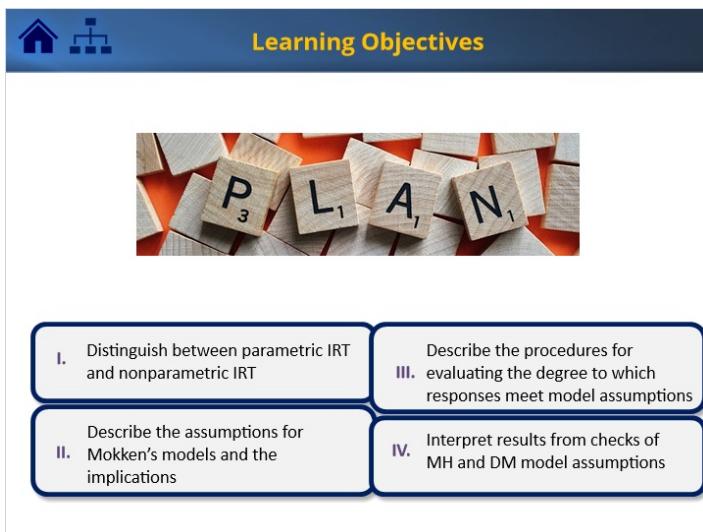
## 2. Section 1: Scaling Principles

### 2.1 Cover: Section 1



Section 1:  
Scaling  
Principles  
[10 Minutes]

### 2.2 Objectives: Section 1

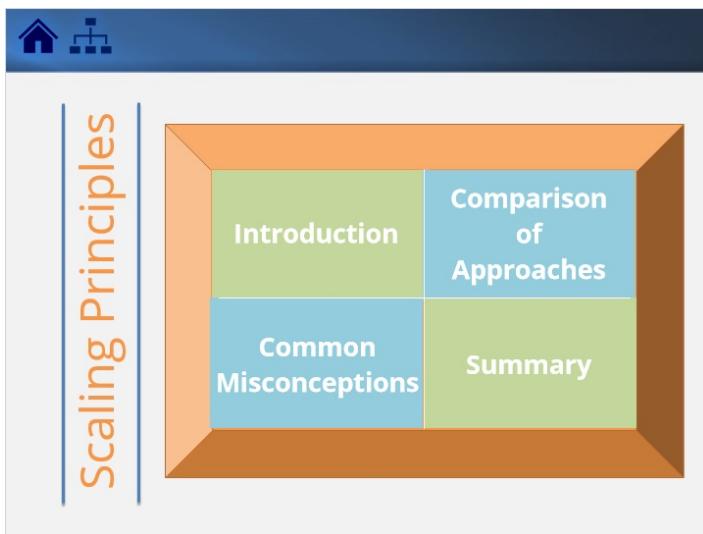


Learning Objectives



- I. Distinguish between parametric IRT and nonparametric IRT
- II. Describe the assumptions for Mokken's models and the implications
- III. Evaluate the degree to which responses meet model assumptions
- IV. Interpret results from checks of MH and DM model assumptions

## 2.3 Topic Selection

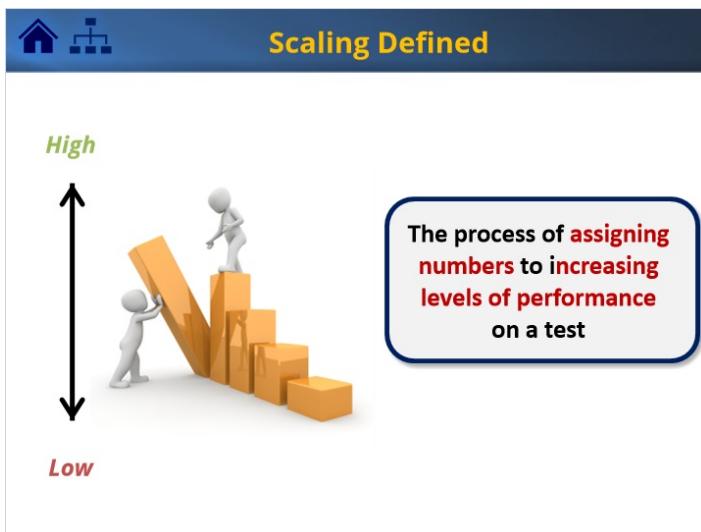


## 2.4 Bookmark: Introduction



## 2.5 Definition: Scaling

**Scaling Defined**



**High**

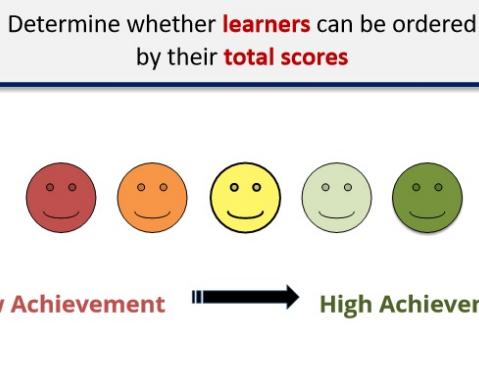
**Low**

The process of **assigning numbers to increasing levels of performance on a test**

## 2.6 Learner Objective

**Objective 1: Learner Ordering**

Determine whether **learners** can be ordered by their **total scores**



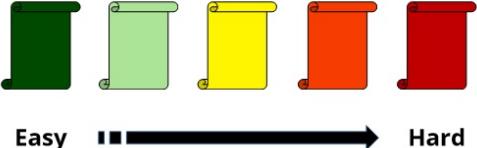
**Low Achievement**  **High Achievement**

## 2.7 Item Objective

### Objective 2: Item Ordering

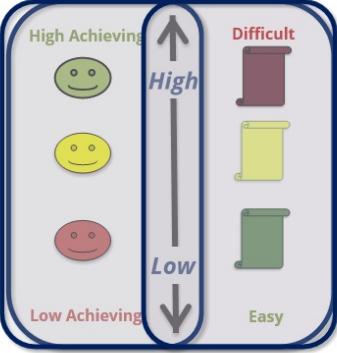
Determine whether **items** can be ordered the **same way** for **all test-takers**



## 2.8 Scaling Approaches

### Response-Centered Approach

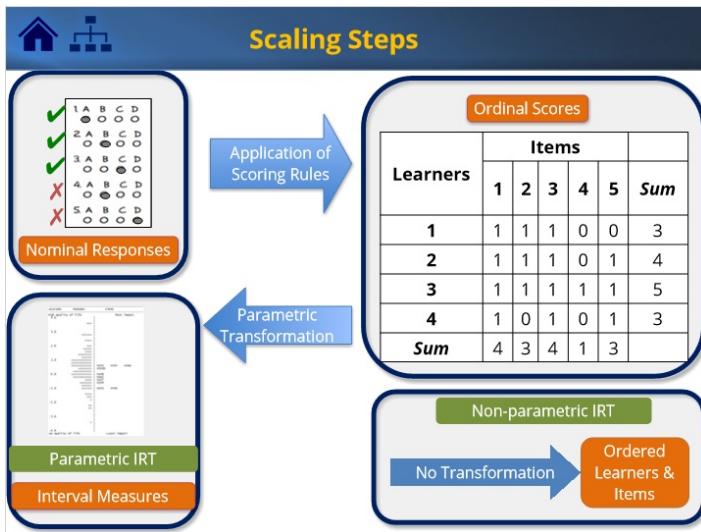


**Subject-Centered Approach**  
Focus is on **locating learners** on a scale that represents a construct

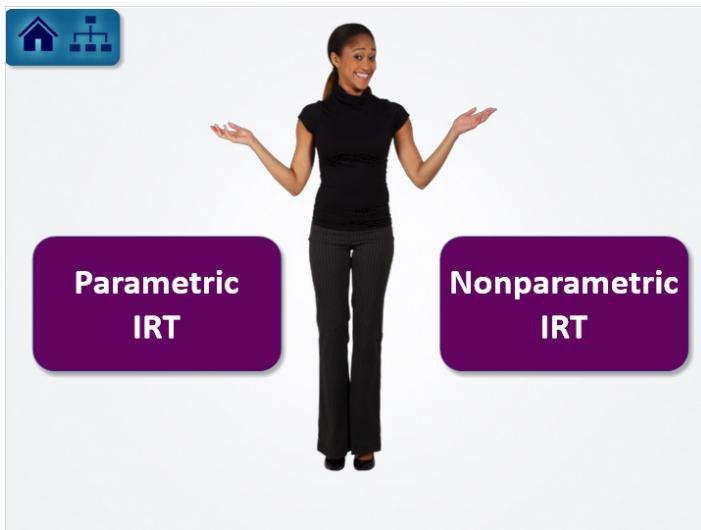
**Item-Centered Approach**  
Focus is on **locating items** on a scale that represents a construct

Focus is on **locating learners and items simultaneously** on a scale

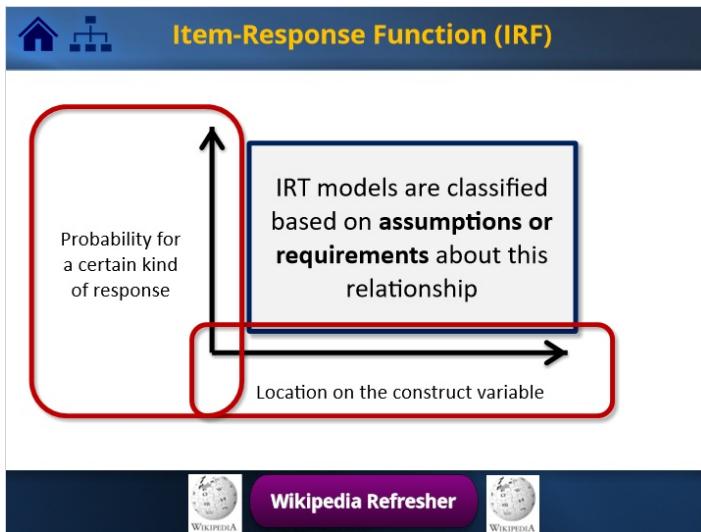
## 2.9 Scaling Steps



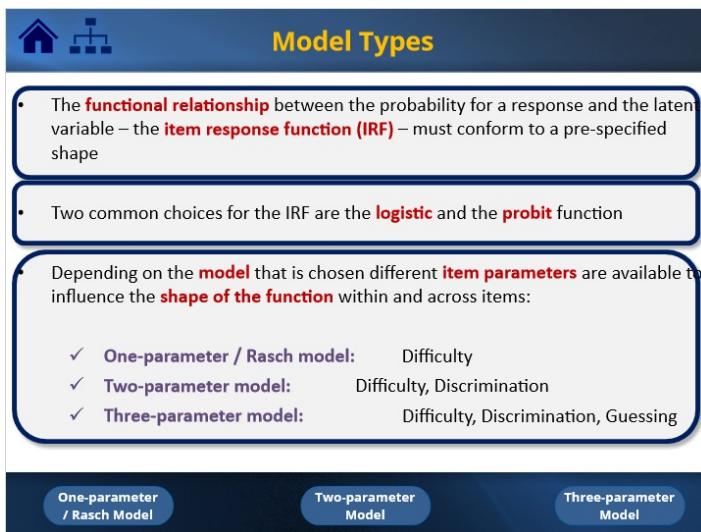
## 2.10 Framework Selection



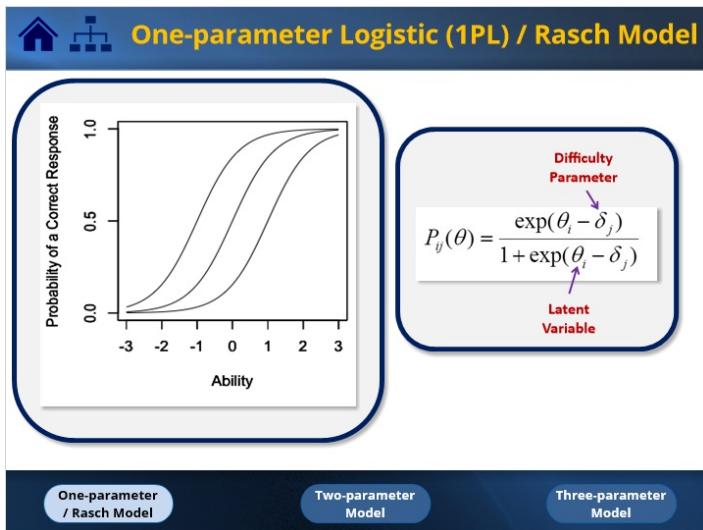
## 2.11 Item Response Function (IRF)



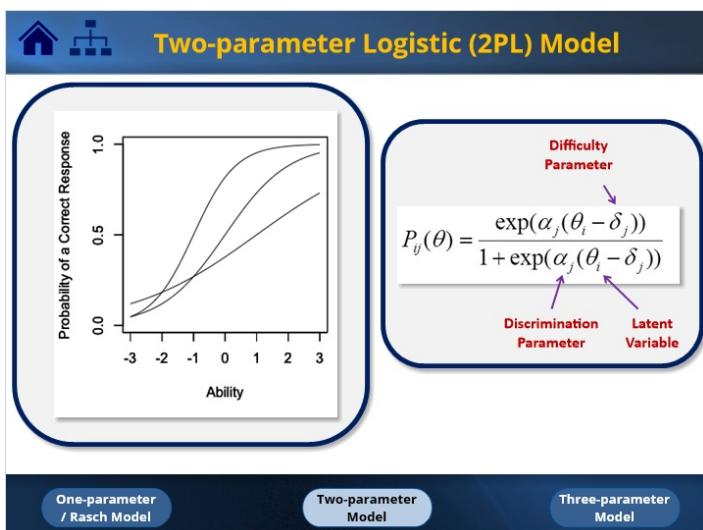
## 2.12 Parametric IRT



## 2.13 One-parameter Model

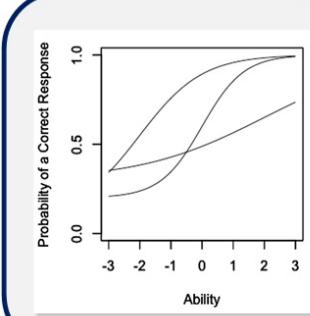


## 2.14 Two-parameter Model



## 2.15 Three-parameter Model

**Three-parameter Logistic (3PL) Model**



**Pseudo-guessing Parameter**

**Difficulty Parameter**

**Discrimination Parameter**

**Latent Variable**

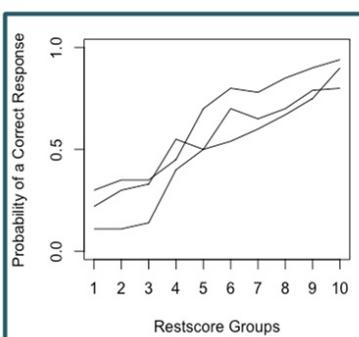
$$P_y(\theta) = \chi_j + (1 - \chi_j) \frac{\exp(\alpha_j(\theta_j - \delta_j))}{1 + \exp(\alpha_j(\theta_j - \delta_j))}$$

One-parameter / Rasch Model    Two-parameter Model    Three-parameter Model

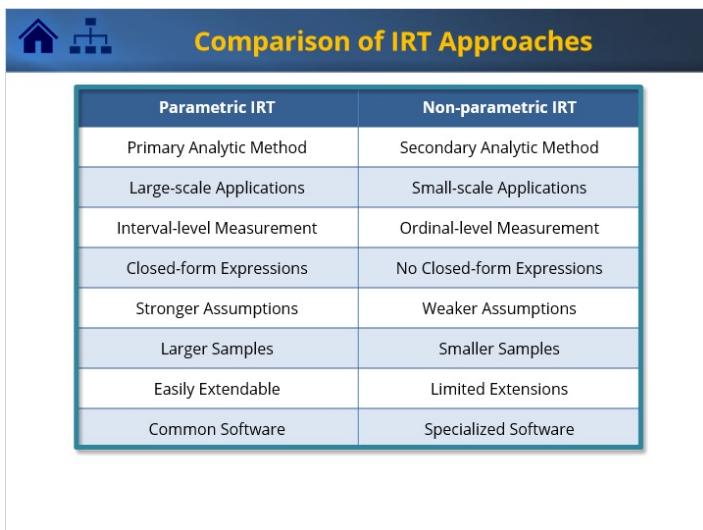
## 2.16 Nonparametric IRT

**Nonparametric Item-Response Theory**

The shape of the item response function is **not specified by parameters** in an algebraic formula



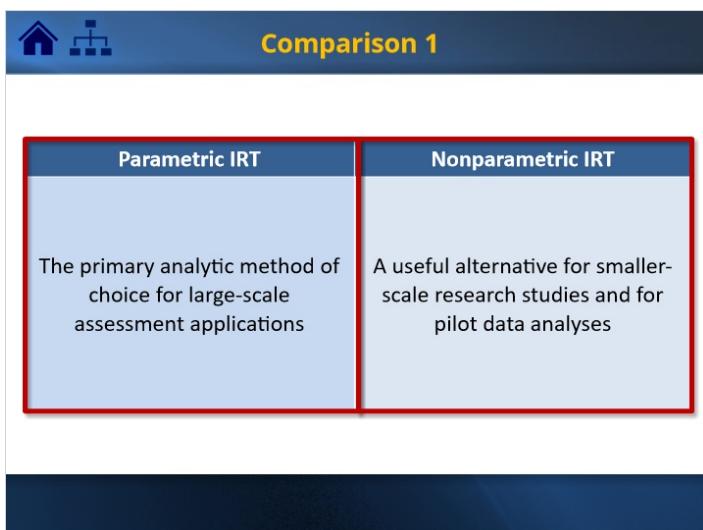
## 2.17 Framework Comparisons



A table titled "Comparison of IRT Approaches" comparing Parametric IRT and Non-parametric IRT. The table has two columns: "Parametric IRT" and "Non-parametric IRT".

Parametric IRT	Non-parametric IRT
Primary Analytic Method	Secondary Analytic Method
Large-scale Applications	Small-scale Applications
Interval-level Measurement	Ordinal-level Measurement
Closed-form Expressions	No Closed-form Expressions
Stronger Assumptions	Weaker Assumptions
Larger Samples	Smaller Samples
Easily Extendable	Limited Extensions
Common Software	Specialized Software

## 2.18 Comparison 1



A table titled "Comparison 1" comparing Parametric IRT and Nonparametric IRT. The table has two columns: "Parametric IRT" and "Nonparametric IRT".

Parametric IRT	Nonparametric IRT
The primary analytic method of choice for large-scale assessment applications	A useful alternative for smaller-scale research studies and for pilot data analyses

## 2.19 Comparison 2

Comparison 2	
Parametric IRT	Nonparametric IRT
Necessary when parameters are needed for form equating, item banking, and adaptive assessment	Can be used in cases when simpler test procedures are sufficient

## 2.20 Comparison 3

Comparison 3	
Parametric IRT	Nonparametric IRT
Necessary when follow-up analyses require interval level of measurement	Can be used when follow-up analyses only require ordinal level of measurement

## 2.21 Comparison 4

Comparison 4	
Parametric IRT	Nonparametric IRT
Provides closed-form expressions for item response curves that yield interpretable parameters for statistical inference	Lack of closed-form expression; highlights adherence to basic measurement properties but without formal parameters

## 2.22 Comparison 5

Comparison 5	
Parametric IRT	Nonparametric IRT
Requires that several stringent assumptions hold for estimates to be reliable and interpretable	Requires less stringent / more flexible assumptions but allows for weaker inference

## 2.23 Comparison 6

Comparison 6	
Parametric IRT	Nonparametric IRT
Often requires large samples of students and items	Can be used with small samples of students and items

## 2.24 Comparison 7

Comparison 7	
Parametric IRT	Nonparametric IRT
Can be easily extended to accommodate multiple dimensions and other complex data properties	Cannot be easily extended to include accommodations for more complex data properties

## 2.25 Comparison 8

Comparison 8	
Parametric IRT	Nonparametric IRT
Can be estimated with common software routines	Requires the use of even more specialized software routines

## 2.26 Bookmark: Misconceptions



## 2.27 Common Misconceptions

The slide is titled "Common Misconceptions" in a yellow header. On the left, there is a sidebar with five purple buttons, each labeled "Misconception 1" through "Misconception 5". The main content area features a woman with her arms outstretched, looking upwards, surrounded by several colorful thought bubbles. In the bottom right corner of the main area, there is a blue button labeled "Back to Topics".

## 2.28 Misconception 1

The slide is titled "Misconception 1" in a yellow header. It features a table with two columns: "Misconception" and "Reality". The "Misconception" column contains the text: "Mokken scaling is a nonparametric version of the Rasch model". The "Reality" column contains the text: "This is only true for the dichotomous double monotonicity model".

Misconception	Reality
Mokken scaling is a nonparametric version of the Rasch model	This is only true for the dichotomous double monotonicity model

## 2.29 Misconception 2

Misconception 2	
Misconception	Reality
All response data fit Mokken models	Mokken models have relatively stringent assumptions that are often violated in practice

## 2.30 Misconception 3

Misconception 3	
Misconception	Reality
Mokken scaling is the only type of nonparametric item response theory	There are other approaches to nonparametric IRT that have different sets of assumptions

## 2.31 Misconception 4

Misconception 4	
Misconception	Reality
Mokken scaling is too simplistic to be useful	Mokken scaling provides a variety of indicators of item quality and person fit

## 2.32 Misconception 5

Misconception 5	
Misconception	Reality
No one uses Mokken scaling anymore	<ul style="list-style-type: none"><li>Methodological &amp; applied Mokken scaling research appears frequently in top educational measurement, psychology, &amp; statistics journals</li><li>R package is available that is frequently updated</li></ul>

## 2.33 Bookmark: Summary



## 2.34 Summary: Section 1

Summary

- MSA is a **probabilistic-nonparametric approach to IRT** that provides a systematic framework for evaluating measurement quality in terms of fundamental measurement properties.
- MSA can be used to **explore fundamental measurement properties**, including invariant person and item ordering, when an ordinal level of measurement is sufficient to inform decisions based on a measurement procedure.
- MSA is an especially **useful approach in contexts in which the underlying response processes are not well understood** such as for measuring affective variables.
- MSA is also useful in contexts in which information about measurement quality and person and item ordering is needed, but **sample sizes are not sufficient to achieve stable estimates** based on parametric IRT models.

## 2.35 Section 1 Bookend



This is the end of this section.

## 3. Section 2: Mokken Models

### 3.1 Cover: Section 2



**Section 2:**  
**Mokken**  
**Models**  
**[10 Minutes]**

### 3.2 Objectives: Section 2

 Learning Objectives



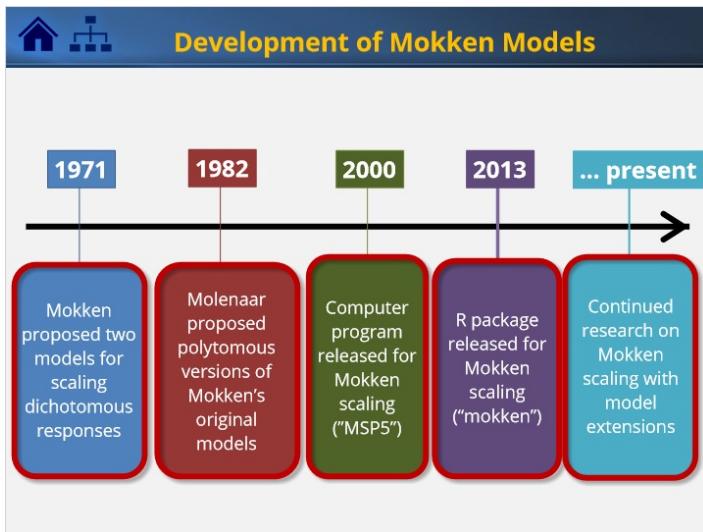
**I.** Understand the history of Mokken scale analysis and its current developments

**II.** Understand the key differences between the Monotone Homogeneity and the Double Monotonicity model

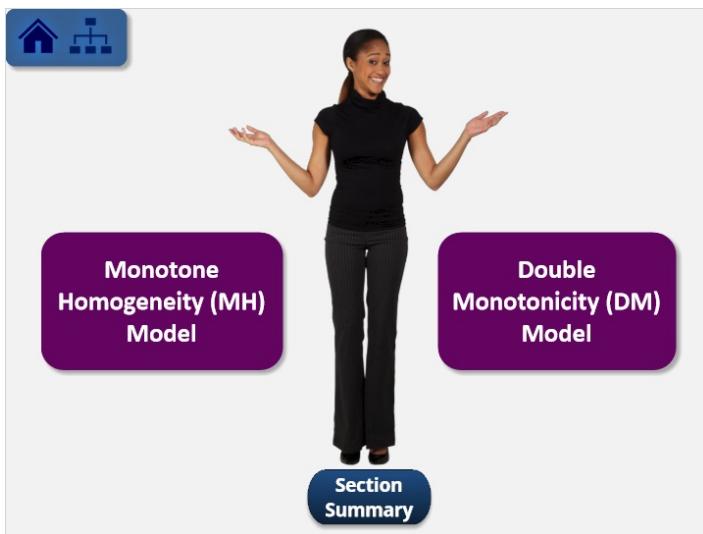
**III.** Understand each of the four key assumptions and their relationships to assumptions in parametric IRT models

**IV.** Understand how each assumption can be visually evaluated or represented

### 3.3 Development of Mokken Models



### 3.4 Model Selection



### 3.5 Bookmark: MH Model



### 3.6 Learner Objective: MH Model

  MH Model: Learner Properties

Determine whether **learners** can be meaningfully ordered by their **total scores**



Low Achievement  High Achievement

### 3.7 MH Model Assumptions

  MH Model: Assumptions



1. Unidimensionality
2. Local Independence
3. Monotonicity

### 3.8 Unidimensionality

  **Assumption 1: Unidimensionality**

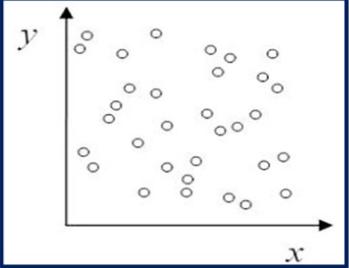
High

Responses reflect **only one major underlying dimension**  
(i.e., construct, latent trait)

Low

### 3.9 Local Independence

  **Assumption 2: Local Independence**

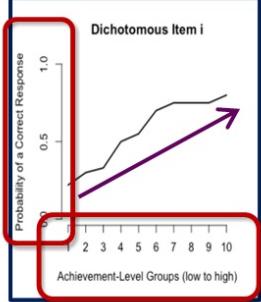


No meaningful (systematic) relationships among responses after controlling for the major underlying dimension

### 3.10 Monotonicity: Dichotomous Items

  Assumption 3: Monotonicity (I)

Dichotomous Item i



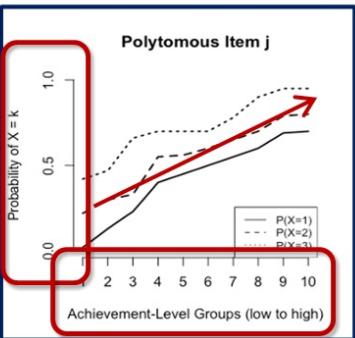
The graph shows the probability of a correct response on the y-axis (ranging from 0.0 to 1.0) versus achievement level groups on the x-axis (ranging from 1 to 10). A solid purple line represents the probability, which starts at approximately 0.1 for group 1 and increases to about 0.8 for group 10. A red box highlights the y-axis label 'Probability of a Correct Response' and the x-axis label 'Achievement-Level Groups (low to high)'. A red arrow points to the right along the curve.

The probability for a **correct response** is **non-decreasing** across **increasing levels** of achievement

### 3.11 Monotonicity: Polytomous Items

  Assumption 3: Monotonicity (II)

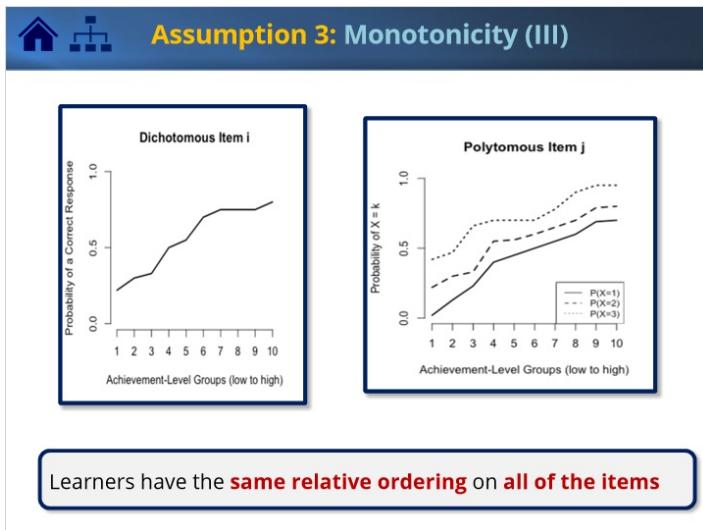
Polytomous Item j



The graph shows the probability of rating  $X = k$  on the y-axis (ranging from 0.0 to 1.0) versus achievement level groups on the x-axis (ranging from 1 to 10). Three curves are shown: a solid red line for  $P(X=1)$ , a dashed black line for  $P(X=2)$ , and a dotted black line for  $P(X=3)$ . All three curves show an increasing trend as achievement level increases. A red box highlights the y-axis label 'Probability of  $X = k$ ' and the x-axis label 'Achievement-Level Groups (low to high)'. A red arrow points to the right along the curves.

The probability for a **rating in a particular category** is **non-decreasing** across **increasing levels** of achievement

### 3.12 Monotonicity: Item Ordering



### 3.13 Bookmark: DM Model



### 3.14 Learner Objective: DM Model

  **DM Model: Learner Properties**

Determine whether **learners** can be meaningfully ordered by their **total scores**

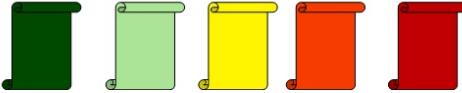


Low Achievement  High Achievement

### 3.15 Item Objective: DM Model

  **DM Model: Item Properties**

Determine whether **items** can be ordered the **same way** over learners



Easy  Hard

### 3.16 DM Model Assumptions

  **DM Model: Assumptions**

- 1. Unidimensionality
- 2. Local Independence
- 3. Monotonicity
- 4. Invariant Item Ordering

**MH Model Assumptions**

**Additional Assumption**

[Back to Model Selection](#)

### 3.17 Unidimensionality

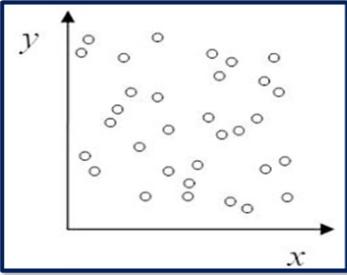
  **Assumption 1: Unidimensionality**

Responses reflect **only one major underlying dimension**  
(i.e., construct, latent trait)



### 3.18 Local Independence

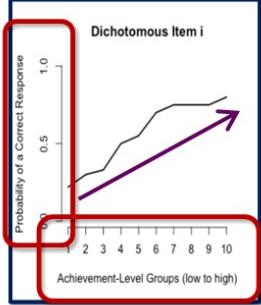
  Assumption 2: Local Independence



No meaningful (systematic) relationships among responses after controlling for the major underlying dimension

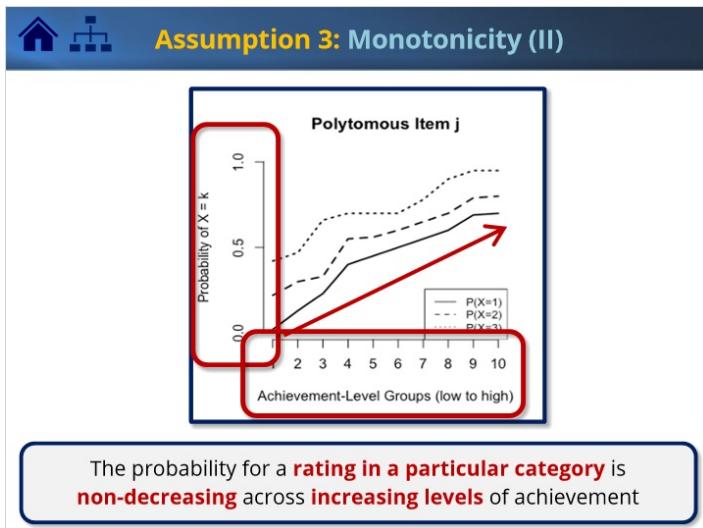
### 3.19 Monotonicity: Dichotomous Items

  Assumption 3: Monotonicity (I)

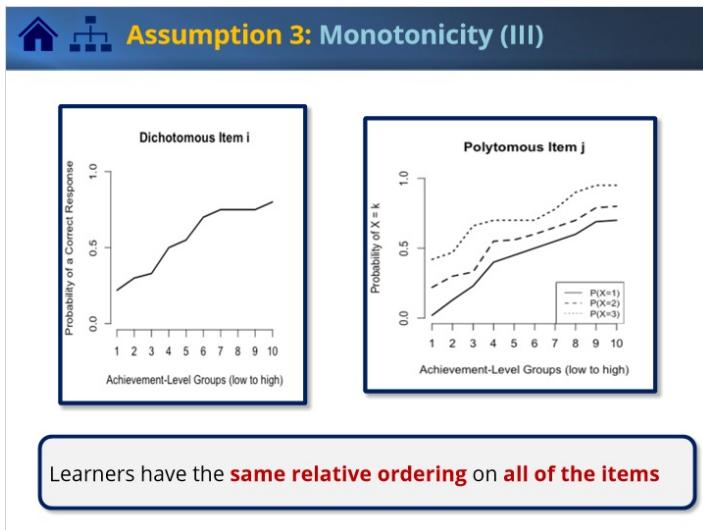


The probability for a **correct response** is **non-decreasing** across **increasing levels** of achievement

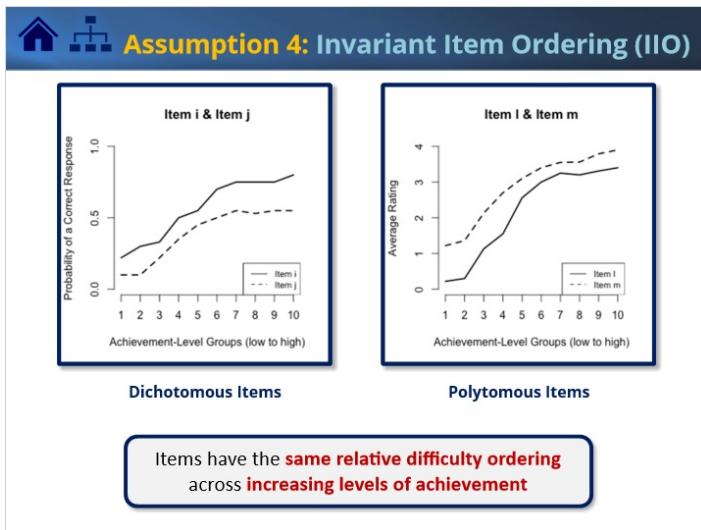
### 3.20 Monotonicity: Polytomous Items



### 3.21 Monotonicity: Item Ordering



### 3.22 Invariant Item Ordering



### 3.23 Bookmark: Summary



### 3.24 Summary: Section 2

**Summary**

- Two key models form the theoretical foundation of Mokken scale analysis:  
**monotone homogeneity (MH) & double monotonicity (DM) model**
- Both models allow for a **rank-ordering of learners** but only the latter allows for an **invariant ordering of items** across the scale.

**Table 1. Alignment Between Mokken Model Assumptions and Indicators of Measurement Quality**

Assumptions	Double Monotonicity Model	Monotone Homogeneity Model	Model-Based Indicators
Monotonicity	✓	✓	(A) Item/Rater monotonicity
Conditional Independence	✓	✓	(B) Item/Rater scalability coefficients
Unidimensionality	✓	✓	(A) Item/Rater monotonicity; (B) Item/Rater scalability
Nonintersecting Response Functions	✓		(C) Invariant item/rater ordering

### 3.25 Outlook

**Outlook**

- Modeling extensions** have been proposed in recent years, which mimic modeling developments in certain areas of parametric IRT.
- Model-data fit** can be evaluated via scaling coefficients at the item, item pair, and scale level for which hypothesis tests are available.
- Specialized software packages** exist, including a package in the freeware suite *R* ("mokken")

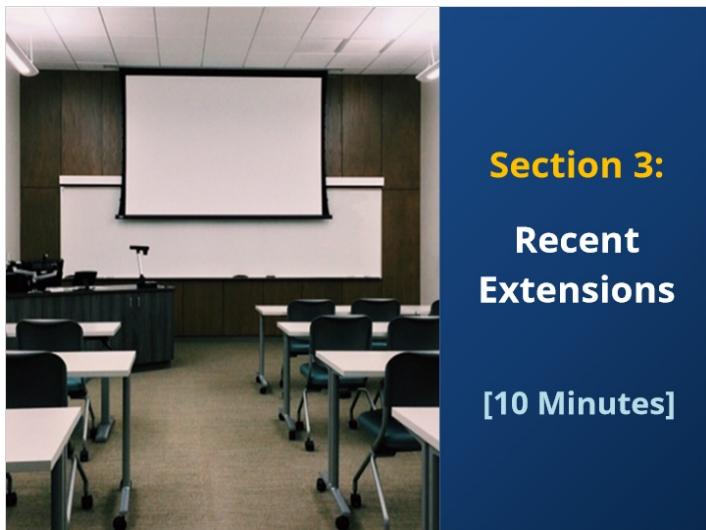


### **3.26 Bookend:Summary**



## **4. Section 3: Recent Extensions**

### **4.1 Cover: Section 3**



## 4.2 Objectives: Section 3

  Learning Objectives



**I.** Understand the idea that Mokken scale analysis is an active area of research

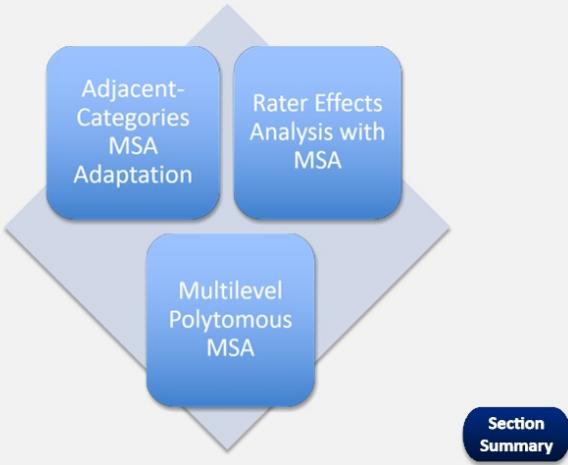
**II.** Understand the key extensions of polytomous models, rater effect models, and multilevel models

**III.** Understand the relationship between the nonparametric models and their parametric analogues

**IV.** Understand how these models can be used in assessment development

## 4.3 Extension Selection



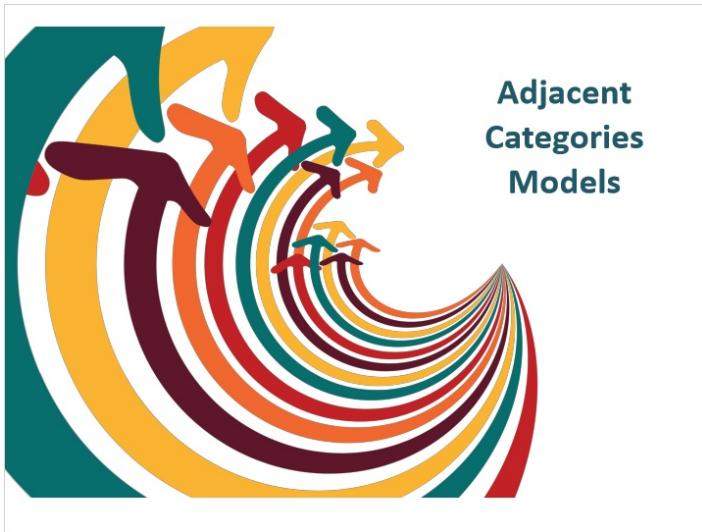
**Adjacent-Categories MSA Adaptation**

**Rater Effects Analysis with MSA**

**Multilevel Polytomous MSA**

**Section Summary**

#### 4.4 Bookmark: Adj Cat Models



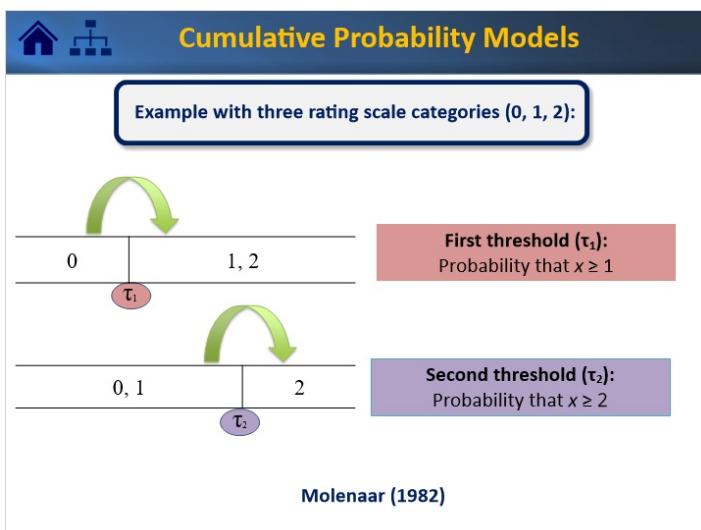
#### 4.5 ACMs: Introduction (I)

A screenshot of a presentation slide. The title bar says 'Adjacent Category Models'. The main content area contains a box with the following text: 'Adaptation of **polytomous Mokken models** with the goal to more closely align Mokken models to **rating scale interpretations** in educational assessments'. Below this, there is a flow diagram with two green boxes connected by a blue arrow. The first box says 'Different category probability formulations' and the second box says 'Different interpretations of rating scale categories'.

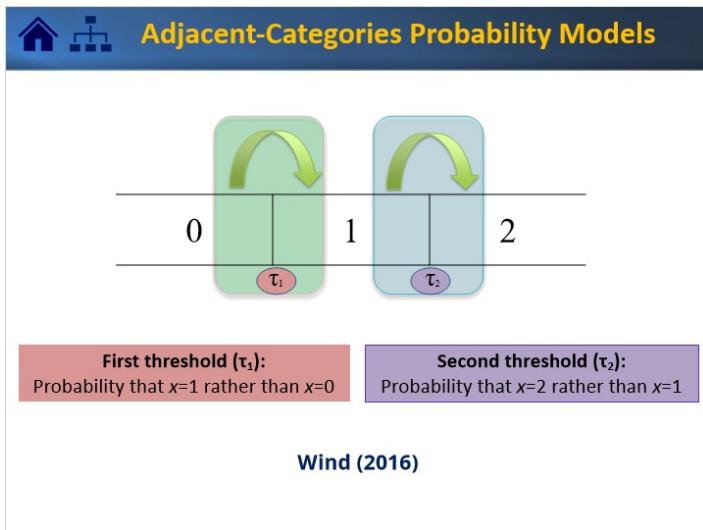
## 4.6 ACMs: Introduction (II)

Comparison with Parametric Methods	
MSA Approach	Parametric Analogue
Adapted polytomous Mokken scaling models with adjacent-categories threshold formulations	Parametric IRT models with adjacent-categories threshold formulations (e.g., Rating Scale Model & Partial Credit Model)

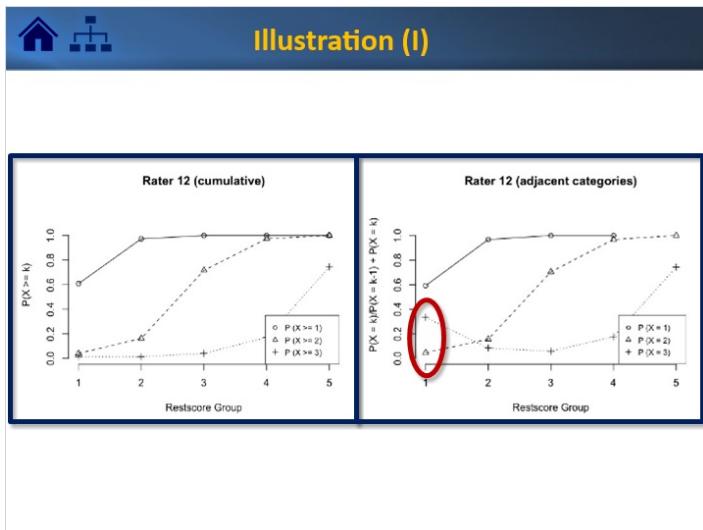
## 4.7 ACMs: Introduction (III)



## 4.8 ACMs: Introduction (V)



## 4.9 ACMs: Illustration (I)



## 4.10 ACMS: Illustration (II)

**Illustration (II)**

**Table 2. Scalability Results.**

Rater	Conventions $H_i$		Organization $H_i$		Sentence Formation $H_i$		Style $H_i$	
	MH-R	ac-MH	MH-R	ac-MH	MH-R	ac-MH	MH-R	ac-MH
1	0.81	0.79	0.80	0.74	0.84	0.20 <sup>a</sup>	0.77	0.77
2	0.78	0.78	0.74	0.71	0.78	0.64 <sup>a</sup>	0.76	0.77
3	0.82	0.82	0.79	0.75	0.81	0.32 <sup>a</sup>	0.78	0.79
4	0.81	0.79	0.81	0.72	0.78	0.30 <sup>a</sup>	0.77	0.72
5	0.83	0.81	0.79	0.73	0.81	0.69 <sup>a</sup>	0.76	0.75
6	0.76	0.75	0.80	0.47	0.78	0.57 <sup>a</sup>	0.74	0.75
7	0.83	0.83	0.79	0.73	0.82	0.66 <sup>a</sup>	0.78	0.77
8	0.83	0.79	0.83	0.75	0.86	0.66 <sup>a</sup>	0.82	0.80
9	0.82	0.81	0.74	0.71	0.82	0.66 <sup>a</sup>	0.78	0.79
10	0.82	0.79	0.76	0.70	0.83	0.68 <sup>a</sup>	0.78	0.79
11	0.81	0.78	0.79	0.73	0.81	0.33 <sup>a</sup>	0.78	0.80
12	0.83	0.83	0.78	0.72	0.80	0.63 <sup>a</sup>	0.78	0.79
13	0.84	0.75	0.78	0.72	0.85	0.42 <sup>a</sup>	0.76	0.76
14	0.79	0.75	0.83	0.74	0.81	0.37 <sup>a</sup>	0.77	0.78
15	0.78	0.77	0.76	0.71	0.81	0.65 <sup>a</sup>	0.78	0.76
16	0.83	0.81	0.78	0.73	0.82	0.68 <sup>a</sup>	0.80	0.80
17	0.77	0.78	0.75	0.71	0.81	0.70 <sup>a</sup>	0.75	0.73
18	0.76	0.78	0.77	0.71	0.77	0.70 <sup>a</sup>	0.76	0.73
19	0.81	0.81	0.78	0.73	0.83	0.66 <sup>a</sup>	0.78	0.79
20	0.78	0.78	0.75	0.67	0.80	0.13 <sup>a</sup>	0.74	0.74
Overall $H$	0.81	0.79	0.78	0.67	0.81	0.67 <sup>a</sup>	0.77	0.77

Note: MH-R = monotone homogeneity for ratings; ac-MH = adjacent-categories monotone homogeneity.  
<sup>a</sup>Indicates change in unit, leading non-monotone classification based on the following criteria:  $H \geq 0.50$ , strong;  $0.40 \leq H < 0.50$ , medium;  $0.30 \leq H < 0.40$ , weak (Mokken, 1971).

## 4.11 ACMS: References

**References: Adjacent-Categories Models**

**Wind (2017)**

**Wind & Schumacker (2017)**

**+ more work in development**

#### 4.12 Bookend: Adj Cat Models



This is the end of this section.

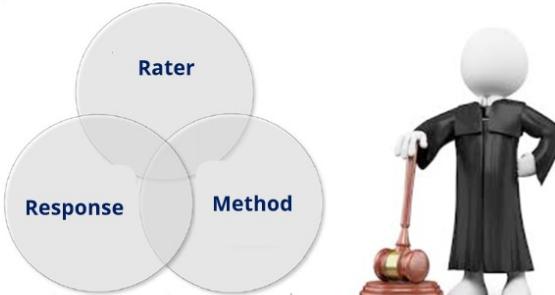
#### 4.13 Bookmark: Rater Models



#### 4.14 RMs: Introduction (I)

**Rater Effects Analysis**

Research that uses Mokken scaling to examine **rater effects** in **performance assessments**



adapted from Wolfe et al. (2016)

#### 4.15 RMs: Introduction (II)

**Definition: Rater Effects**

Raters' scoring tendencies that result in **inappropriate ratings assigned to learner performances** given the quality of the learner's response.

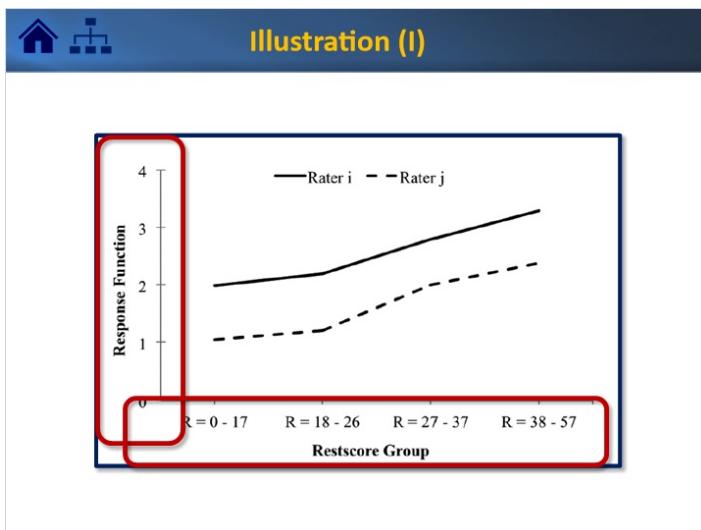
Common effects include **leniency / severity, centrality, halo effects, inaccuracy, and differential dimensionality**. There is a rich body of literature on human rating processes, in particular for constructed responses such as essays.



#### 4.16 RMs: Introduction (III)

Comparison with Parametric Methods	
MSA Approach	Parametric Analogue
<p>Nonparametric approach allows for the examination of rater effects via graphical displays of idiosyncratic rater behavior</p> <p>Scalability coefficients for individual raters can be used to quantify their relative fit with the model assumptions</p>	<p>Parametric IRT models include parameters for rater characteristics</p> <p>Inclusion of rater parameters improves score precision for learners and allows for diagnostic information about raters</p>

#### 4.17 RMs: Illustration (I)



#### 4.18 ACMs: Introduction (IV)

**Model Criticism**

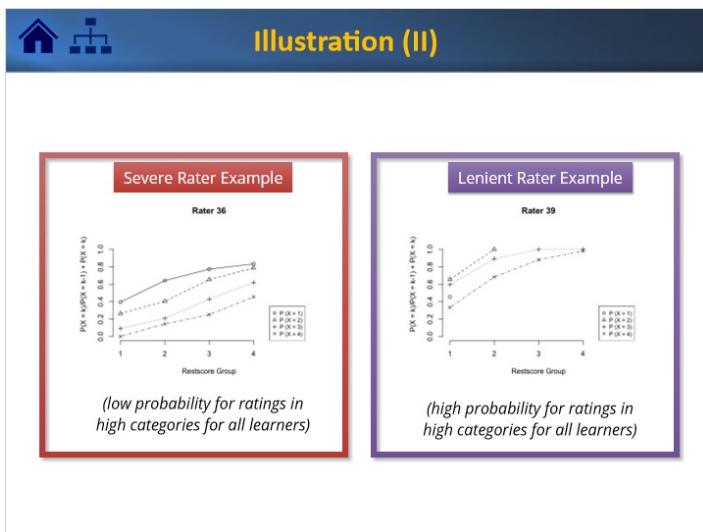


The model specifies the probability that a person will be classified **above any threshold....**

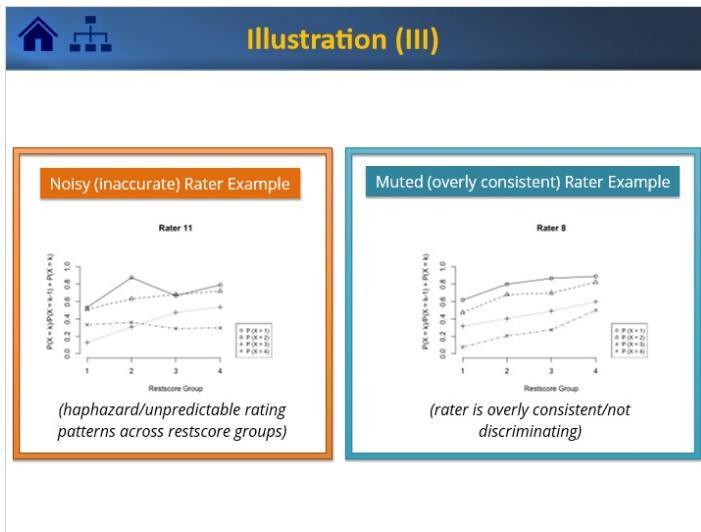
This does **not seem consistent** with performance assessment—judges locate a performance **in one of the categories, not in and beyond** any particular category

(Andrich, 2015, p. 6)

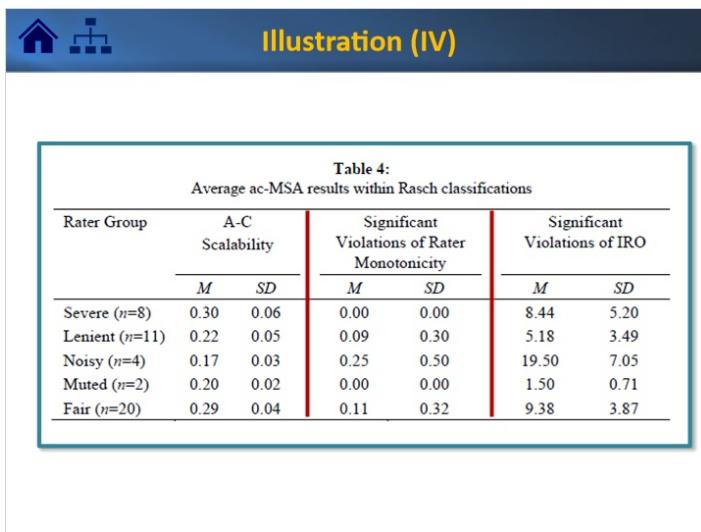
#### 4.19 RMs: Illustration (II)



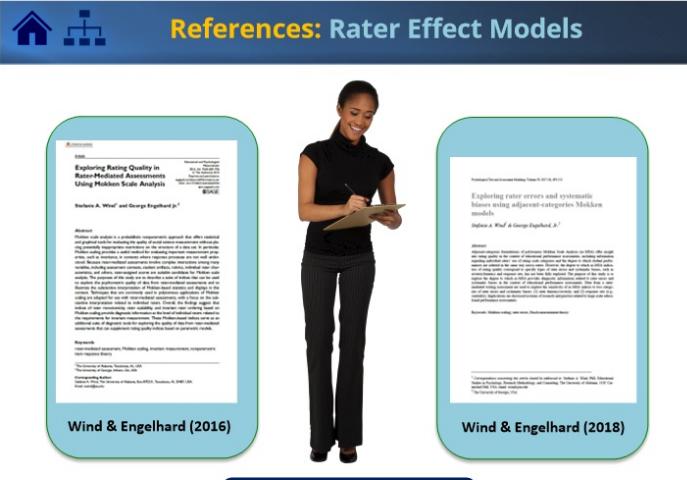
## 4.20 RMs: Illustration (III)



## 4.21 RMs: Illustration (IV)



## 4.22 RMs: References



**References: Rater Effect Models**

**Wind & Engelhard (2016)**

**Exploring Rating Quality in Rater-Moderated Assessments Using Mokken Scale Analysis**

Stephan A. Wind\* and George Engelhard, Jr.<sup>1</sup>

**Abstract**

Rating scale analysis is a probability approach that often provides useful information about the quality of rating scales. However, rating scales are often used in rater-moderated assessments, such as in the case of students writing prompts, which makes rating scale analysis a poor choice for these types of assessments. This study explores the use of Mokken scale analysis to examine the quality of rating scales in rater-moderated assessments. Mokken scale analysis is a nonparametric approach that is based on the assumption that the data are generated by a single latent variable, such as a rating scale. This study shows that Mokken scale analysis can be used to examine the quality of rating scales in rater-moderated assessments and that it can be used to identify rating scales that are not suitable for use in these types of assessments. The results of this study indicate that Mokken scale analysis can be used to identify rating scales that are not suitable for use in rater-moderated assessments, such as in the case of students writing prompts, and that it can be used to identify rating scales that are suitable for use in these types of assessments.

**Keywords**

rating scale analysis, rater-moderated assessments, student writing, student improvement, rater moderation

<sup>1</sup>Department of Psychology, Florida Institute of Technology, Melbourne, FL, USA

Corresponding author: Stephan A. Wind, Department of Psychology, Florida Institute of Technology, Melbourne, FL, USA.

**Wind & Engelhard (2018)**

**Exploring rater errors and systematic biases using adjacent categories Mokken models**

Stephan A. Wind\* and George Engelhard, Jr.<sup>1</sup>

**Abstract**

Rating scale analysis is a probability approach that often provides useful information about the quality of rating scales. However, rating scales are often used in rater-moderated assessments, such as in the case of students writing prompts, which makes rating scale analysis a poor choice for these types of assessments. This study explores the use of Mokken scale analysis to examine the quality of rating scales in rater-moderated assessments. Mokken scale analysis is a nonparametric approach that is based on the assumption that the data are generated by a single latent variable, such as a rating scale. This study shows that Mokken scale analysis can be used to examine the quality of rating scales in rater-moderated assessments and that it can be used to identify rating scales that are not suitable for use in these types of assessments. The results of this study indicate that Mokken scale analysis can be used to identify rating scales that are not suitable for use in rater-moderated assessments, such as in the case of students writing prompts, and that it can be used to identify rating scales that are suitable for use in these types of assessments.

**Keywords**

rating scale analysis, rater-moderated assessments, student writing, student improvement, rater moderation

<sup>1</sup>Department of Psychology, Florida Institute of Technology, Melbourne, FL, USA

Corresponding author: Stephan A. Wind, Department of Psychology, Florida Institute of Technology, Melbourne, FL, USA.

**+ more work in development**

## 4.23 Bookend: Rater Models

#### 4.24 Bookmark: Multilevel Models



#### 4.25 MLMs: Introduction (I)

**Multilevel Polytomous Models**

Approach to Mokken scaling that takes into account  
**nested structures in item response data**

(e.g., learners nested within schools, responses nested  
within raters, items nested within stimuli / testlets)

Rater 1      Rater 2      Rater 3

```
graph TD; R1[Rater 1] --- S11(( )); R1 --- S12(( )); R1 --- S13(( )); R2[Rater 2] --- S21(( )); R2 --- S22(( )); R2 --- S23(( )); R3[Rater 3] --- S31(( )); R3 --- S32(( )); R3 --- S33(( ));
```

## 4.26 MLMs: Introduction (II)

  Comparison with Parametric Methods

MSA Approach	Parametric Analogue
Multilevel Mokken Scaling	Multilevel Modeling (Hierarchical Linear Models)

## 4.27 MLMs: Introduction (III)

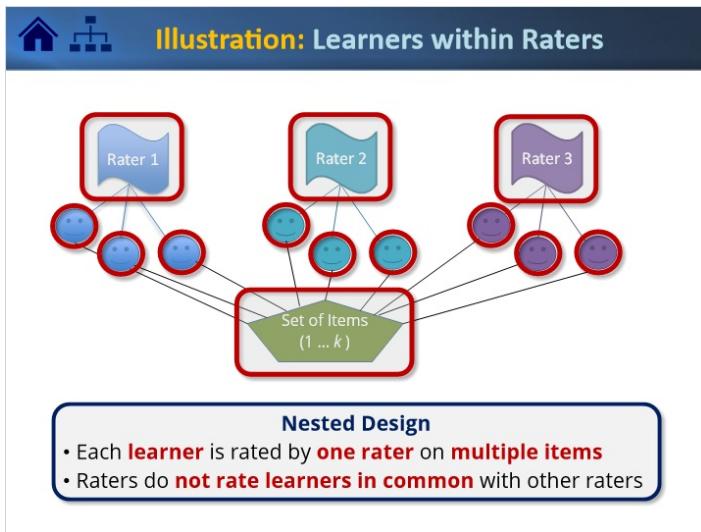
  Multilevel vs. Single-level

Multilevel MSA models include the **same major indicators of measurement quality** as single-level MSA models:

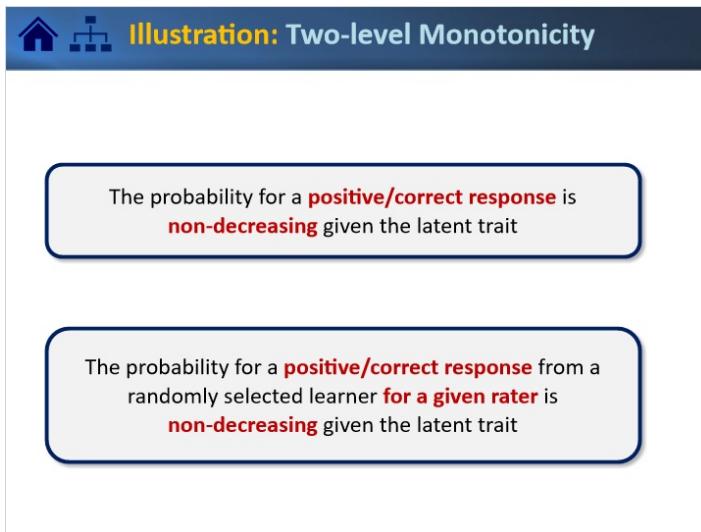
Monotonicity      Scalability      Invariant Ordering      }

Calculated for  
**all levels of the design**  
(e.g., within and between raters)

#### 4.28 MLMs: Illustration: Learners within Raters



#### 4.29 MLMs: Illustration: Two-level Monotonicity



#### 4.30 MLMs: Illustration: Two-level Scalability

  **Illustration: Two-level Scalability**

Scalability coefficients are calculated **within and between nested objects** (raters in our example)

The **ratio of within-rater scalability to between-rater scalability coefficients** reveals the extent to which responses reflect **learner variability** or **rater variability**

#### 4.31 MLMs: Illustration: Two-level Invariant Ordering

  **Illustration: Two-level Invariant Ordering**

Item ordering is **equal for all learner locations** on the latent trait

Item ordering for a randomly selected learner **for a given rater** is **equal for all locations** on the latent trait

## 4.32 MLMs: References

The screenshot shows a web page with a dark blue header containing a house icon and the text 'References: Multilevel Models'. Below the header, there are two side-by-side panels. The left panel is titled 'Scalability Coefficients for Two-Level Polychoric Item Scores: An Introduction and an Application' and is authored by Crisan, van de Pol, and van der Ark (2016). The right panel is titled 'Weighted Guttman Errors: Handling Tie and Two-Level Data' and is authored by Lutz Koopman, Bas de Zeeuw, and L. van der Ark. A central figure of a woman in a black turtleneck and black pants is standing between the two panels, holding a clipboard. At the bottom of the page is a blue button with the text '+ more work in development'.

## 4.33 Bookend: Multilevel Models



#### 4.34 Bookmark: Summary



#### 4.35 Summary: Section 3

• Nonparametric IRT, including Mokken scale analysis, is still an **active area of research**

• **Three key extensions** of the basic Mokken models include models for **polytomous data, rater effects, and multilevel data**

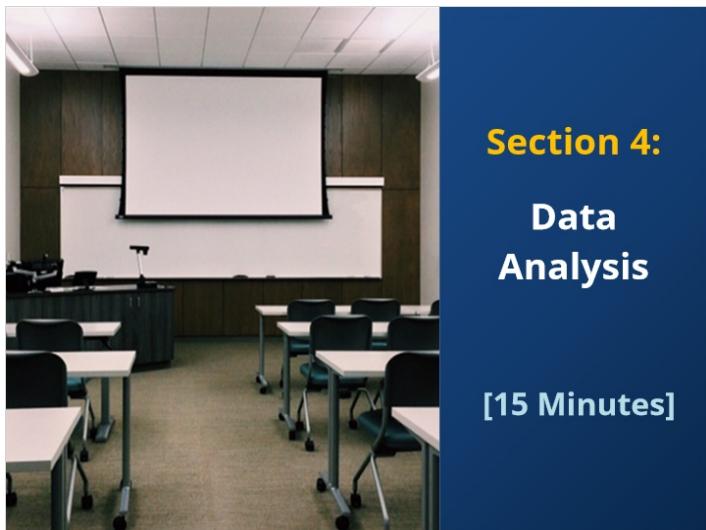
• All of these models have **analogues in parametric IRT** where **individual model parameters** are used to capture these effects

### 4.36 Bookend: Summary



## 5. Section 4: Data Analysis

### 5.1 Cover: Section 4



## 5.2 Objectives: Section 4

 **Learning Objectives**



**I.** Describe the general procedure for conducting a Mokken Scale Analysis      **III.** Describe and interpret procedures for calculating scalability coefficients

**II.** Describe and interpret graphical and statistical procedures for evaluating item monotonicity      **IV.** Describe and interpret graphical and statistical procedures for evaluating invariant item ordering

## 5.3 General Ideas

 **General Ideas**

- **Best practices** in item analysis under a nonparametric IRT approach mimic those under a parametric IRT approach.
- **Items are first developed** under best practices for assessment design and then tried out with a representative sample of learners from the target population.
- **Response data are then analyzed** to evaluate item properties (e.g., difficulty, discrimination) as well as scale properties (e.g., rank-ordering of learners with the items, score precision).
- **Misfitting items need to be revised and / or removed** from the scale depending on the assessment development stage.
- **Graphical and descriptive approaches** dominate nonparametric IRT but select inferential procedures are available.

## 5.4 Activity Selection

Click on one of the four areas to learn more.

**1. Import Data Matrix**

Items ( $i$ )

Students ( $n$ )

Ordinal Scored Responses:

$$X_{ni}$$

- Dichotomous (0,1)
- or
- Polytomous (0, ...,  $k$ )

**2. Analyze Items**

**A. Monotonicity**

Response functions within items

(Dichotomous & Polytomous)

(Polytomous)

Hypothesis Tests & Confidence Intervals

**B. Scalability**

Individual items:

$$H_i = 1 - \left( \sum_{j=1}^k F_{ij} / \sum_{j=1}^k E_{ij} \right)$$

Pairs of items:

$$H = F_{ij} / E_{ij}$$

All items:

$$H = 1 - \left( \sum_{i=1}^L \sum_{j=1}^k F_{ij} / \sum_{i=1}^L \sum_{j=1}^k E_{ij} \right)$$

Joint Response Functions for Pairs of Items:

(Dichotomous & Polytomous)

Hypothesis Tests & Confidence Intervals

**3. Interpret Results within Context**

Assessment consequences

Intended interpretation & use

Practical considerations

Content coverage

Findings for individual items:

- Monotonic/ Non-monotonic
- Scalable / Unscalable
- Invariant Order / Variant Order

**4. Modify Items**

Revise, Remove, or Replace misfitting items as appropriate, given interpretation within context of the assessment

Continue Iterating as Needed

## 5.5 Step 1

Click on one of the four areas to learn more.

**1. Import Data Matrix**

Items ( $i$ )

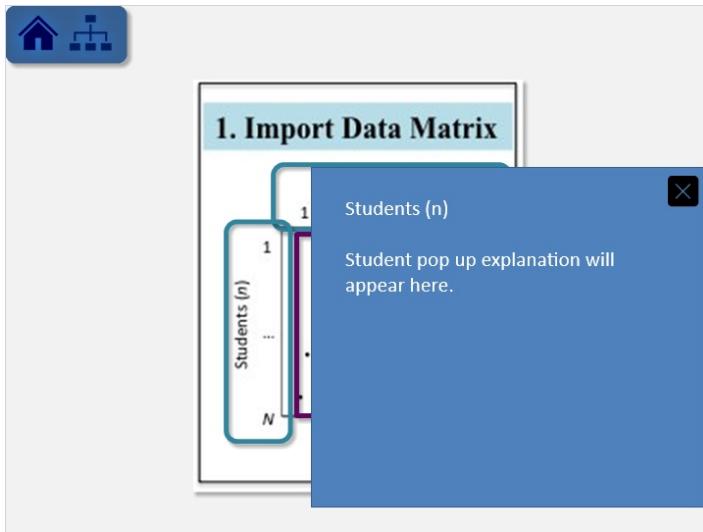
Students ( $n$ )

Ordinal Scored Responses:

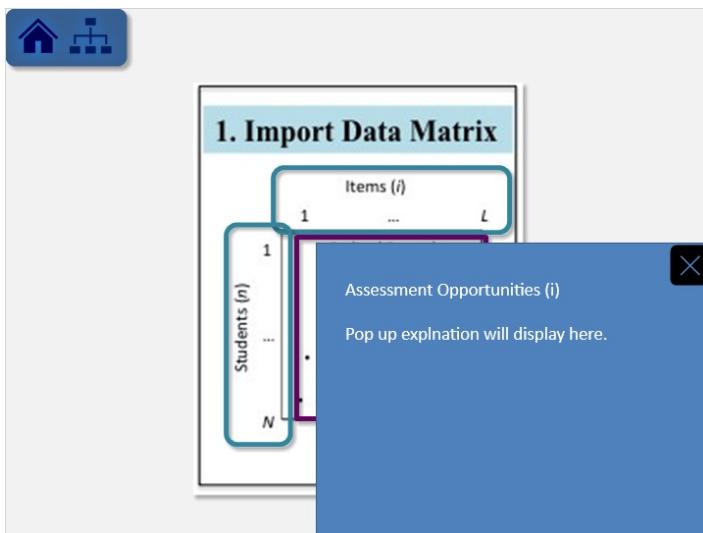
$$X_{ni}$$

- Dichotomous (0,1)
- or
- Polytomous (0, ...,  $k$ )

## Student Pop (Slide Layer)



## AO Pop (Slide Layer)



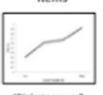
## 5.6 Step 2: Topic Selection



### 2. Analyze Items

#### A. Monotonicity

Response functions within items



(Dichotomous & Polytomous)

Response functions for rating scale categories



(Polytomous)

+ Hypothesis Tests & Confidence Intervals

#### B. Scalability

- Individual items:  
$$H_j = 1 - \left( \sum_{i \neq j} F_{ij} / \sum_{i \neq j} E_{ij} \right)$$

- Pairs of items:

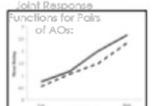
$$H_{ij} = F_{ij} / E_{ij}$$

- All items:  
$$H = 1 - \left( \sum_{i=1}^{k-1} \sum_{j=i+1}^k F_{ij} / \sum_{i=1}^{k-1} \sum_{j=i+1}^k E_{ij} \right)$$

+ Hypothesis Tests & Confidence Intervals

#### C. Invariant Ordering

Joint Response Functions for Pairs of items:



(Joint Response Functions for Pairs of Items)

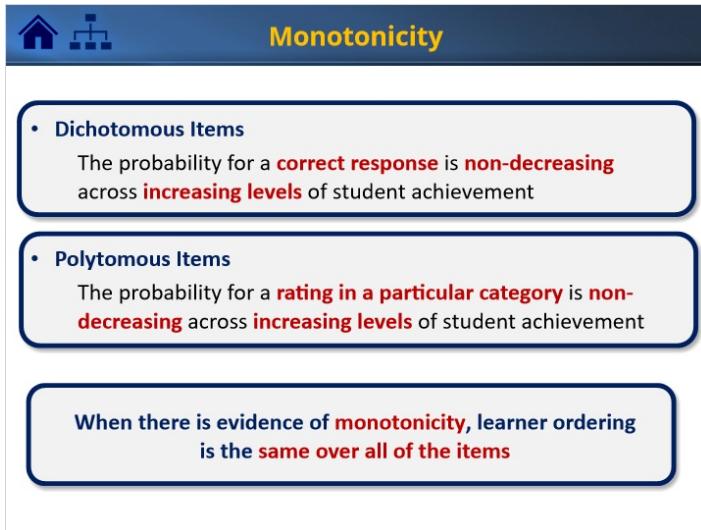
(Dichotomous & Polytomous)

+ Hypothesis Tests & Confidence Intervals

## 5.7 Bookmark: Monotonicity



## 5.8 Monotonicity (I)

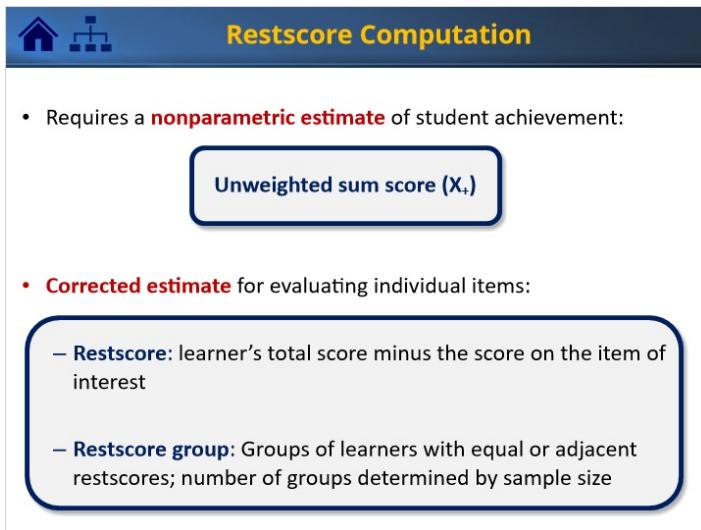


The slide has a blue header with a house icon and the word "Monotonicity". The content is organized into three main sections: "Dichotomous Items", "Polytomous Items", and a summary statement.

- Dichotomous Items**  
The probability for a **correct response** is **non-decreasing** across **increasing levels** of student achievement
- Polytomous Items**  
The probability for a **rating in a particular category** is **non-decreasing** across **increasing levels** of student achievement

When there is evidence of **monotonicity**, learner ordering is the **same over all of the items**

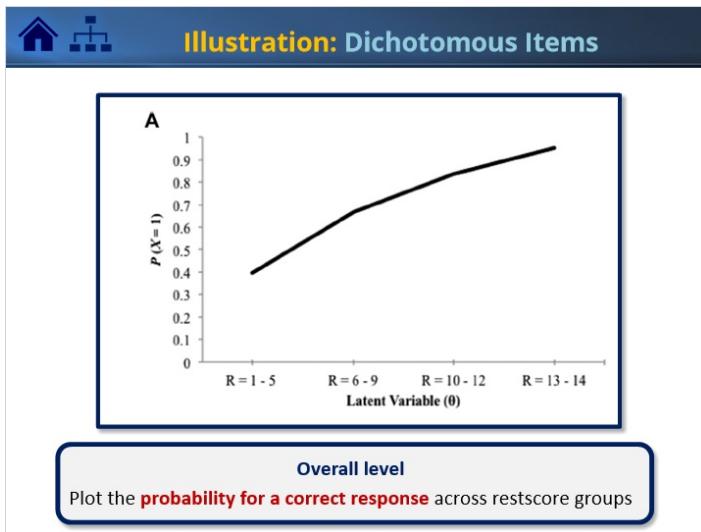
## 5.9 Monotonicity (II)



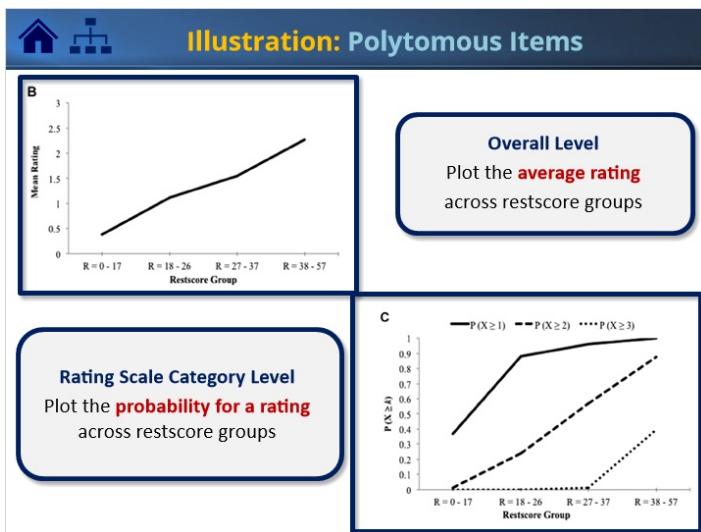
The slide has a blue header with a house icon and the words "Restscore Computation". The content is organized into three main sections: a general requirement, a corrected estimate for individual items, and two specific terms.

- Requires a **nonparametric estimate** of student achievement:  
**Unweighted sum score ( $X_i$ )**
- Corrected estimate** for evaluating individual items:
  - Restscore**: learner's total score minus the score on the item of interest
  - Restscore group**: Groups of learners with equal or adjacent restscores; number of groups determined by sample size

## 5.10 Monotonicity (III)



## 5.11 Monotonicity (IV)



## 5.12 Monotonicity (V)



### Hypothesis Testing (I)

- Statistical tests are one-sided, one-sample Z tests  
Available for both **dichotomous** and **polytomous items** and based on comparisons of **adjacent restscore groups** (lower vs. higher)
- Violations of monotonicity are identified with significant Z statistics
  - (a) **Dichotomous items**  
Statistically significant Z statistics indicate a **higher probability** for a **correct response** in the **lower restscore group**
  - (b) **Polytomous items**  
Statistically significant Z statistics indicate a **higher average rating** in the **lower restscore group**

## 5.13 Monotonicity (VI)



### Hypothesis Testing (II)



For a pair of adjacent restscore groups:

**H<sub>0</sub>:** the probability for a correct response is **equal** across the two restscore groups

**H<sub>A</sub>:** the probability for a correct response is **lower** in the higher restscore group

**Test Statistic**

## Test Statistic: Dichotomous Items (Slide Layer)

### Test Statistic: Dichotomous Items

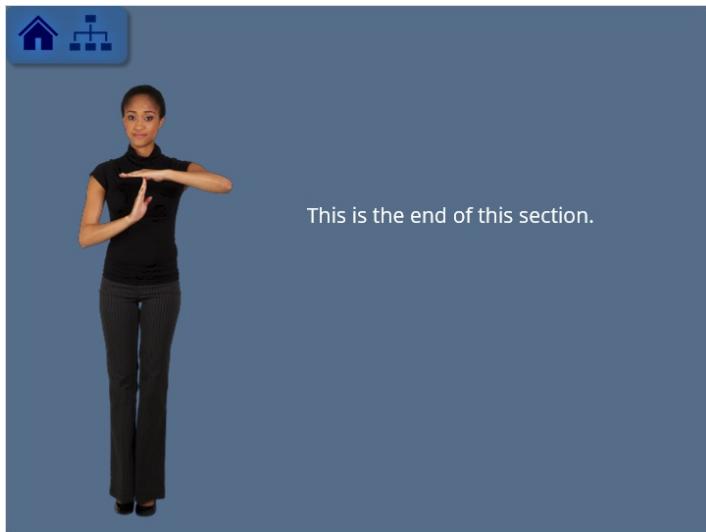
Statistical test is based on a **normal approximation to the hypergeometric distribution** with the marginals observed in a **2\*2 table for the restscore groups**:

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

see Molenaar and Sijtsma (2000, pp. 71-72)

[Back to Slide](#)

## 5.14 Bookend: Monotonicity



## 5.15 Bookmark: Scalability

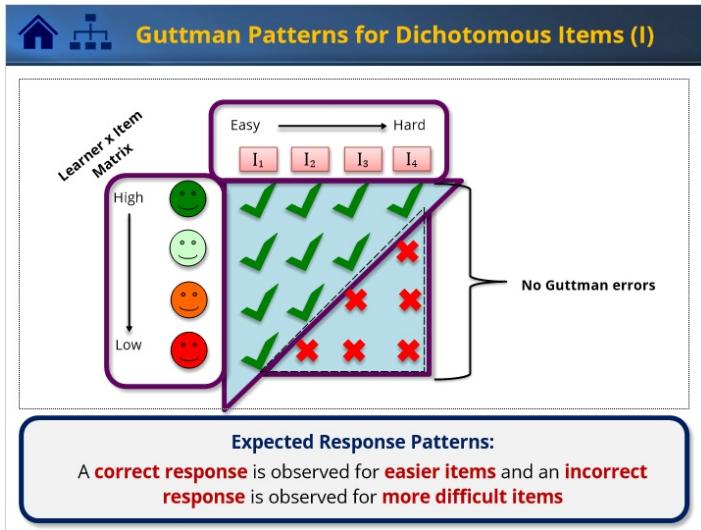


## 5.16 Scalability (I)

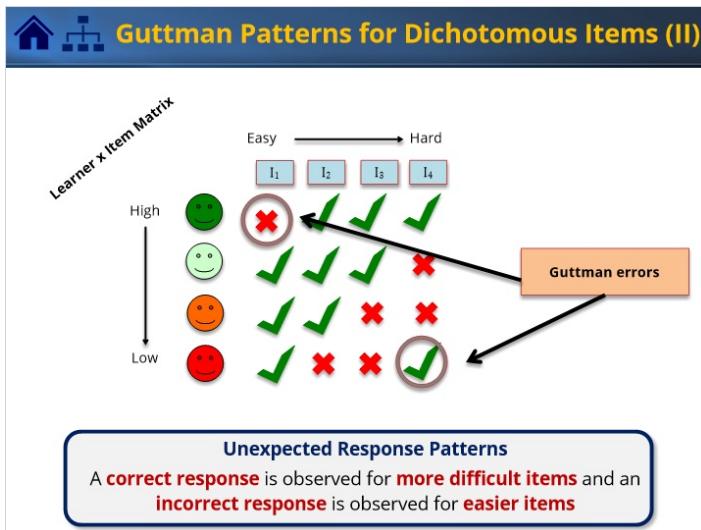
**Scalability Coefficients**

- Describe the degree to which **individual items, pairs of items, and overall sets of items** form a scale that can be used to order learners on a construct
- Summarize the influence of **Guttman errors** on a measurement procedure, where **fewer Guttman errors** mean **stronger evidence** for a meaningful interpretation of total scores
- Are **adapted** from Loevinger's (1948) **scalability coefficients**

## 5.17 Scalability (II)



## 5.18 Scalability (III)



## 5.19 Scalability (IV)

Panel A: Responses Contain No Guttman Errors

Students		Items				
		Easy		Difficult		
		Item 1	Item 2	Item 3	Item 4	Item 5
High	Student 1	1	1	1	1	1
↓	Student 2	1	1	1	1	0
	Student 3	1	1	1	0	0
	Student 4	1	1	0	0	0
Low	Student 5	1	0	0	0	0

Panel B: Responses Contain Two Guttman Errors

Students		Items				
		Easy		Difficult		
		Item 1	Item 2	Item 3	Item 4	Item 5
High	Student 1	0*	1	1	1	1
↓	Student 2	1	1	1	1	0
	Student 3	1	1	1	0	0
	Student 4	1	1	0	0	0
Low	Student 5	1	0	0	0	1*

## 5.20 Scalability (V)

  Illustration (I)

Step 1: Identify empirical item category order

Item <i>i</i>	Item <i>j</i>			
	0	1	2	3
0	(0, 0)	(0, 1)*	(0, 2)*	(0, 3)*
1	(1, 0)	(1, 1)*	(1, 2)*	(1, 3)*
2	(2, 0)	→(2, 1)	→(2, 2)	(2, 3)*
3	(3, 0)*	(3, 1)*	(3, 2)	→(3, 3)

Cell entries show item responses in the form (i, j)

## 5.21 Scalability (VI)

  **Illustration (II)**

**Step 2: Use item category order to identify Guttman-expected ratings**

Joint Rating (Item i, Item j)	Ordered Rating Scale Categories (Easy → Difficult)							
	Item i = 0	Item j = 0	Item i = 1	Item i = 2	Item i = 1	Item j = 2	Item j = 3	Item i = 3
0, 0	1	1	0	0	0	0	0	0
1, 0	1	1	1	0	0	0	0	0
2, 0	1	1	1	1	0	0	0	0
2, 1	1	1	1	1	1	0	0	0
2, 2	1	1	1	1	1	1	0	0
3, 2	1	1	1	1	1	1	1	0
3, 3	1	1	1	1	1	1	1	1

*Cell entries show recoded item responses where 0 = fail, 1 = pass*

## 5.22 Scalability (VII)

  **Illustration (III)**

**Step 3: Use item category order to identify Guttman errors**

Joint Rating (Item i, Item j)	Ordered Rating Scale Categories (Easy → Difficult)							
	Item i = 0	Item j = 0	Item i = 1	Item i = 2	Item i = 1	Item j = 2	Item j = 3	Item i = 3
0, 1	1	1	0	0	1*	0	0	0
0, 2	1	1	0	0	1*	1*	0	0
0, 3	1	1	0	0	1*	1*	0	1*
1, 1	1	1	1	0	1*	0	0	0
1, 2	1	1	1	0	1*	1	0	0
1, 3	1	1	1	1	1	1	0	1*
2, 3	1	1	1	1	1	0	0	1*
3, 0	1	0	1*	1	0	1*	1	0
3, 1	1	1	0	0	1*	0	0	0

*Cell entries show recoded item responses where 0 = fail, 1 = pass*

## 5.23 Scalability (VIII)

  General Form of Scalability Coefficients

$$H = 1 - \frac{F}{E}$$

Observed frequency of Guttman errors at level of analysis

Expected frequency of Guttman errors at level of analysis

## 5.24 Scalability (IX)

Individual Items

$$H_i = 1 - \left( \sum_{i \neq j} F_{ij} / \sum_{i \neq j} E_{ij} \right)$$

Pairs of Items

$$H_{ij} = F_{ij} / E_{ij}$$

All Items / Scale

$$H = 1 - \left( \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^k F_{ij}}{\sum_{i=1}^{k-1} \sum_{j=i+1}^k E_{ij}} \right)$$

**F** = Observed Guttman Errors, **E** = Expected Guttman Errors

## 5.25 Scalability ( $X$ )

  **Worked Example: Dichotomous Item Pair**

	Item $j = 0$	Item $j = 1$	Total
Item $i = 0$	66	16*	82
Item $i = 1$	89	97	186
Total	155	113	268

\* Guttman Error cell:

$$H_{ij} = 1 - \frac{16}{34.57} = 0.54$$

**Computation of  $E_{ij}$**

## Computation (Slide Layer)

  **Computation**

Expected error cell frequency ( $E_{ij}$ ) = (Row total \* Column total) /  $N$

$$E_{ij} = \frac{82 * 113}{268} = 34.57$$

Also: item  $i$  is more difficult than item  $j$  (113 vs. 186 correct responses)

**Back to Slide**

## 5.26 Scalability (XI)

  Interpreting Scalability Coefficients

- **Statistical Inference**

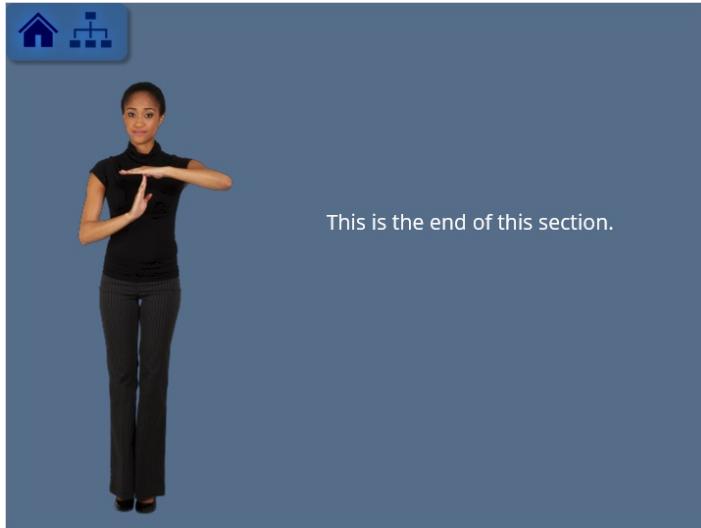
Standard errors can be used for **statistical inference** using **hypothesis tests** and **confidence intervals** with proper **sampling distributions** (e.g., Kuijpers, van der Ark, & Croon, 2013)

Confidence intervals help evaluate whether scalability coefficients are different from **known values** or to **compare scalability coefficients**.
- **Interpretational Guidance**

Usually, **values  $\geq 0.30$**  are considered "good enough," but these values have **not really been studied empirically**, especially not for polytomous items

Scalability coefficients are on a **continuous scale** and have **ordinal interpretations** (i.e., larger is better) with relative magnitudes informed by historical benchmarks and stakes of the assessment

## 5.27 Bookend: Scalability



## 5.28 Bookmark: Invariant Item Ordering



## 5.29 Invariant Ordering (I)

  **Invariant Ordering**

- Invariant ordering of learners across items

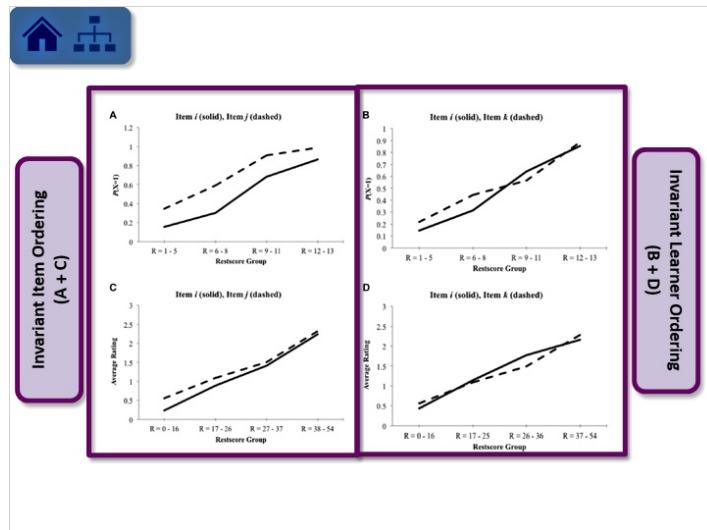
Learners have the **same relative order across items** with **different levels of difficulty**. Item response functions should be **non-decreasing**

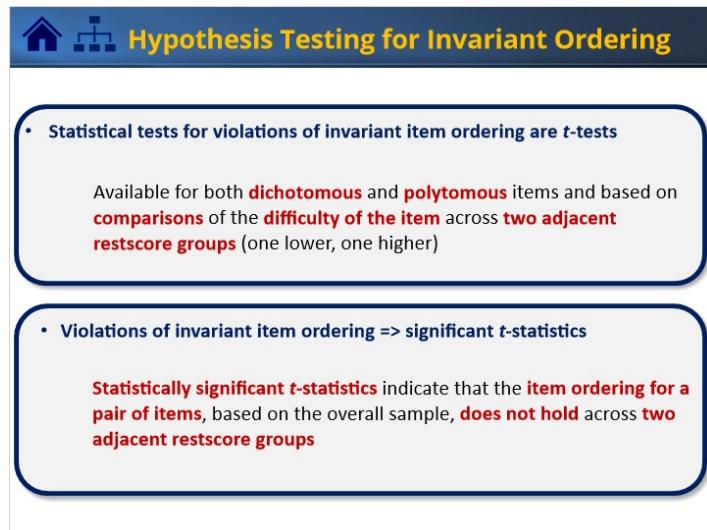
- Invariant ordering of items across learners

Items have the **same relative ordering** across **levels of the latent variable**, which is evaluated using **restscore groups**; item response functions should be **non-intersecting**

## 5.30 Invariant Ordering (II)



## 5.31 Invariant Ordering (III)



## 5.32 Invariant Ordering (IV)

 **Hypothesis Testing (I): Dichotomous Items**

**Conceptually:**

For a **pair of items ordered  $i < j$** , the **null hypothesis** that the probability for a correct response is **equal across the two items** is evaluated against the **alternative hypothesis** that the **item order is reversed ( $j < i$ )**, which would be a **violation** of invariant item ordering.

**Symbolically:**

**H<sub>0</sub>:**  $P(X_i = 1 | R = r) = P(X_j = 1 | R = r)$   
**H<sub>a</sub>:**  $P(X_i = 1 | R = r) > P(X_j = 1 | R = r)$

**Test Statistic**

## Test Statistic (Slide Layer)

 **Test Statistic: Dichotomous Items**

Statistical test is used to evaluate the extent to which **differences in the order of items** can be explained by **random variation in the sample / sampling error**. Create a **2-by-2 table of the frequency of 0s and 1s** and use this formula:

$$z = \sqrt{(2k + 2 + b)} - \sqrt{(2n)(2k + b)}$$
$$b = (2k + 1)(n) - (0n)(12n)$$

**$k$  = the smallest of the two frequencies of 1s between the items**

**$n$  = sum of the two frequencies of 1s between the items**

**Back to Slide**

### 5.33 Invariant Ordering (V)

  Hypothesis Testing (II): Polytomous Items

If the average ratings on item  $i$  and item  $j$  can be ordered such that  $X_i < X_j$ , a violation of this ordering is observed for a particular restscore group  $r$  when this ordering is reversed.

**Symbolically:**

**H<sub>0</sub>:**  $(X_i | R = r) = (X_j | R = r)$

**H<sub>a</sub>:**  $(X_i | R = r) > (X_j | R = r)$

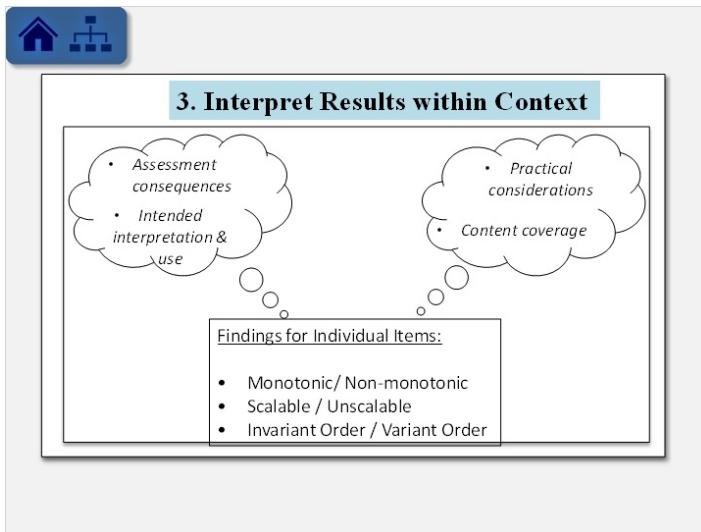
### 5.34 Bookend: Invariant Item Ordering

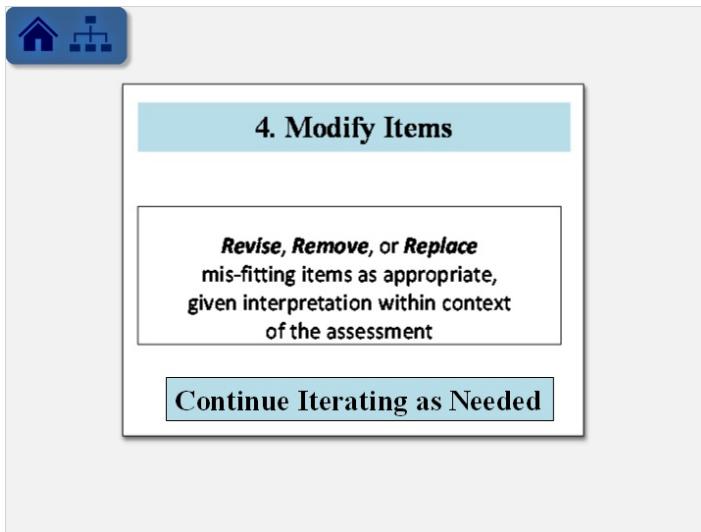


This is the end of this section.

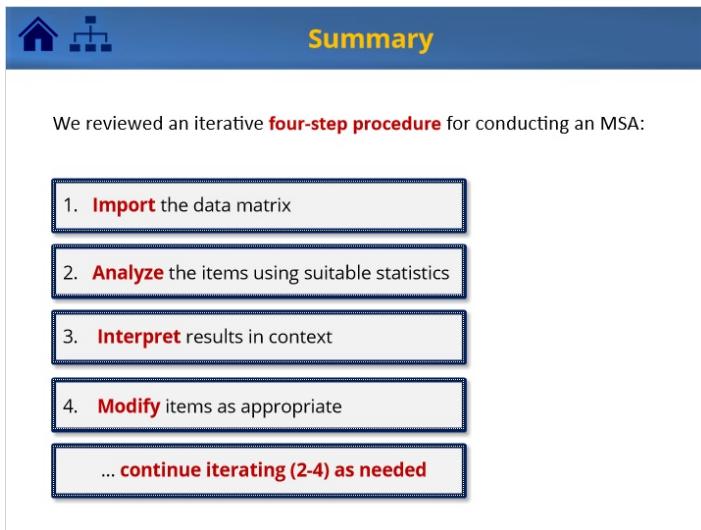
### 5.35 Step 3



### 5.36 Step 4



### 5.37 Summary: Section 4



The slide has a blue header bar with a house icon and the word "Summary" in yellow. The main content area is white with a blue border around the list. The text in the list is red.

We reviewed an iterative **four-step procedure** for conducting an MSA:

1. **Import** the data matrix
2. **Analyze** the items using suitable statistics
3. **Interpret** results in context
4. **Modify** items as appropriate

... continue iterating (2-4) as needed

### 5.38 Bookend: Section 4



The slide has a blue header bar with a house icon. The main content area is white with a blue border around the text and image. A woman in a black top and pants is making a "stop" hand gesture. The text "This is the end of this section." is centered below her.

This is the end of this section.

## 5.39 Module Cover (END)

