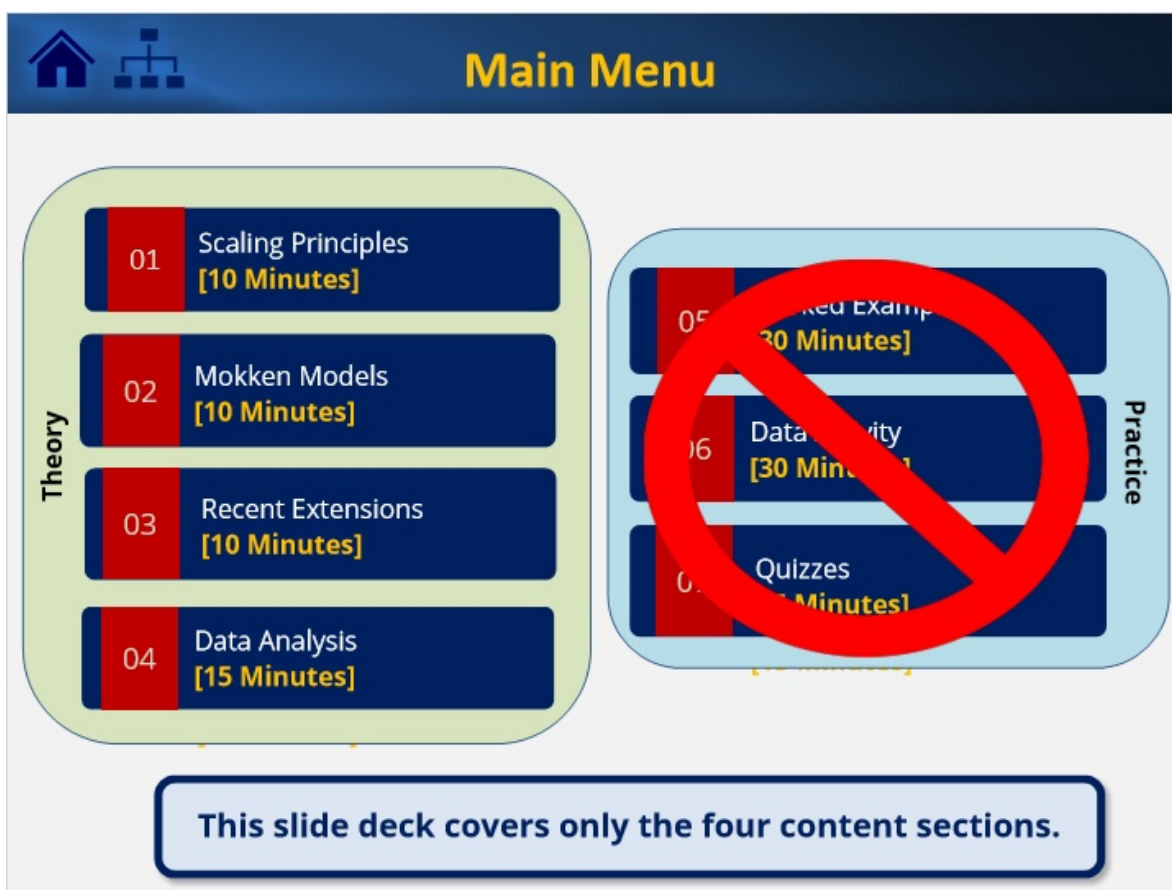


ITEMS Digital Module 03: Nonparametric Item Response Theory

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:



André A. Rupp
ETS



Xi Lu
Florida State
University

Special thanks:



ETS
NCME
National Council on
Measurement in Education

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)

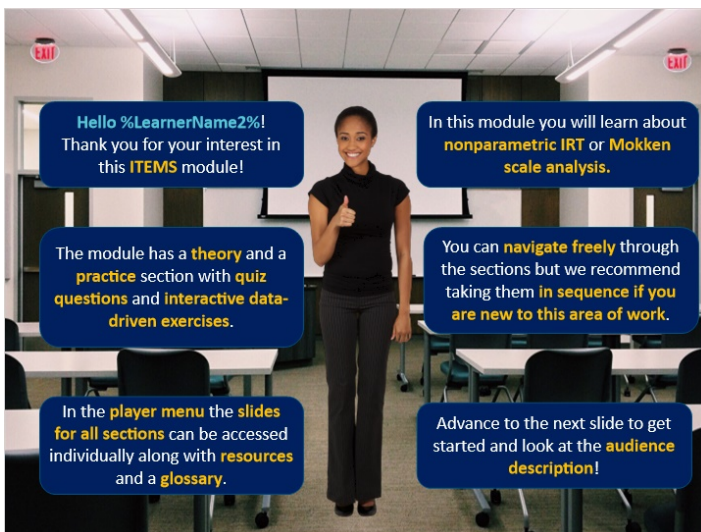


**Welcome to the
ITEMS Module!**

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!

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

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

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

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

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

1.6 Target Audience

Target Audience

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

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



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


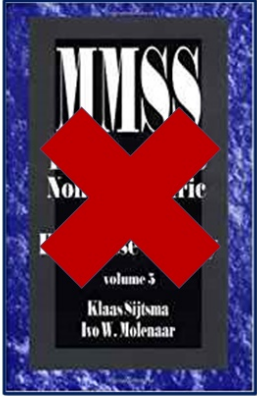
1.7 Expecations (I)



Let's discuss expectations....


1.8 Expectations (II)

ITEMS Modules in Context



1.9 Learning Objectives

Learning Objectives



I Understand the relationship between parametric and nonparametric IRT models	IV Perform real-data analyses with computational routines in R
II Understand the key differences between the two core nonparametric models	V Estimate scaling coefficients and conduct hypothesis tests with them
III Understand the key ideas of extensions of nonparametric models	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

Theory

- 01 Scaling Principles [10 Minutes]
- 02 Mokken Models [10 Minutes]
- 03 Recent Extensions [10 Minutes]
- 04 Data Analysis [15 Minutes]

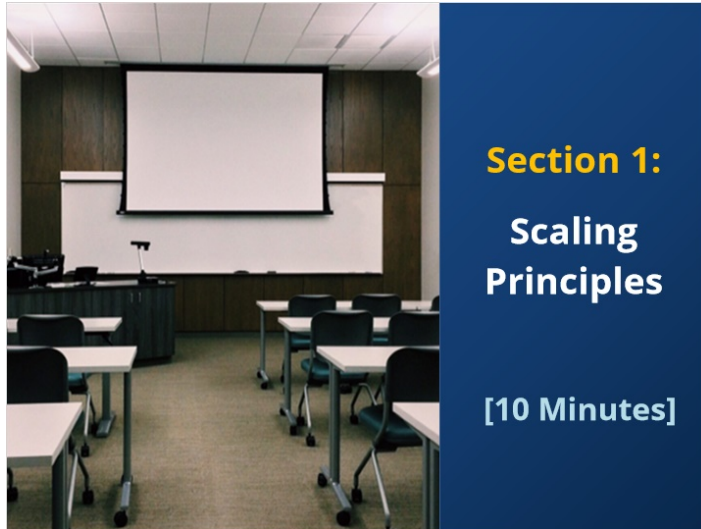
Practice

- 05 Mock Exam [30 Minutes]
- 06 Data Entry [30 Minutes]
- 07 Quizzes [15 Minutes]



This slide deck covers only the four content sections.

2. Section 1: Scaling Principles


2.1 Cover: Section 1



2.2 Objectives: Section 1

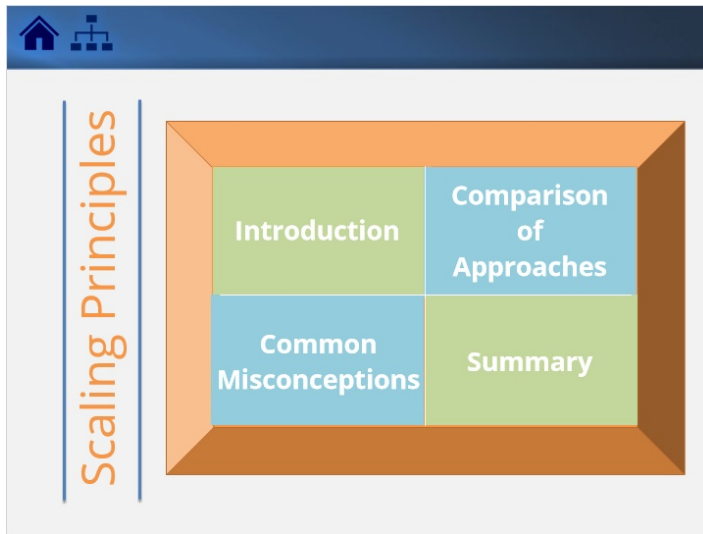


Learning Objectives



I. Distinguish between parametric IRT and nonparametric IRT	III. Describe the procedures for evaluating the degree to which responses meet model assumptions
II. Describe the assumptions for Mokken's models and the implications	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




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

2.6 Learner Objective



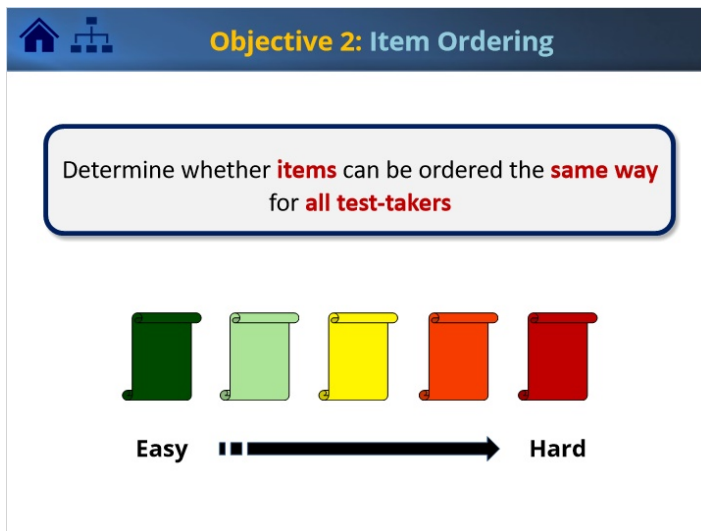
Objective 1: Learner Ordering

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

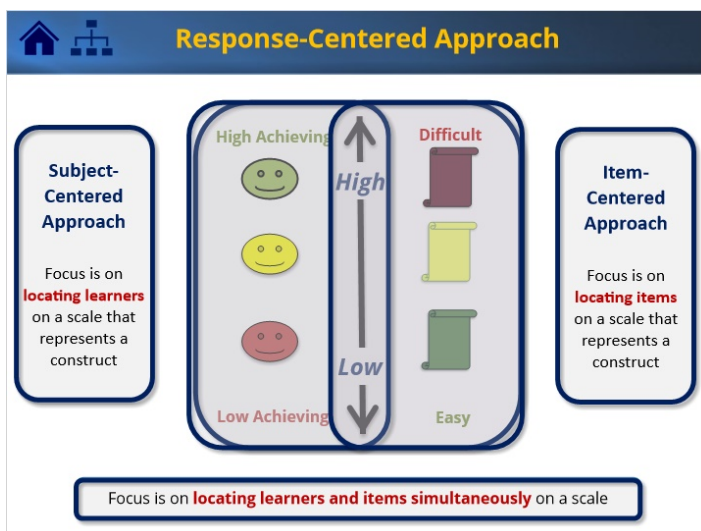


Low Achievement → **High Achievement**

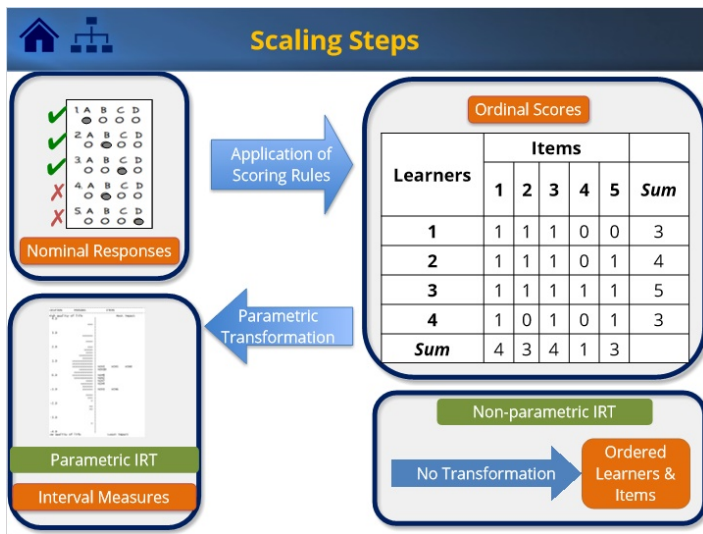
2.7 Item Objective



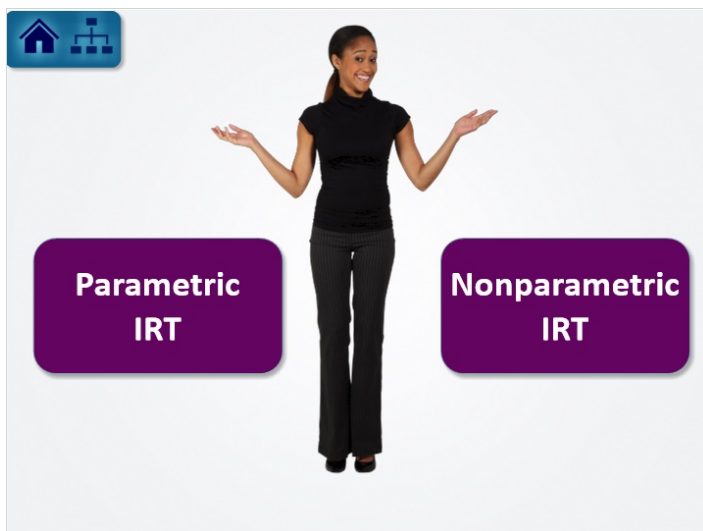
2.8 Scaling Approaches



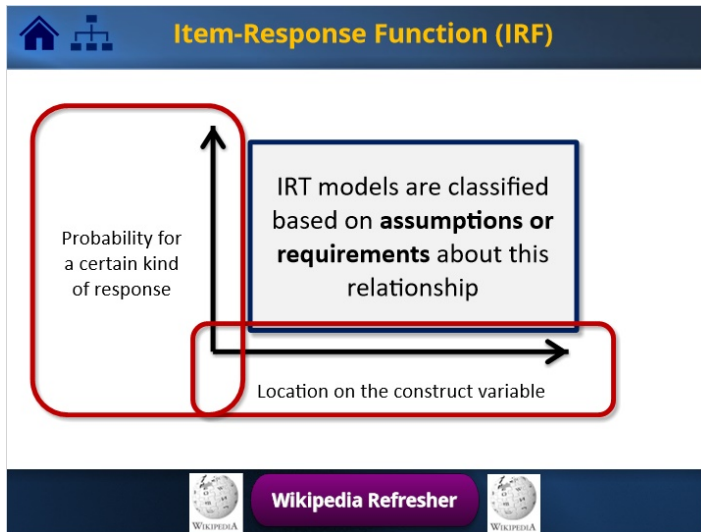
2.9 Scaling Steps



2.10 Framework Selection



2.11 Item Response Function (IRF)



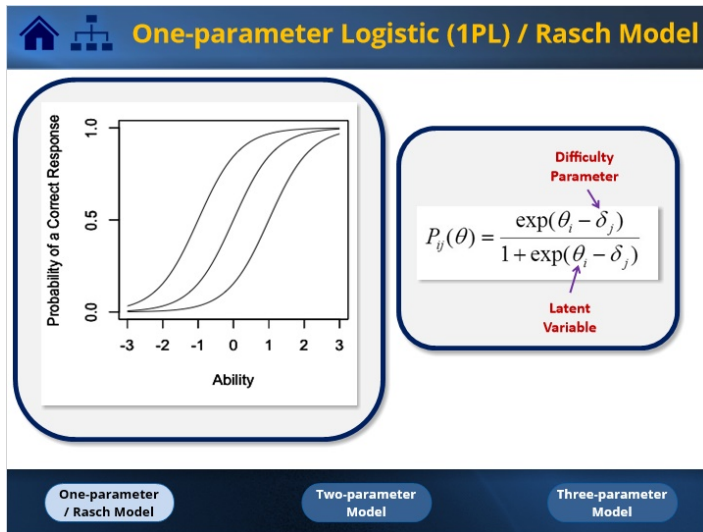
2.12 Parametric IRT

The diagram illustrates the Parametric IRT Model Types. It features a blue header with a home icon and the title "Model Types". The content is organized into three main sections:

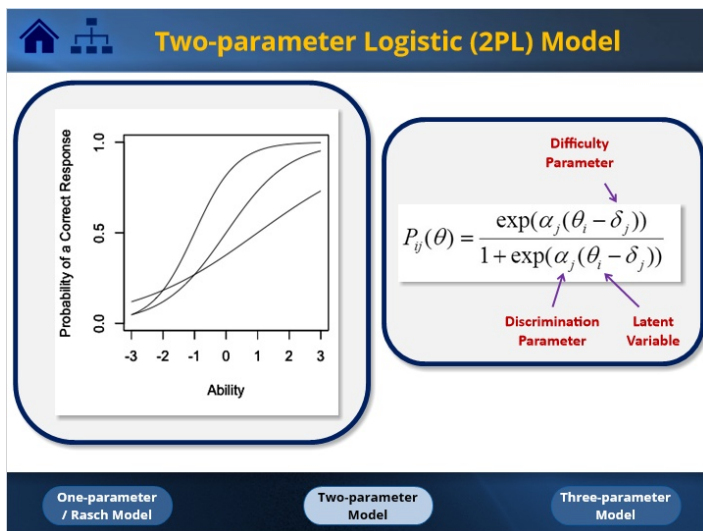
- The **functional relationship** between the probability for a response and the latent variable – the **item response function (IRF)** – must conform to a pre-specified shape
- Two common choices for the IRF are the **logistic** and the **probit** function
- Depending on the **model** that is chosen different **item parameters** are available to influence the **shape of the function** within and across items:
 - ✓ **One-parameter / Rasch model:** Difficulty
 - ✓ **Two-parameter model:** Difficulty, Discrimination
 - ✓ **Three-parameter model:** Difficulty, Discrimination, Guessing

The bottom of the slide features three buttons: "One-parameter / Rasch Model", "Two-parameter Model", and "Three-parameter Model".

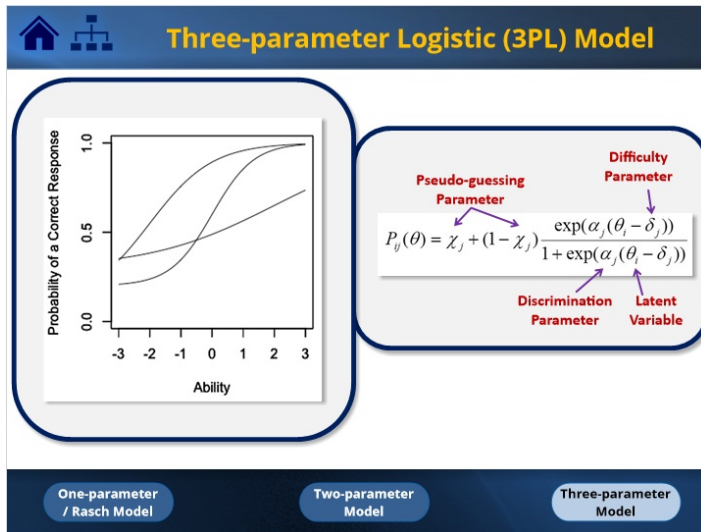
2.13 One-parameter Model



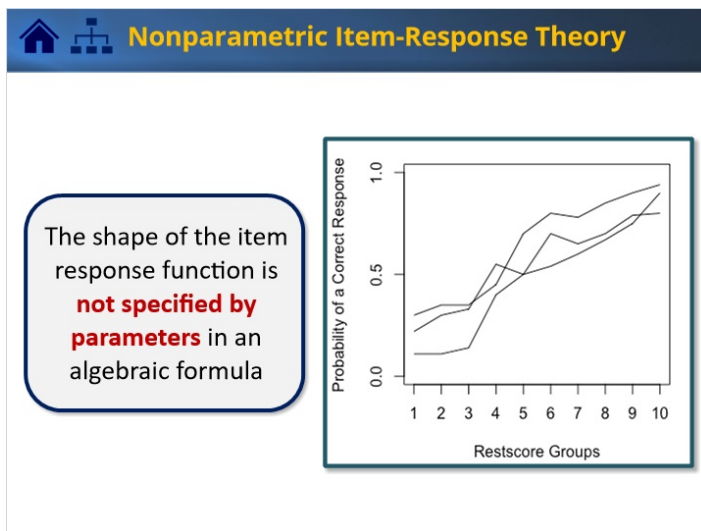
2.14 Two-parameter Model




2.15 Three-parameter Model




2.16 Nonparametric IRT




2.17 Framework Comparisons

 Comparison of IRT Approaches	
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

 Comparison 1	
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



Common Misconceptions

Misconception 1

Misconception 2

Misconception 3



Misconception 4

Misconception 5



Back to Topics


2.28 Misconception 1




Misconception 1

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

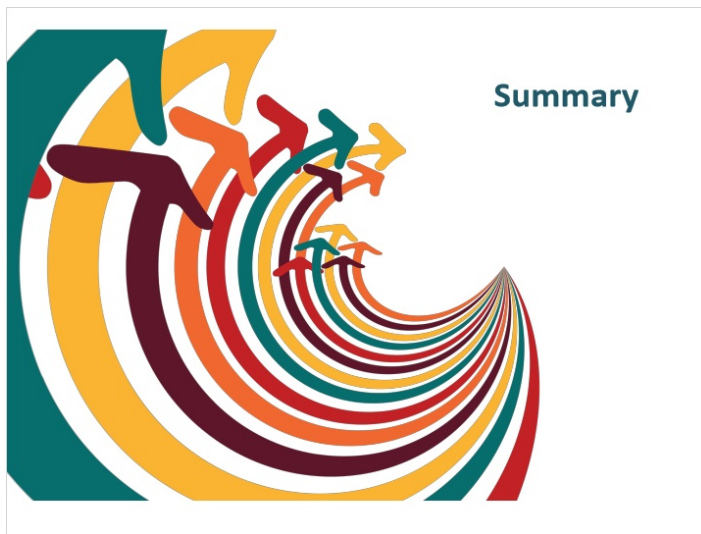
2.31 Misconception 4

 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">• Methodological & applied Mokken scaling research appears frequently in top educational measurement, psychology, & statistics journals• R package is available that is frequently updated

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





3. Section 2: Mokken Models


3.1 Cover: Section 2



3.2 Objectives: Section 2

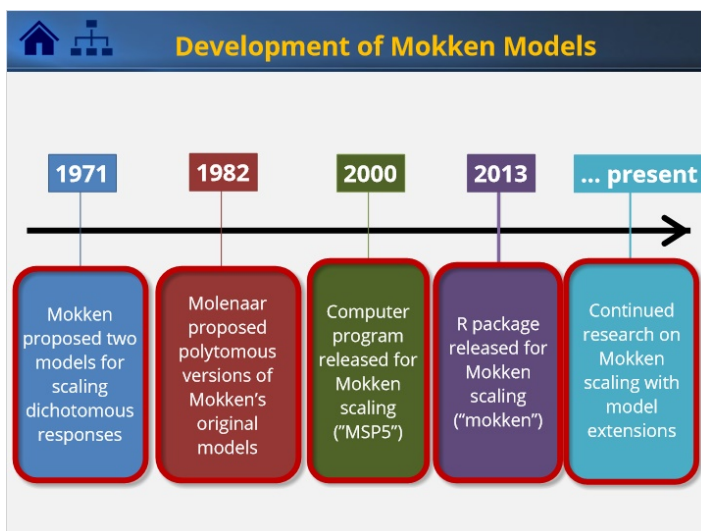


Learning Objectives

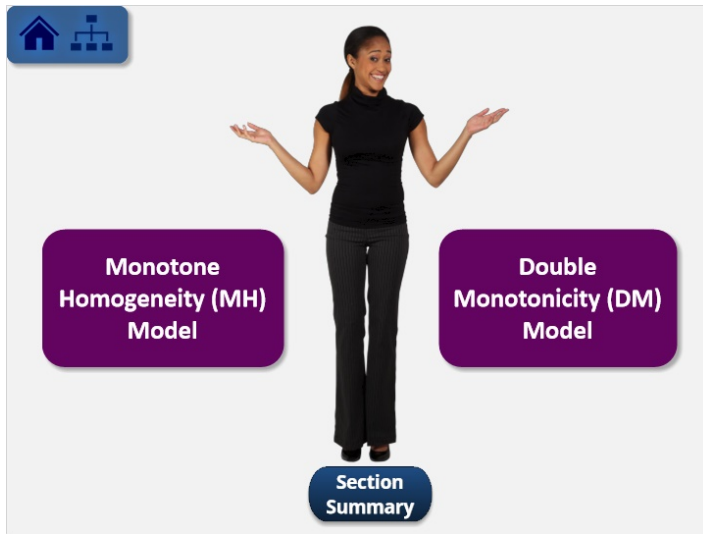


I. Understand the history of Mokken scale analysis and its current developments	III. Understand each of the four key assumptions and their relationships to assumptions in parametric IRT models
II. Understand the key differences between the Monotone Homogeneity and the Double Monotonicity model	IV. Understand 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**

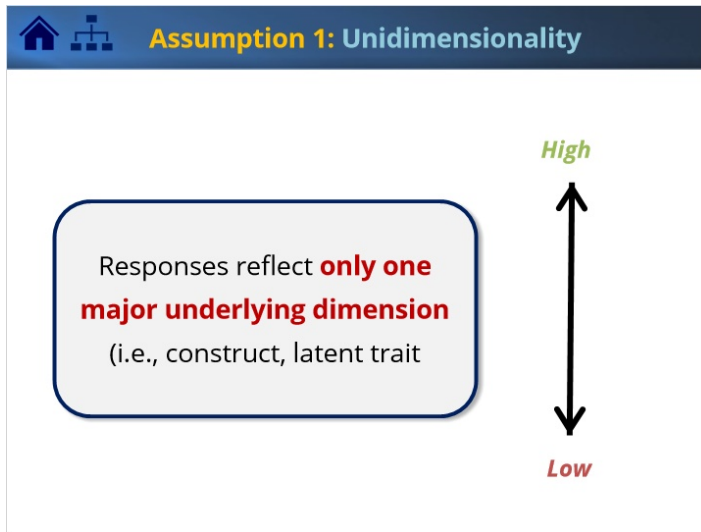
3.7 MH Model Assumptions

  **MH Model: Assumptions**

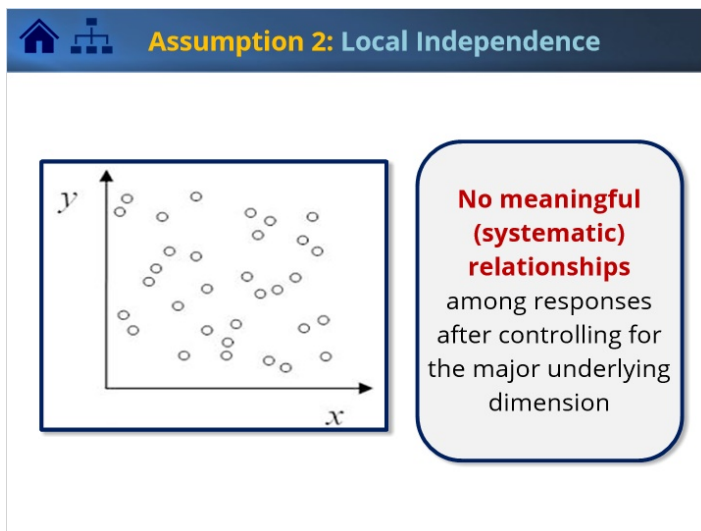


1. Unidimensionality
2. Local Independence
3. Monotonicity

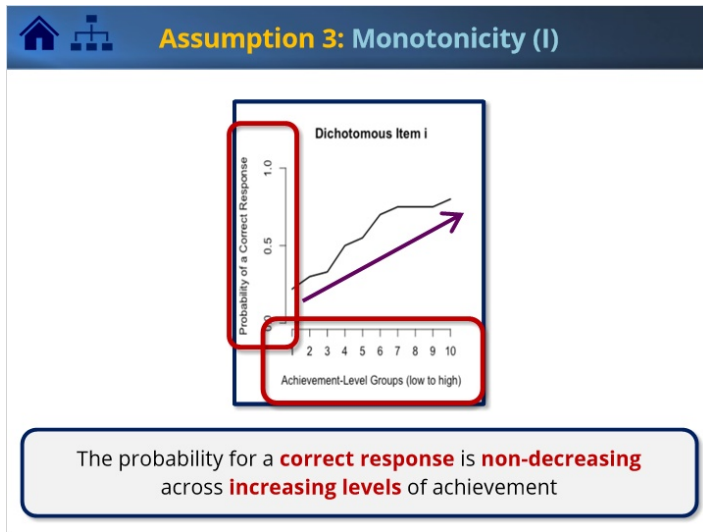
3.8 Unidimensionality



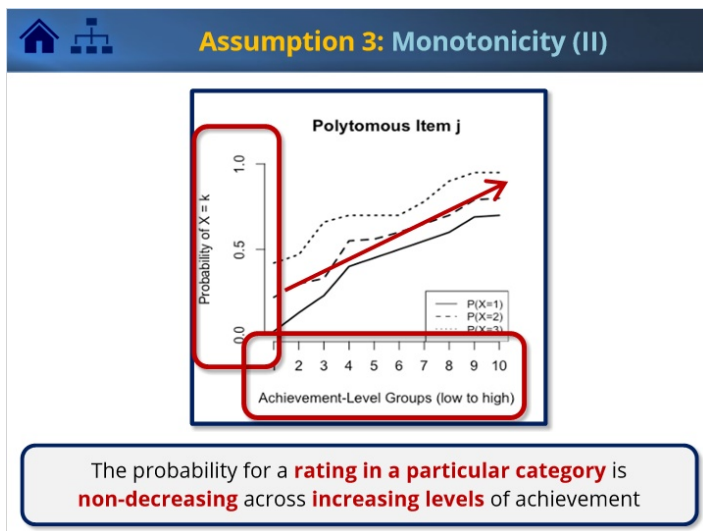
3.9 Local Independence



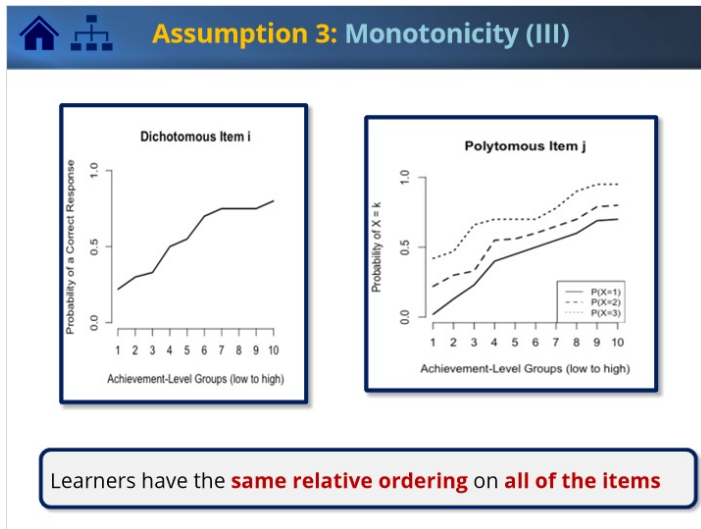
3.10 Monotonicity: Dichotomous Items



3.11 Monotonicity: Polytomous Items



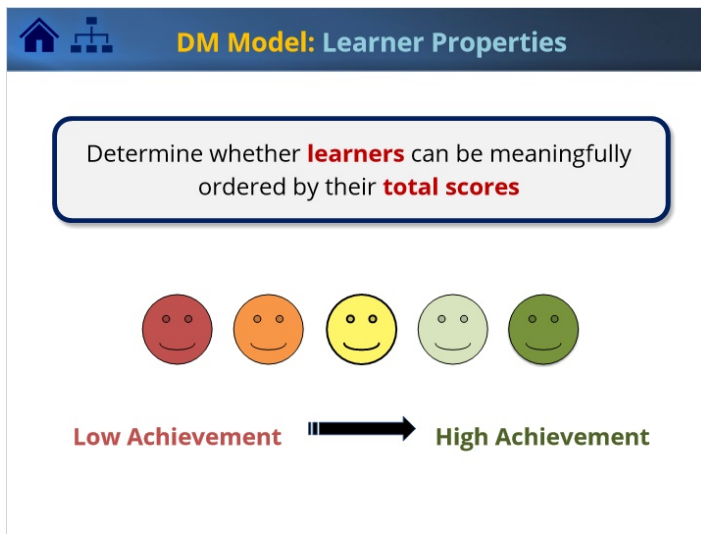
3.12 Monotonicity: Item Ordering



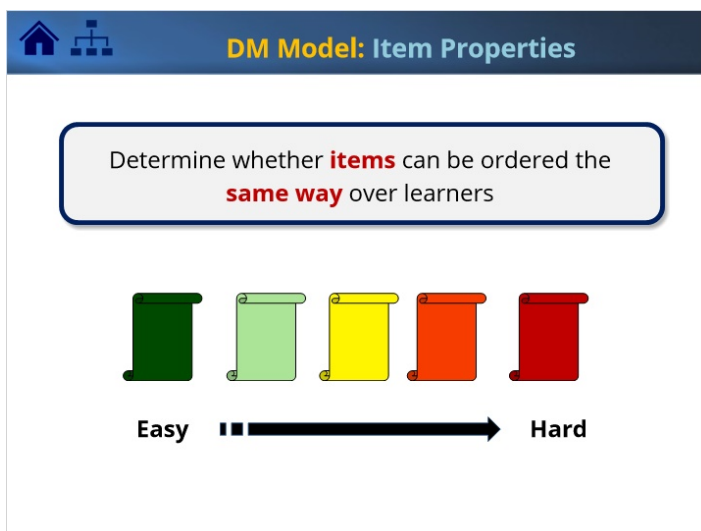
3.13 Bookmark: DM Model



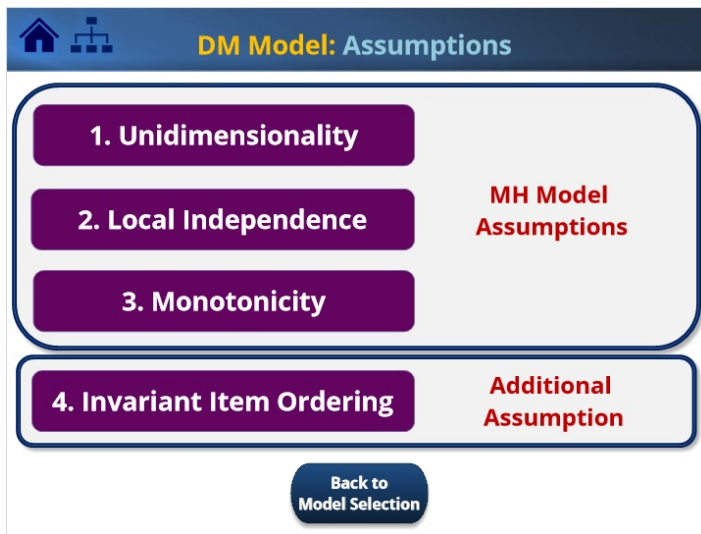
3.14 Learner Objective: DM Model



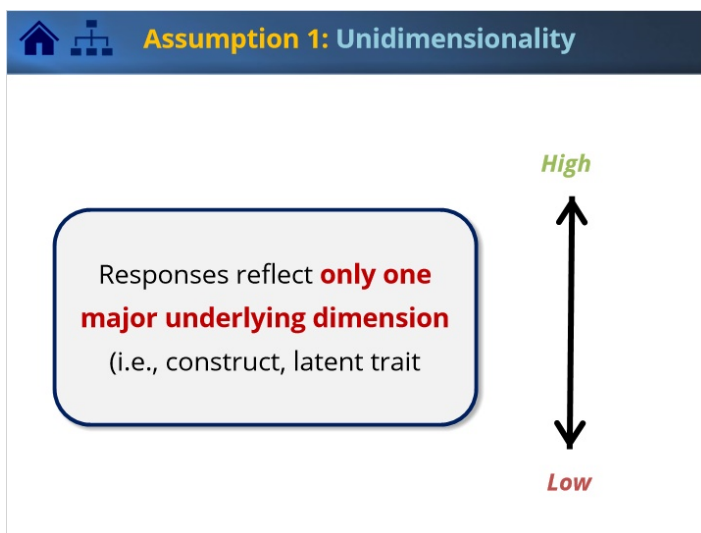
3.15 Item Objective: DM Model



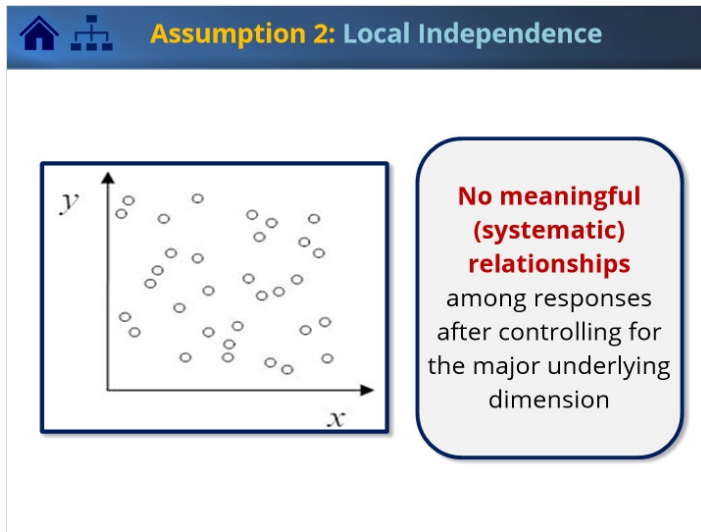
3.16 DM Model Assumptions



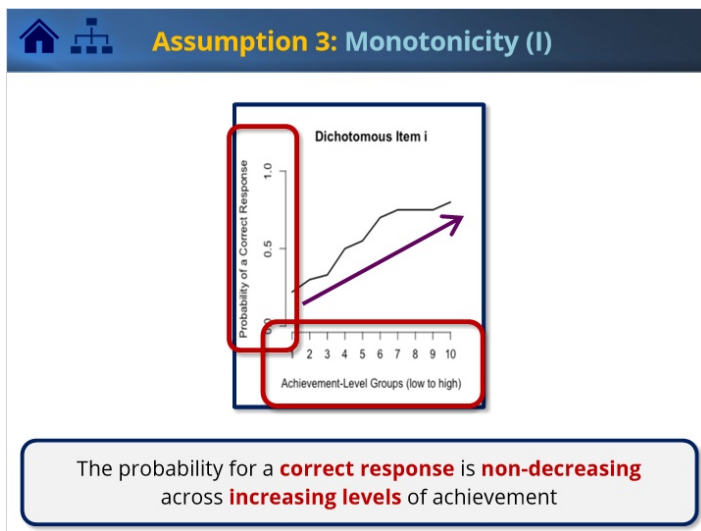
3.17 Unidimensionality



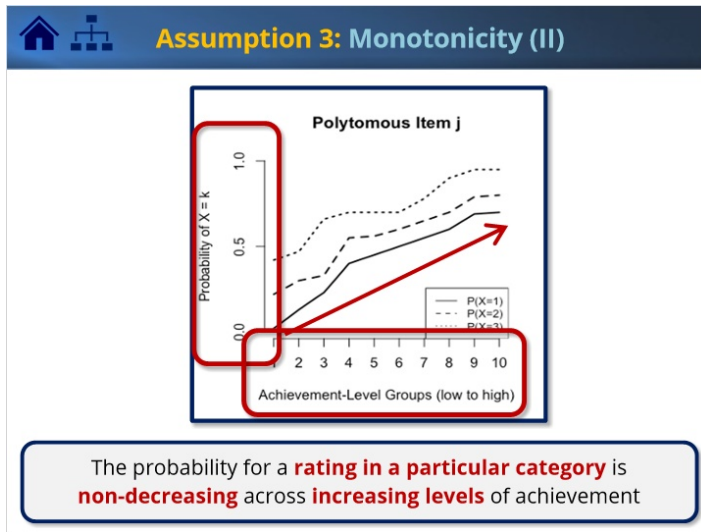
3.18 Local Independence



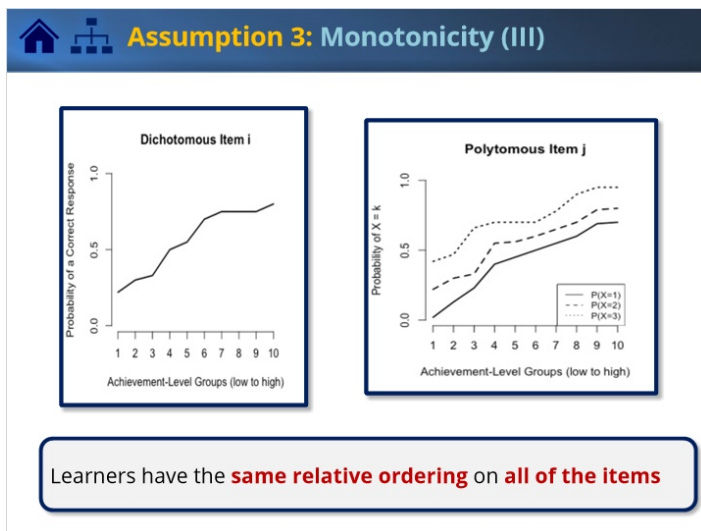
3.19 Monotonicity: Dichotomous Items



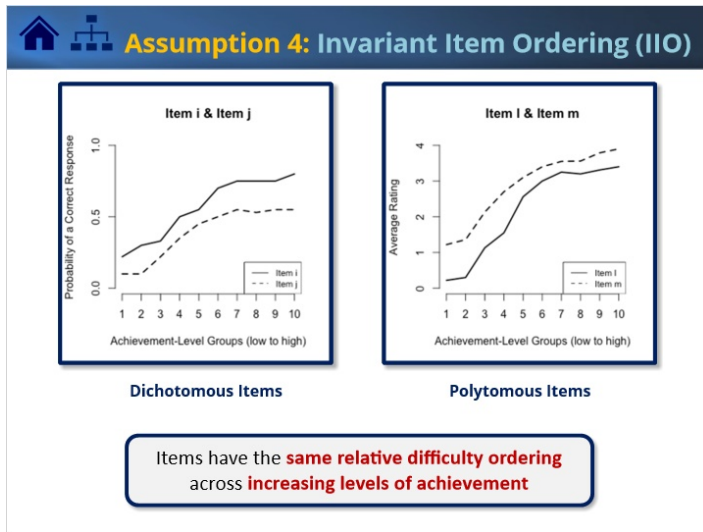
3.20 Monotonicity: Polytomous Items



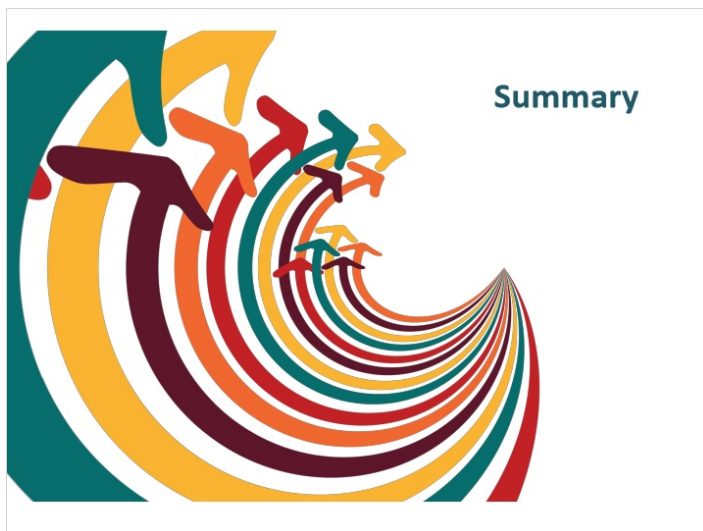
3.21 Monotonicity: Item Ordering



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.

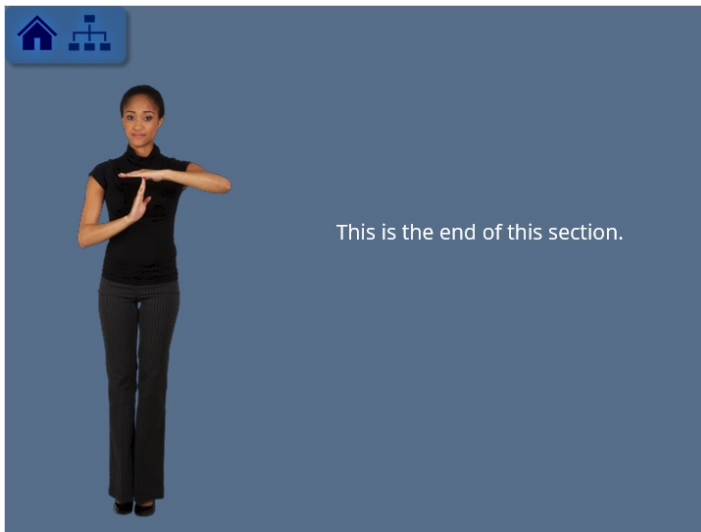
Assumptions	Double Monotonicity Model	Monotone Homogeneity Model	Model-Based Indicators
Monotonicity	✓	✓	(A) Item/Rater monotonicity
Conditional Independence	✓	✓	(B) Item/Rater scalability coefficients
Unidimensionality	✓	✓	(A) Item/Rater monotonicity; (B) Item/Rater scalability
Nonintersecting Response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

3.25 Outlook

Outlook

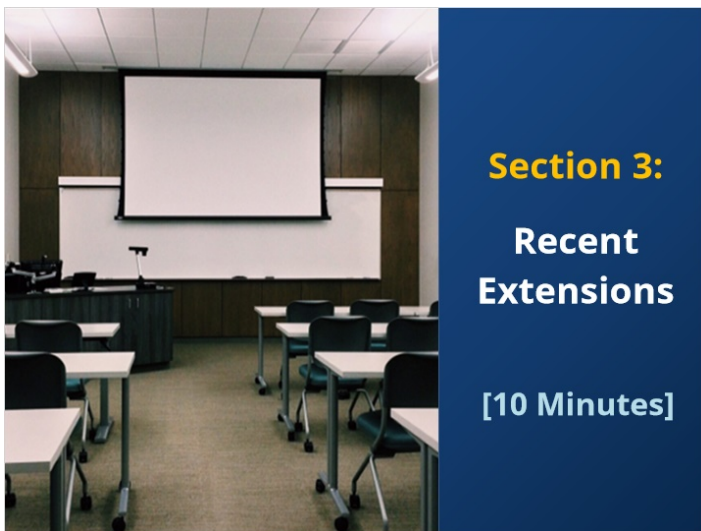
- Modeling extensions** have been proposed in recent years, which mimic modeling developments in certain areas of parametric IRT.
- Model-data fit** can be evaluated via scaling coefficients at the item, item pair, and scale level for which hypothesis tests are available.
- Specialized software packages**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

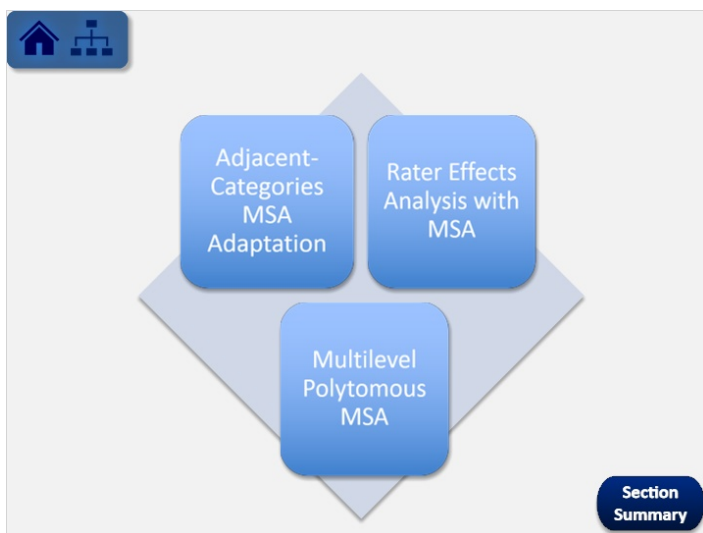
 

Learning Objectives



I. Understand the idea that Mokken scale analysis is an active area of research	III. Understand the relationship between the nonparametric models and their parametric analogues
II. Understand the key extensions of polytomous models, rater effect models, and multilevel models	IV. Understand how these models can be used in assessment development

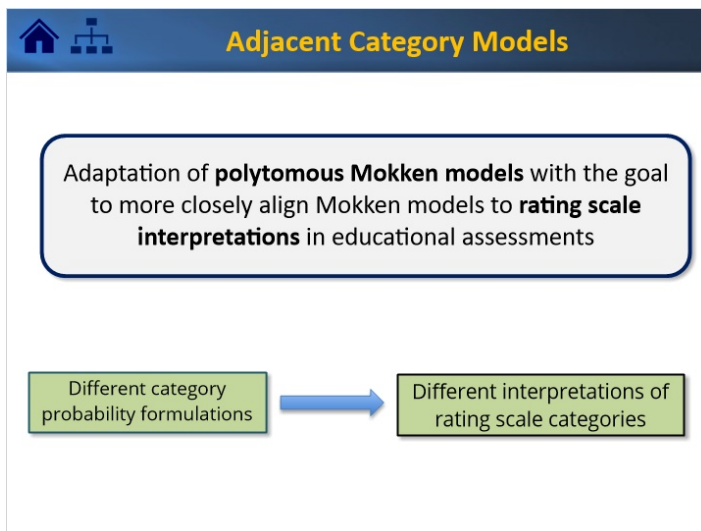
4.3 Extension Selection



4.4 Bookmark: Adj Cat Models



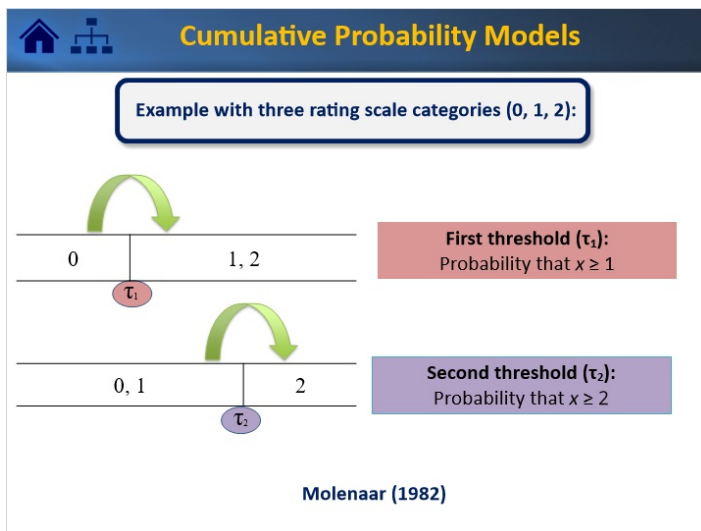
4.5 ACMs: Introduction (I)



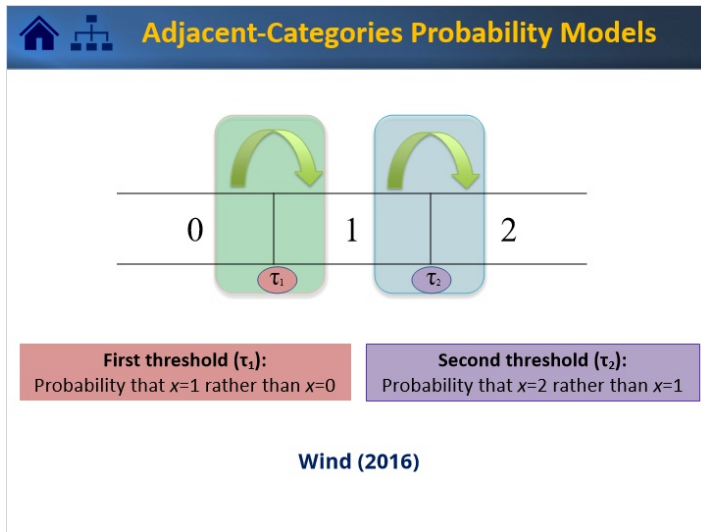
4.6 ACMs: Introduction (II)

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)

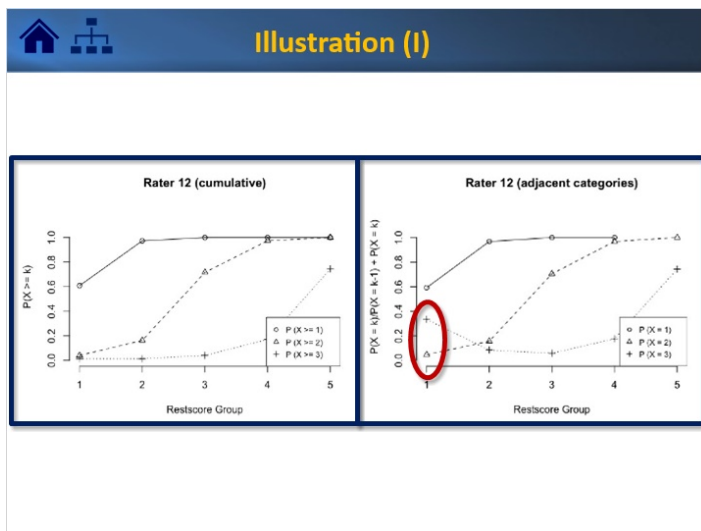
4.7 ACMs: Introduction (III)



4.8 ACMs: Introduction (V)



4.9 ACMs: Illustration (I)






Illustration (II)

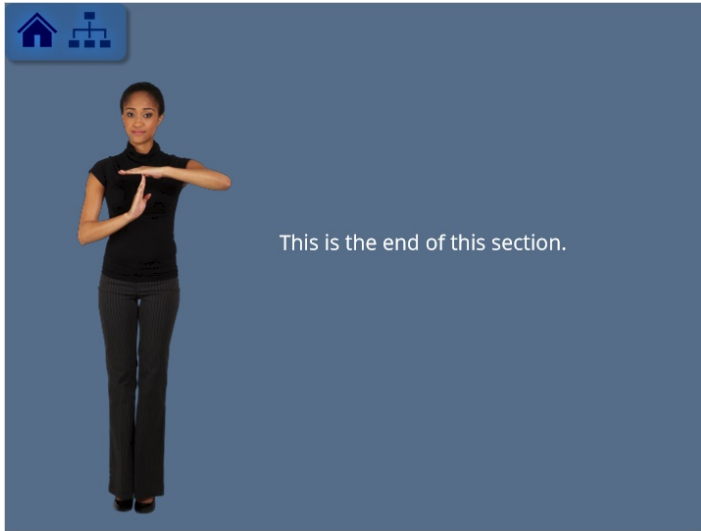
Table 2. Scalability Results.

Rater	Conventions H_1		Organization H_1		Sentence Formation H_1		Style H_1	
	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	0.30*	0.77	0.77
2	0.78	0.78	0.74	0.71	0.78	0.64	0.76	0.77
3	0.82	0.82	0.79	0.75	0.81	0.32*	0.78	0.79
4	0.81	0.79	0.81	0.72	0.78	0.30*	0.77	0.72
5	0.83	0.81	0.79	0.73	0.81	0.69	0.76	0.75
6	0.76	0.75	0.80	0.47	0.78	0.37	0.74	0.75
7	0.83	0.83	0.79	0.73	0.82	0.66	0.78	0.77
8	0.83	0.79	0.83	0.75	0.86	0.66	0.82	0.80
9	0.82	0.81	0.74	0.71	0.82	0.66	0.78	0.79
10	0.82	0.79	0.76	0.70	0.83	0.68	0.78	0.79
11	0.81	0.78	0.79	0.73	0.81	0.33*	0.78	0.80
12	0.83	0.83	0.78	0.72	0.80	0.63	0.78	0.79
13	0.84	0.75	0.78	0.72	0.83	0.42*	0.76	0.76
14	0.79	0.75	0.83	0.74	0.81	0.37*	0.77	0.78
15	0.78	0.77	0.76	0.71	0.81	0.65	0.78	0.76
16	0.83	0.81	0.78	0.73	0.82	0.68	0.80	0.80
17	0.77	0.78	0.75	0.71	0.81	0.70	0.75	0.73
18	0.75	0.78	0.77	0.71	0.77	0.70	0.76	0.73
19	0.81	0.81	0.78	0.73	0.83	0.66	0.78	0.79
20	0.78	0.78	0.75	0.38*	0.80	0.13*	0.74	0.74
Overall H	0.81	0.79	0.78	0.67	0.81	0.67	0.77	0.77

Note. MH-R = monotone homogeneity for rating; ac-MH = adjacent-categories monotone homogeneity.
 *Indicates change in scalability classification based on the following criteria: $H_1 \geq 0.50$, strong; $0.40 \leq H_1 < 0.50$, medium; $0.30 \leq H_1 < 0.40$, weak (Mölkien, 1971).

[illegible]



4.12 Bookend: Adj Cat Models



4.13 Bookmark: Rater Models

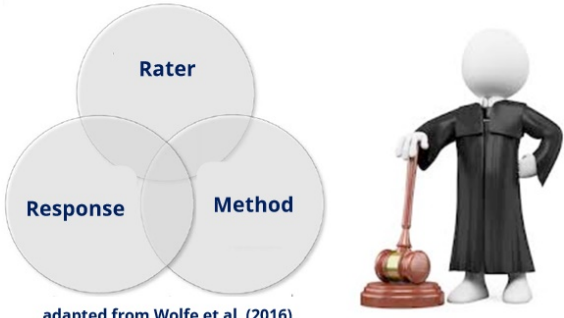


4.14 RMs: Introduction (I)



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.

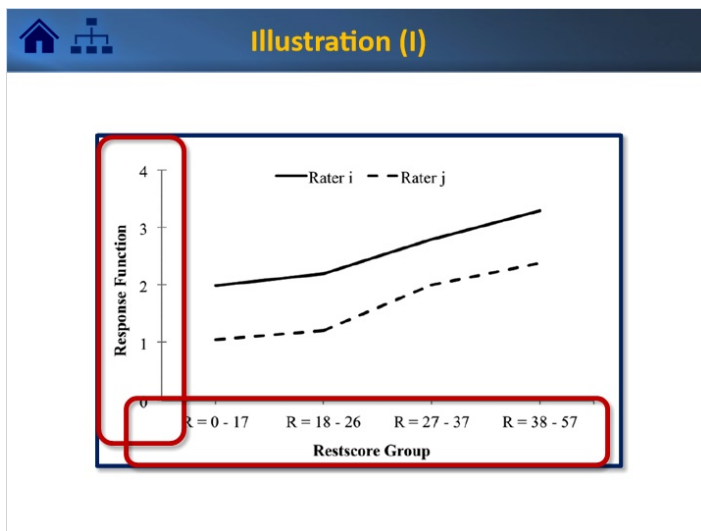
Common effects include leniency / severity, centrality, halo effects, inaccuracy, and differential dimensionality. There is a rich body of literature on human rating processes, in particular for constructed responses such as essays.




4.16 RMs: Introduction (III)

Home	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

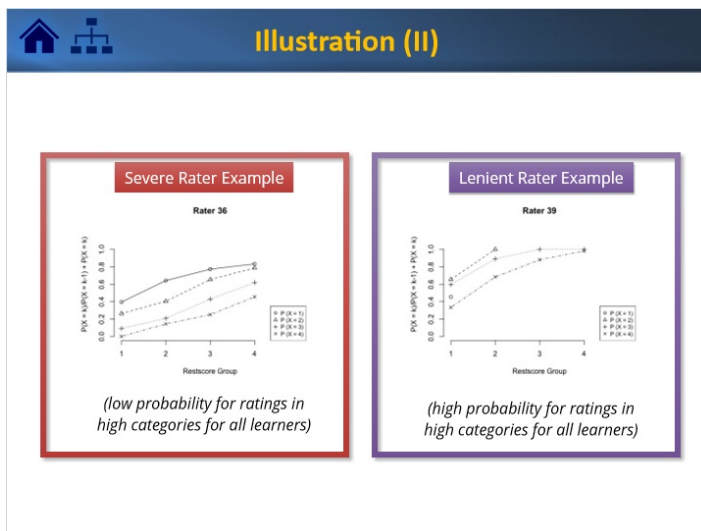


The model specifies the probability that a person will be classified **above any threshold....**

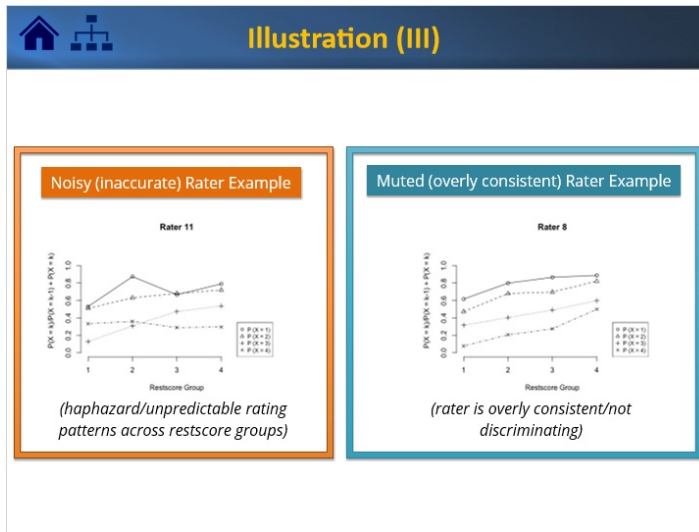
This does **not seem consistent** with performance assessment—judges locate a performance **in one of the categories, not in and beyond** any particular category

(Andrich, 2015, p. 6)

4.19 RMs: Illustration (II)



4.20 RMs: Illustration (III)





4.21 RMs: Illustration (IV)

Illustration (IV)


Table 4:
Average ac-MSA results within Rasch classifications

Rater Group	A-C Scalability		Significant Violations of Rater Monotonicity		Significant Violations of IRO	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Severe (<i>n</i> =8)	0.30	0.06	0.00	0.00	8.44	5.20
Lenient (<i>n</i> =11)	0.22	0.05	0.09	0.30	5.18	3.49
Noisy (<i>n</i> =4)	0.17	0.03	0.25	0.50	19.50	7.05
Muted (<i>n</i> =2)	0.20	0.02	0.00	0.00	1.50	0.71
Fair (<i>n</i> =20)	0.29	0.04	0.11	0.32	9.38	3.87


4.22 RMs: References




References: Rater Effect Models



Wind & Engelhard (2016)








Wind & Engelhard (2018)

+ more work in development

4.23 Bookend: Rater Models







This is the end of this section.

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


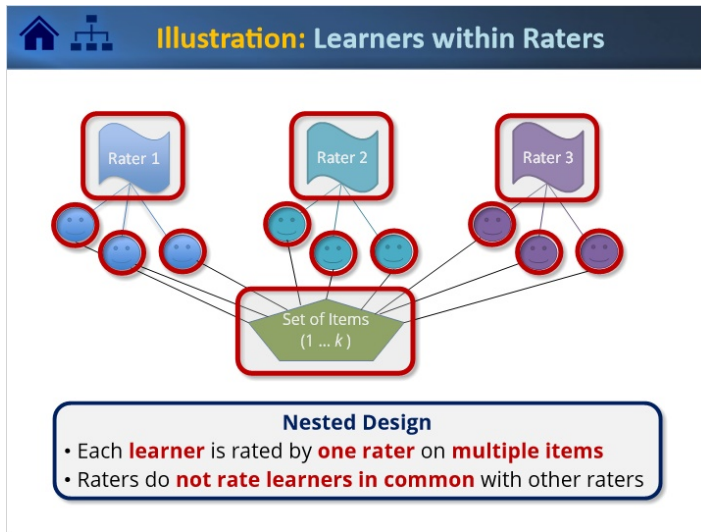
4.26 MLMs: Introduction (II)

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

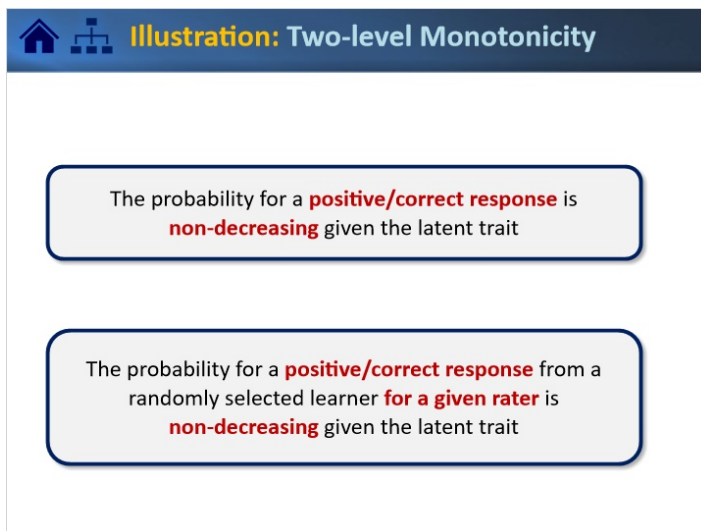
4.27 MLMs: Introduction (III)

  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



References: Multilevel Models

**Scalability Coefficients for Two-Level
Performance Item Scores: An Introduction
and an Application**

Rooske M. Crisan, Jeroen K. van de Pol, and L. J. J. van der Ark

Abstract: This article introduces the concept of scalability coefficients for two-level performance item scores. The article discusses the theoretical background of these coefficients and their application in the context of a large-scale assessment. The article also presents a practical example of how to calculate these coefficients using the software program *MLM*.

Keywords: Scalability coefficients, two-level performance item scores, large-scale assessment

1. Introduction

One of the main goals of a large-scale assessment is to provide a reliable and valid measure of the performance of a group of individuals. To achieve this goal, it is essential to use a set of items that are well-suited to the purpose of the assessment. One way to ensure that the items are well-suited is to use a set of items that have high scalability coefficients. Scalability coefficients are a measure of the degree to which the items in a set are related to each other. The higher the scalability coefficient, the more related the items are to each other. In this article, we introduce the concept of scalability coefficients for two-level performance item scores. We discuss the theoretical background of these coefficients and their application in the context of a large-scale assessment. We also present a practical example of how to calculate these coefficients using the software program *MLM*.

2. Scalability Coefficients

The concept of scalability coefficients is based on the idea of a common factor. A common factor is a factor that is shared by two or more variables. In the context of a large-scale assessment, a common factor is a factor that is shared by two or more items. The scalability coefficient is a measure of the degree to which the items in a set are related to each other. The higher the scalability coefficient, the more related the items are to each other. In this article, we introduce the concept of scalability coefficients for two-level performance item scores. We discuss the theoretical background of these coefficients and their application in the context of a large-scale assessment. We also present a practical example of how to calculate these coefficients using the software program *MLM*.

3. Application


In this section, we present a practical example of how to calculate scalability coefficients for two-level performance item scores using the software program *MLM*. We use a set of 10 items that are related to the concept of 'mathematics'. We calculate the scalability coefficients for each item and find that the highest scalability coefficient is 0.85. This indicates that the items are well-suited to the purpose of the assessment.

4. Conclusion

In conclusion, scalability coefficients are a useful tool for ensuring that the items in a set are well-suited to the purpose of a large-scale assessment. By using items with high scalability coefficients, we can ensure that the assessment is reliable and valid. In this article, we have introduced the concept of scalability coefficients for two-level performance item scores. We have discussed the theoretical background of these coefficients and their application in the context of a large-scale assessment. We have also presented a practical example of how to calculate these coefficients using the software program *MLM*.

References

Crisan, van de Pol, & van der Ark (2016)



**Weighted Common Errors: Handling Two-
and Three-Level Data**

Liesje Koopman, Jeroen K. van de Pol, and L. J. J. van der Ark

Abstract: This article introduces the concept of weighted common errors for two- and three-level data. The article discusses the theoretical background of these errors and their application in the context of a large-scale assessment. The article also presents a practical example of how to calculate these errors using the software program *MLM*.

Keywords: Weighted common errors, two-level data, three-level data, large-scale assessment

1. Introduction

One of the main goals of a large-scale assessment is to provide a reliable and valid measure of the performance of a group of individuals. To achieve this goal, it is essential to use a set of items that are well-suited to the purpose of the assessment. One way to ensure that the items are well-suited is to use a set of items that have high weighted common errors. Weighted common errors are a measure of the degree to which the items in a set are related to each other. The higher the weighted common errors, the more related the items are to each other. In this article, we introduce the concept of weighted common errors for two- and three-level data. We discuss the theoretical background of these errors and their application in the context of a large-scale assessment. We also present a practical example of how to calculate these errors using the software program *MLM*.

2. Weighted Common Errors

The concept of weighted common errors is based on the idea of a common factor. A common factor is a factor that is shared by two or more variables. In the context of a large-scale assessment, a common factor is a factor that is shared by two or more items. The weighted common errors are a measure of the degree to which the items in a set are related to each other. The higher the weighted common errors, the more related the items are to each other. In this article, we introduce the concept of weighted common errors for two- and three-level data. We discuss the theoretical background of these errors and their application in the context of a large-scale assessment. We also present a practical example of how to calculate these errors using the software program *MLM*.

3. Application

In this section, we present a practical example of how to calculate weighted common errors for two- and three-level data using the software program *MLM*. We use a set of 10 items that are related to the concept of 'mathematics'. We calculate the weighted common errors for each item and find that the highest weighted common errors are 0.85. This indicates that the items are well-suited to the purpose of the assessment.

4. Conclusion


In conclusion, weighted common errors are a useful tool for ensuring that the items in a set are well-suited to the purpose of a large-scale assessment. By using items with high weighted common errors, we can ensure that the assessment is reliable and valid. In this article, we have introduced the concept of weighted common errors for two- and three-level data. We have discussed the theoretical background of these errors and their application in the context of a large-scale assessment. We have also presented a practical example of how to calculate these errors using the software program *MLM*.


References

Koopman, Zijlstra, & van der Ark (2016)

+ more work in development

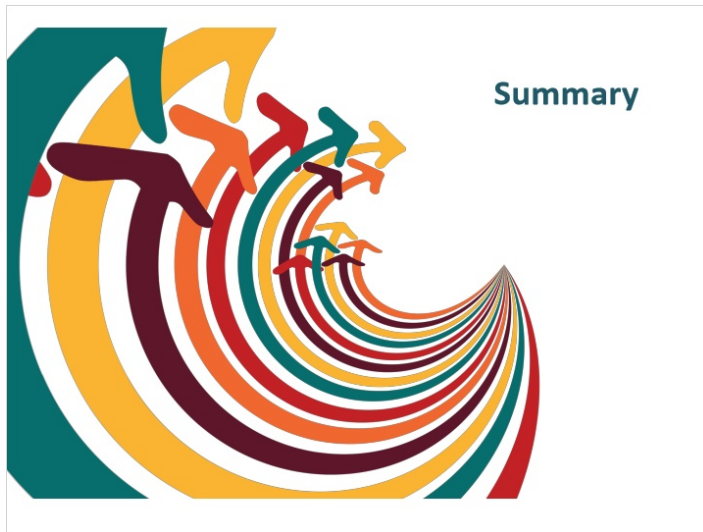
4.33 Bookend: Multilevel Models







This is the end of this section.

4.34 Bookmark: Summary



4.35 Summary: Section 3

 **Summary**

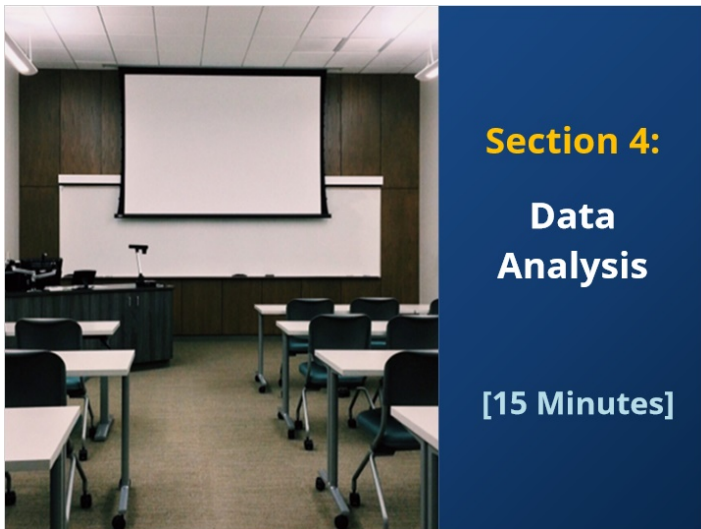
- **Nonparametric IRT**, including Mokken scale analysis, is still an **active area of research**
- **Three key extensions** of the basic Mokken models include models for **polytomous data**, **rater effects**, and **multilevel data**
- All of these models have **analogues in parametric IRT** where **individual 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)

1 ... L

Ordinal Scored Responses:

X_{ni}

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

Students (n)

1 ... N

2. Analyze Items

A. Monotonicity

Response functions for rating scale categories

(Dichotomous & Polytomous)

Hypothesis Tests & Confidence Intervals

B. Scalability

Individual Items:

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

Pairs of Items:

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

All Items:

$$H = 1 - \left(\sum_{i=1}^L F_i / \sum_{i=1}^L E_i \right)$$

Hypothesis Tests & Confidence Intervals

C. Invariant Ordering

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

1. Import Data Matrix

Items (i)

1 ... L

Ordinal Scored Responses:

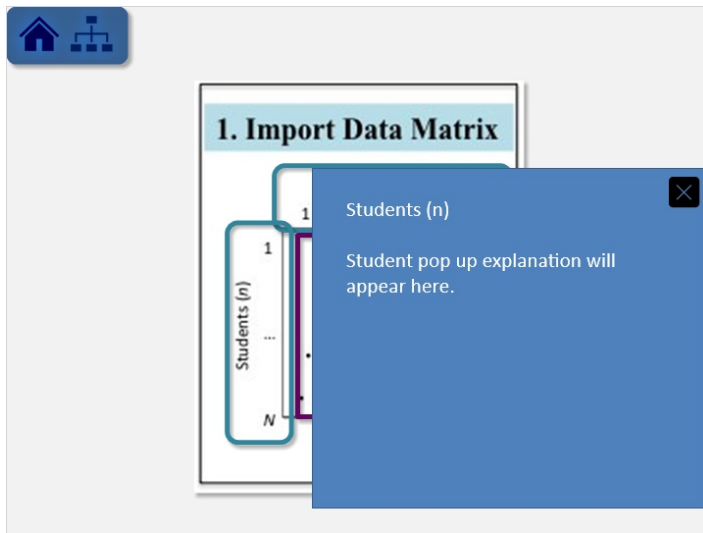
X_{ni}

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

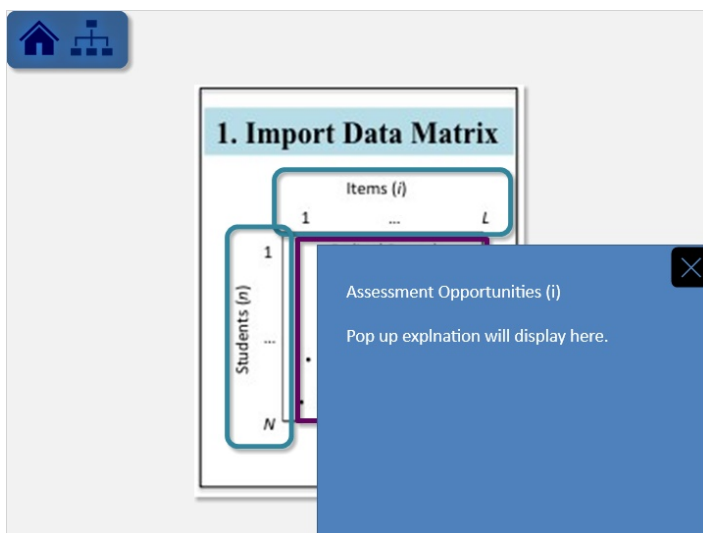
Students (n)

1 ... N


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)

• Hypothesis Tests & Confidence Intervals

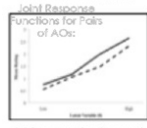
B. Scalability

- Individual items:
$$H_i = 1 - \left(\sum_{j \in J} F_{ij} / \sum_{j \in J} E_{ij} \right)$$
- Pairs of items:
$$H_{ij} = F_{ij} / E_{ij}$$
- All items:
$$H = 1 - \left(\sum_{i=1}^{I-1} \sum_{j=i+1}^I F_{ij} / \sum_{i=1}^{I-1} \sum_{j=i+1}^I E_{ij} \right)$$

• Hypothesis Tests & Confidence Intervals

C. Invariant Ordering

Joint Response Functions for Pairs of Items:





(Dichotomous & Polytomous)

• Hypothesis Tests & Confidence Intervals

5.7 Bookmark: Monotonicity



5.8 Monotonicity (I)





Monotonicity

- **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)

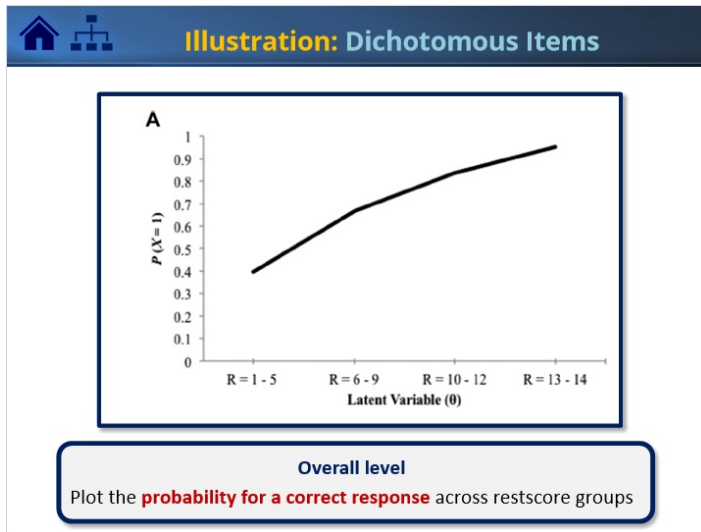


Restscore Computation

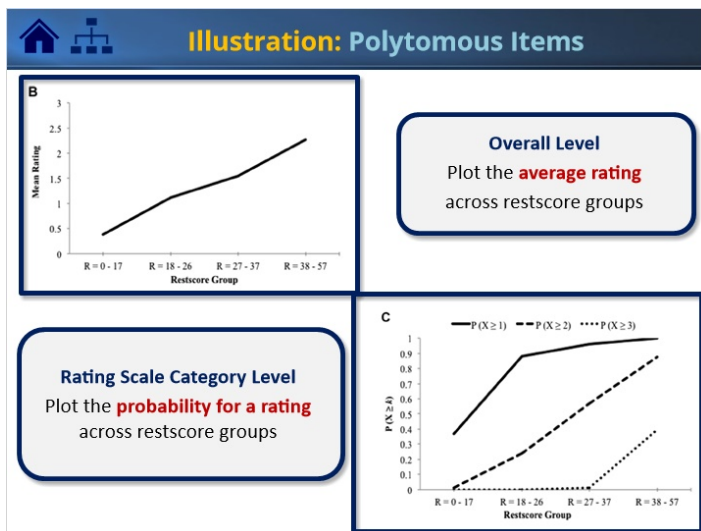
- Requires a **nonparametric estimate** of student achievement:

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_0 : the probability for a correct response is **equal** across the two restscore groups

H_A : 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**:

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

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

[Back to Slide](#)



5.14 Bookend: Monotonicity



5.15 Bookmark: Scalability




5.16 Scalability (I)

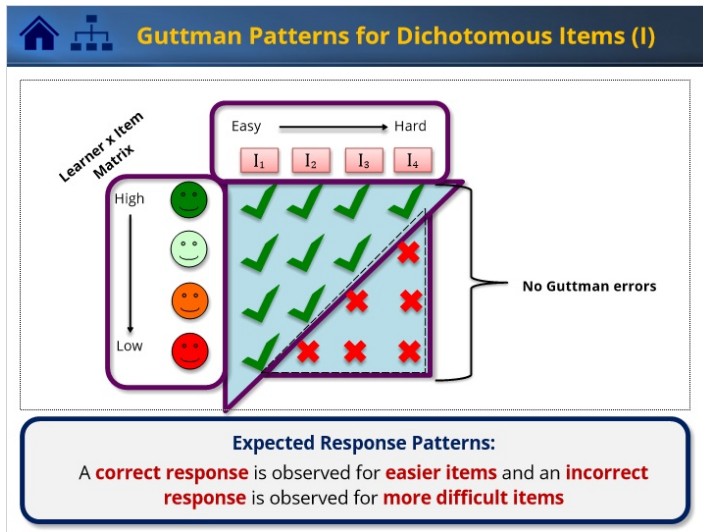


Scalability Coefficients

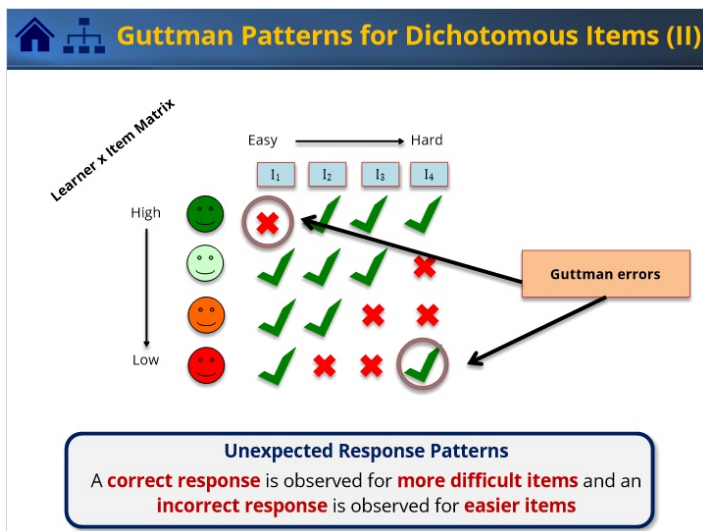
- Describe the degree to which **individual items, pairs of items, and overall sets of items** form a scale that can be used to order learners on a construct
- Summarize the influence of **Guttman errors** on a measurement procedure, where **fewer Guttman errors** mean **stronger evidence** for a meaningful 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
<div style="display: flex; flex-direction: column; align-items: center;"> <div>High</div> <div style="margin: 5px 0;">↓</div> <div>Low</div> </div>	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
	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
<div style="display: flex; flex-direction: column; align-items: center;"> <div>High</div> <div style="margin: 5px 0;">↓</div> <div>Low</div> </div>	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
	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	<u>(0, 0)</u>	(0, 1)*	(0, 2)*	(0, 3)*
1	<u>(1, 0)</u>	(1, 1)*	(1, 2)*	(1, 3)*
2	<u>(2, 0)</u>	<u>(2, 1)</u>	<u>(2, 2)</u>	(2, 3)*
3	(3, 0)*	(3, 1)*	<u>(3, 2)</u>	<u>(3, 3)</u>

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 j = 1	Item j = 2	Item i = 3	Item j = 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 j = 1	Item j = 2	Item i = 3	Item j = 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

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

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**

Table 2. Observed Joint Frequencies of Item i and Item j



	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 ≥ 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







This is the end of this section.

5.28 Bookmark: Invariant Item Ordering



5.29 Invariant Ordering (I)



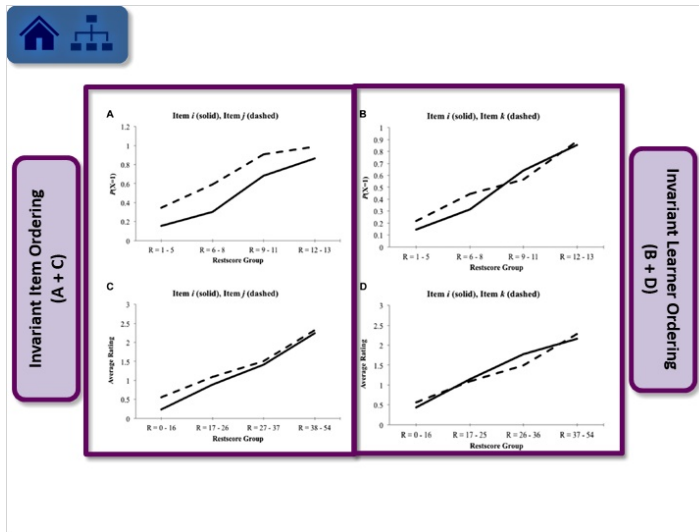
Invariant Ordering

- 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)

Hypothesis Testing for Invariant Ordering

- Statistical tests for violations of invariant item ordering are *t*-tests
 - Available for both **dichotomous** and **polytomous** items and based on **comparisons** of the **difficulty of the item** across **two adjacent restscore groups** (one lower, one higher)
- Violations of invariant item ordering => significant *t*-statistics
 - Statistically significant *t*-statistics indicate that the **item ordering for a pair of items**, based on the overall sample, **does not hold** across **two adjacent restscore groups**

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:

H₀: $P(X_i = 1 | R = r) = P(X_j = 1 | R = r)$

H_a: $P(X_i = 1 | R = r) > P(X_j = 1 | R = r)$

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:

H₀: $(X_i | R = r) = (X_j | R = r)$

H_a: $(X_i | R = r) > (X_j | R = r)$

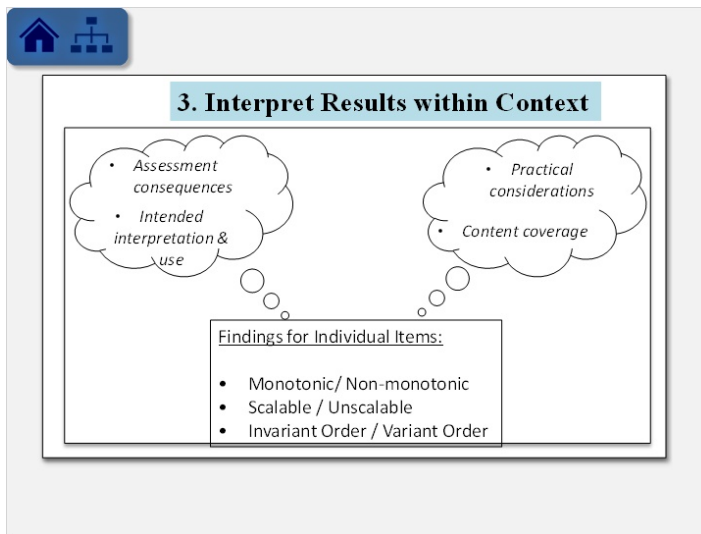
5.34 Bookend: Invariant Item Ordering



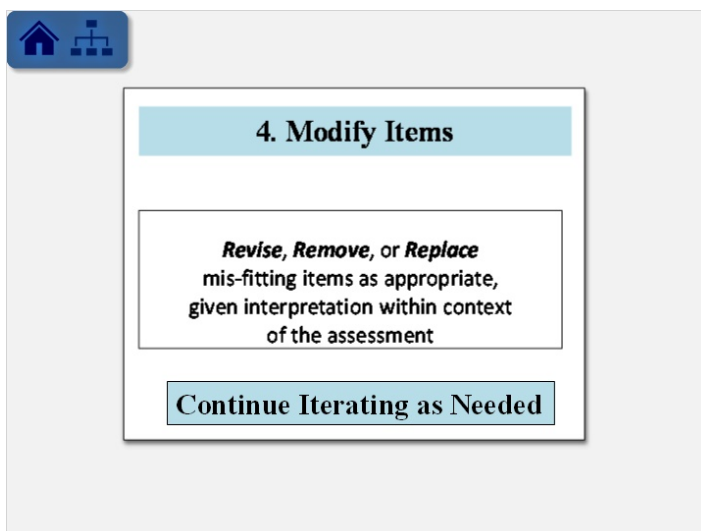


This is the end of this section.



5.35 Step 3



5.36 Step 4



5.37 Summary: Section 4





Summary


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





This is the end of this section.

5.39 Module Cover (END)

