

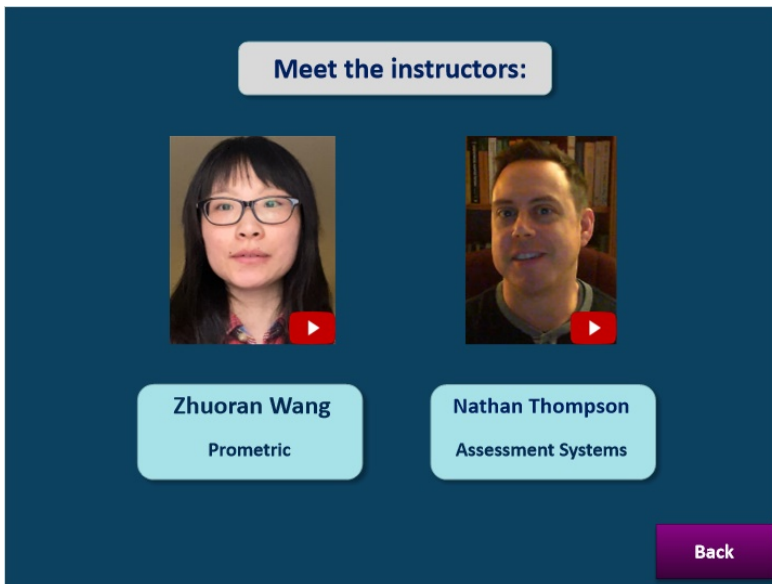
# DM19 SLIDES (IRT Estimation, Version 1.0)

## 1. Module Overview

### 1.1 Module Cover (START)





### 1.2 Instructors



### 1.3 Designers

Meet the designers:




Jon Lehrfeld  
ETS

André A. Rupp  
Mindful Measurement

Back

### 1.4 Welcome



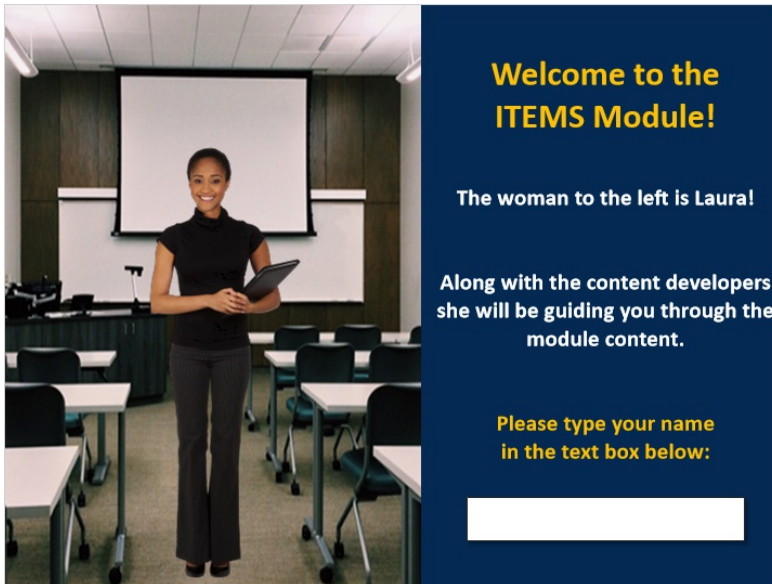
**Welcome to the  
ITEMS Module!**

The woman to the left is Laura!

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

Please type your name  
in the text box below:

## Untitled Layer 1 (Slide Layer)



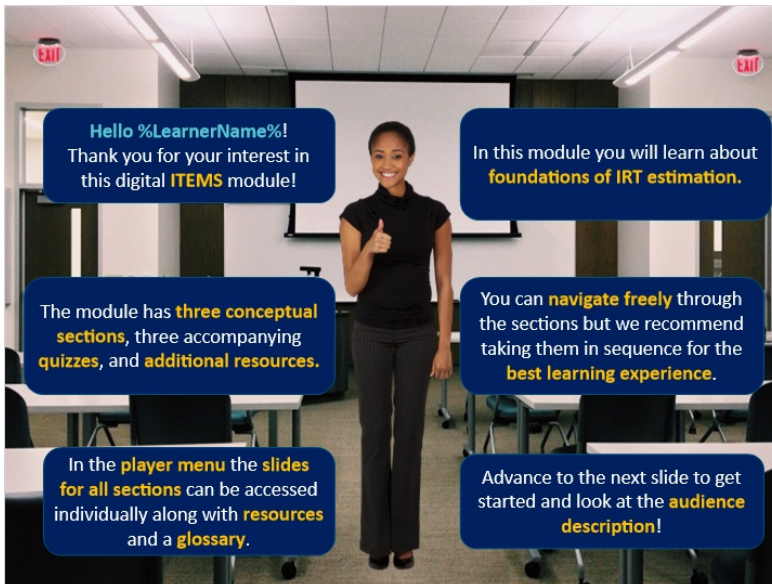
**Welcome to the ITEMS Module!**

The woman to the left is Laura!

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

Please type your name in the text box below:

## 1.5 Overview



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

In this module you will learn about **foundations of IRT estimation**.

The module has **three conceptual sections**, three accompanying **quizzes**, and **additional resources**.

You can **navigate freely** through the sections but we recommend taking them in sequence for the **best learning experience**.

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!

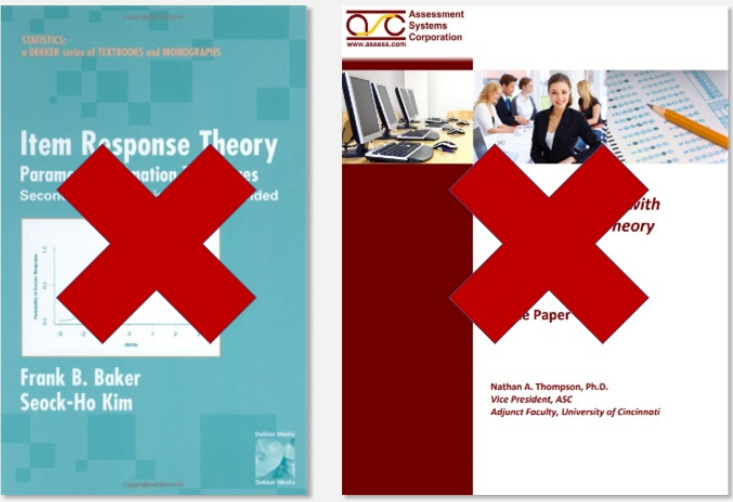
## 1.7 Expectations (I)



Let's discuss expectations....

## 1.8 Expectations (II)


**ITEMS Modules in Context**



The image shows two book covers side-by-side. The left cover is for 'Item Response Theory: Parameter Estimation and Scoring' by Frank B. Baker and Seock-Ho Kim, published by Dekker. The right cover is for 'Item Response Theory: A Practical Approach' by Nathan A. Thompson, published by Assessment Systems Corporation. Both covers have a large red 'X' over them, indicating they are not to be used.

## 1.9 Learning Objectives

**Learning Objectives**



1. Demonstrate a preliminary understanding of item calibration and examinee scoring algorithms
2. Conduct item calibration and examinee scoring using R
3. Read and interpret item calibration and examinee scoring results
4. Evaluate estimation accuracy
5. Identify the potential need to increase estimation accuracy
6. Use multiple strategies to facilitate estimation

## 1.10 Prerequisites

**Prerequisites**


- **Foundational IRT concepts:**
  - ✓ Item response functions for 1/2/3PL models
  - ✓ Goal and logic of item calibration and examinee scoring
  - ✓ Parameter recovery using response matrix
  
- **Probability**
  - ✓ Prior and posterior distributions
  
- **R statistical software**
  - ✓ Useful for Section 3: Factors Affecting Estimation Accuracy

## 1.11 Resources


**Resources**

**Module Citation**

Wang, Z., & Thompson, N. (2020). Foundations of IRT estimation (Digital ITEMS Module 19). *Educational Measurement: Issues and Practice*, 39(4), 134-135.



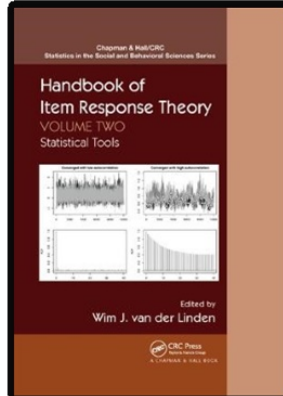
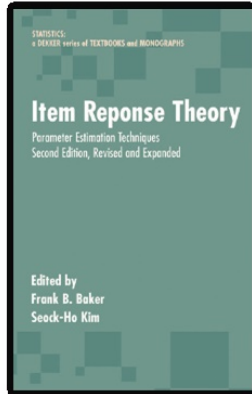
**Sample R Code**



**Additional Resources**

## References (Slide Layer)

### Resources



Click on the images to go to the publisher websites.



Back

## 1.12 Main Menu

### Main Menu

The Main Menu is divided into two sections: Theory and Practice.

**Theory**

- 01 IRT Basics [10 Minutes]
- 02 Calibration and Scoring [15 Minutes]
- 03 Factors Affecting Estimation Accuracy [20 Minutes]

**Practice**

- 04 Example using [10 Minutes] (external website)
- 05 Quizzes [10 Minutes]

A red circle with a diagonal slash is overlaid on the Practice section, indicating that the content is not available or is restricted.

## 2. Section 1: IRT Basics

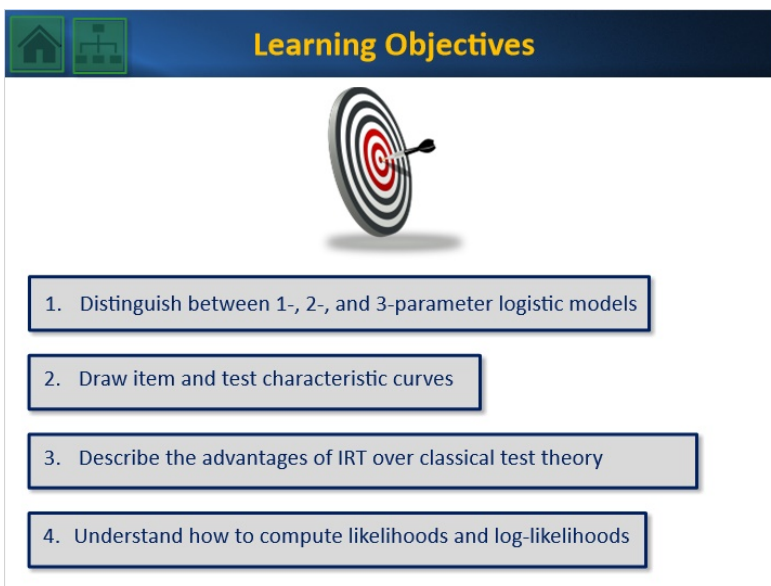
### 2.1 Cover: IRT Basics




Section 1:  
IRT Basics

[10 minutes]

### 2.2 IRT Basics: Learning Objectives





Learning Objectives



1. Distinguish between 1-, 2-, and 3-parameter logistic models
2. Draw item and test characteristic curves
3. Describe the advantages of IRT over classical test theory
4. Understand how to compute likelihoods and log-likelihoods



## 2.3 Item Response Theory



### Item Response Theory

- Theoretical approach of item response theory (IRT)
  - Relates test performance to examinee ability levels
  - Focuses on item-level performance
  - Models the performance of examinees at each ability level to each item on the test

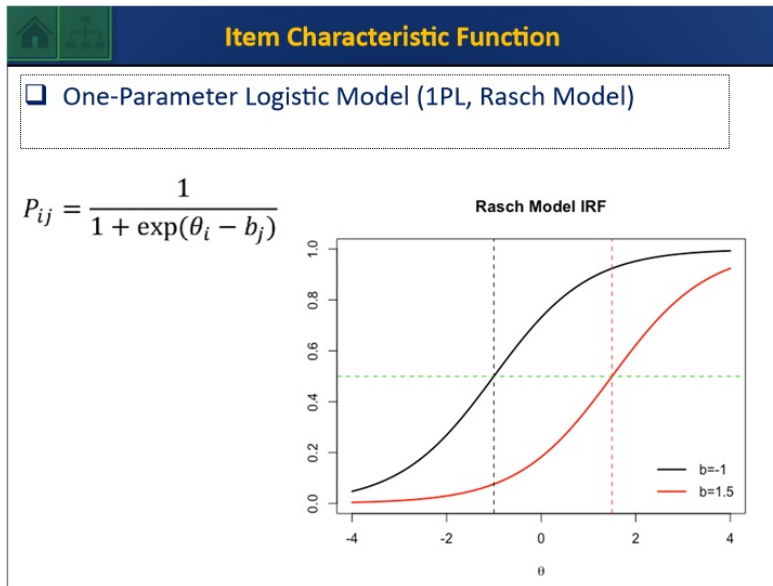
## 2.4 Advantages of IRT



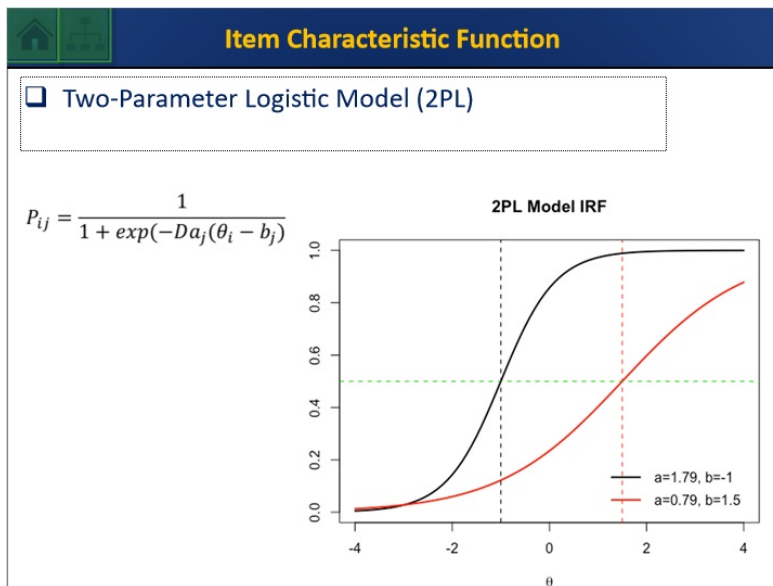
### Advantages of IRT

- IRT overcomes certain issues in classical test theory (CTT)
  - Examinees and items are on the same scale (independent of sample)
  - Once linking of item parameters takes place, equating automatically occurs
  - Individualized standard error of measurement
  - Base of advanced testing schemes (e.g., computerized adaptive testing, multistage testing)

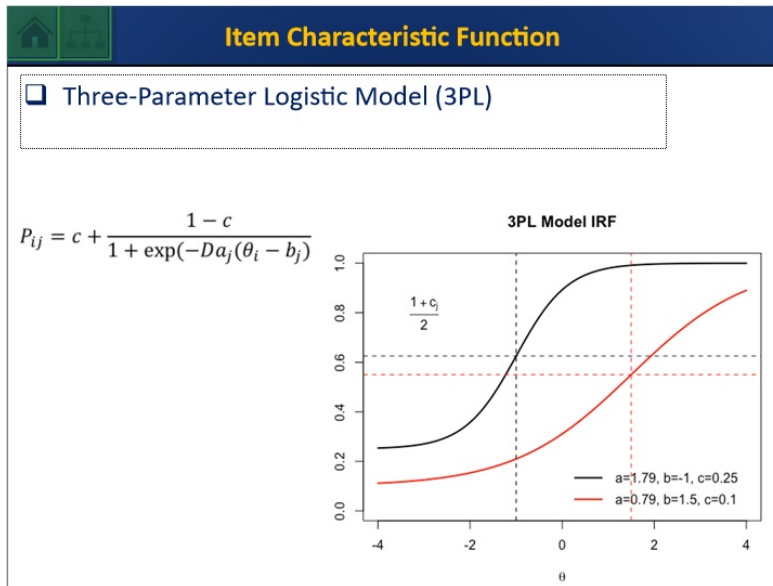
## 2.5 Item Characteristic Functions I



## 2.6 Item Characteristic Functions II



## 2.7 Item Characteristic Functions III



## 2.8 Simulation Example

**Simulation Example**

2PL Example

- 5 items
- $a \sim \text{Lognormal}(0,1)$
- $b \sim \text{Normal}(0,1)$
- $\theta \sim \text{Normal}(0,1)$
- Response for examinee #1 is (0, 1, 0, 0, 1) ["Mixed" response]

## 2.9 Item Information

🏠 📄 **Item Information**

IRT advances the concept of item and test information to replace reliability.

1PL:  $I_{ij} = P_{ij}Q_{ij}$

2PL:  $I_{ij} = a_j^2 P_{ij}Q_{ij}$

3PL:  $I_{ij} = a_j^2 \frac{(P_{ij}-c_j)^2 Q_{ij}}{(1-c_j)^2 P_{ij}}$

where  $Q_{ij} = 1 - P_{ij}$

**Item Information**

The graph shows three bell-shaped curves representing item information for different models. The x-axis is theta (ranging from -4 to 4) and the y-axis is Item Information (ranging from 0.0 to 0.8). A vertical dashed line is at theta = -1. The 1PL curve (black) has a peak of approximately 0.25 at theta = -1. The 2PL curve (red) has a peak of approximately 0.8 at theta = -1. The 3PL curve (green) has a peak of approximately 0.5 at theta = -1. The legend indicates: 1PL: b=-1; 2PL: a=1.79, b=-1; 3PL: a=1.79, b=-1, c=0.25.

## 2.10 Test Information

🏠 📄 **Test Information**

Test information is the sum of all item information

**Test Information**

The graph shows the test information curve (black) and five individual item information curves (red, green, blue, cyan, magenta). The x-axis is theta (ranging from -4 to 4) and the y-axis is Test Information (ranging from 0.0 to 1.0). The test information curve is the sum of the five item information curves, peaking at 1.0 at theta = -1. The legend indicates: Test Information (black), Item 1 (red), Item 2 (green), Item 3 (blue), Item 4 (cyan), Item 5 (magenta).

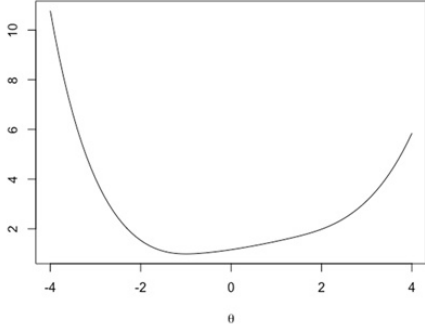
## 2.11 Standard Error of Measurement

**Standard Error of Measurement**

□ Standard error of measurement (SE) is the reciprocal of the test information at a given ability level.

$$SE_i = \frac{1}{\sqrt{I_i}}$$

SE



theta	SE
-4	10
-2	2.5
0	1.5
2	2.5
4	6

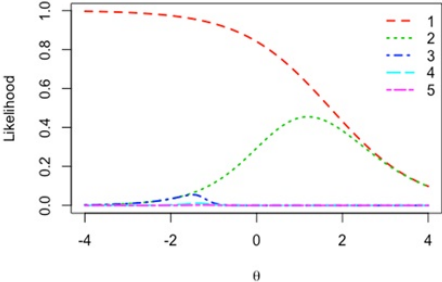
## 2.12 Likelihood

**Likelihood**

□ Likelihood is the plausibility of a response pattern

$$L_i = \prod_{j=1}^J P_{ij}^{y_{ij}} Q_{ij}^{1-y_{ij}} = Q_{i1} \times P_{i2} \times Q_{i3} \times P_{i4} \times P_{i5}$$

Likelihood of 1 to 5 items (0,1,0,1,1)



where  $y_{ij}$  is examinee  $i$ 's response on item  $j$

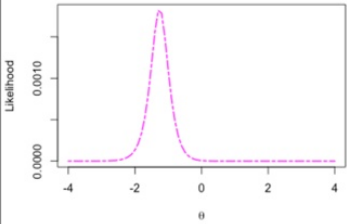
## 2.13 Log-Likelihood

**Log-likelihood**

Likelihoods approach 0 and log transformation can stretch it to more widely spread negative numbers

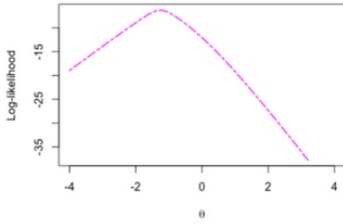
$$ll_j = \sum_{i=1}^N y_{ij} \log(P_{ij}) + (1 - y_{ij}) \log(Q_{ij})$$

**Likelihood with 5 items**



The graph shows a very narrow, sharp peak of the likelihood function centered at 0. The y-axis is labeled 'Likelihood' and ranges from 0.0000 to 0.0010. The x-axis is labeled 'θ' and ranges from -4 to 4.

**Log-likelihood with 5 items**



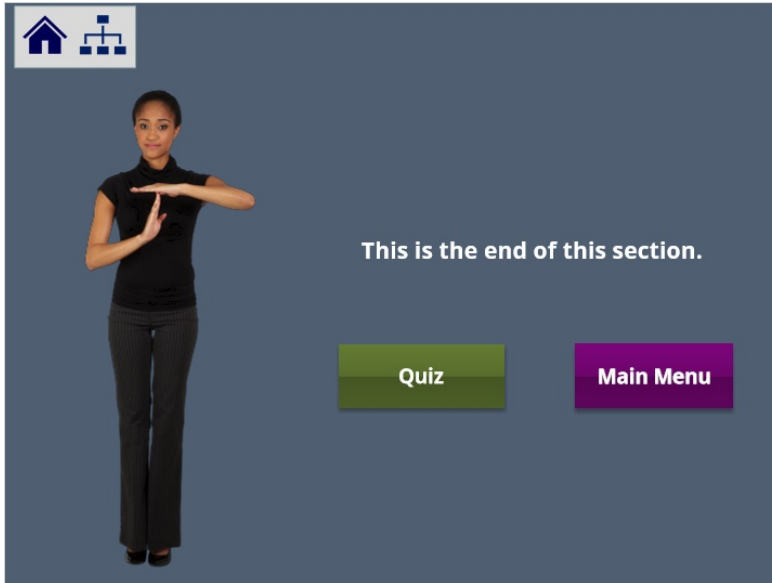
The graph shows a much wider and flatter curve of the log-likelihood function centered at 0. The y-axis is labeled 'Log-likelihood' and ranges from -35 to -15. The x-axis is labeled 'θ' and ranges from -4 to 4.

## 2.14 Summary

**Summary**

- IRT puts item difficulty and examinee ability on the same scale, thus exerts advantages over CTT
- With item parameters, item response functions as well as likelihood functions can be derived
- Log-likelihood function is used more in computation than likelihood function, as it has wider spread

## 2.15 Bookend: IRT Basics






## 3. Section 2: Calibration and Scoring

### 3.1 Cover: Calibration and Scoring





### 3.2 Estimation Algorithms: Learning Objectives


  **Learning Objectives**



1. Understand the difference between scoring and calibration
2. Learn about different scoring and calibration procedures
3. Learn how to find the maximum of a function
4. Understand the logic behind the EM algorithm

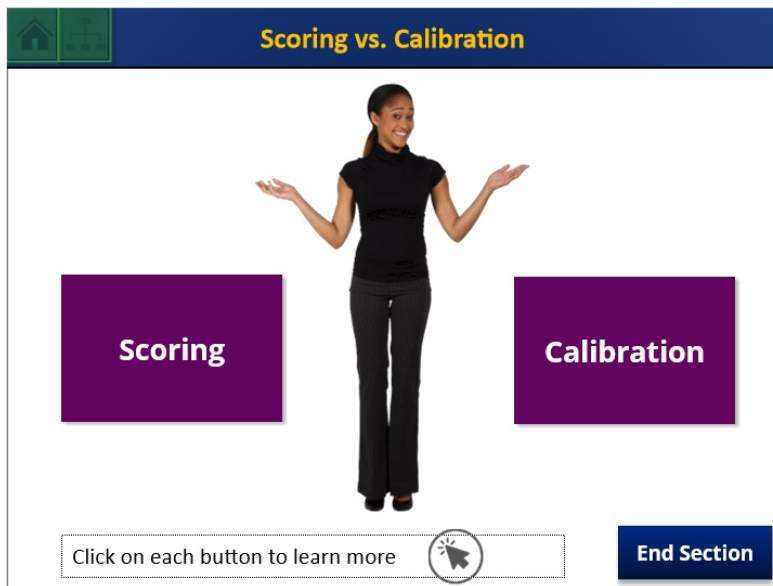
### 3.3 Scoring vs. Calibration I

  **Scoring vs. Calibration**




- In IRT, calibration is usually referred to the estimation of item parameters
- On the other hand, scoring is the estimation of person parameters

### 3.4 Topic Selection



Scoring vs. Calibration

Scoring Calibration

Click on each button to learn more 

End Section

A woman in a black top and pants stands with her arms outstretched between two purple buttons labeled 'Scoring' and 'Calibration'. The slide has a dark blue header with a home icon and a title. At the bottom, there is a navigation bar with a 'Click on each button to learn more' button containing a mouse cursor icon, and an 'End Section' button.

### 3.5 Bookmark: Scoring



Scoring

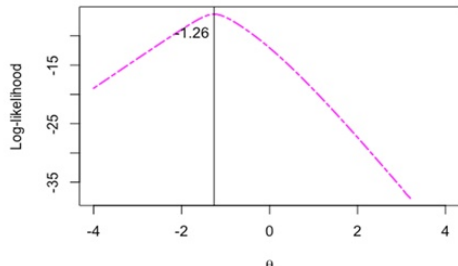
A graphic featuring a series of colorful, curved arrows pointing towards the right, set against a background of teal and yellow. The word 'Scoring' is written in blue text in the upper right corner. The slide includes a home icon and a tree icon in the top left corner.

### 3.6 Scoring: MLE

**Scoring**

- Scoring is always performed when the item parameters are known. When they are unknown, calibration is performed first.
- Using likelihood  $\rightarrow$  Maximum likelihood estimation (MLE)

**Maximum Likelihood Estimation (MLE)**

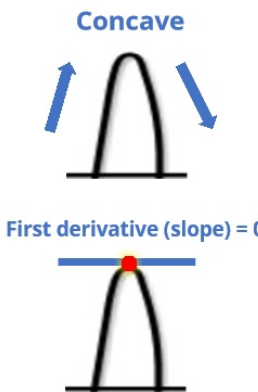
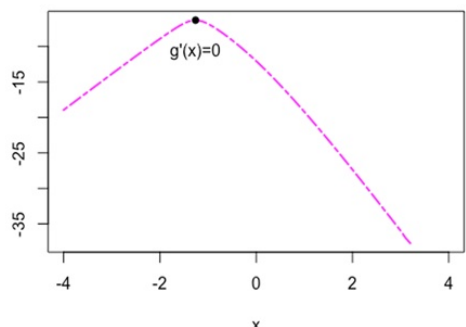


The graph shows a downward-opening parabola representing the log-likelihood function. The vertical axis is labeled 'Log likelihood' and has tick marks at -35, -25, and -15. The horizontal axis has tick marks at -4, -2, 2, and 4, with a '0' centered below the axis. A vertical line is drawn from the peak of the curve down to the x-axis, where the value is labeled as -1.26.

### 3.7 Finding a Function Maximum

**Finding a Function Maximum**

- How to find maximum of a function  $g(x)$ ?  
 $g'(x) = 0$



The graph shows a downward-opening parabola representing the function  $g(x)$ . The vertical axis is labeled 'g(x)' and has tick marks at -35, -25, and -15. The horizontal axis is labeled 'x' and has tick marks at -4, -2, 0, 2, and 4. A point is marked on the peak of the curve with the label  $g'(x)=0$ .

The diagram illustrates the concept of a concave function. The top part shows a downward-opening curve with two blue arrows pointing outwards from the peak, labeled "Concave". The bottom part shows the same curve with a horizontal blue line tangent to the peak, labeled "First derivative (slope) = 0".

### 3.8 Newton-Raphson Algorithm

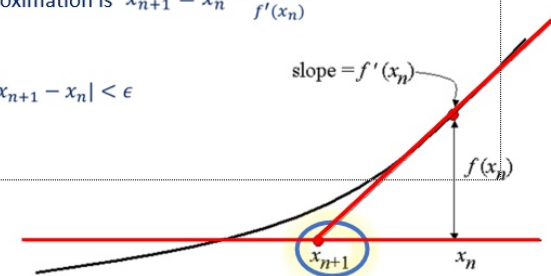
**Newton-Raphson Algorithm**

How to find  $g'(x) = 0$ ?

Newton Raphson iteration

$f(x) = g'(x)$

1. Start with initial value  $x_0$  (becomes  $x_n$  on first iteration)
2. A better approximation is  $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$
- ...
3. Stop, when  $|x_{n+1} - x_n| < \epsilon$



The diagram illustrates the Newton-Raphson method. A black curve represents the function  $f(x)$ . A red tangent line is drawn at the point  $(x_n, f(x_n))$ . The slope of this tangent line is labeled as  $f'(x_n)$ . The next iteration point  $x_{n+1}$  is the x-intercept of the tangent line, where it crosses the horizontal axis. The current point  $x_n$  is marked on the x-axis, and  $f(x_n)$  is marked on the y-axis. The point  $x_{n+1}$  is circled in blue.

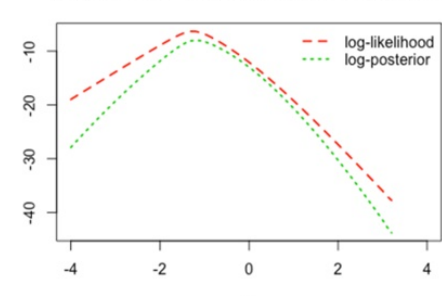
### 3.9 Posterior Distributions

**Posterior Distributions**

When response pattern consists of all 0 or all 1, likelihood function doesn't have a maximum, thus MLE does not work.



Using posterior distribution can solve this problem

**Log-likelihood and log-posterior comparison**



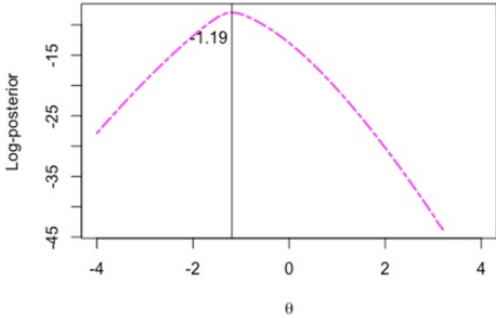
The graph shows two curves: a dashed red line for log-likelihood and a dotted green line for log-posterior. Both curves are downward-opening parabolas. The log-likelihood curve has its maximum at  $x=0$ . The log-posterior curve has its maximum at approximately  $x=1.5$ . The x-axis ranges from -4 to 4, and the y-axis ranges from -40 to -10.

### 3.10 Maximum A Posteriori



  **Maximum A Posteriori**

Using posterior distribution  
Maximum a posteriori (MAP) (only use the maximum)

**Maximum A Posteriori (MAP)**

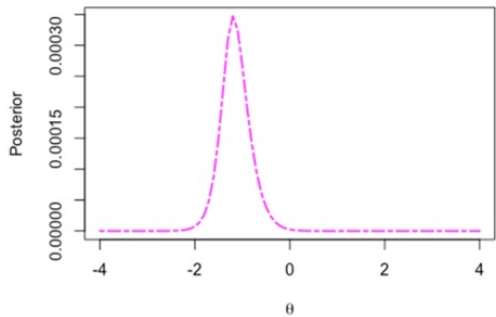


### 3.11 Expected A Posteriori I

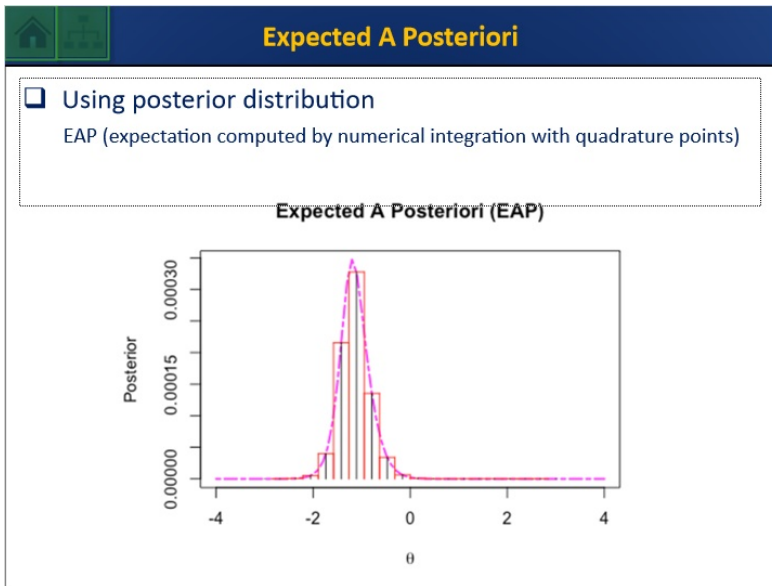
  **Expected A Posteriori**

Using posterior distribution  
Expected a posteriori (EAP) (use the entire distribution)

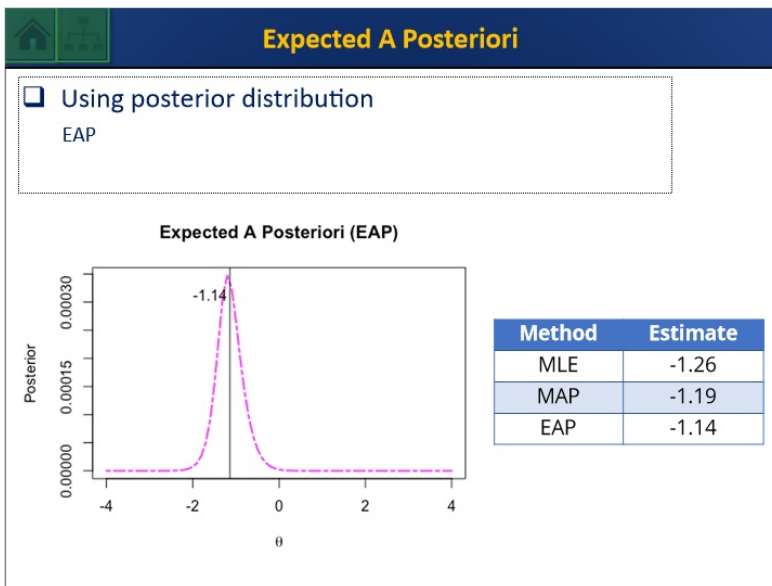
**Posterior**



### 3.12 Expected A Posteriori II



### 3.13 Expected A Posteriori III



### 3.14 Bookend: Scoring



### 3.15 Bookmark: Calibration



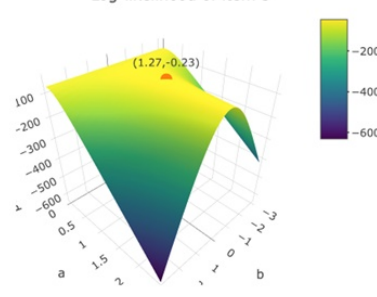
### 3.16 Calibration with Unknown Item Parameters

**Calibration with Known Examinee Parameters**

If examinee parameters are known, MLE is used

$$ll_j = \sum_{i=1}^N y_{ij} \log(P_{ij}) + (1 - y_{ij}) \log(Q_{ij})$$

Log-likelihood of item 5



A 3D surface plot showing the log-likelihood function for item 5. The vertical axis represents the log-likelihood value, ranging from -600 to 100. The horizontal axes are labeled 'a' and 'b', with 'a' ranging from 0 to 2 and 'b' ranging from 0 to 1.5. A color bar on the right indicates the log-likelihood values, with a gradient from purple (-600) to yellow (100). The surface is a smooth, bell-shaped curve peaking at the coordinates (1.27, -0.23).

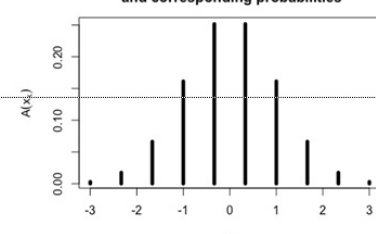
### 3.17 Calibration with Unknown Item and Person Parameters

**Calibration with Unknown Item and Person Parameters**

Marginal maximum likelihood estimation (MMLE)

- Although examinee parameters are unknown, we can assume a known population distribution and replace the person parameters with quadrature points.
- $x_k$  are quadrature points and  $A(x_k)$  are corresponding quadrature weights/probabilities under the population distribution
- The quadrature weights sum to 1

Quadrature points and corresponding probabilities



A bar chart showing the quadrature points  $x_k$  on the x-axis and the corresponding probabilities  $A(x_k)$  on the y-axis. The x-axis ranges from -3 to 3, and the y-axis ranges from 0.00 to 0.20. The bars represent the probabilities for each quadrature point, with the highest probability around  $x_k = 0$ .

### 3.18 MMLE Log-Likelihood

MMLE Log-Likelihood

MMLE

- Replace single response  $y_{ij}$  with  $\bar{r}_{kj}$  which is number of correct answer in group  $k$ ; replace  $1-y_{ij}$  with  $\bar{n}_k - \bar{r}_{kj}$ , where  $\bar{n}_k$  is the size of group  $k$ .
- $\bar{n}_k$  and  $\bar{r}_{kj}$  are called artificial data
- The log-likelihood function for item  $j$  becomes:

$$ll_j = \sum_{i=1}^N \bar{r}_{kj} \log(P_{kj}) + (\bar{n}_k - \bar{r}_{kj}) \log(Q_{kj})$$

### 3.19 MMLE Log-Likelihood II



MMLE Log-Likelihood

MMLE

- $L_{ik} = \prod_{j=1}^J P_{kj}^{y_{ij}} Q_{kj}^{1-y_{ij}}$  is the likelihood of an examinee  $i$  with ability  $x_k$
- Compute  $\bar{n}_k$  and  $\bar{r}_{jk}$  based on known item parameters and responses



$$\bar{n}_k = \frac{\sum_{i=1}^N L_{ik} A(x_k)}{\sum_{k=1}^K L_{ik} A(x_k)}$$
$$\bar{r}_{jk} = \frac{\sum_{i=1}^N y_{ij} L_{ik} A(x_k)}{\sum_{k=1}^K L_{ik} A(x_k)}$$


### 3.20 EM Algorithm

  **Expectation-Maximization (EM) Algorithm**

- In reality, item parameters are usually unknown
- The EM algorithm iteratively estimates item parameters
- Begin by assigning starting values for item parameters, then cycle between E-steps and M-steps until the estimates converge
  - E-step: assume item parameters are known and compute  $\bar{n}_k$  and  $\bar{r}_{kj}$
  - M-step: using previously calculated values of  $\bar{n}_k$  and  $\bar{r}_{kj}$ , and compute MLE of item parameters

### 3.21 Bookend: Calibration

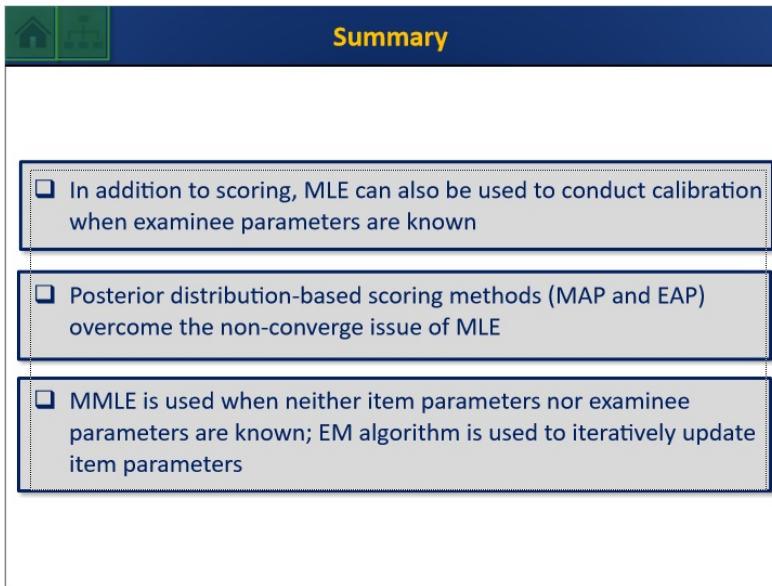
 



This is the end of this part.

**Topic Selection**

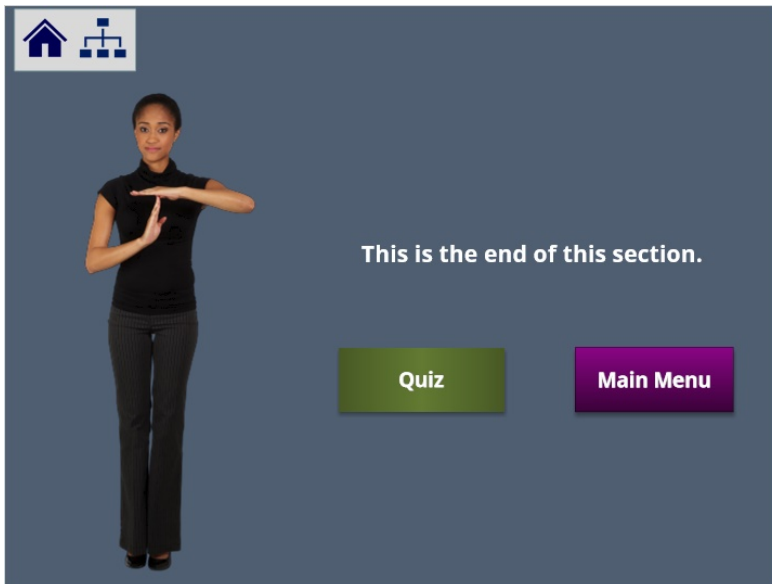
### 3.22 Summary






**Summary**

- In addition to scoring, MLE can also be used to conduct calibration when examinee parameters are known
- Posterior distribution-based scoring methods (MAP and EAP) overcome the non-converge issue of MLE
- MMLE is used when neither item parameters nor examinee parameters are known; EM algorithm is used to iteratively update item parameters

### 3.23 Bookend: IRT Basics





This is the end of this section.

[Quiz](#) [Main Menu](#)

## 4. Section 3: Factors Affecting Estimation Accuracy

### 4.1 Cover: Factors Affecting Estimation Accuracy



### 4.2 Factors Affecting Estimation Accuracy: Learning Objectives

Learning Objectives

1. Understand how estimation accuracy is quantified
2. Recognize common factors affecting estimation accuracy
3. Figure out strategies to improve estimation accuracy
4. Understand how multiple factors can affect each other

### 4.3 Estimation Accuracy Evaluation Criteria

**Estimation Accuracy Evaluation Criteria**

- Bias and root mean square error (RMSE) are used to evaluate the estimation accuracy of calibration and scoring
- $\pi$  is the true value of a parameter and  $\hat{\pi}$  is the estimated value of this parameter

**Bias:** 
$$\frac{\sum_{i=1}^n (\hat{\pi}_i - \pi_i)}{n}$$

**RMSE:** 
$$\sqrt{\frac{\sum_{i=1}^n (\hat{\pi}_i - \pi_i)^2}{n}}$$

- The *proportional* bias and RMSE are computed as the proportion of bias and RMSE of a parameter over the SD of that parameter.

### 4.4 Identify the Need to Improve Estimation Accuracy

**Identify the Need to Improve Estimation Accuracy**

- How to identify the need of improving estimation accuracy?
  - Look at population size.
    - > 500 for 2PL
    - > 1000 for 3PL
  - Look at SE of parameter estimation (especially  $a$  parameter)


**Estimation Accuracy**

The graph plots Standard Error (SE) on the y-axis (0.1 to 0.6) against various parameters on the x-axis. Four lines represent different conditions: 'original' (red), 'short test' (green), 'skewed population' (blue), and 'small population' (cyan). For parameters 'a', 'b', and 'c', SE is generally lower, with 'small population' showing the highest SE. For 'ML', 'skewed population' has the highest SE (~0.6). For 'ML\_est', 'EAP', 'EAP\_est', 'MAP', and 'MAP\_est', SE values are relatively stable and low, around 0.3-0.4.

Parameter	original	short test	skewed population	small population
a	0.25	0.35	0.45	0.55
b	0.20	0.30	0.40	0.50
c	0.15	0.25	0.35	0.45
ML	0.35	0.60	0.45	0.40
ML_est	0.35	0.35	0.35	0.35
EAP	0.35	0.35	0.35	0.35
EAP_est	0.35	0.35	0.35	0.35
MAP	0.35	0.35	0.35	0.35
MAP_est	0.35	0.35	0.35	0.35

## 4.5 Topic Selection

Factors Affecting Estimation Accuracy

Click on each image to learn more 

End Section

## 4.6 Bookmark: Methodological Choices

Methodological Choices

## 4.7 Scoring Method

**Scoring Method**

The utilization of MLE, MAP, and EAP affects the accuracy of scoring:

- Example with 40 3PL items and 1000 normally distributed examinees
- Use the real item parameters in scoring
- Use standard normal distribution is used as prior for MAP and EAP

**Estimation Accuracy**

The graph displays Bias and RMSE for three scoring methods: MLE, MAP, and EAP. The y-axis represents Bias/RMSE from 0.0 to 0.4. The x-axis represents the Scoring method. The RMSE (solid red line) starts at approximately 0.45 for MLE and drops to about 0.33 for MAP and EAP. The Bias (dashed red line) remains very low, near 0.0, across all methods.

Scoring method	Bias	RMSE
MLE	~0.01	~0.45
MAP	~0.01	~0.33
EAP	~0.01	~0.33

## 4.8 Calibration Error

**Calibration Error**

Calibration error enlarges scoring error

**Estimation Accuracy**

The graph compares Bias and RMSE for three scoring methods (MLE, MAP, EAP) with and without calibration. The y-axis represents Bias/RMSE from 0.0 to 0.5. The x-axis represents the Scoring method. The RMSE (solid red line) starts at approximately 0.45 for MLE and drops to about 0.33 for MAP and EAP. The RMSE with calibration (solid green line) starts at approximately 0.50 for MLE and drops to about 0.33 for MAP and EAP. The Bias (dashed red line) remains very low, near 0.0, across all methods. The Bias with calibration (dashed green line) also remains very low, near 0.0, across all methods.

Scoring method	Bias	RMSE	Bias-with calibration	RMSE-with calibration
MLE	~0.01	~0.45	~0.01	~0.50
MAP	~0.01	~0.33	~0.01	~0.33
EAP	~0.01	~0.33	~0.01	~0.33

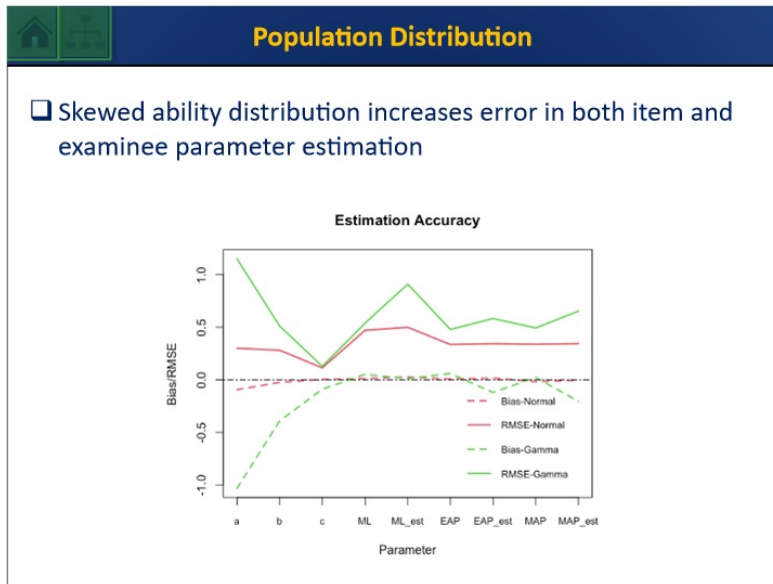
## 4.9 Bookend: Calibration



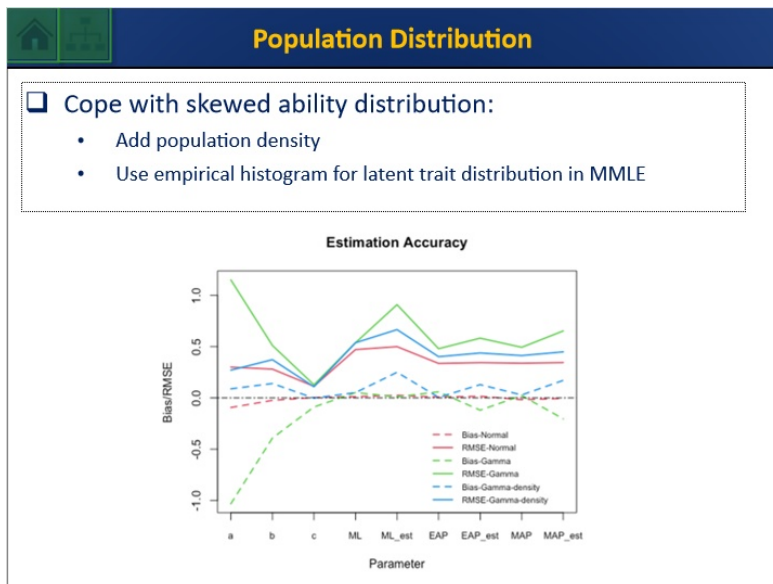
## 4.10 Bookmark: Population Characteristics



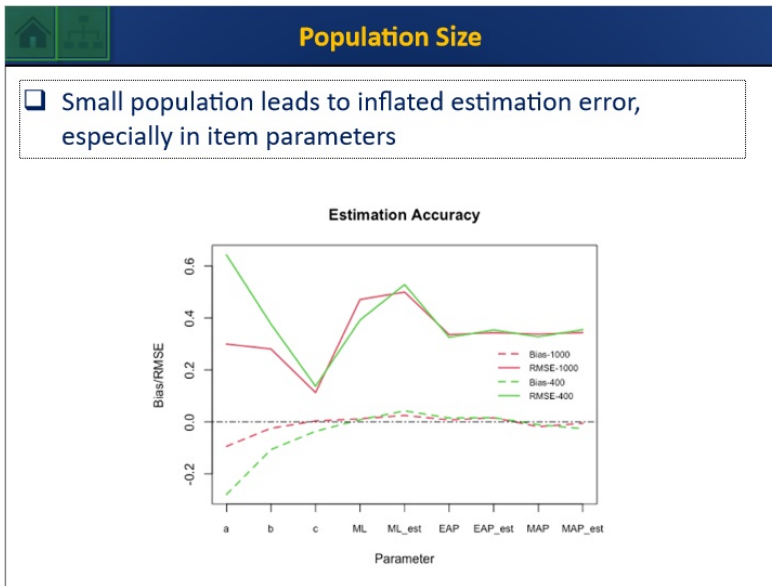
## 4.11 Population Distribution I



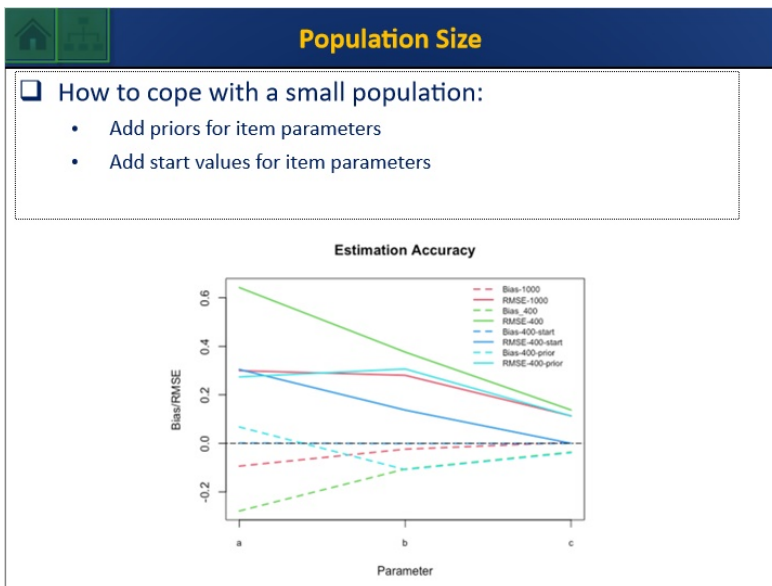
## 4.12 Population Distribution II



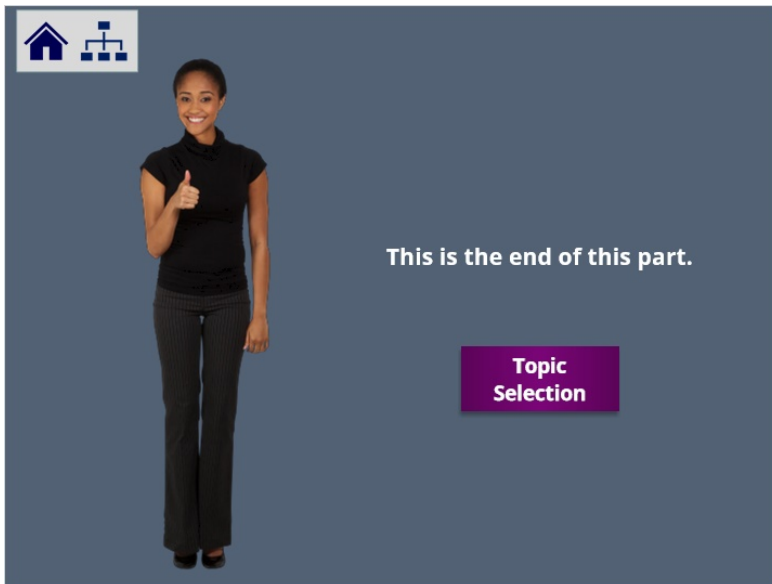
### 4.13 Population Size I



### 4.14 Population Size II



#### 4.15 Bookend: Calibration



#### 4.16 Bookmark: Item and Test Characteristics



## 4.17 Item Parameter Types

🏠 📄 Item Parameter Type

In 3PL,  $b$  parameters are easier to calibrate, thus have relatively smaller estimation error

↑

↑

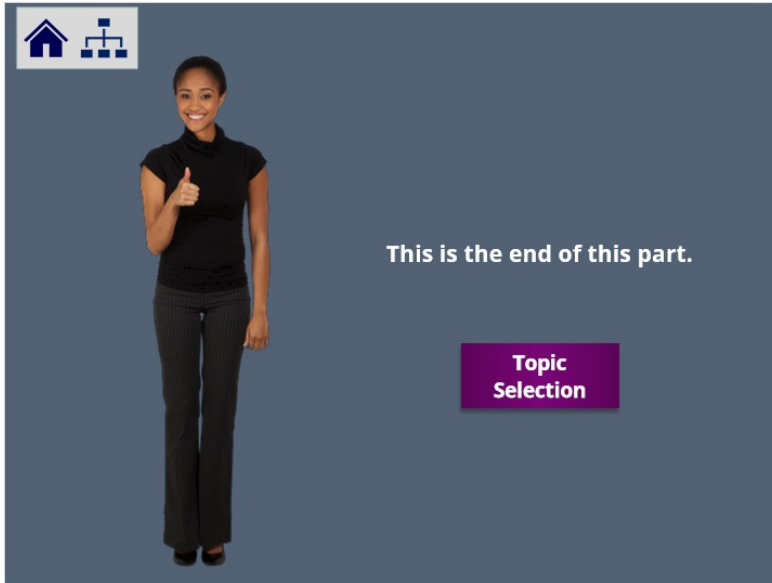
## 4.18 Test Length

🏠 📄 Test Length

Short tests increase error in both item and person parameter estimation

The advantage of MAP and EAP is greater for short tests

## 4.19 Bookend: Calibration

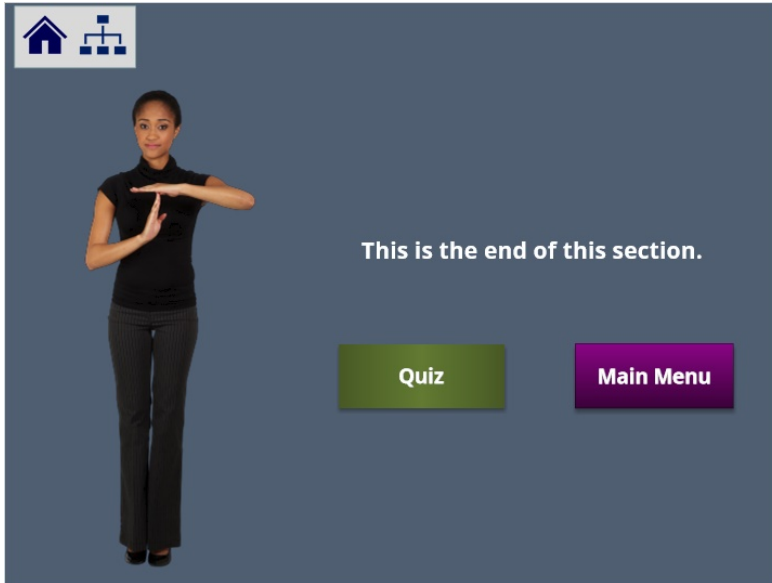


## 4.20 Summary

Summary

- Various factors can influence estimation accuracy
- SE of parameter estimation is a good resource to identify potential problems in estimation
- Potential remedies for estimation issues:
  - Adding population density prior
  - Adding item parameter prior
  - Adding item parameter start values
  - Using empirical histogram for latent trait in MMLE

## 4.21 Bookend: IRT Basics



## 4.22 Module Cover (START)

