DM19 SLIDES (IRT Estimation, Version 1.0)

1. Module Overview

1.1 Module Cover (START)

1.2 Instructors

Meet the instructors:

Zhuoran Wang
Prometric

Nathan Thompson
Assessment Systems
1.3 Designers

Meet the designers:

Jon Lehrfeld  
ETS

André A. Rupp  
Mindful Measurement

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:
1.5 Overview

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:

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

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

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

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)

1.9 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

**Module Citation**

1.12 Main Menu

- 01 IRT Basics [10 Minutes]
- 02 Calibration and Scoring [15 Minutes]
- 03 Factors Affecting Estimation Accuracy [20 Minutes]
- 04 Example using software [20 Minutes]
- 05 Quizzes [20 Minutes]
2. Section 1: IRT Basics

2.1 Cover: IRT Basics

2.2 IRT Basics: 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

- 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

- 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

One-Parameter Logistic Model (1PL, Rasch Model)

$$P_{ij} = \frac{1}{1 + \exp(\theta_i - b_j)}$$

Rasch Model IRF

2.6 Item Characteristic Functions II

Two-Parameter Logistic Model (2PL)

$$P_{ij} = \frac{1}{1 + \exp(-Da_i(\theta_i - b_j))}$$

2PL Model IRF
2.7 Item Characteristic Functions III

Three-Parameter Logistic Model (3PL)

\[ P_{ij} = c + \frac{1 - c}{1 + \exp(-D_{ij}(\theta_i - b_j))} \]

2.8 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

- IRT advances the concept of item and test information to replace reliability.
  
  \[
  \begin{align*}
  &1PL: \quad I_{ij} = P_{ij}Q_{ij} \\
  &2PL: \quad I_{ij} = a_j^2 P_{ij}Q_{ij} \\
  &3PL: \quad I_{ij} = a_j^2 \left( \frac{p_{ij} - c_j}{1-c_j} \right)^2 \frac{q_{ij}}{P_{ij}} \\
  \text{where} \quad Q_{ij} &= 1 - P_{ij}
  \end{align*}
  \]

2.10 Test Information

- Test information is the sum of all item information

- [Graph showing test information for different items]
2.11 Standard Error of Measurement

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

\[ SE_i = \frac{1}{\sqrt{I_i}} \]

2.12 Likelihood

- Likelihood is the plausibility of a response pattern:

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

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

\[ L_I = \sum_{i=1}^{N} y_{ij} \log(P_{ij}) + (1 - y_{ij}) \log(Q_{ij}) \]

2.14 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

This is the end of this section.

Quiz  Main Menu

3. Section 2: Calibration and Scoring

3.1 Cover: Calibration and Scoring

Section 2:
Calibration and Scoring

[15 minutes]
3.2 Estimation Algorithms: Learning Objectives

- Understand the difference between scoring and calibration
- Learn about different scoring and calibration procedures
- Learn how to find the maximum of a function
- Understand the logic behind the EM algorithm

3.3 Scoring vs. Calibration I

- 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

3.5 Bookmark: Scoring
3.6 Scoring: MLE

- 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)

3.7 Finding a Function Maximum

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

- Concave
  - First derivative (slope) = 0
3.8 Newton-Raphson Algorithm

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

Newton Raphson iteration

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

1. Start with initial value $x_0$ (becomes $x_1$ on first iteration)

2. A better approximation is $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$

... 

3. Stop, when $|x_{n+1} - x_n| < \epsilon$

3.9 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

![Graph showing log-likelihood and log-posterior comparison](image)
3.10 Maximum A Posteriori

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

3.11 Expected A Posteriori I

- Using posterior distribution
  Expected a posteriori (EAP) (use the entire distribution)
3.12 Expected A Posteriori II

- Using posterior distribution
  - EAP (expectation computed by numerical integration with quadrature points)

3.13 Expected A Posteriori III

- Using posterior distribution
  - EAP

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>-1.26</td>
</tr>
<tr>
<td>MAP</td>
<td>-1.19</td>
</tr>
<tr>
<td>EAP</td>
<td>-1.14</td>
</tr>
</tbody>
</table>
3.14 Bookend: Scoring

This is the end of this part.

3.15 Bookmark: Calibration
3.16 Calibration with Unknown Item Parameters

If examinee parameters are known, MLE is used

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

3.17 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
3.18 MMLE Log-Likelihood

$ll_{ij} = \sum_{i=1}^{N} \bar{r}_{kj} \log(p_{kj}) + (\bar{n}_k - \bar{r}_{kj}) \log(q_{kj})$

3.19 MMLE Log-Likelihood II

$L_{ik} = \prod_{j=1}^{J} p_{kj}^{y_{ij}} q_{kj}^{1-y_{ij}}$ is the likelihood of a examinee $i$ with ability $x_k$

Compute $\bar{n}_k$ and $\bar{r}_{kj}$ based on known item parameters and responses

$$\bar{n}_k = \sum_{i=1}^{N} \sum_{k=1}^{K} l_{ik} A(x_k)$$

$$\bar{r}_{kj} = \sum_{i=1}^{N} \frac{y_{ij} l_{ik} A(x_k)}{\sum_{k=1}^{K} l_{ik} A(x_k)}$$
3.20 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 $\overline{p}_k$ and $\overline{r}_{kj}$
  - M-step: using previously calculated values of $\overline{p}_k$ and $\overline{r}_{kj}$, and compute MLE of item parameters

3.21 Bookend: Calibration

This is the end of this part.
3.22 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.
4. Section 3: Factors Affecting Estimation Accuracy

4.1 Cover: Factors Affecting Estimation Accuracy

4.2 Factors Affecting Estimation Accuracy: 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

- Bias and root mean square error (RMSE) are used to evaluate the estimation accuracy of calibration and scoring.

\[
\text{Bias: } \frac{\sum_{i=1}^{n} (\hat{\pi}_i - \pi_i)}{n}
\]

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

- 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

- 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)
4.5 Topic Selection

4.6 Bookmark: Methodological Choices
4.7 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

4.8 Calibration Error

Calibration error enlarges scoring error
4.9 Bookend: Calibration

4.10 Bookmark: Population Characteristics
4.11 Population Distribution I

- Skewed ability distribution increases error in both item and examinee parameter estimation

![Population Distribution I graph]

4.12 Population Distribution II

- Cope with skewed ability distribution:
  - Add population density
  - Use empirical histogram for latent trait distribution in MML

![Population Distribution II graph]
4.13 Population Size I

- Small population leads to inflated estimation error, especially in item parameters

4.14 Population Size II

- How to cope with a small population:
  - Add priors for item parameters
  - Add start values for item parameters
4.15 Bookend: Calibration

This is the end of this part.

Topic Selection

4.16 Bookmark: Item and Test Characteristics
4.17 Item Parameter Types

- In 3PL, \( b \) parameters are easier to calibrate, thus have relatively smaller estimation error

4.18 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

- 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

This is the end of this section.

Quiz  Main Menu

4.22 Module Cover (START)