Introduction to Item Response Theory

Adam Wyse and Ji Zeng

Psychometricians

Michigan Department of Education

Office of Educational Assessment and Accountability

MICHIGAN Department of Education

Focus of this session

- The other session discussed Classical Test Theory (CTT).
- The focus of this session is on Item Response Theory (IRT) and how IRT is used at MDE.

Basic Differences between CTT and IRT

- Focus on item performance (IRT) versus Total Test performance (CTT).
- Population dependent statistics (CTT) versus population independent statistics (IRT).
- Test specific statistics (CTT) versus Test independent statistics (IRT).
- Definition, which cannot be tested (CTT), versus a model, which can be tested (IRT).
- Few assumptions (CTT) versus several assumptions (IRT).

What is IRT?

- Relates student ability and item characteristics to the probability of obtaining a particular score on an item.
- Many IRT models exist, including models for multiple-choice, short answer, and constructed response items.
- Models differ in how probabilities are related to student ability and item characteristics.

IRT assumptions

- Monotonicity: A more able person has a higher probability of responding correctly to an item than a less able person.
- Local independence: the response to one item is independent of and does not influence your probability of responding correctly to another item after controlling for ability.
- Item and person parameters do not change across populations.

Unidimensionality

- Models used by MDE also assume unidimensionality.
  - A single underlying construct measured by the assessment (i.e. mathematics achievement, reading achievement, etc.)

Common IRT Models

- Multiple-Choice and Short Answer Items
  - Rasch Model (MEAP, MEAP-Access, MI-Access FI, ELPA)
  - 2 PL Model (NAEP)
  - 3 PL Model (MME, NAEP)
- Constructed Response Items
  - Partial Credit Model (MEAP Writing, MEAP-Access Writing, MI-Access FI Expressing Ideas, ELPA)
  - Generalized Partial Credit Model (MME Writing, NAEP)

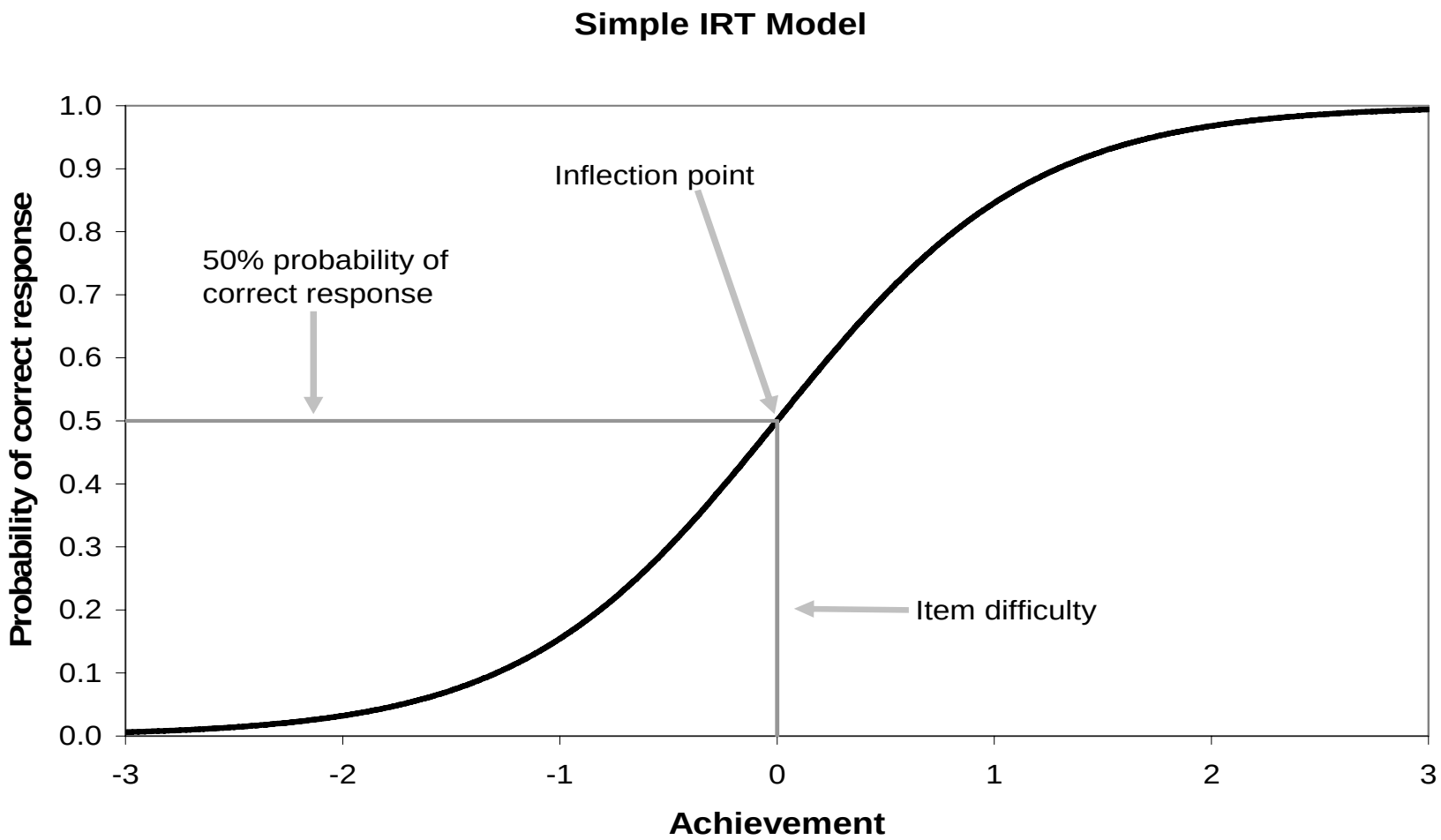
Note: NAEP is not analyzed or administered by MDE. It is a test administered by the federal government!

The Rasch Model (sometimes called the 1 PL Model)

$$P(\theta) = \frac{e^{(\theta - b)}}{1 + e^{(\theta - b)}}$$

The Rasch Model

- An item characteristic curve for a sample MEAP item



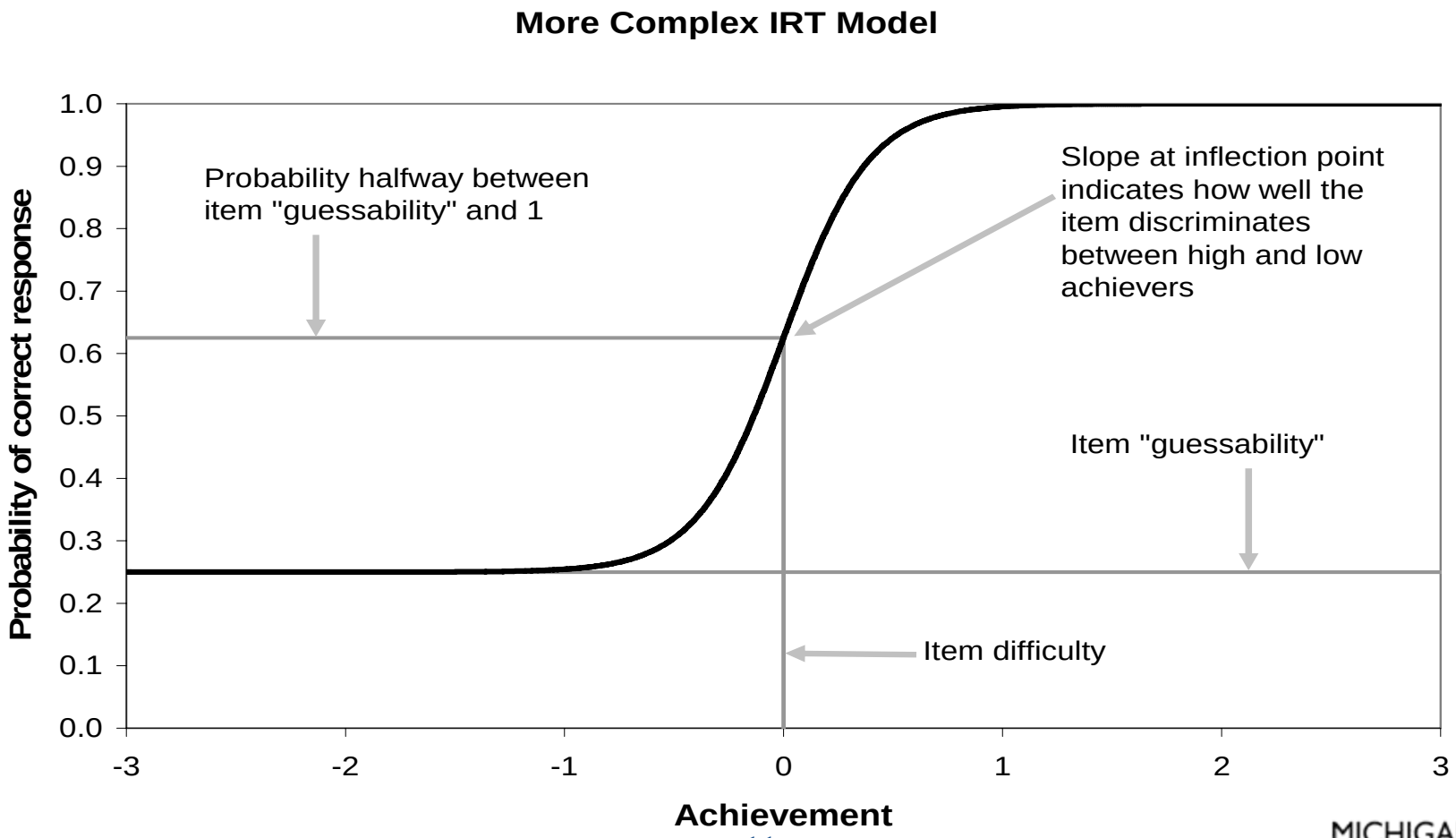
MICHIGAN Department of Education

The 3 PL Model

$$P(\theta) = c + \frac{(1-c)e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}}$$

The 3 PL Model

- An item characteristic curve for a sample MME item.



MICHIGAN Department of Education

Rasch vs. 3 PL

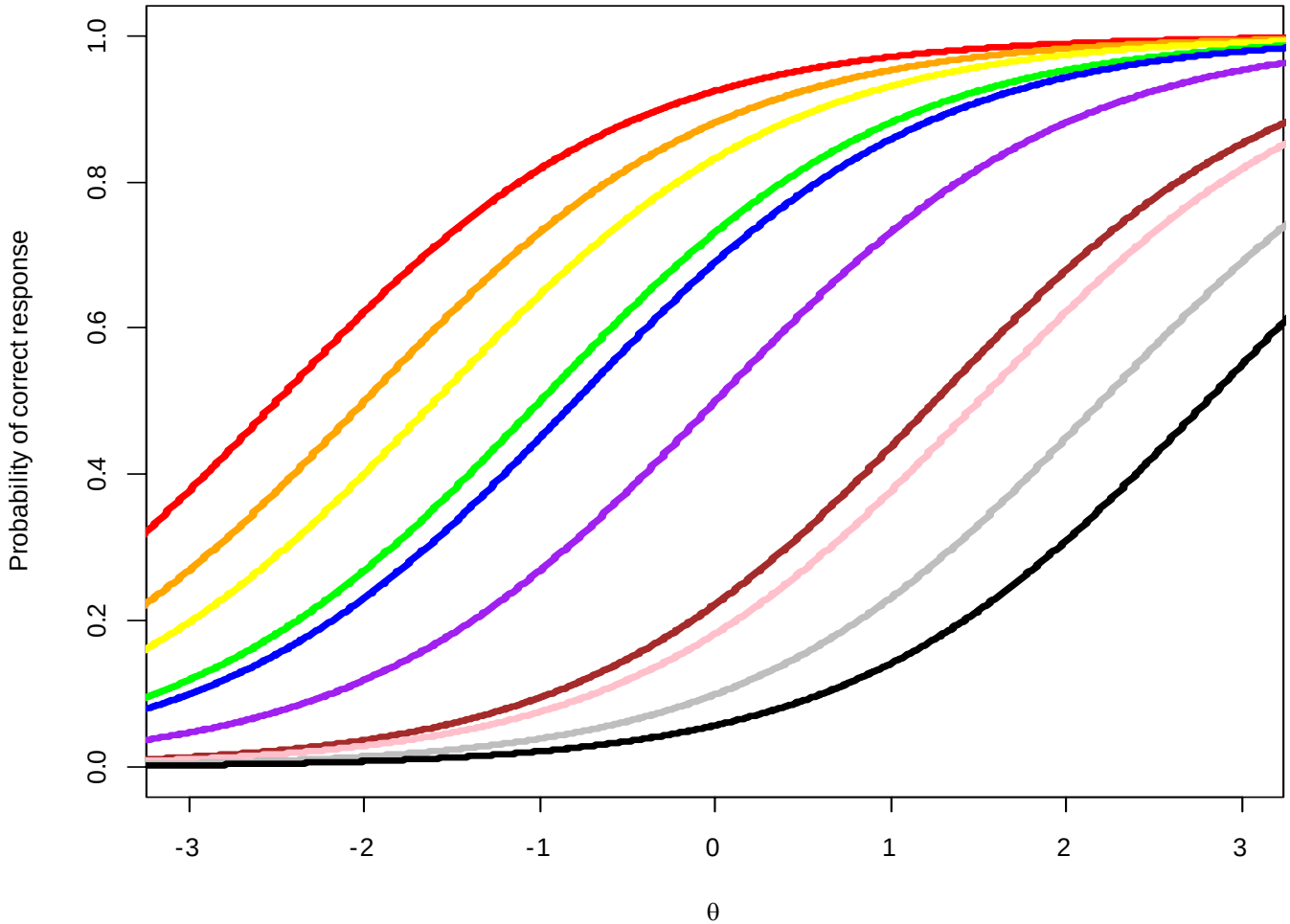
What features do the Rasch and 3PL model have in common?

What features of the Rasch and 3 PL Model are different?

Rasch vs. 3PL

- In the Rasch model, the item difficulty parameter and its difference from student ability drives the probability of a correct response. All other elements are constants in the equation.
  - Therefore, when you see the plots of multiple items, they only differ by a constant in terms of their location on the scale (shown in diagram on next slide).

10 Rasch item characteristic curves

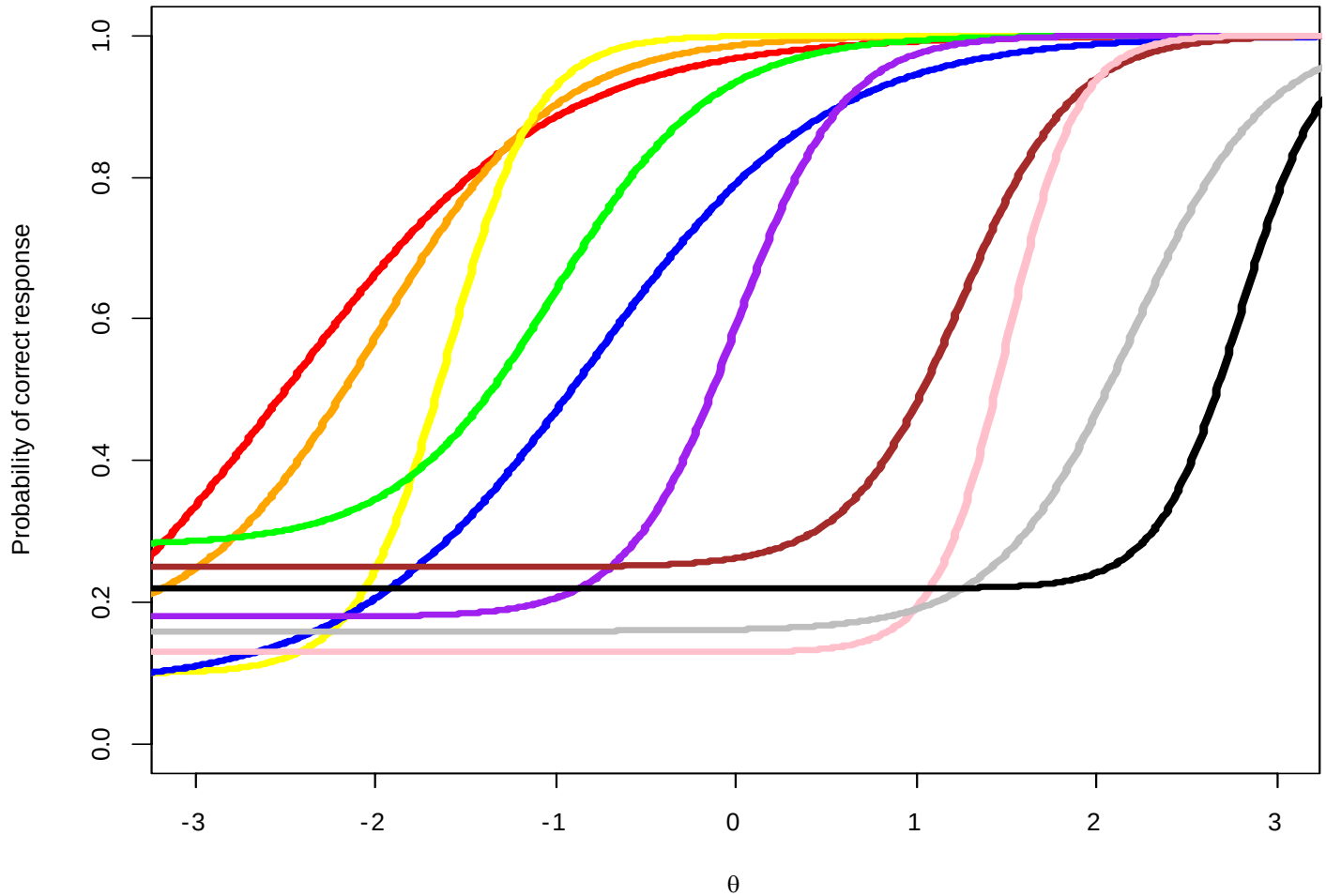


MICHIGAN Department of Education

Rasch vs. 3PL

- In 3 PL model, the difference between ability and difficulty is still the critical piece. However, the discrimination parameter changes the influence of the difference between ability and difficulty for each item. Furthermore, the minimum possible result for the equation is influenced by the 'c' parameter.
  - If $c > 0.00$, the probability of correct response is greater than 0.
  - Item characteristic curves will vary by location on the scale as well as lower asymptote (c parameter) and slope (a parameter).
  - Knowing how difficult an item is compared to another is still relevant but is not the only piece of information that leads to differences in items.

10 3 PL item characteristic curves



MICHIGAN Department of Education

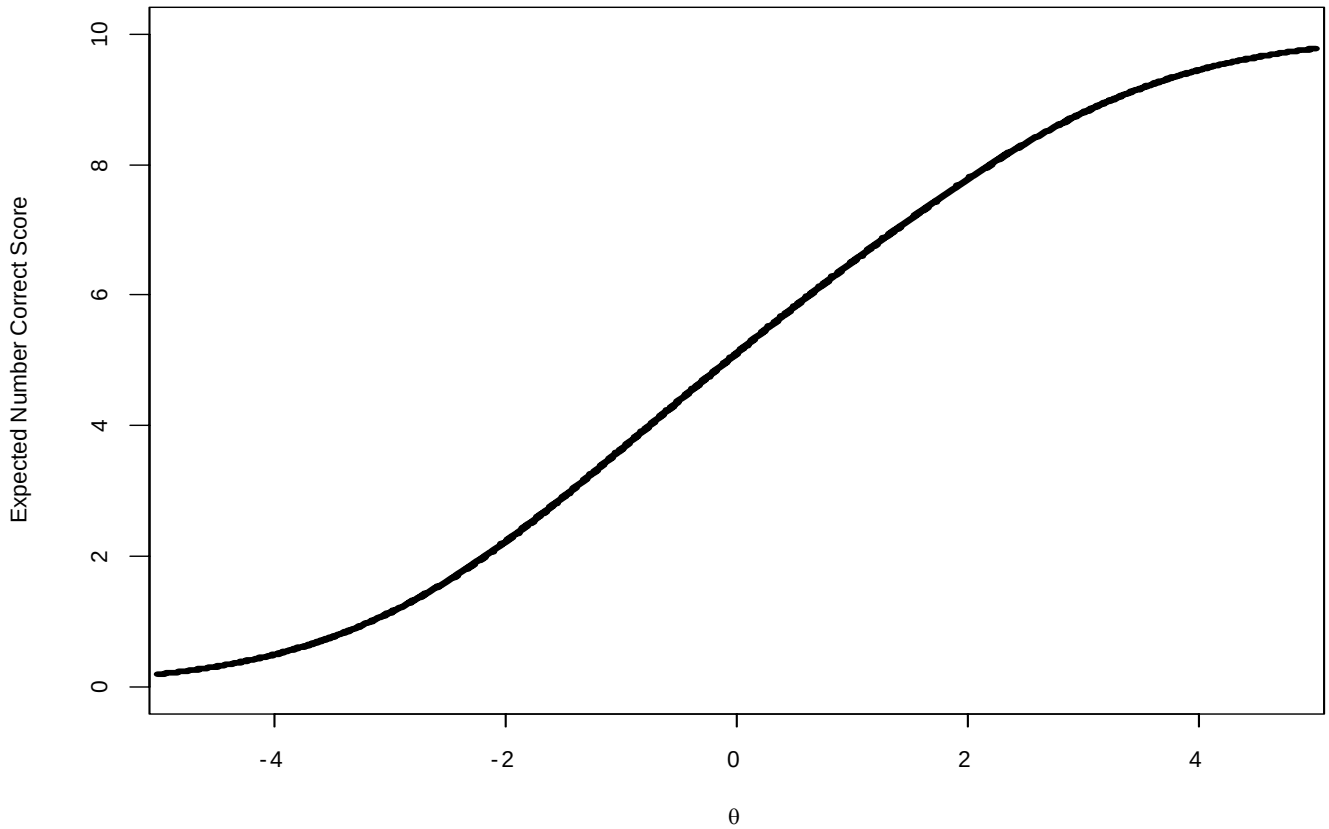
2 PL Model

- How do you end up with the 2 PL model?

Test Characteristic Curves

- Relates achievement to scores examinees are expected to receive on the assessment.
- Sum of Item Characteristic Curves in IRT.
- Defined the same way for Rasch and 3 PL models.

Example of Test Characteristic Curve for 10 Rasch Items



MICHIGAN Department of Education

Partial Credit Model (PCM)

$$P_{ix}(\theta) = P(X_i = x|\theta) = \frac{\exp\sum_{k=0}^{x}(\theta - b_{ik})}{\sum_{h=0}^{m_i}\exp\sum_{k=0}^{h}(\theta - b_{ik})}$$

Generalized Partial Credit Model (GPCM)

$$P_{ix}(\theta) = P(X_i = x|\theta) = \frac{\exp\sum_{k=0}^{x} a_i(\theta - b_{ik})}{\sum_{h=0}^{m_i}\exp\sum_{k=0}^{h} a_i(\theta - b_{ik})}$$

How do we get there?

- IRT models depend on item and person parameters.
- Item and person parameters have to be estimated.
- Person by item matrix is needed to begin the process.

MME Science
01010010111100011110 1
11010111100010110011 0
11001101100001110100 1
01101011101100101001 1

IRT Estimation

- Person by item matrix input into an IRT estimation program.
- Program uses an estimation algorithm (a set of mathematical rules) to come up with a solution.
- The end products are best estimates of the item parameters and person ability estimates.
  - Item parameters are the 'guessability', discrimination and difficulty parameters
  - Person parameters are the ability estimates we use to create a student's scale score.

Estimating Ability

- For the 3PL/GPCM, people who share the same response string (same pattern of correct and incorrect responses/ same score on constructed response items) will have the same ability estimate.
  - It is possible for people with the same raw score to end up with different ability estimates.
- In the Rasch/PCM, the raw score is used to derive the abilities.
  - Each person with the same raw score will have the same estimate of ability.

Common IRT Software Packages

- Rasch/PCM:
  - WINSTEPS
  - FACETS
  - CONQUEST
- 3PL/GPCM:
  - PARSCALE
  - MULTILOG
  - BILOG-MG (cannot be used for constructed response items)

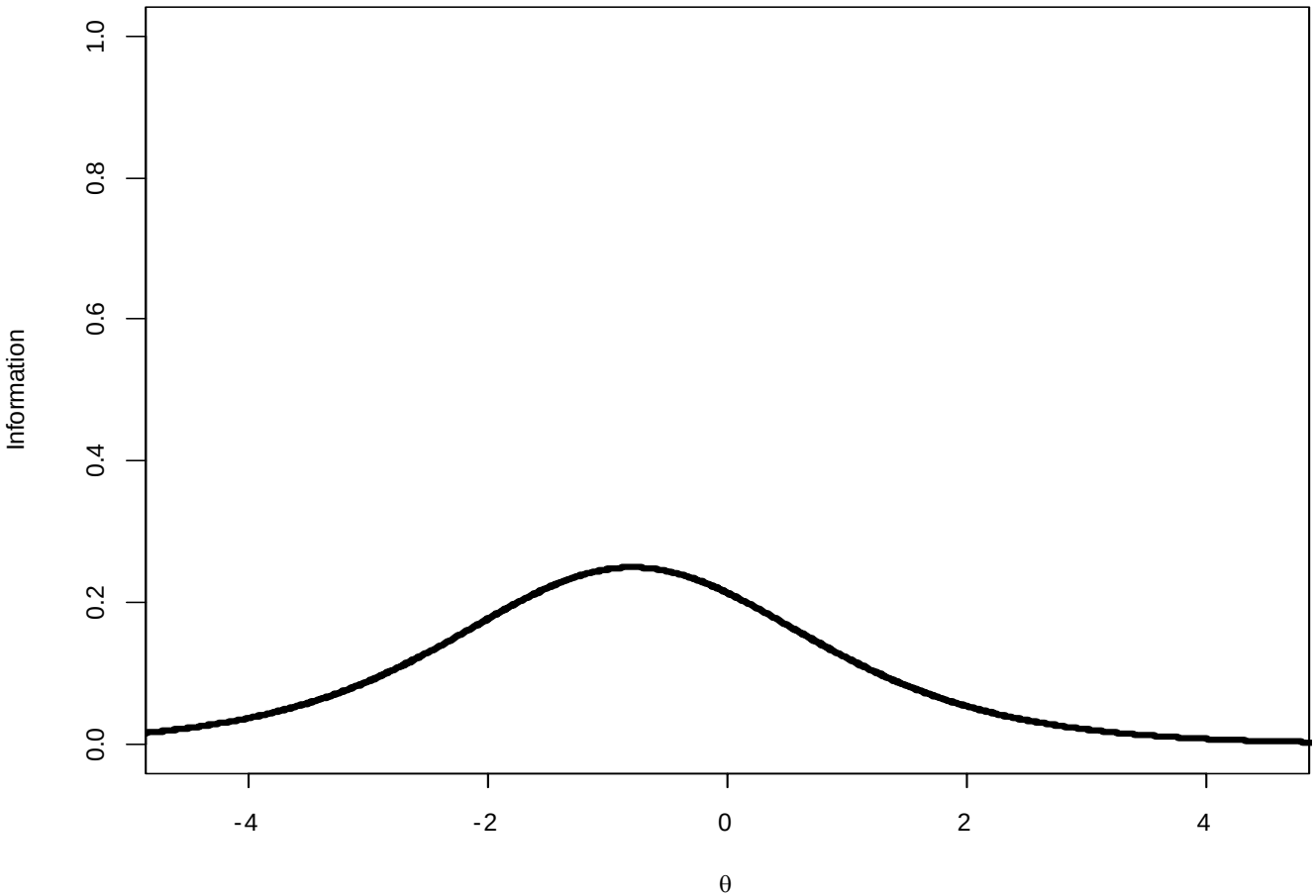
Uses of IRT

- Item/Test Information
- Conditional Standard Error of Measurement
- Creation of Scale Scores
- Standard Setting
- Equating/Linking
- Test assembly/Test Design
- Differential Item Functioning (DIF)

Item/Test Information

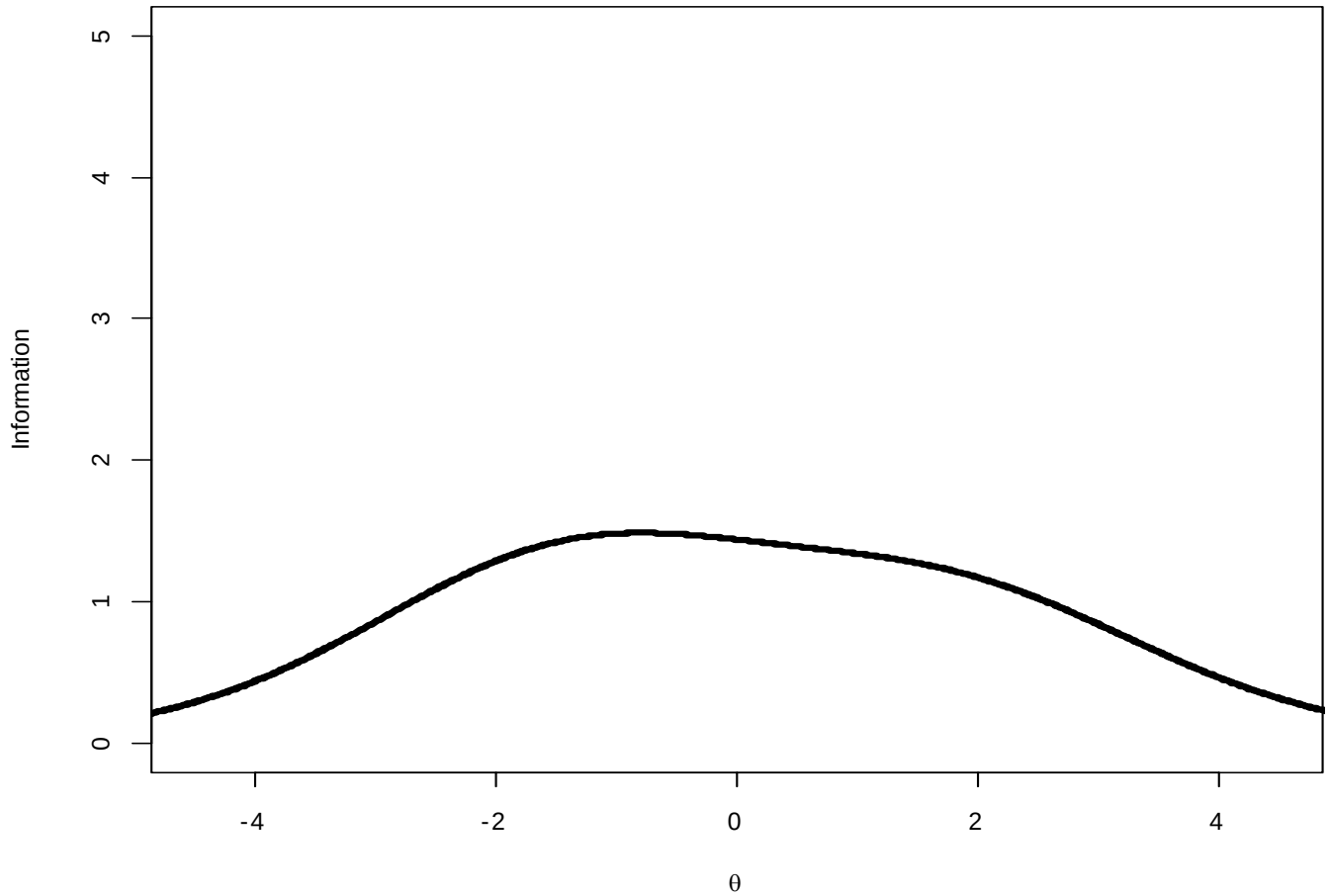
- Each IRT model has an item information function.
- Item information provides an indicator of the accuracy of ability estimates at each location.
- Test information is the sum of item information over items.

Item Information (Rasch item)



MICHIGAN Department of Education

Test information (10 Rasch items)

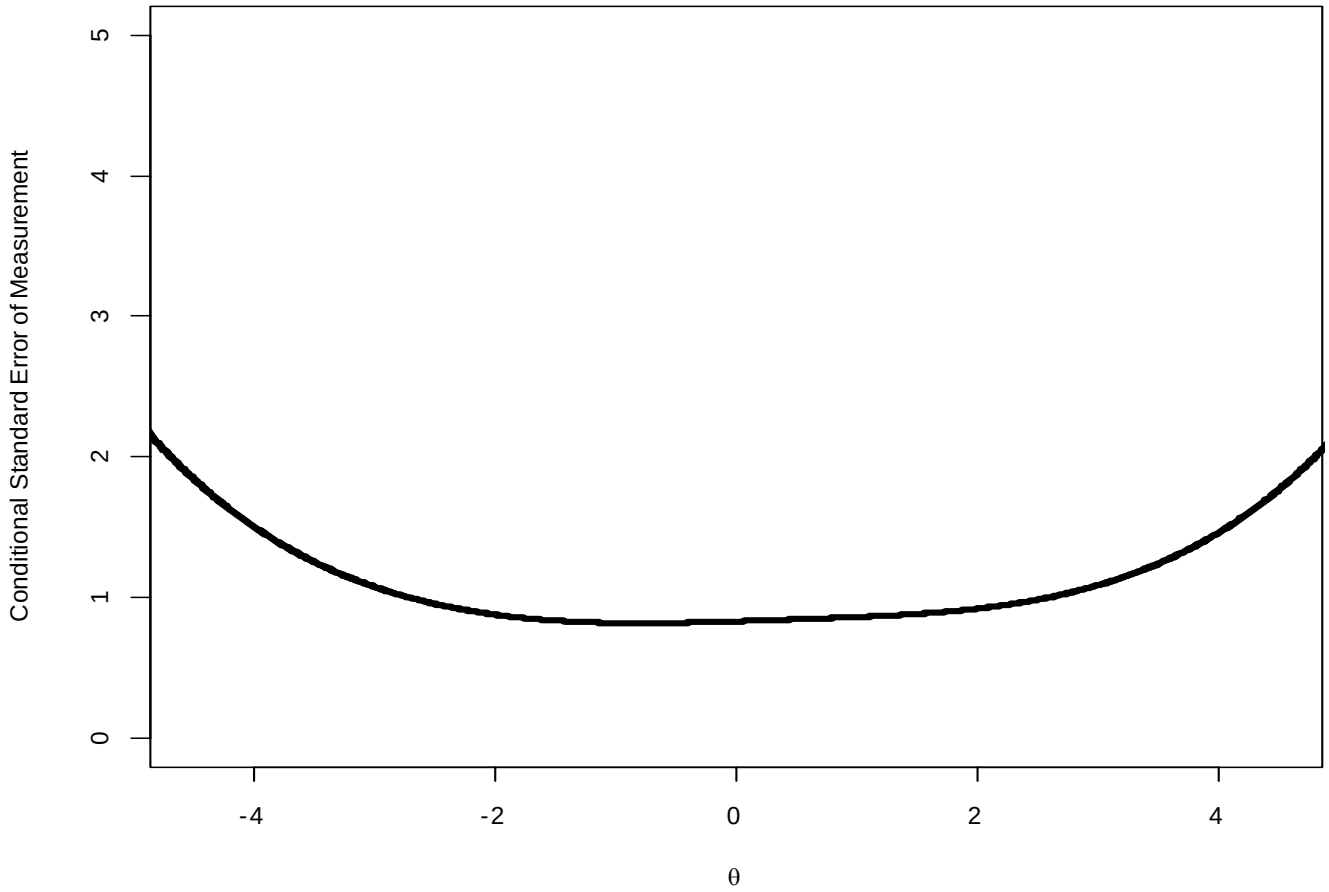


MICHIGAN Department of Education

Conditional Standard Error of Measurement

- Equal to reciprocal of the square root of the Test Information Function.
- Provides indicator of assessment accuracy at each ability level.
- MDE reports conditional standard error of measurement for student's scale scores.

Conditional Standard Error of Measurement



MICHIGAN Department of Education

Theta to scale score transformation

- Remember the linear equation?
  - $y = mx + b$
- MDE uses linear equations to transform θ (Ability) to scale scores.
- Different transformation for each grade, content area, and assessment.
- Performance levels are determined by the student's scale score.

Example of a Raw to Scale Score Table

Raw Score	Scale Score	PL	SE
0	385	4	44
1	414	4	25
2	432	4	18
…	…	…	…
9	478	4	10
10	482	3	10
…	…	…	…
14	497	3	9
15	500	2	9
…	…	…	…
23	534	2	11
24	540	1	12
…	…	…	…

Standard Setting

- Process of establishing cut scores on the score scale of an assessment.
- Involves groups of teachers, administrators, and content experts who make cut score recommendations.
- Recommendations are based on panelists' understanding of students and content as well as assessment characteristics.
- MDE has applied IRT based standard setting methods (e.g. Bookmark and Body of Work).
- State Board of Education sets final cut scores after considering panelists' cut score recommendations.

Equating/Linking

- Process of placing scores from different test administrations onto a common scale so that scores can be used interchangeably.
- Equating adjusts for differences in difficulty between test forms.
- IRT facilitates equating/linking by assuming item parameters for common items do not change over time.
- Many IRT linking methods exist for creating a common scale once this assumption is made.
- MDE uses the Stocking-Lord procedure for MME and the fixed parameter method for the other assessments.

Test Design/Assembly

- MDE checks item and content characteristics when creating new test forms.
- Make test information as large as possible near the cut scores to make performance level classifications as accurate as possible.
- Make sure that the IRT test information and test characteristic curves for alternate test versions are as close to each other as possible.
- Why do we want the test information functions and test characteristic curves to be as similar as possible?

Differential Item Functioning (DIF)

- DIF refers to the situation where examinees with the same ability differ on average in their item performance depending on subgroup membership.
- MDE checks for DIF for each subgroup (e.g. males vs. females) that it is tested on the assessment that has a large enough sample size.
- Items identified as exhibiting DIF are reviewed by a panel of teachers and content experts to make sure that they are fair to all subgroups of examinees being tested.

Summary

- You were introduced to IRT models and how they are used by MDE.
- Goal is that you leave with a greater understanding of how MDE assessments are scored, scaled, and interpreted.
- In addition, you now should have some ‘tools’ that can assist you in your own analyses.

Contact Information

Adam Wyse
(517)-373-2435
WyseA@Michigan.gov

Ji Zeng
(517)-241-3517
ZengJ@Michigan.gov

Please feel free to contact us if you have any questions ☺

MICHIGAN Department of Education