IRT Model Specifications and Scale Characteristics

Lecture #3

ICPSR Item Response Theory Workshop

Lecture Overview

- IRT model specifications
- Scaling characteristics

Purpose of IRT

- The main purpose of IRT is to create a ***scale*** for the interpretation of tests with useful properties

- Some of the properties of IRT will allow us to describe the characteristics of that scale in a more meaningful way

Model Specifications

- Logistic models are used to link person trait and item response probabilities
- The probability of a correct response is a monotonically increasing function of the trait being measured, theta
- Conditional probability of item performance is available all along the scale of the trait being measured

Scale Characteristics

- As with many (most?) metrics, the scale itself in IRT is arbitrarily chosen

- Once a scale is chosen, the model has some very useful properties:
  - Test items and person trait levels are referenced to the same interval scale
  - Person and item statistics are not dependent on one another

Dichotomous models

- These models are used when test items are binary
  - Scored as either incorrect or correct, “0” or “1”

- The three-parameter logistic (3-PL) model describes the relationship between examinee ability and the probability of a correct response with 3 parameters: difficulty, discrimination, and guessing

3-PL IRT Model

$$P(Y_{si} = 1|\theta_s)$$

$$= c_i + (1 - c_i)\ \frac{\exp(1.7a_i(\theta_s - b_i))}{1 + \exp(1.7a_i(\theta_s - b_i))}$$

Other Models

- 2-PL :: no guessing (***c***) parameter
  - It is assumed that guessing is not a factor in responding to an item

- 1-PL :: no ***c*** or slope (***α***) parameter
  - It is also assumed that all items are equally discriminating
  - A.K.A. “the Rasch model”

2-PL IRT Model

$$P(Y_{si} = 1|\theta_s) = \frac{\exp\left(1.7a_i(\theta_s - b_i)\right)}{1 + \exp\left(1.7a_i(\theta_s - b_i)\right)}$$

2-PL :: c=0

1-PL IRT Model

$$P(Y_{si} = 1|\theta_s) = \frac{\exp\left(1.7a(\theta_s - b_i)\right)}{1 + \exp\left(1.7a(\theta_s - b_i)\right)}$$

1-PL :: a=common, c=0,

IRT Parameters

- Ability (θ): generally scaled with a mean of 0 and SD of 1 (like a z-score)

- The effective range of θ is therefore from about -4 to +4

- This scale is arbitrary, but once chosen it:
  - Is used to identify the model
  - Determines the scale of the item parameters

Item Parameters

- b – difficulty or location
  - Same scale as θ, generally $-4 \leq b \leq +4$
- a – discrimination or slope
  - Often bounded by 0, generally $a \leq 2.0$
- c – guessing or lower asymptote
  - Bounded by 0 & 1, generally $c \leq 0.25$

Important Assumptions

- Unidimensionality of the Test
- Local Independence
- Nature of the ICC
- Parameter Invariance

Arbitrariness of the scale

- Parameters in an IRT model are invariant, but also ***scale indeterminate***

- A scale must be chosen to identify an IRT model
  - That scale is only defined up to a linear transformation

- Choosing a mean of 0 and SD of 1 for θ identifies a scale for interpretation, and determines the scale of item parameters

- Any linear transformation of θ, with a corresponding transformation for item parameters, would provide the same ICCs

Parameter Invariance

- This assumption states that parameters are invariant ***up to a linear transformation***
  - Accounts for the arbitrariness of the scale chosen to identify the model

- Once the scale is chosen, this assumption can be tested

Ability Scale

- Because response probabilities (ICCs) are maintained through a linear transformation, the ability scale can be (and often is) transformed after calibration to create a more convenient scale for interpretation, usage, and score reporting

- Example: GRE ($\mu = 500$, $\sigma = 100$)

if $\theta_{new} = x\theta + y$

then $b_{new} = xb + y$

$$a_{new} = \frac{a}{x}$$

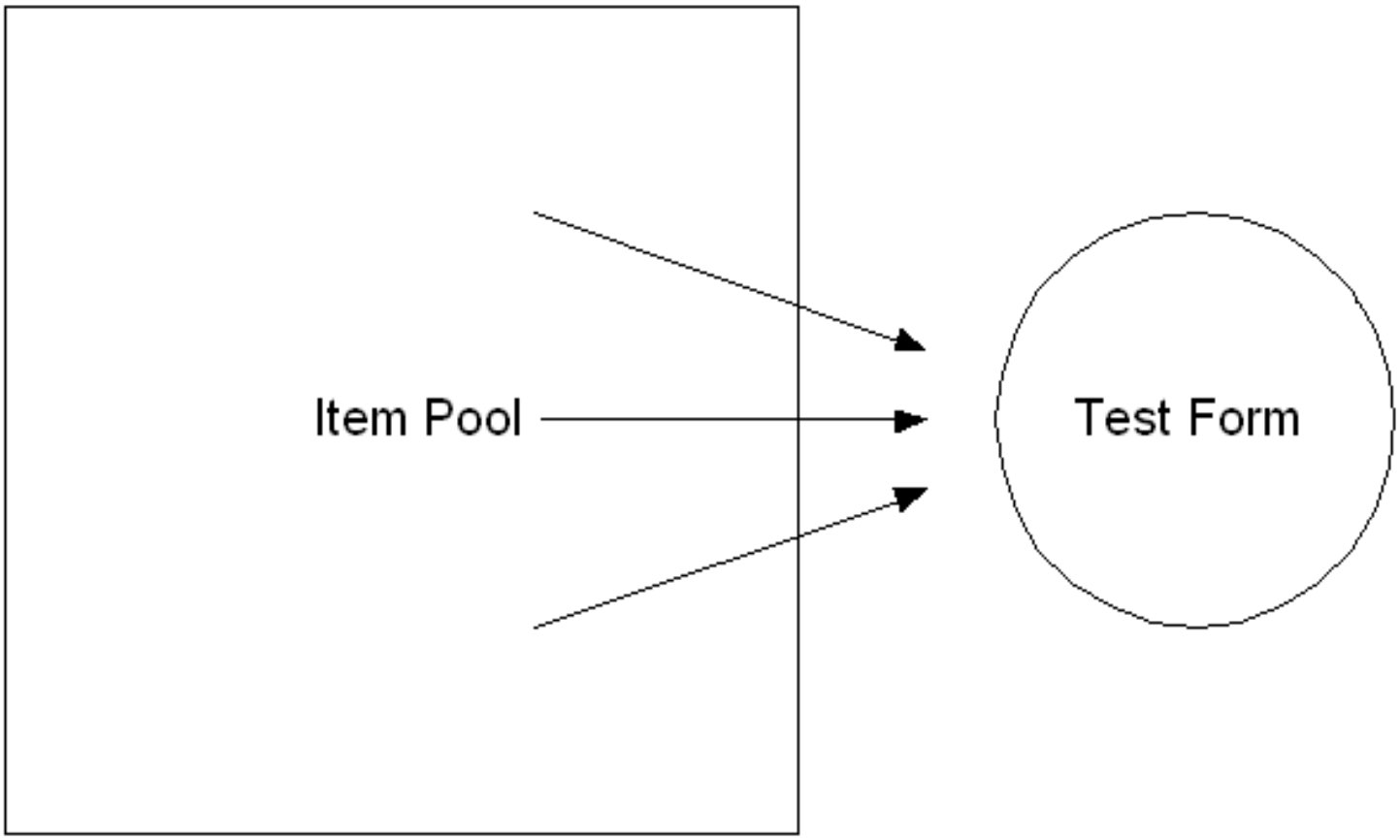
$$c_{new} = c$$

These transformation preserve the probability:

$$P(Y_{si} = 1|\theta_{new}) = P(Y_{is} = 1|\theta_s)$$

“Ability” Scores

- “Ability” is often the label used to describe what the test measures in educational contexts.
- A more general term would be “Trait” which would also encompass psychological (non-cognitive) measures
- The “Trait” or “Ability” is used to define what is being measured by the pool of items from which the test items were drawn



We can’t actually sample the entire universe of possible test items, so we are often interested in adding meaning the trait scale to get a better understanding of the construct being measured

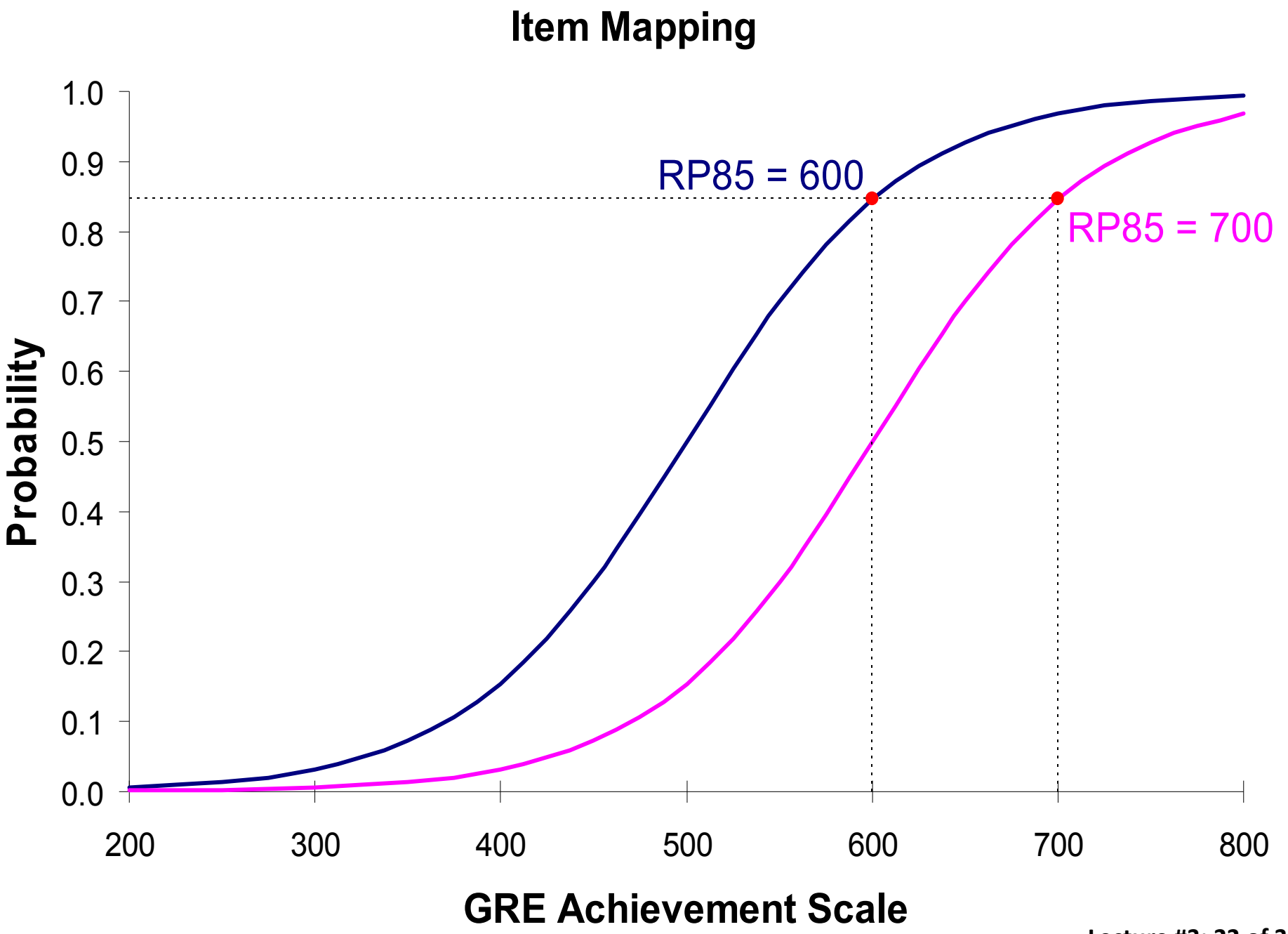
Adding Meaning to "Ability"

- IRT allows us to increase the meaning and interpretability of scaled scores through:
  - Item Mapping
    - Identifying ability levels that correspond to particular levels of item performance
  - Benchmarking
    - Determining "anchor points" that give meaning to the scale

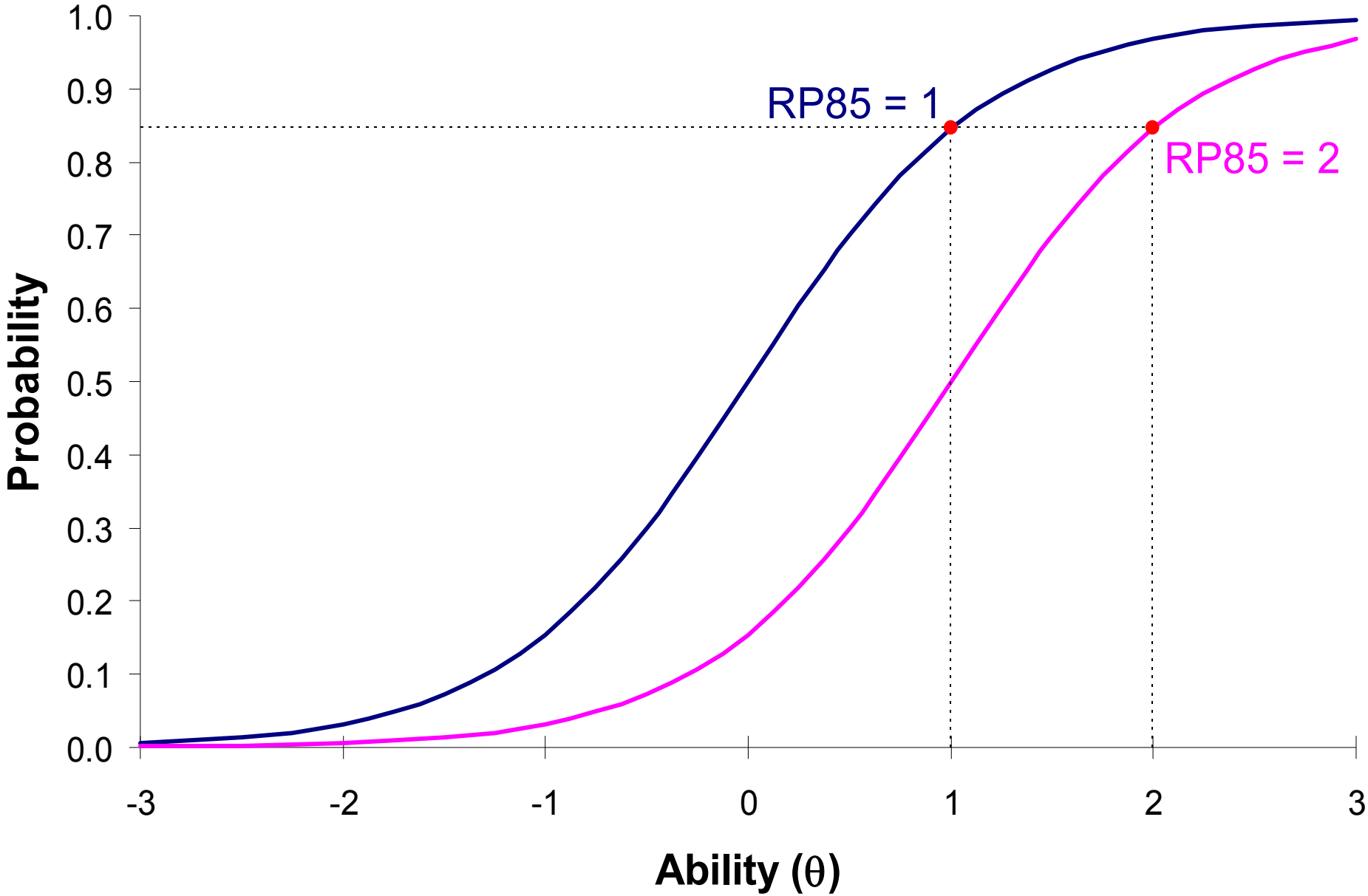
Item Mapping

- Determine a particular level of Response Probability (RP) that represents “mastery” and map the ability level that corresponds to this RP value for each item

- Examples: RP50, RP65, RP85


Item Mapping
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Probability
RP85 = 600
RP85 = 700
200
300
400
500
600
700
800
GRE Achievement Scale

Item Mapping



Item Mapping

- Through examination of the item itself, a test maker may then relate the particular RP value representing "mastery" to a given θ level

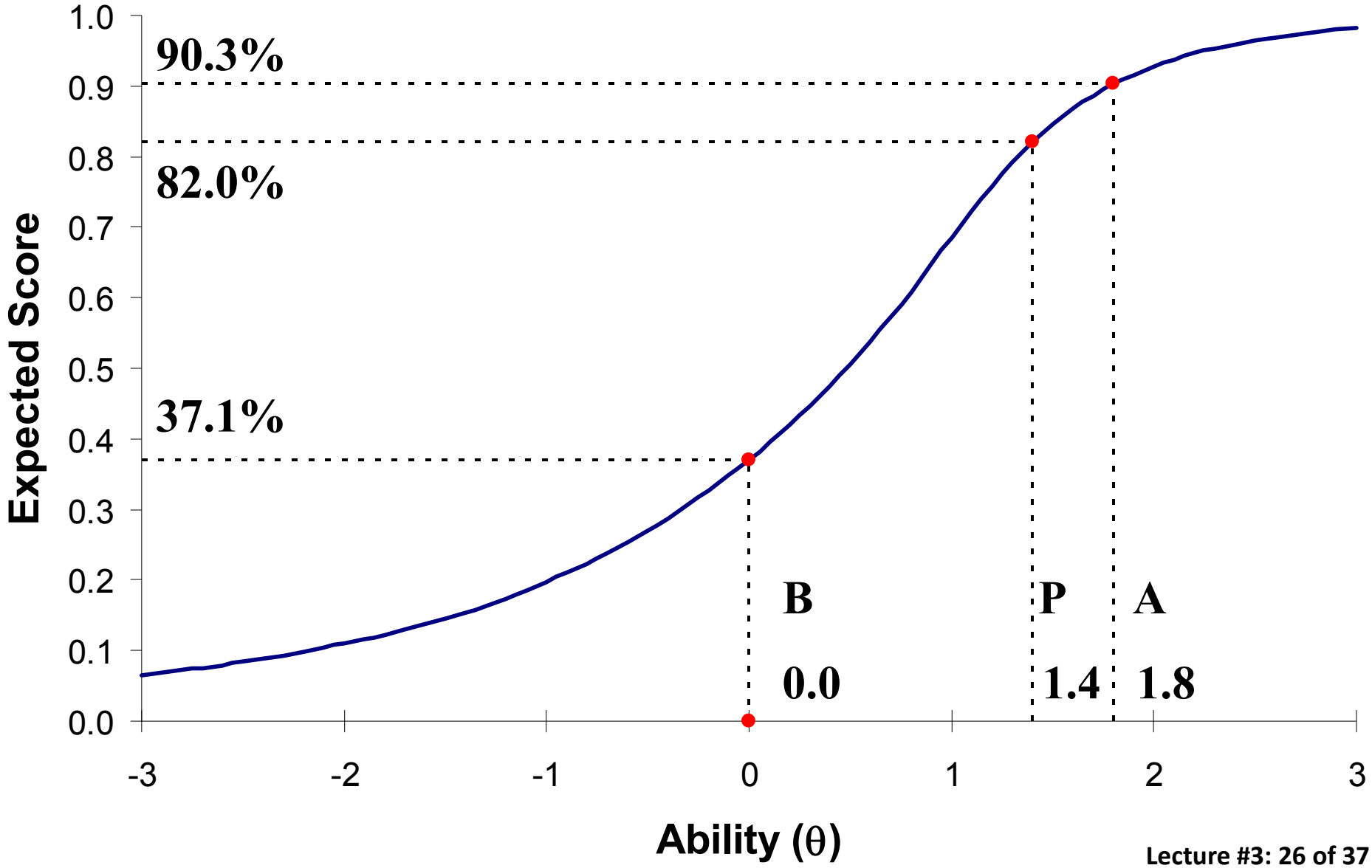
- "Here is the kind of item that someone with a 600 GRE score has mastered…"

Anchor Points

- Determine test score levels that correspond to meaningful categorizations of “Ability” (e.g., Basic, Proficient, Advanced)

- Other examples of Benchmarks:
  - Last year’s average score
  - Location of best or worst schools
  - Location of average student’s score

Anchor Points for a TCC



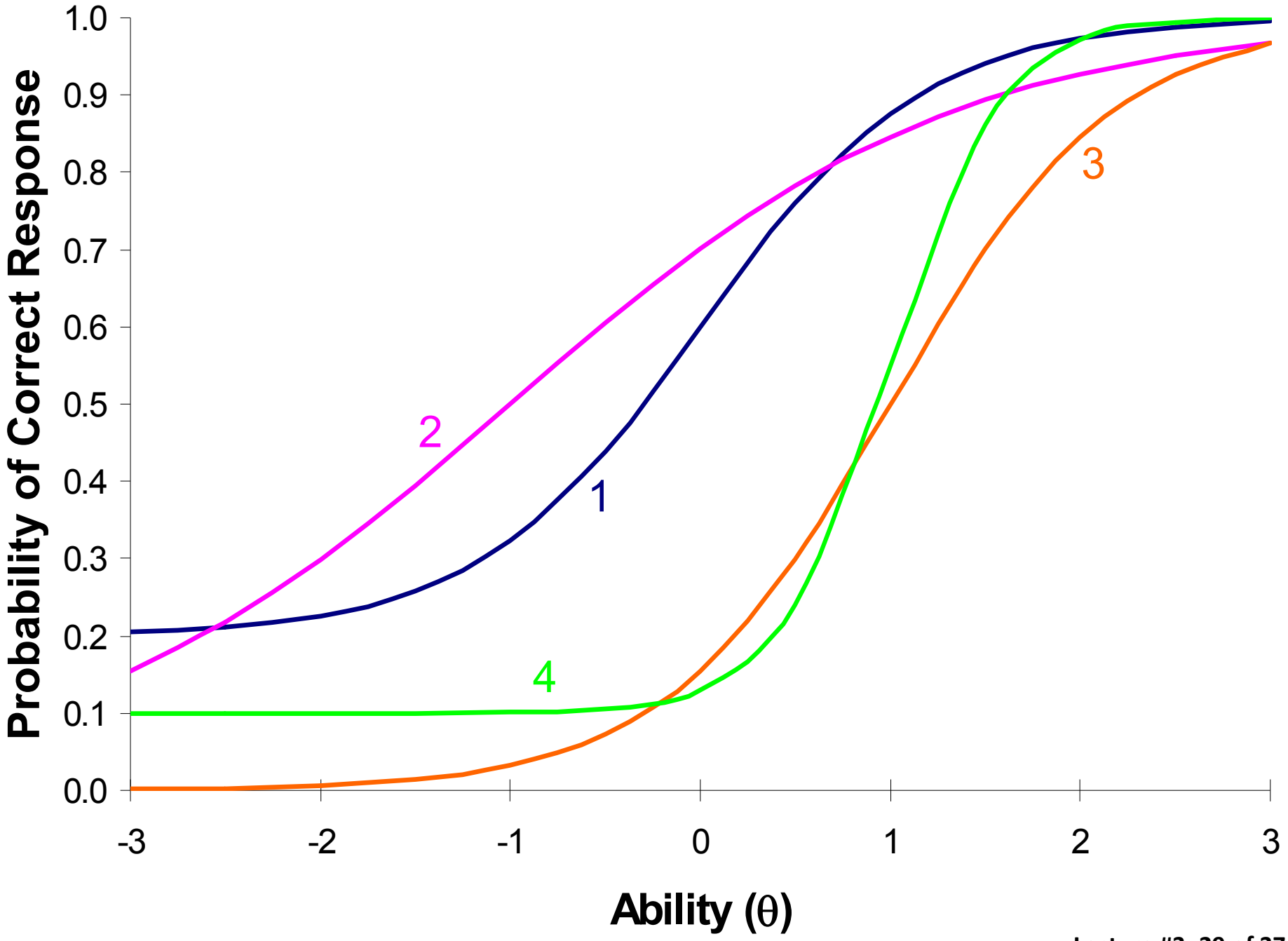
Excel Spreadsheet Demo

- Show Excel Spreadsheet containing four items, their ICCs, and the associated TCC

- Specify different item parameters and determine how changes affect the resulting ICCs

Example Items

Parameter	Item 1	Item 2	Item 3	Item 4
b	0.0	-1.0	1.0	1.0
a	1.0	0.5	1.0	2.0
c	0.2	0.0	0.0	0.1


Probability of Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Ability (θ)
1
2
3
4

Test Characteristic Curve

- A test characteristic curve (TCC) is created by summing each ICC across the ability continuum

- The vertical axis now reflects the expected score *on the test* for an examinee with a given ability level

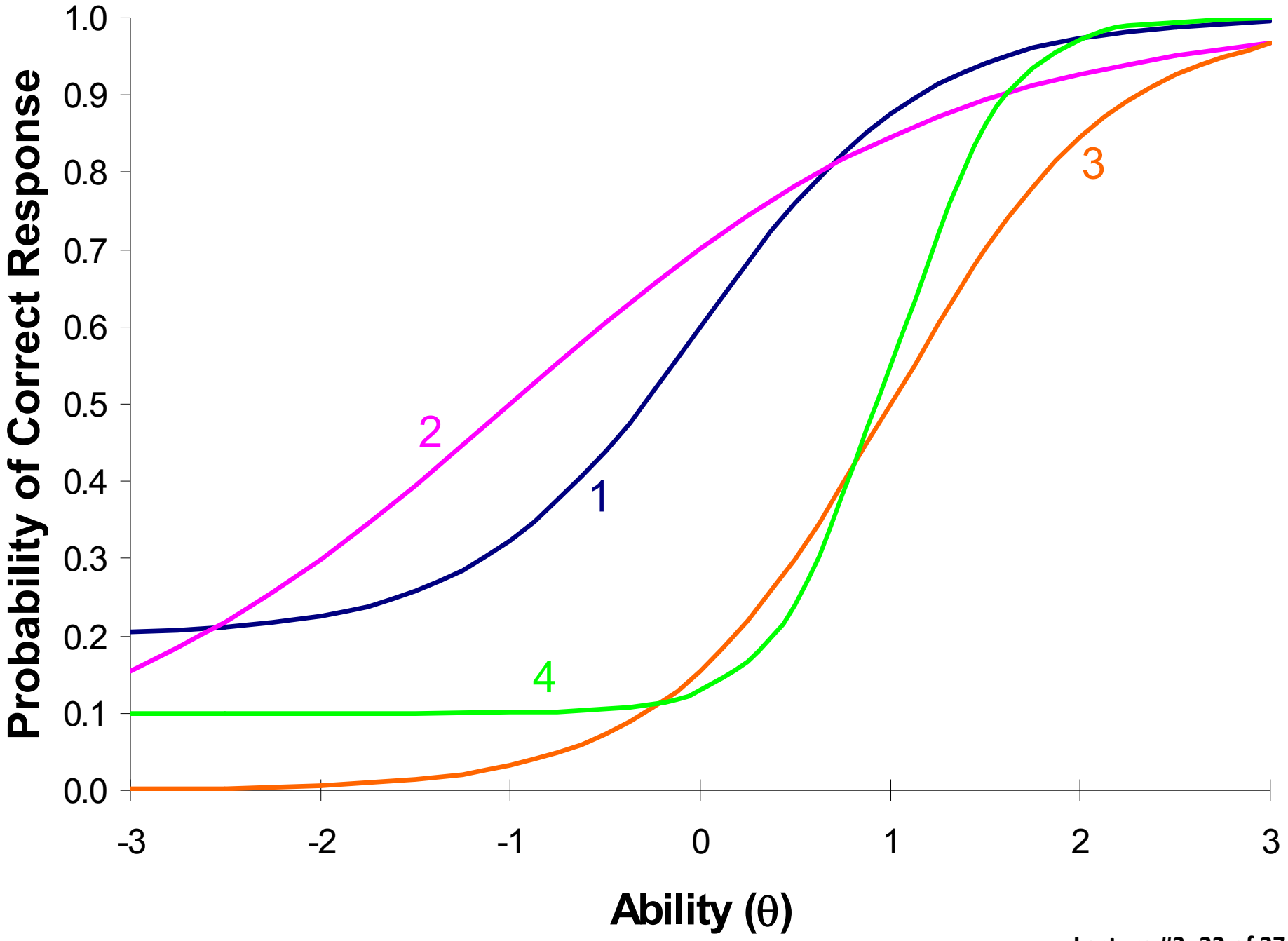
Test Characteristic Curve

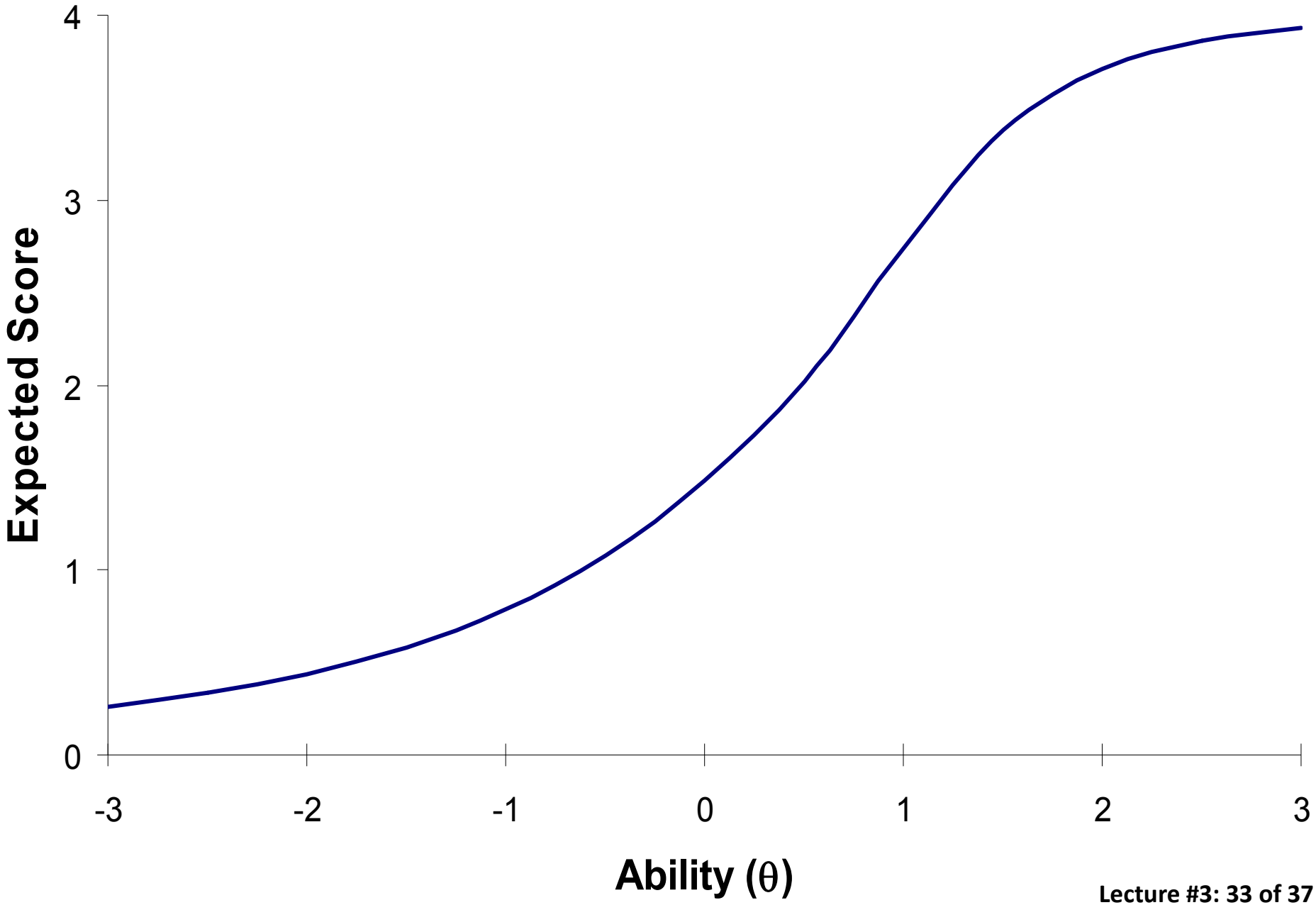
- A test characteristic curve (TCC) is created by summing each ICC across the ability continuum

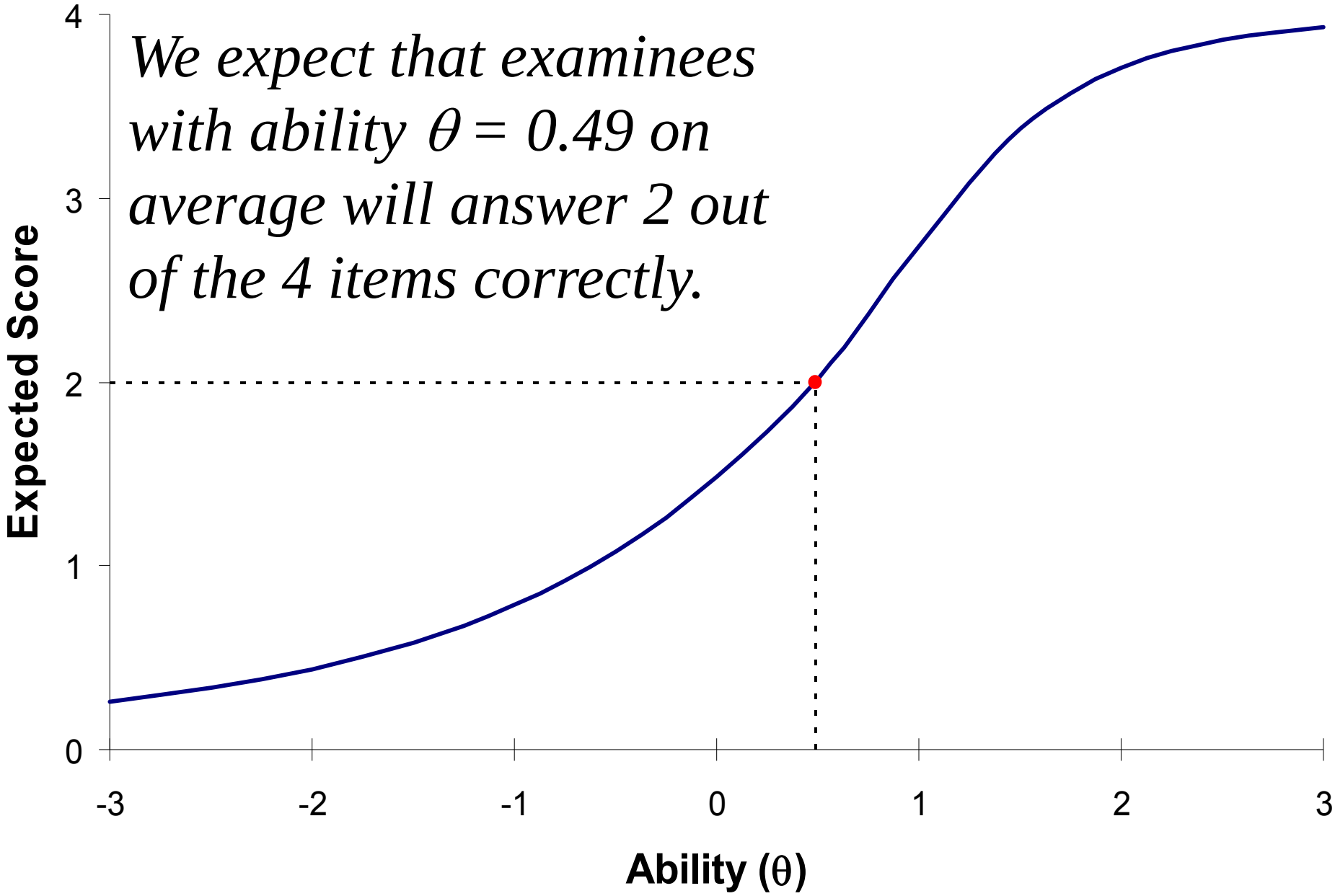
$$TCC(\theta_s) = \sum_{i=1}^{I} P(Y_{si} = 1|\theta_s)$$

- The vertical axis now reflects the expected score *on the test* for a subject with a given ability level

- Since $P(Y_{si} = 1|\theta_s)$ is the expected score for the item, the TCC is the expected score, E(Y), for the test
  - How many items we expect a subject with a particular ability level to answer correctly


Probability of Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Ability (θ)
1
2
3
4


Expected Score
4
3
2
1
0
-3
-2
-1
0
1
2
3
Ability (θ)


We expect that examinees with ability θ = 0.49 on average will answer 2 out of the 4 items correctly.
Expected Score
4
3
2
1
0
-3
-2
-1
0
1
2
3
Ability (θ)

CONCLUDING REMARKS

Wrapping Up

- Item response theory is a powerful method that can be used to build and assess scales
- The method is flexible and accommodate many types of testing items and situations
- Today was the introduction to IRT concepts
  - Over the rest of the week we will expand upon each of these

Next…Lab

- Today - computer time: Introduction to Mplus
  - 1PL and 2PL examples with syntax

- Tomorrow morning:
  - Polytomous Data
    - We will discuss what happens we have items scored with more than two categories
    - The IRT models used will be generalizations of the dichotomous models presented here