Item Response Theory

PSYC3302: Psychological Measurement and Its Applications

Mark Hurlstone
Univeristy of Western Australia

Week 11

Learning Objectives

- Item Response Theory (IRT)
- Factors affecting responses to test items:
  - trait level
  - item difficulty
  - item discrimination
  - guessing
- IRT Models:
  - one-, two-, and three-parameter logistic models
- Item characteristic curves
- IRT and reliability
- Applications of IRT

Classical Test Theory

- So far in this unit, the content has focussed upon Classical Test Theory (CTT)
- CTT incorporates terms such as "observed scores" and "true scores"
- There is a substantial emphasis on the description and estimation of the reliability of scores
- Additionally, observed scores are considered a function of the sum of true scores and error scores

CTT

- In CTT, a person's observed score on a test is that person's true score, plus error, which can be expressed as:

$$X_o = X_t + X_e \tag{10}$$

- Where X_o represents a person's observed score, X_t represents a person's true score, and X_e represents error

Item Response Theory

- **Item Response Theory (IRT)** is a contemporary alternative to CTT
- It has emerged recently as an alternative approach to measurement in the behavioural sciences
- IRT is more complex than CTT, but its proponents suggest that this complexity is offset by several important advantages
- In this lecture, I will outline the conceptual basis of IRT

IRT

- According to IRT a person's response to a particular test item is influenced by two factors:
  1. qualities of the individual
  2. qualities of the item
- There are three well-established models of IRT
- In the most basic model of IRT the only item characteristic taken into consideration is:
  - item difficulty (the probability a person will answer a question correctly)

Conceptual Example

- Suppose a person takes a five-item test of mathematical ability
- According to the most basic model of IRT, the likelihood the person will respond correctly to any given item on the 5-item test will be affected by two things:
  1. the person's level of mathematical ability
  2. the item's difficulty

IRT

- Thus, in the most basic IRT model a person's response to an item is influenced by the individual's trait level (e.g., level of mathematical ability) and the item's difficulty level
- More complex forms of IRT include additional factors (or parameters) affecting a person's responses to items:
  - item discrimination
  - guessing

IRT and Self-Report Questionnaires

- IRT is typically used in the context of intelligence testing or achievement testing
  - where questions can be answered correctly or not
- However, the basic IRT model can be extended to personality type questions
- The principles are effectively the same:
  - how much of the trait does the person possess?
  - how likely is it that someone would endorse or agree with the item?

Factors Affecting Responses to Test Items

- Trait level
- Item difficulty
- Item discrimination
- Guessing

Factors Affecting Responses to Test Items

- Trait level
- Item difficulty
- Item discrimination
- Guessing

Trait Level

- One factor affecting a person's probability of responding in a particular way to an item is the individual's **trait level**
- This is the person's level on the psychological trait being measured by the item
- For example, a person with a high level of mathematical ability will be more likely to respond correctly to a math item than a person with a low level of mathematical ability
- Similarly, a person with a high level of extraversion will be more likely to endorse or agree with an item that measures extraversion than will a person with a low level of extraversion

Factors Affecting Responses to Test Items

- Trait level
- Item difficulty
- Item discrimination
- Guessing

Factors Affecting Responses to Test Items

- Trait level
- Item difficulty
- Item discrimination
- Guessing

Item Difficulty

- Another factor affecting a person's probability of responding in a particular way is **item difficulty**
- Arithmetic example:
  - What is 2+2? (high probability)
  - What is the square root of 10,000? (low probability)
- Extraversion example:
  - "I enjoy socialising with groups of people" (high probability)
  - "I enjoy speaking before large audiences" (low probability)
- Job satisfaction example:
  - "My job is okay" (high probability)
  - "My job is the best thing in my life" (low probability)

Trait Level and Item Difficulty

- Although they are separate issues in an IRT analysis, trait level and item difficulty are intrinsically connected concepts
- In fact, item difficulty is conceived in terms of trait level:
  - a difficult item requires a relatively high trait level to be answered correctly
  - an easy item requires only a low trait level to be answered correctly
- For example (Arithmetic):
  - What is 2+2? (second grade mathematical ability)
  - What is the square root of 10,000? (ninth grade mathematical ability)

IRT Metric

- In an IRT analysis, trait levels and item difficulties are usually scored on a standardised metric:
  - Mean = 0
  - Standard deviation = 1
- Thus, a person who has a trait level of 0 has an average level of that trait
  - a person who has a trait level of 1.5 has a trait level that is 1.5 standard deviations above the mean
- Similarly, an item with a difficulty level of 0 is an average item
  - an item with a difficulty level of 1.5 is a relatively difficult item

Item Difficulty

- An item's difficulty is defined as: *the trait level required for participants to have a .50 probability of answering the item correctly*
- Thus, if an item has a difficulty level of 0, then a person with an average trait level (i.e., a person with a trait level of 0) will have a 50% chance of responding to the item correctly
- If an item has a difficulty level of 1.5, then it will take a trait level of 1.5 to have a 50% chance of responding to the item correctly
- If an item had a difficulty level of –1.5, then a person with a trait level of 1.0 would have a much greater than 50% chance of answering the question correctly

Factors Affecting Responses to Test Items

- Trait level
- Item difficulty
- Item discrimination
- Guessing

Factors Affecting Responses to Test Items

- Trait level
- Item difficulty
- Item discrimination
- Guessing

Item Discrimination

- **Item discrimination** refers to the degree to which items on a test can differentiate individuals who have high trait levels from individuals who have low trait levels
- An item's discrimination value indicates the relevance of the item to the trait being measured by the test:
  - An item with a large and positive discrimination value (e.g., 3.5) is highly consistent with the underlying trait
  - An item with a discrimination value of 0 is unrelated to the underlying trait
  - An item with a negative discrimination value is inversely related to the underlying trait
- It is preferable for items to have a large positive discrimination value

Item Discrimination

- Why would some items have good discrimination and others have poor discrimination?
- Consider the following two items that might be written for a mathematics test:
  - How many pecks are in three bushels? (a) 12 (b) 24
  - What is 10 times 10? (a) 10 (b) 100
- Both items require the ability to perform multiplication
- However, the first item also requires knowledge of the number of pecks in a bushel—construct irrelevant content
- Thus, this item would likely have a low discrimination value, as it is only weakly related to the underlying trait being measured

Factors Affecting Responses to Test Items

- Trait level
- Item difficulty
- Item discrimination
- Guessing

Factors Affecting Responses to Test Items

- Trait level
- Item difficulty
- Item discrimination
- Guessing

Guessing

- A third item property that might affect participant's responses to some types of test items is **guessing**
- IRT models can include a guessing component to account for this possibility
- It reflects the probability that participants will answer an item correctly purely on the basis of chance (e.g., 50% for true/false items)
- This property is mainly relevant for items that are scored as correct or incorrect (tests of knowledge, skill, ability, or achievement)
- It is less relevant for tests of attitudes or personality

IRT Models

- A variety of models have been developed from the IRT perspective
- The main way the models differ from each other is with respect to the nature and number of item **parameters** they include
- There are only three commonly used IRT models:
  1. one-parameter logistic model
  2. two-parameter logistic model
  3. three-parameter logistic model
- All of these models are designed for items with binary outcomes (i.e., right/wrong, true/false) as the response option

One-Parameter Logistic Model

- The simplest IRT model is often called the **one-parameter logistic model (1PL)** or **Rasch model**
- A person's response to a binary item is determined by the individual's trait level and only a single item characteristic or parameter:
  1. the difficulty of the item

One-Parameter Logistic Model

- The Rasch model is the probability that a person of a specific trait level will correctly answer an item of a given difficulty:

$$P(X_{is} = 1|\theta_s, \beta_i) = \frac{e^{(\theta_s - \beta_i)}}{1 + e^{(\theta_s - \beta_i)}}, \tag{23}$$

- P is the "conditional probability"
- X_{is} refers to a particular response (X) made by subject s to item i (X_{is} = 1 refers to a "correct" response or an endorsement of the item)
- θ_s is the trait level of subject s
- β_i is the difficulty of item i
- e is the exponential constant 2.7182818 ...

One-Parameter Logistic Model

- What is the probability a person with above-average math ability (θ_s = 1) will correctly answer an item with a low level of difficulty (β_i = –.5)?

$$P(X_{is} = 1|1, -.5) = \frac{e^{(1-(-.5))}}{1 + e^{(1-(-.5))}} = \frac{e^{(1.5)}}{1 + e^{(1.5)}} = \frac{4.48}{1 + 4.48} = .82$$

- Thus, there is a .8 probability that the person will answer the item correctly
- This makes sense, because the individual's trait level is markedly higher than the item's difficulty level

One-Parameter Logistic Model

- What is the probability a person with below-average math ability (θ_s = –1.39) will correctly answer an item with a low level of difficulty (β_i = –1.61)?

$$P(X_{is} = 1| -1.39, -1.61) = \frac{e^{(-1.39-(-1.61))}}{1 + e^{(-1.39-(-1.61))}} = \frac{e^{(.22)}}{1 + e^{(.22)}}$$

$$= \frac{1.25}{1 + 1.25} = .56$$

- Thus, there is a .56 probability that the person will answer the item correctly
- This makes sense, because the individual's trait level is only slightly higher than the item's difficulty level

Two-Parameter Logistic Model

- A slightly more complex IRT model is called the **two-parameter logistic model (2PL)**
- In addition to an individual's trait level, a person's response to a binary item is influenced by two item parameters:
  1. the difficulty of the item
  2. the discrimination of the item
- The difference between the one and two-parameter models is the latter includes item discrimination information
- Not surprisingly, the two-parameter model is much more useful than the one-parameter model
- It is probably the most commonly used IRT model

Two-Parameter Logistic Model

- The two-parameter logistic model can be expressed as:

$$P(X_{is} = 1|\theta_s, \beta_i, \alpha_i) = \frac{e^{(\alpha_i(\theta_s - \beta_i))}}{1 + e^{(\alpha_i(\theta_s - \beta_i))}}, \tag{24}$$

- α_i is the discrimination of item i

Two-Parameter Logistic Model

- Suppose two items have the equal difficulty ($\beta = -.5$), but one has a low discrimination value ($\alpha_1 = .5$) and the other has a high discrimination value ($\alpha_2 = 2$)
- What is the probability that a person with an above-average trait level ($\theta = 1$) will correctly answer item 1?

$$P(X_{is} = 1|1, -.5, .5) = \frac{e^{(.5(1-(-.5))}}{1 + e^{(.5(1-(-.5))}} = \frac{e^{(.75)}}{1 + e^{(.75)}} = \frac{2.12}{1 + 2.12} = .68$$

- What about a person with an average trait level ($\theta = 0$)?

$$P(X_{is} = 1|0, -.5, .5) = \frac{e^{(.5(0-(-.5))}}{1 + e^{(.5(0-(-.5))}} = \frac{e^{(.25)}}{1 + e^{(.25)}} = \frac{1.28}{1 + 1.28} = .56$$

Two-Parameter Logistic Model

- Now consider the second item with a high discrimination value ($\alpha_2 = 2$)
- What is the probability that a person with an above-average trait level ($\theta = 1$) will correctly answer item 2?

$$P(X_{is} = 1|1, -.5, 2) = \frac{e^{(2(1-(-.5))}}{1 + e^{(2(1-(-.5))}} = \frac{e^{(3)}}{1 + e^{(3)}} = \frac{20.09}{1 + 20.09} = .95$$

- What about a person with an average trait level ($\theta = 0$)?

$$P(X_{is} = 1|0, -.5, 2) = \frac{e^{(2(0-(-.5))}}{1 + e^{(2(0-(-.5))}} = \frac{e^{(1)}}{1 + e^{(1)}} = \frac{2.72}{1 + 2.72} = .73$$

Three-Parameter Logistic Model

- The **three-parameter logistic model (3PL)** adds yet another item parameter
- In addition to an individual's trait level, it includes three item parameters:
  1. the difficulty of the item
  2. the discrimination of the item
  3. the probability with which the question can be answered by guessing
- It can be useful in multiple-choice tests
- It is not commonly used, perhaps because there is rarely much benefit to including this third parameter

Binary and Polytomous Items

- IRT is typically discussed in the context of binary items
- However, there are IRT models which can be applied to **polytomous items**:
  - Graded Response Model
  - Partial Credit Model
  - Nominal Response Model
- These are items with three or more response options (e.g., *strongly disagree, disagree, neutral, agree, strongly agree*)
- Use same general principles as binary response models
- Relevant for modelling inventories such as personality questionnaires

Getting Parameter Estimates: A 1PL Example

Table: IRT Example: Item Difficulty Estimates and Person Trait-Level Estimates

ID	Item 1	Item 2	Item 3	Item 4	Item 5	p(Correct)
1	1	0	0	0	0	0.20
2	1	1	0	1	0	0.60
3	1	1	1	0	0	0.60
4	1	1	0	1	0	0.60
5	1	1	1	0	1	0.80
6	0	0	1	0	0	0.20
p(Correct)	0.83	0.67	0.50	0.33	0.17	

Getting Parameter Estimates: A 1PL Example

- To obtain estimates of respondents trait levels, we can use the following formula:

$$\theta_s = \ln\left(\frac{P_s}{1 - P_s}\right), \tag{25}$$

- Where P_s is the proportion of correct responses for respondent s
- Suppose we want to estimate θ for respondent 5 (p(correct) = 0.80):

$$\theta_5 = \ln\left(\frac{.80}{1 - .80}\right) = \ln(4) = 1.39$$

Getting Parameter Estimates: A 1PL Example

- To obtain estimates of item difficulty levels, we can use the following formula:

$$\beta_i = \ln\left(\frac{1 - P_i}{P_i}\right), \tag{26}$$

- Where P_i is the proportion of correct responses for item i
- Suppose we want to estimate β for item 1 (p(correct) = 0.83):

$$\beta_i = \ln\left(\frac{1 - .83}{.83}\right) = \ln(.20) = -1.61$$

Getting Parameter Estimates: A 1PL Example

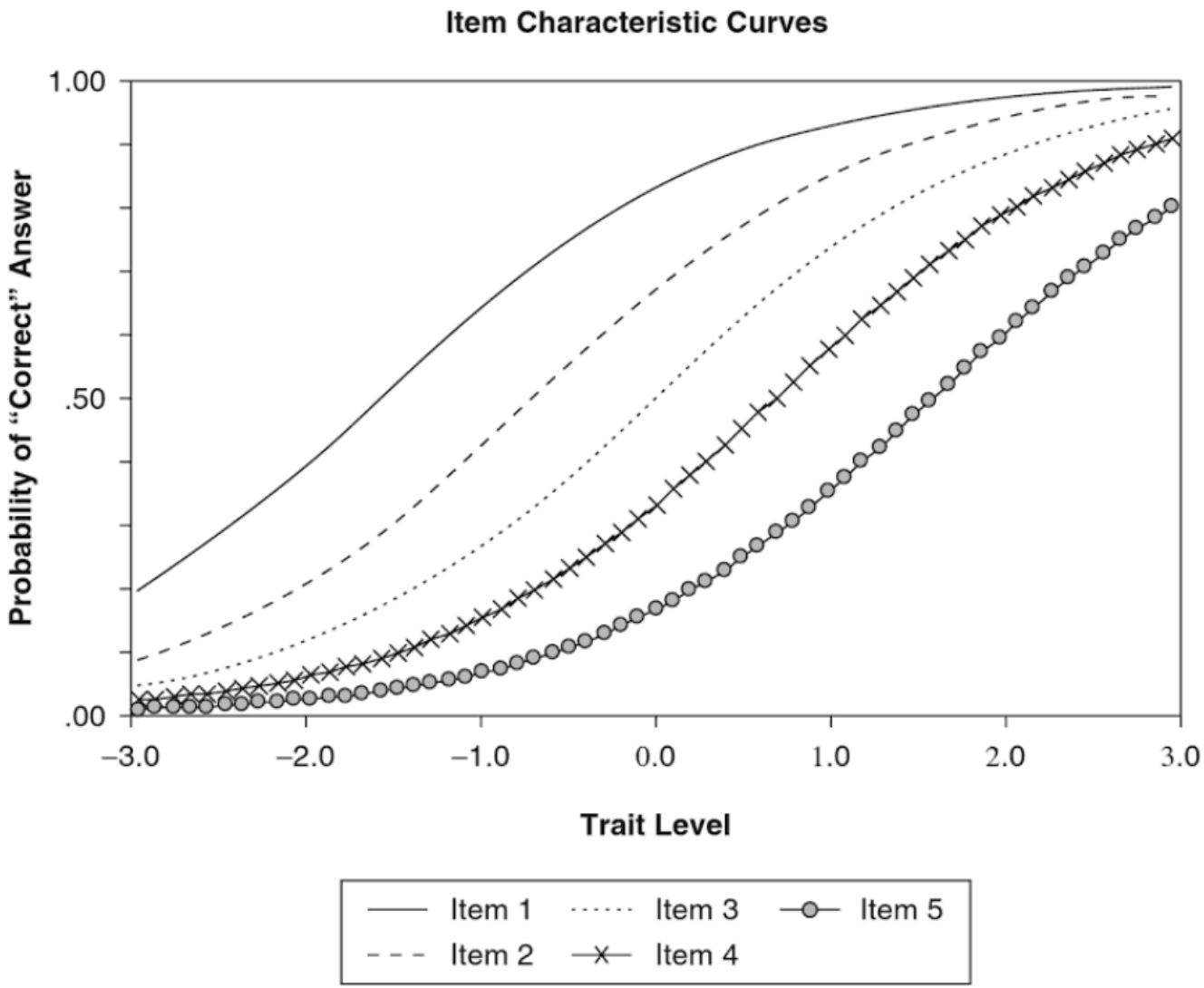
Table: IRT Example: Item Difficulty Estimates and Person Trait-Level Estimates

ID	Item 1	Item 2	Item 3	Item 4	Item 5	p(Correct)	Trait Level (θ)
1	1	0	0	0	0	0.20	–1.39
2	1	1	0	1	0	0.60	0.41
3	1	1	1	0	0	0.60	0.41
4	1	1	0	1	0	0.60	0.41
5	1	1	1	0	1	0.80	1.39
6	0	0	1	0	0	0.20	–1.39
p(Correct)	0.83	0.67	0.50	0.33	0.17		
Difficulty (β)	–1.61	–0.69	0.00	0.69	1.61		

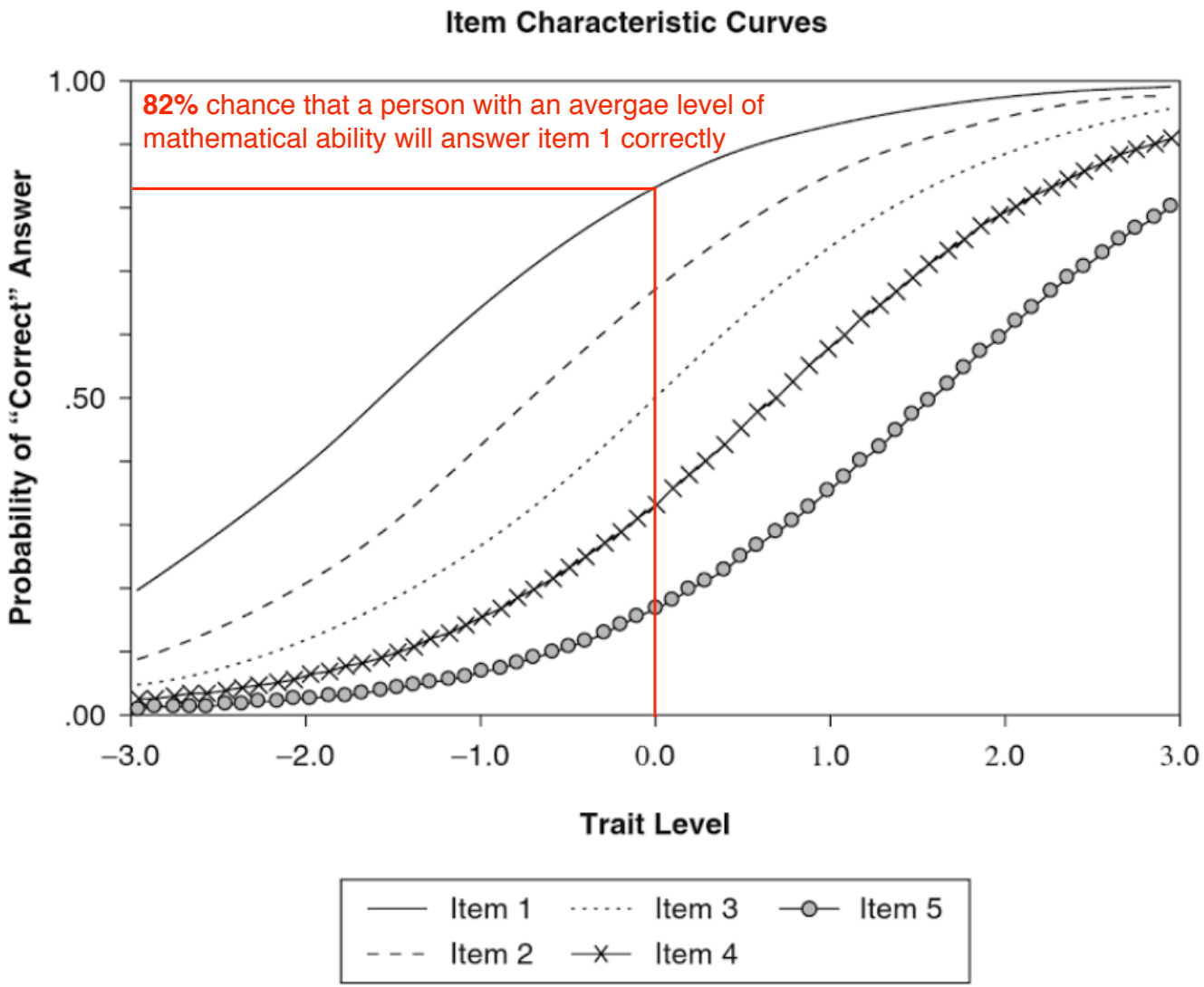
Item Characteristic Curves

- Psychometricians who use IRT often evaluate the characteristics of the items on a test using a graph known as an **item characteristic curve (ICC)**
- An ICC reflects the probabilities with which individuals across a range of trait levels are likely to answer each item correctly
- The logistic formula and the parameters included in the model are used to predict the probabilities
    - much like we could use a regression equation to predict a person's Y score from X
- In an ICC, the x-axis reflects a wide range of trait levels, and the y-axis reflects probability ranging from 0 to 1

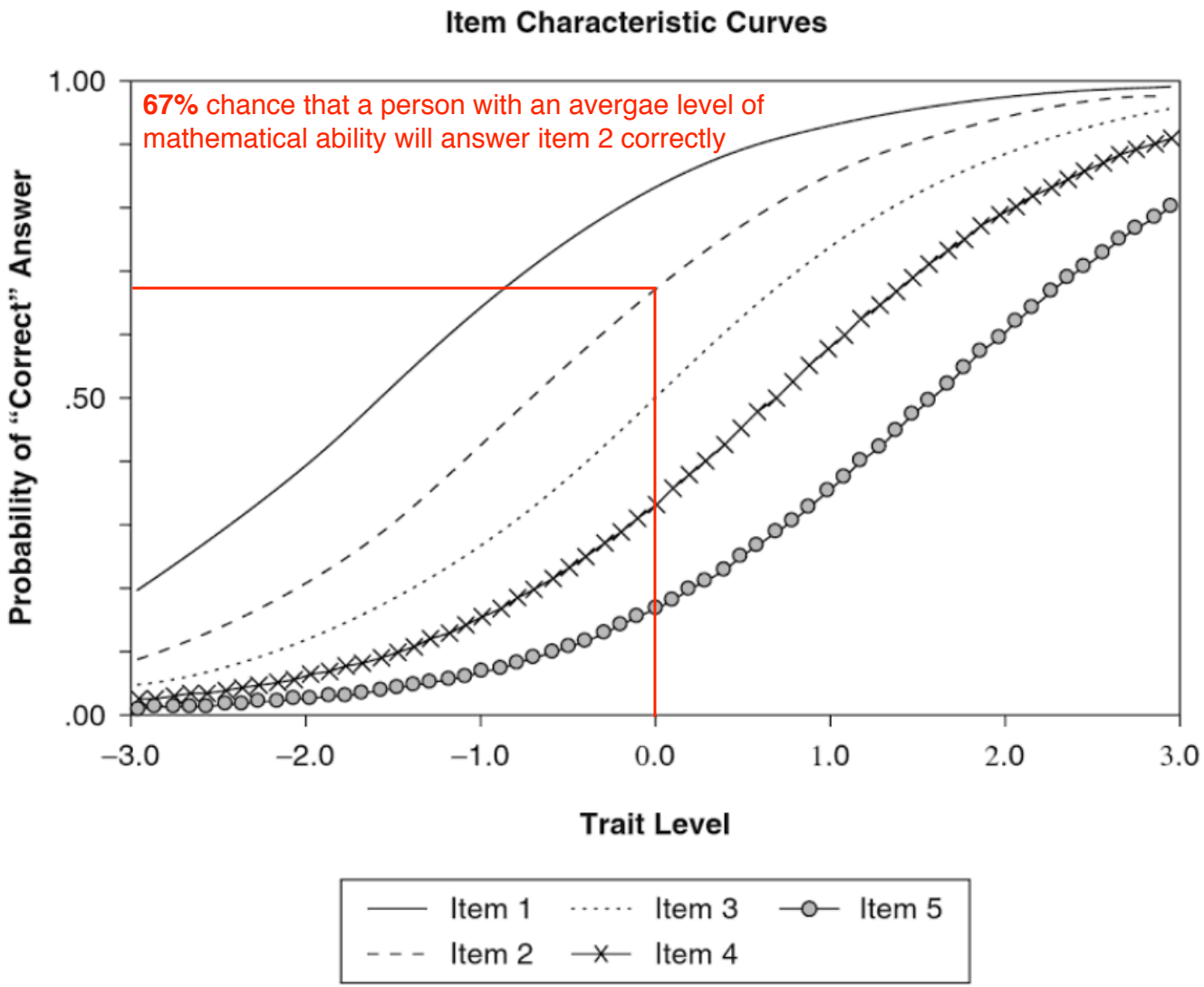
Item Characteristic Curves



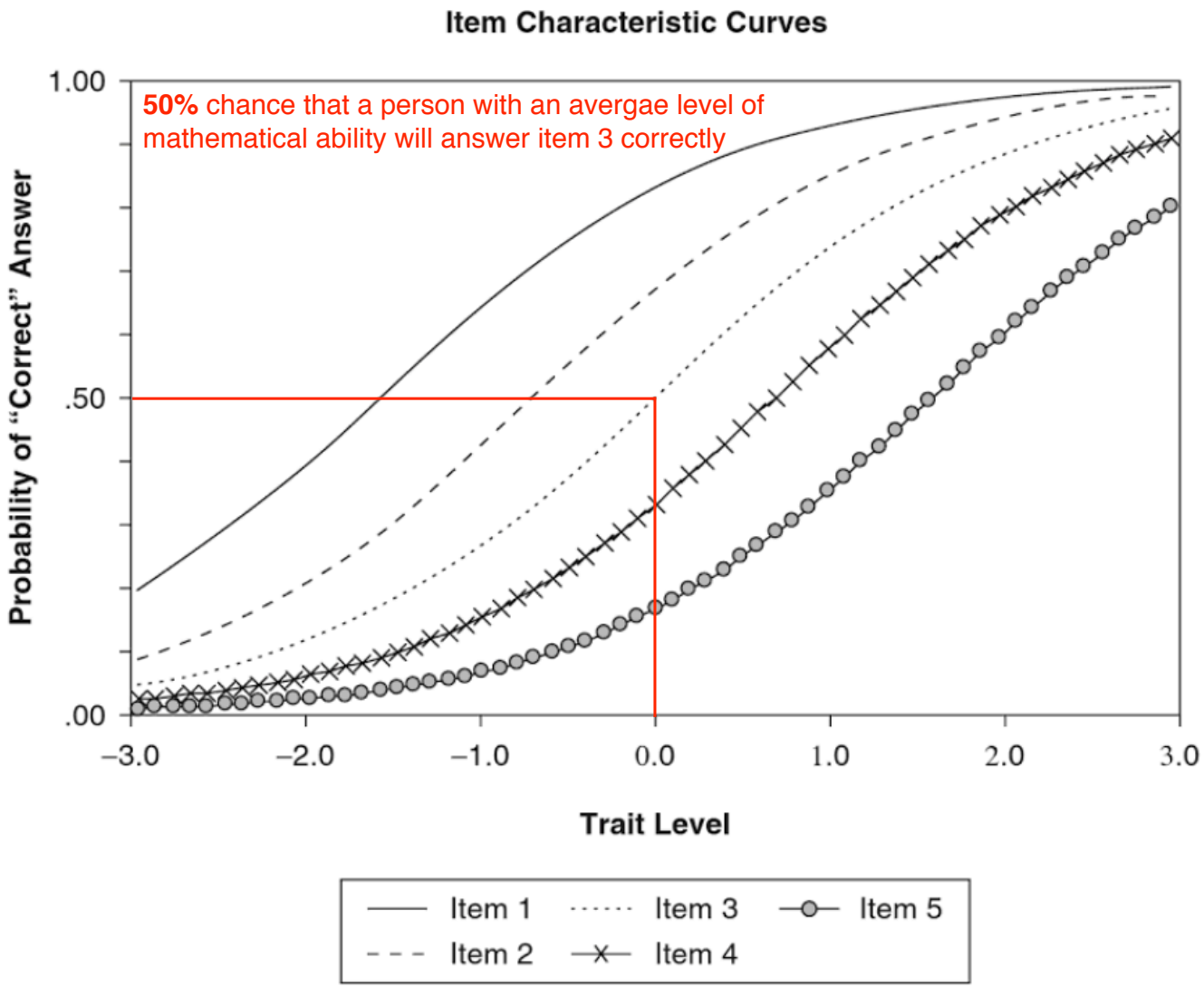
Item Characteristic Curves



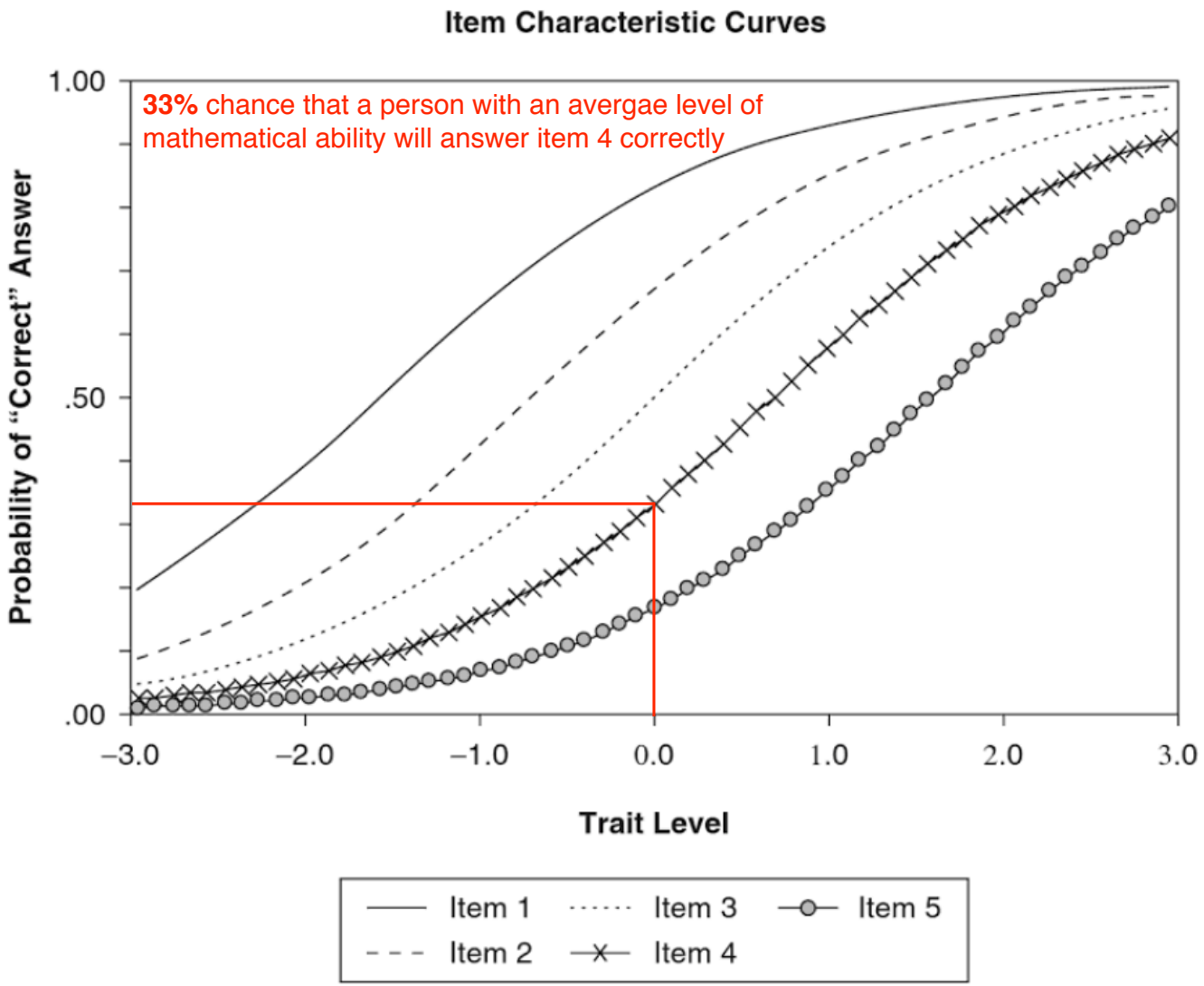
Item Characteristic Curves



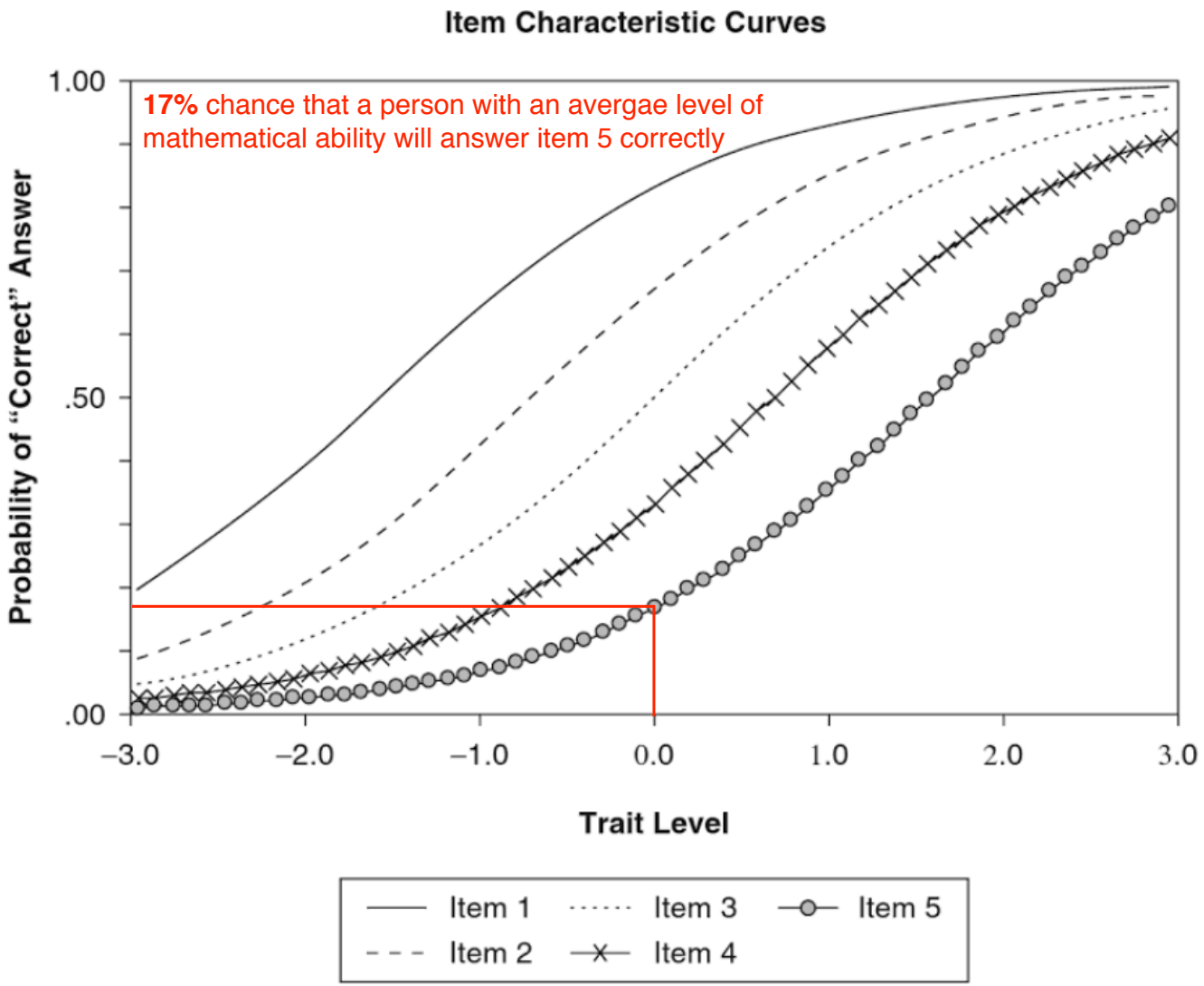
Item Characteristic Curves



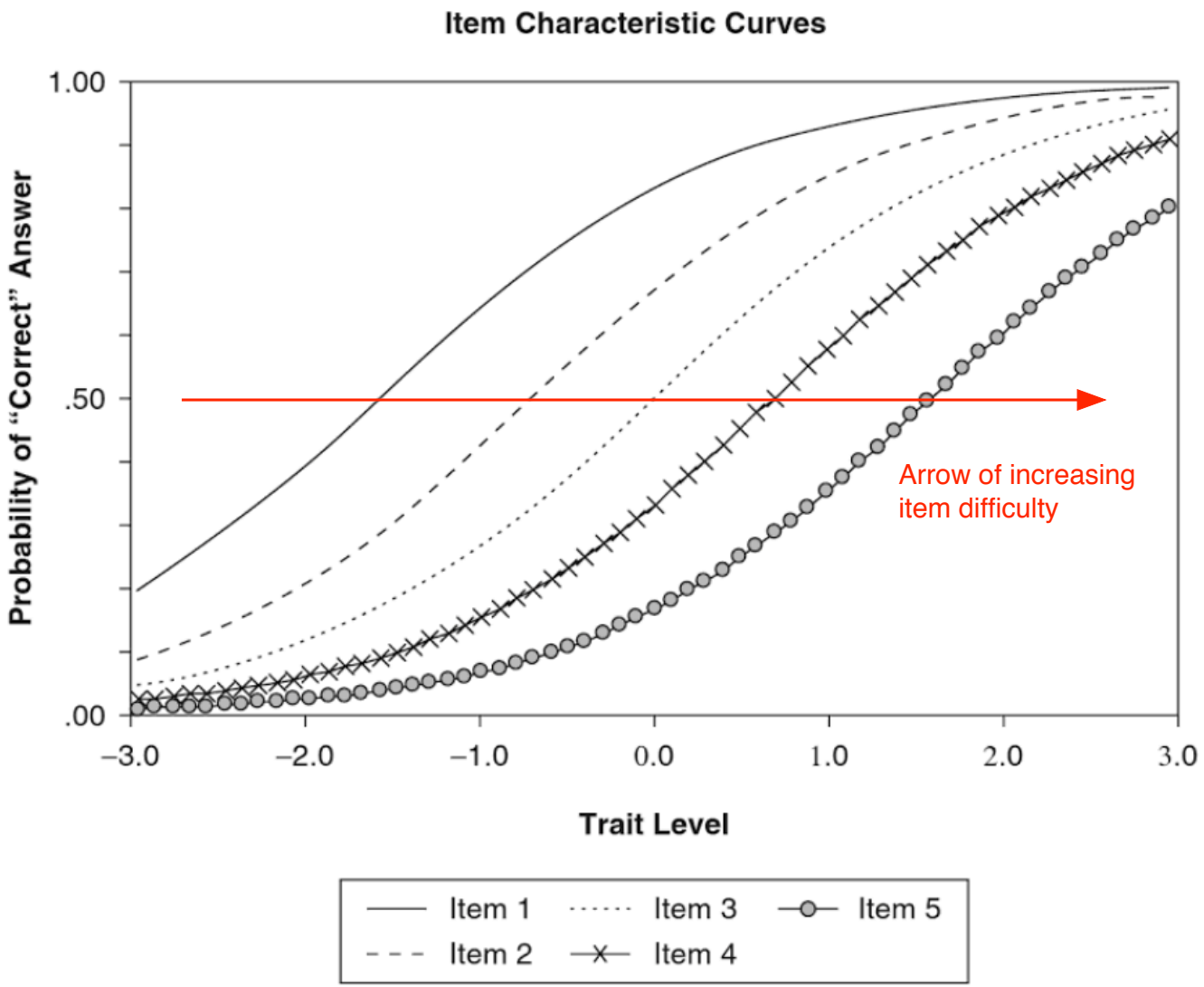
Item Characteristic Curves



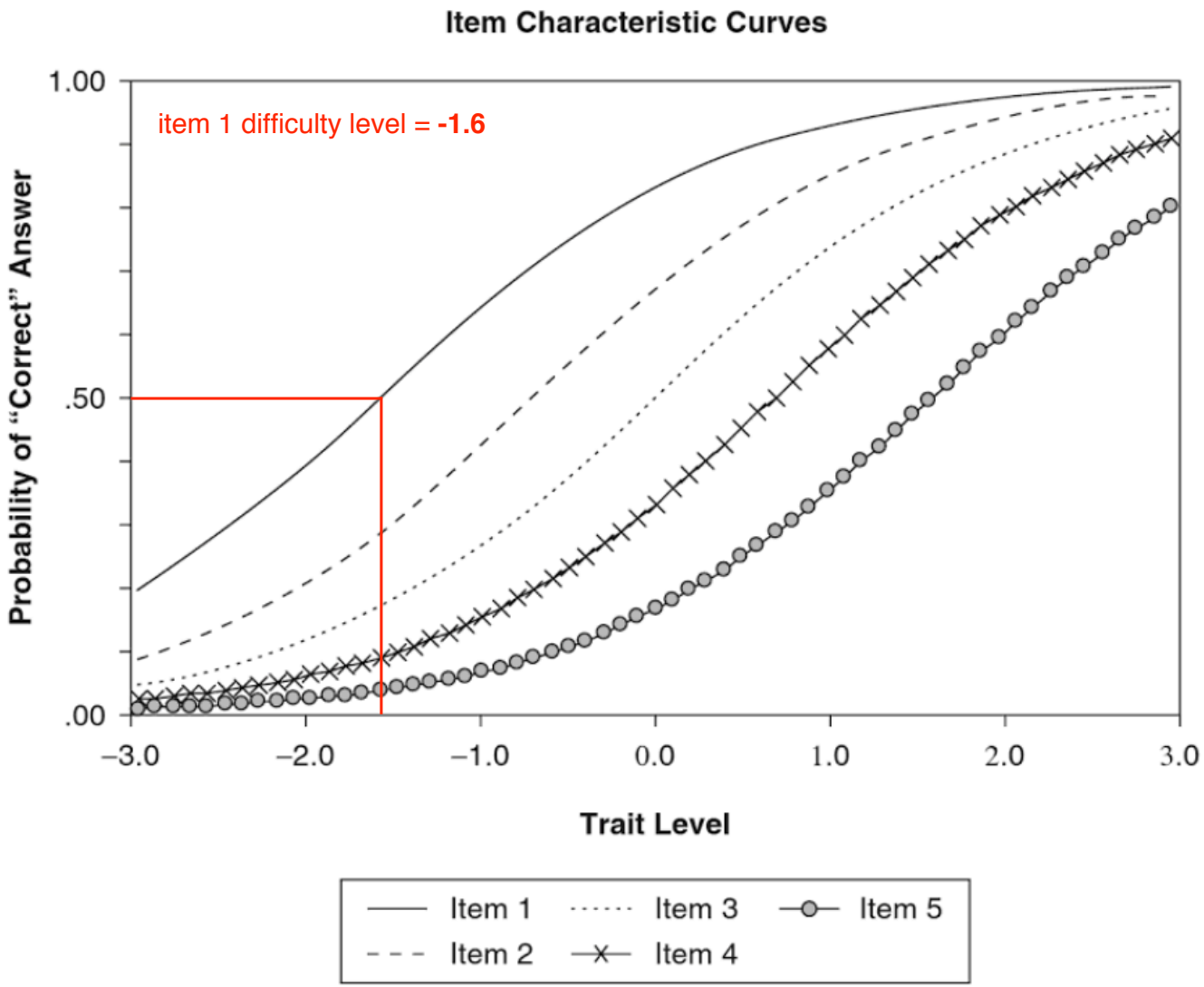
Item Characteristic Curves



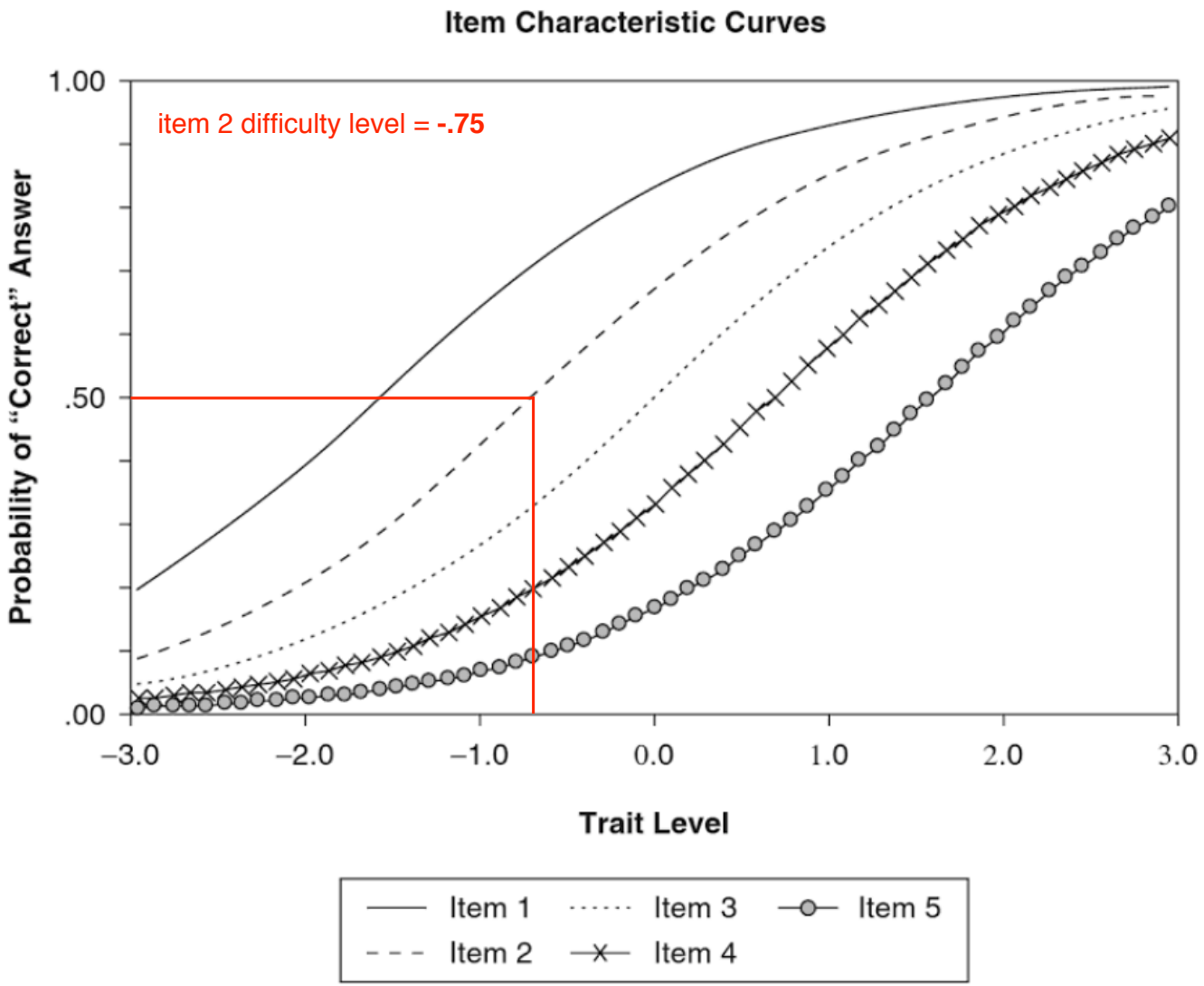
Item Characteristic Curves



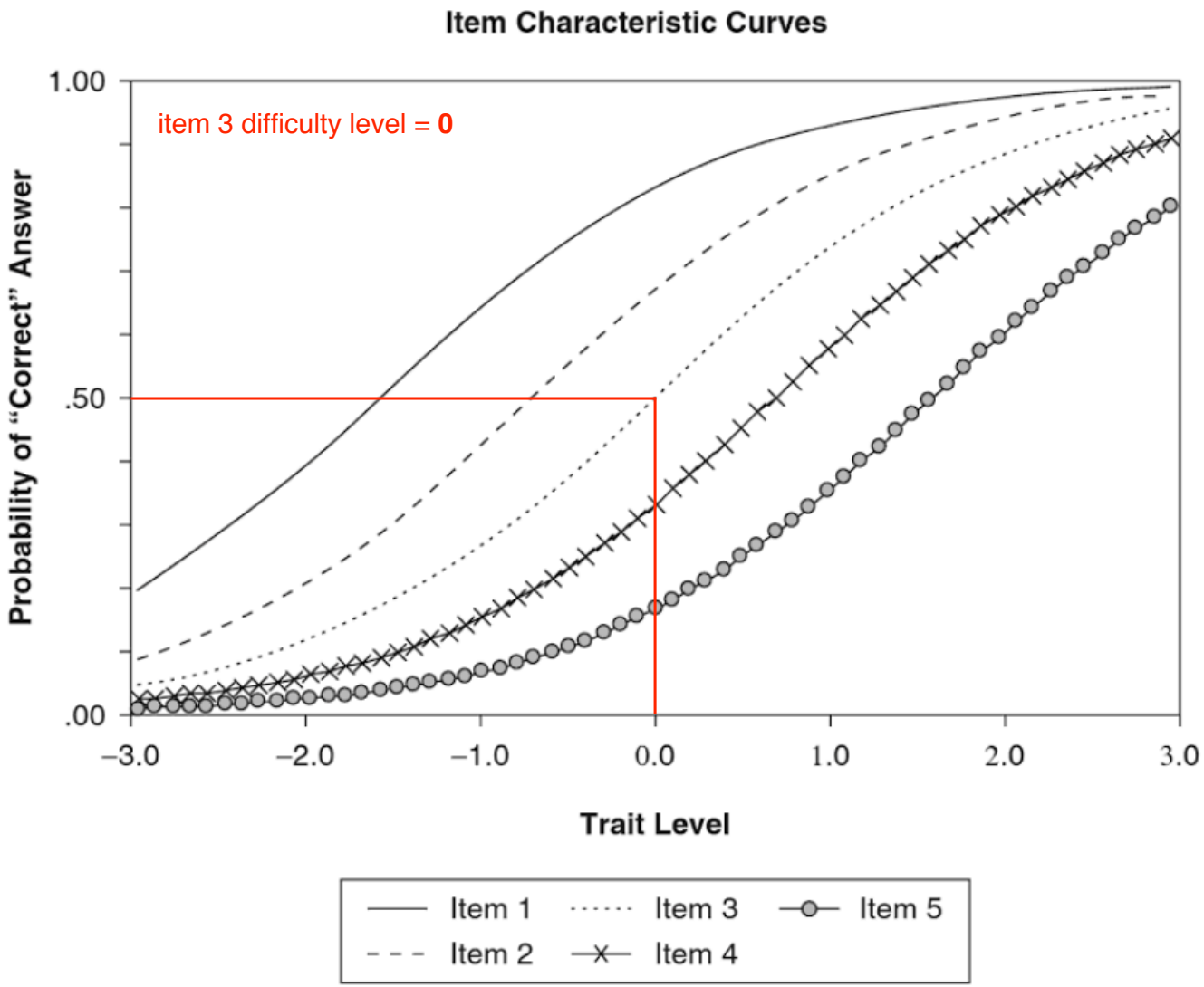
Item Characteristic Curves



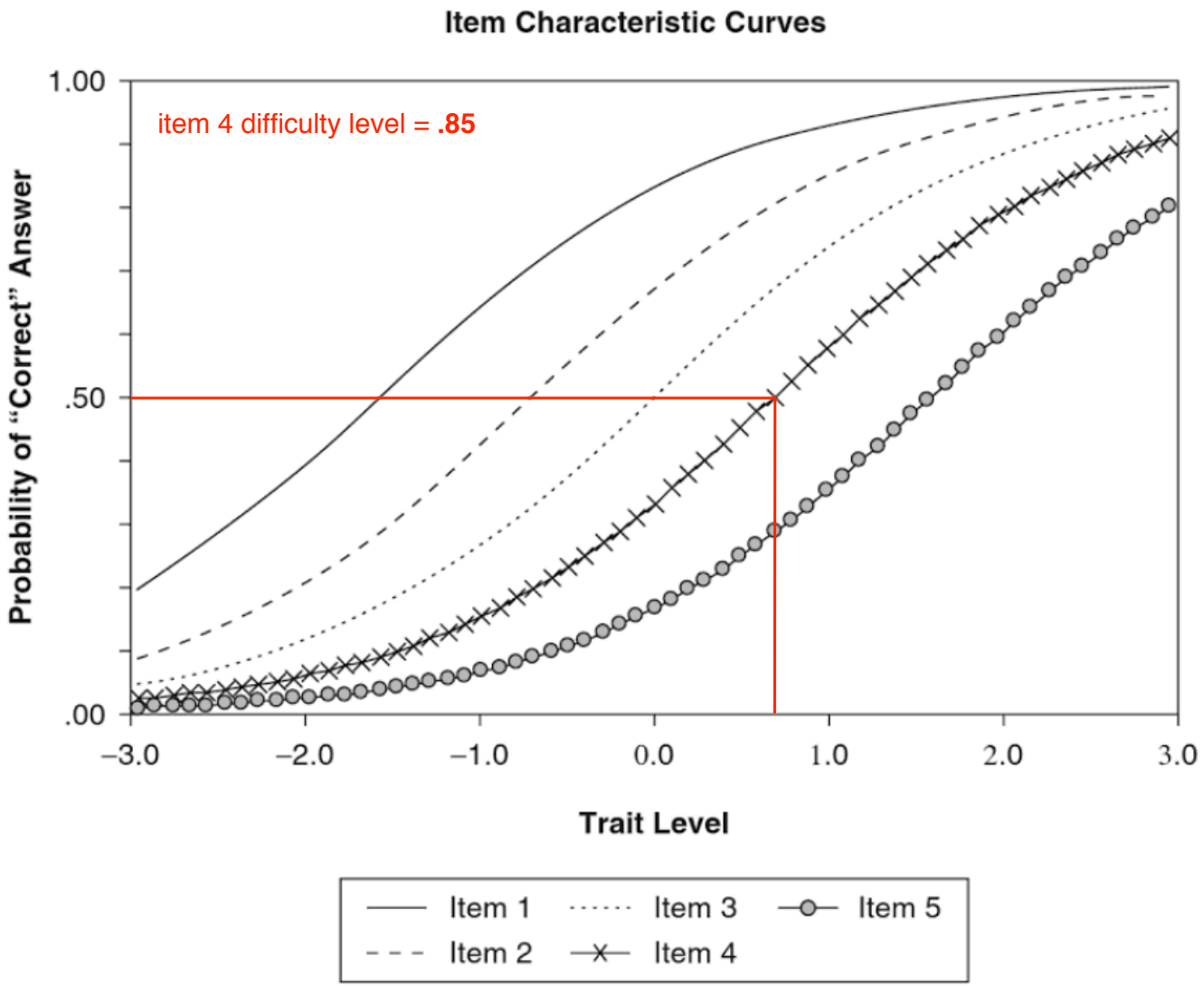
Item Characteristic Curves



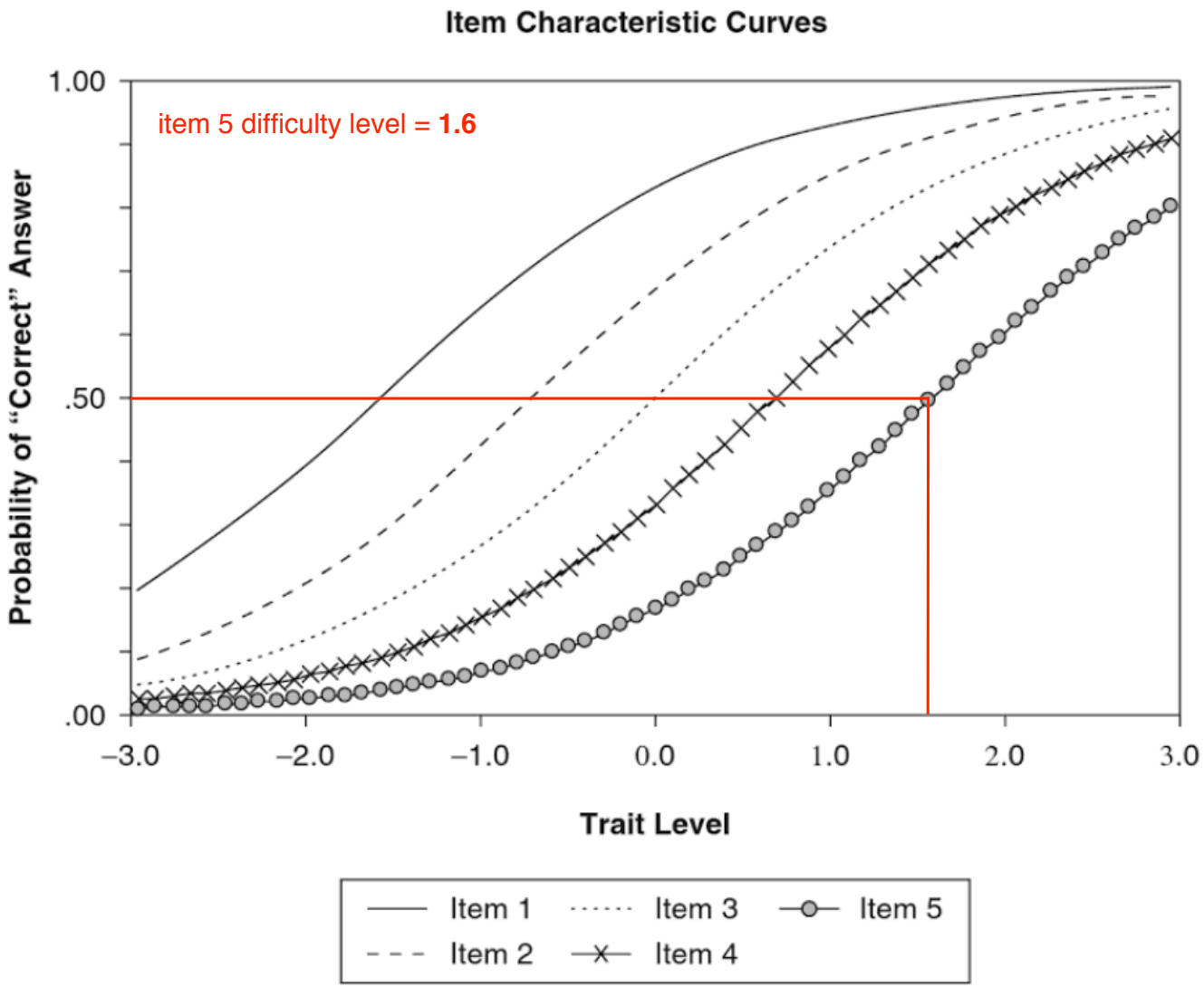
Item Characteristic Curves



Item Characteristic Curves



Item Characteristic Curves

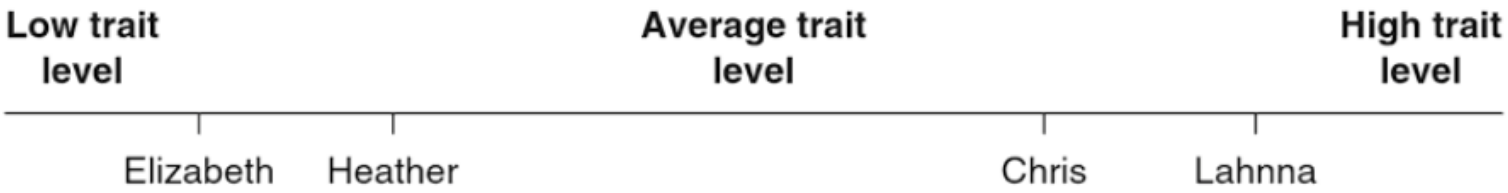


IRT and Reliability

- Under CTT, we might compute Coefficient α to estimate a test's reliability
- We would compute only *one* reliability estimate for a test, and this estimate would indicate the extent to which observed scores are correlated with true scores
- In IRT, a test does not have a single "reliability"
- Instead, a test might have stronger psychometric quality for some people than for others
- That is, a test might provide better information at some trait levels than at other trait levels

IRT and Reliability

- From an IRT perspective, a test has good "psychometric quality" when it can accurately detect differences between individuals at different trait levels



- The psychometric properties of a test may differ across trait levels due to the nature of the items on the test

Applications of IRT

- The textbook mentions three applications of IRT:
  1. Test Development and Improvement
  2. Differential Item Functioning
  3. Computerised Adaptive Testing

Applications of IRT

- The textbook mentions three applications of IRT:
  1. Test Development and Improvement
  2. Differential Item Functioning
  3. Computerised Adaptive Testing

Test Development and Improvement

- A key application of IRT is the evaluation and improvement of the psychometric properties of items and tests
- Using information about item properties, test developers can select items that:
  1. reflect an appropriate range of trait levels
  2. have high discriminatory power
- Guided by IRT analyses these item selections can produce a test with strong psychometric properties over a range of trait levels

Applications of IRT

- The textbook mentions three applications of IRT:
  1. Test Development and Improvement
  2. Differential Item Functioning
  3. Computerised Adaptive Testing

Applications of IRT

- The textbook mentions three applications of IRT:
  1. Test Development and Improvement
  2. Differential Item Functioning
  3. Computerised Adaptive Testing

Differential Item Functioning

- This is a sophisticated approach for detecting response bias which was discussed in our Week 7 lecture on responses biases
- You don't need to know anything more about Differential Item Functioning than what is reported in that lecture

Applications of IRT

- The textbook mentions three applications of IRT:
  1. Test Development and Improvement
  2. Differential Item Functioning
  3. Computerised Adaptive Testing

Applications of IRT

- The textbook mentions three applications of IRT:
  1. Test Development and Improvement
  2. Differential Item Functioning
  3. Computerised Adaptive Testing

CTT: Testing Example

- Generate 30 items
- Order the items in terms of item difficulty
- Administer the items to participants in order of item difficulty
- This is a so-called **static** approach to psychometric testing
- The same questions are asked of everyone regardless of their answers

Computerised Adaptive Testing

- **Computerised adaptive testing (CAT)** is an interactive, computer administered test-taking process wherein items presented to the test taker are based in part on the test taker's performance on previous items
- Accordingly, CAT is a **dynamic** approach to psychometric testing
- CAT provides an accurate and very efficient assessment of an individual's trait level
- This accuracy and efficiency is achieved by giving different tests to different individuals

Computerised Adaptive Testing

- CAT works by using a very large item pool for which IRT has been used to obtain information about the psychometric properties of the items
- For example, we might assemble a pool of 300 items and perform research to estimate the difficulty level of each item
- The information about item difficulties is then entered into a computerised database

Computerised Adaptive Testing

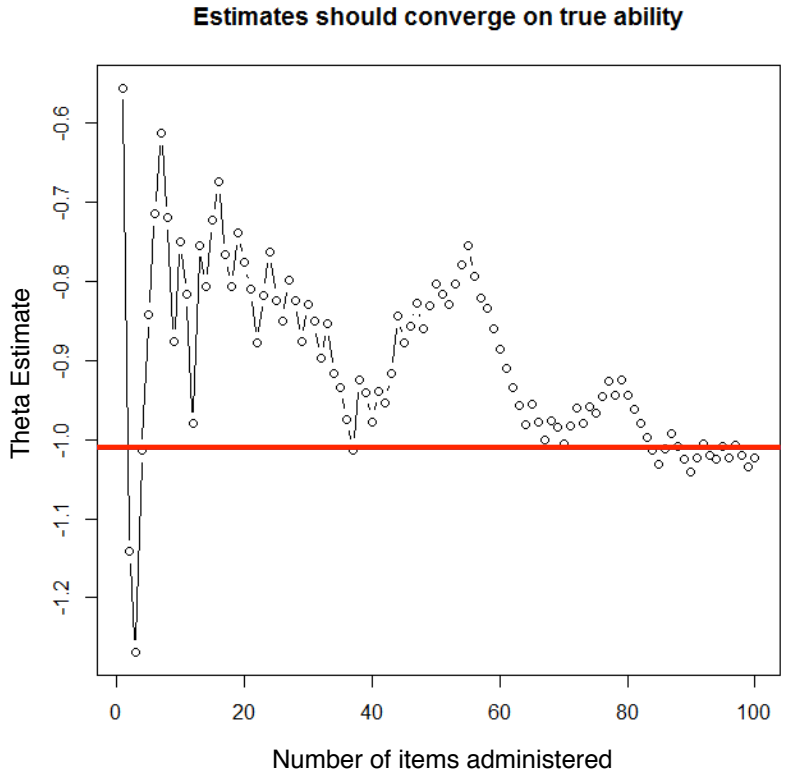
- As a person begins the test, the computer presents items with difficulty levels targeted at an average trait level (i.e., difficulty levels near zero)
- From this point, the computer adapts the test to match the individual's apparent trait level
- If the individual starts the test with several correct answers, then the computer searches its database and selects items with difficulty levels that are a bit above average
- These relatively difficult items are then presented to the individual

Computerised Adaptive Testing

- By contrast, if the individual starts the test with several incorrect answers, then the computer searches its database and selects items with difficulty levels that are a bit below average
- These relatively easy items are then presented to the individual
- Note that the adaptive nature of the CAT algorithm means that different individuals might respond to tests that are almost completely different

Computerised Adaptive Testing

- As the individual continues taking the test, the computer continues to select items that target the individual's trait level
- The computer tracks the individual's responses to specific items with known difficulty levels
- As the individual answers some items correctly and others incorrectly, the computer continually reestimates the individual's trait level
- The computer terminates the test when it has presented enough items to provide a robust estimate of the individual's trait level



The number of items that must be presented to reach the "termination" point will differ across individuals

Advantages of CAT

- Takes as much as 50% less time for testing
  - each individual is presented with only as many items as are required to estimate his or her trait level
- More accurate scores, because respondents answer more items in "their area of difficulty"
  - in CTT testing, many respondents waste time answering very easy or very hard items
  - in CTT testing, typically there are only handful of items that are in the respondents "area of difficulty"
- Disadvantages?
  - more time/money to develop
  - participants don't "trust" it

Next week ...

- Personality and its assessment