Basic IRT Concepts, Models, and Assumptions

Lecture #2
ICPSR Item Response Theory Workshop

Lecture #2 Overview

- Background of IRT and how it differs from CFA
- Creating a scale
- An introduction to common IRT models
  - Item Characteristic Curves
  - Expected Scores
  - Test Characteristic Curves

ICPSR Item Response Theory Workshop

AN INTRODUCTION TO IRT

Item Response Theory

- ***Item Response Theory*** is a psychometric **Theory** and family of associated mathematical models that relate latent trait(s) of interest to the probability of **Responses** to **Items** on the assessment

- IRT is very general method, permitting:
  - One or more traits
  - Various (testable) model assumptions
  - Binary or polytomous response data

A Brief Review of Classical Test Theory

- CTT models the total score: $Y_s = T_s + e_s$
  - Items are assumed exchangeable, and are not part of the model for creating a latent trait estimate
  - **The latent trait estimate is the total score**, which is problematic for making comparisons across different test forms
    - Item difficulty = mean of item (is sample-dependent)
    - Item discrimination = item-total correlation (is sample-dependent)
  - Estimates of reliability assume (without testing) unidimensionality and tau-equivalence (alpha) or parallel items (Spearman-Brown)
    - Measurement error is assumed constant across the trait level

- How do you make your test more reliable?
  - Get more items.
  - What kind of items? More.

A Brief Review of Confirmatory Factor Analysis

- CFA models the ITEM response: $Y_{is} = \mu_i + \lambda_i F_s + e_{is}$
  - Linear regression relating continuous Y to latent predictor F
  - Both items and subjects matter in predicting responses
    - Item difficulty = intercept μ_i (in theory, sample independent)
    - Item discrimination = factor loading λ_i (in theory, sample independent)
  - Factors are estimated as separate entities that predict the observed covariances among items – factors represent testable assumptions
    - local independence :: Items are unrelated after controlling for factors

- Because item responses are modeled:
  - Items can vary in discrimination and difficulty
  - To make your test more reliable, you need items more highly related to the latent trait(s)

- Measurement error is still assumed constant across the latent trait

Similarities of IRT and CFA

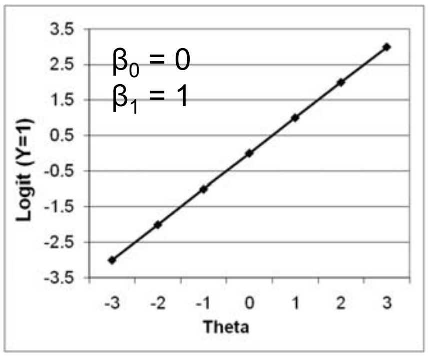
- **IRT** is a model-based measurement model in which latent trait estimates depend on both persons' responses and the properties of the items
  - Like CFA**, both items and persons matter**, and thus properties of both are included in the measurement model
    - Items differ in difficulty and discrimination as in CFA (sample-independent)
- After controlling for a person's latent trait score (now called θ), *the item responses should be uncorrelated*
  - The ONLY reason item responses are correlated is Theta
  - In other words, we typically assume items are unidimensional
    - If this is unreasonable, we can fit multidimensional models instead, and then responses are independent after controlling for ALL Thetas
  - This is the same "**local independence**" assumption as in CFA
    - Can be violated by unaccounted for multidimensionality (i.e., really need multiple Thetas) or other kinds of dependency (e.g., common stem testlets)

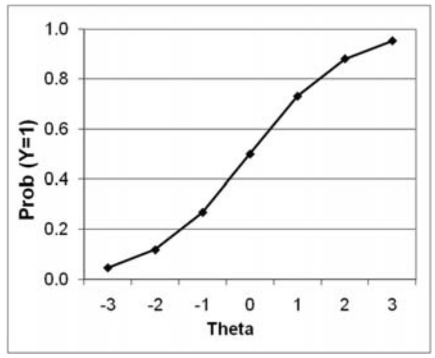
Differences of IRT and CFA

- **IRT specifies a nonlinear relationship between binary, ordinal, or categorical item responses and the latent trait (Theta)**
  - Probability is bounded at 0 and 1, so the effect (slope) of Theta must be nonlinear, so it will shut off at the extremes of Theta (S-shaped curve)
  - Errors cannot have constant variance across Theta or be normal
  - The family of non-linear measurement models for binary and categorical outcomes are called "item response models (IRT)"
    - Or "item response theory" or "latent trait theory"
- IRT uses same family of link functions (transformations) as in generalized models, it's just that the predictor isn't measured directly
  - IRT is logistic regression on latent trait instead of linear regression in CFA
  - Predictor is the latent factor in IRT ("Theta") and still predicts the common variance across item responses just like in CFA

Nonlinearity in IRT

- The relationship between θ and the probability of response=1 is **“nonlinear”**
  - **Linear** with respect to the **logit**, **nonlinear** with respect to **probability**
  - **An s-shaped logistic curve** whose shape and location are dictated by the estimated item parameters





ICPSR Item Response Theory Workshop

THE PURPOSE OF IRT: CREATING A SCALE

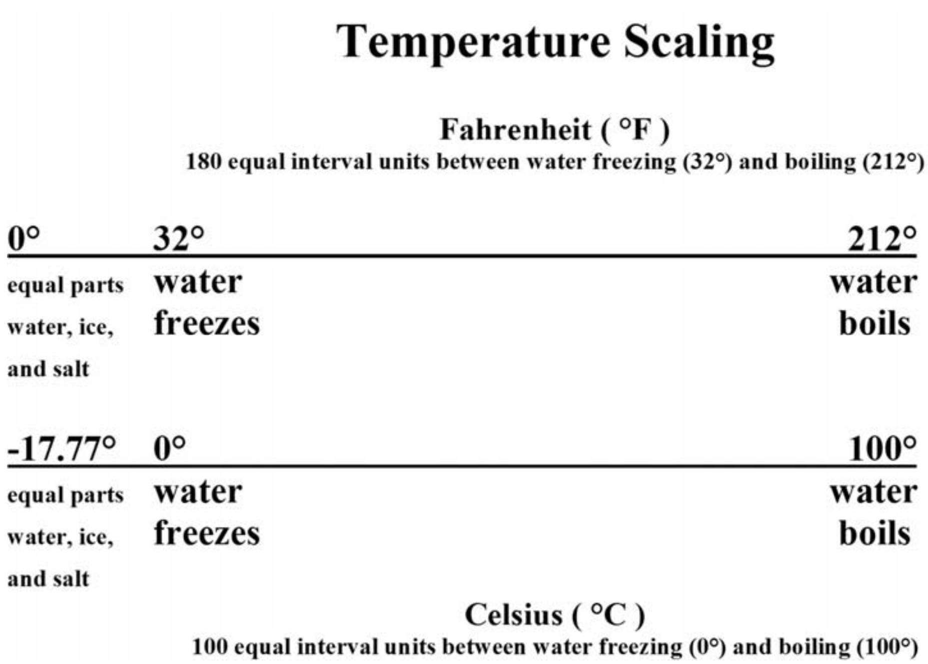
IRT Purpose

- The main purpose of IRT is to create a ***<u>scale</u>*** for the interpretation of assessments with useful properties
  - “Scaling” refers to the process by which we choose a set of rules for measuring a phenomenon

- Creating a “metric” or “scale” for a variable is to systematically assign values to different levels

- Choosing a scale generally involves two important steps:
  - Identifying anchor points
  - Choosing the size of a unit (i.e., a meaningful distance)

Scale Example



IRT Scaling

- IRT proceeds in much the same way
  - A meaningful scale is chosen in order to measure subject "ability" or "trait level"
  - The scale can then be interpreted with reference to the characteristics of test items

- Very important result from IRT: subject traits and item characteristics are referenced to the ***same scale***

Fundamentals of IRT Scaling

- **In CTT**, scores have meaning relative to the persons in the same sample, and thus **sample norms** are needed to interpret a given person's score
  - "I got a 12. Is that good?"
    "Well, that puts you into the 90th percentile."
  - "I got a 12. Is that good?"
    "Well, that puts you into the 10th percentile."
  - Same score in both cases, but different reference group!

- **In IRT**, the properties of items and persons are placed along the same underlying continuous latent metric, called "**conjoint scaling**"
  - This concept can be illustrated using **construct maps** that order both persons in terms of ability and items in terms of difficulty

A Construct Map Example

Theta θ_s = Item difficulty level at which one has a 50% probability of response=1

A Latent Continuum of Disney Princesses

Theta θ_s is interpreted relative to items, not group norms

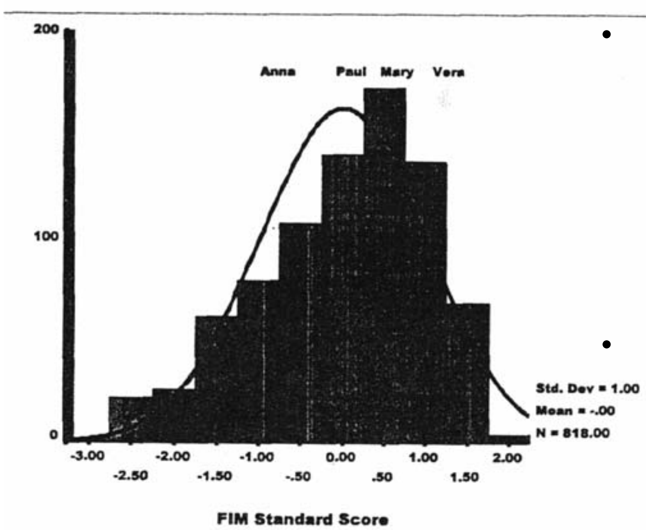
Person Side	**Item Side**
Daphne (daughter)	Megara
.	.
My Sisters	Rapunzel
.	.
My Mom	Aurora
.	.
Average Adult	Cinderella
.	.
Me	Minnie Mouse?

Persons are ordered by Theta ability/severity

Items are ordered by difficulty/severity

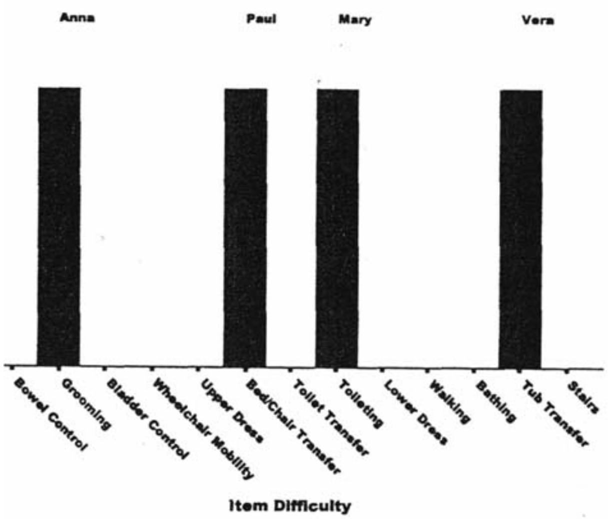
Person Theta and item difficulty share the same latent metric

Norm-referenced Measurement of CTT



- In CTT, the ability level of each person is relative to the abilities of the rest of the test sample
- Here, we would say that Anna is functioning relatively worse than Paul, Mary, and Vera, who are each above average (which is 0)

Item-Referenced Measurement in IRT



- Each person's Theta score reflects the level of activity they can do on their own **50% of the time**

- The model predicts the probability of accomplishing each task given Theta

Features of IRT Models

- Person and item statistics are not dependent on one another

- Conditional probability of item performance is available all along the scale of the trait being measured

- An estimate of the amount of error in each trait estimate, called the **conditional SE of measurement**, is available

- Test items and examinee trait levels are referenced to the same interval scale
  - Although in reality, a true interval scale is difficult to achieve

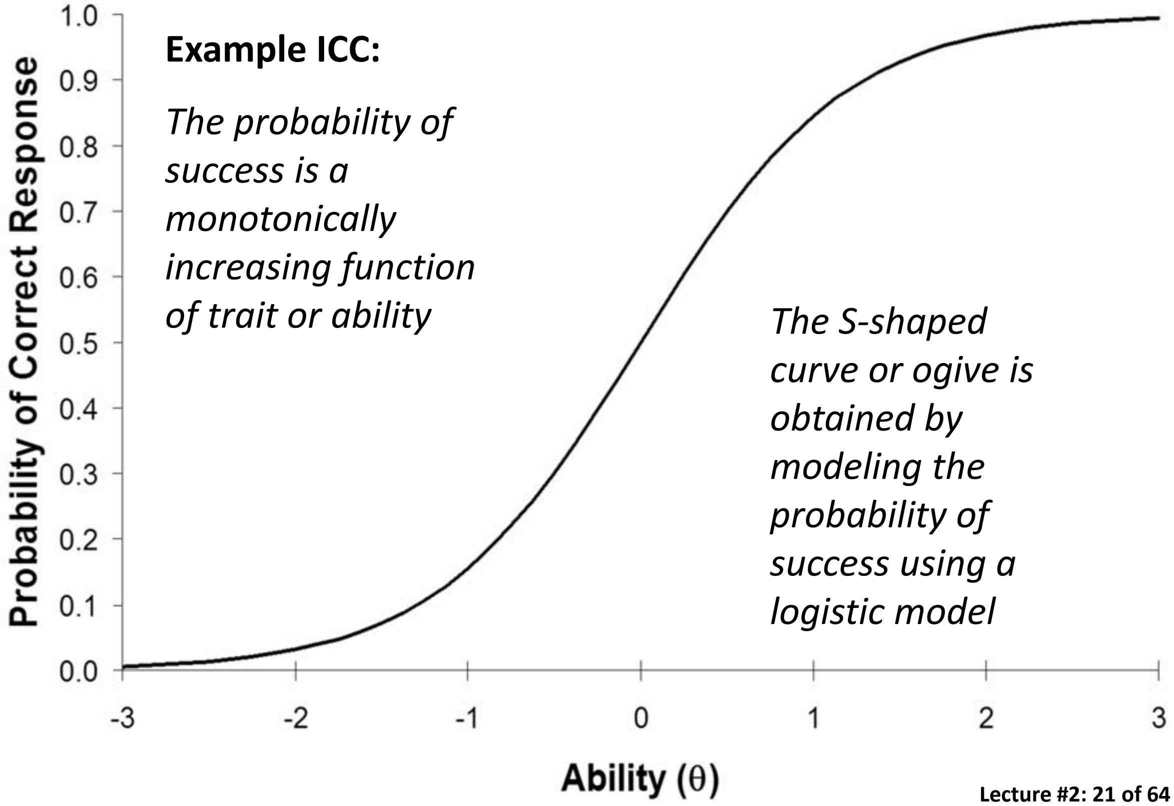
IRT MODEL CHARACTERISTICS AND ASSUMPTIONS

Item Characteristic Curve

- The **Item Characteristic Curve (ICC)** is the primary concept in IRT

- An ICC is a mathematical expression that connects or links a subject's probability of success on an item to the trait measured by the set of test items

- The ICC is a non-linear (logistic) regression line, with item performance regressed on examinee ability



Important Assumptions in IRT

- IRT is based on a set of fairly strong (but testable) assumptions

- If not met, the usefulness or validity of the IRT estimates is severely compromised

- Assumptions:
  - Dimensionality of the Test
    - We will assume one dimension until Friday
  - Local Independence
  - Nature of the ICC
  - Parameter Invariance

Assumption of Unidimensionality

- **Unidimensionality** states that the test measures only ONE construct (e.g., math proficiency, verbal ability)
  - Common to educational testing
  - Less common in psychological (non-cognitive) tests
  - We will use unidimensional models throughout the week to provide a basis for understanding IRT

- <u>Question of interest</u>: Does is make sense to report a single score for an subject's performance on the test?

- The items in a test are considered to be <u>unidimensional</u> when a single factor or trait accounts for a substantial portion of the total test score variance

Assumption of Local Independence

- **Local Independence** assumes that item responses are independent given a subject's latent trait value
  - Related to unidimensionality
  - If only ONE trait determines success on each item, then subject theta is the ONLY thing that systematically affects item performance

- Once you know a subject's theta level, his/her responses to items are independent of one another
  - Important in estimation:: how IRT likelihood function is constructed

What Local Independence Provides

- Conditional independence provides us with statistically independent probabilities for item responses (for items i and i'):

$$P(Y_{is} = 1, Y_{i's} = 1|\theta_s) = P(Y_{is} = 1|\theta_s)\, P(Y_{i's} = 1|\theta_s)$$

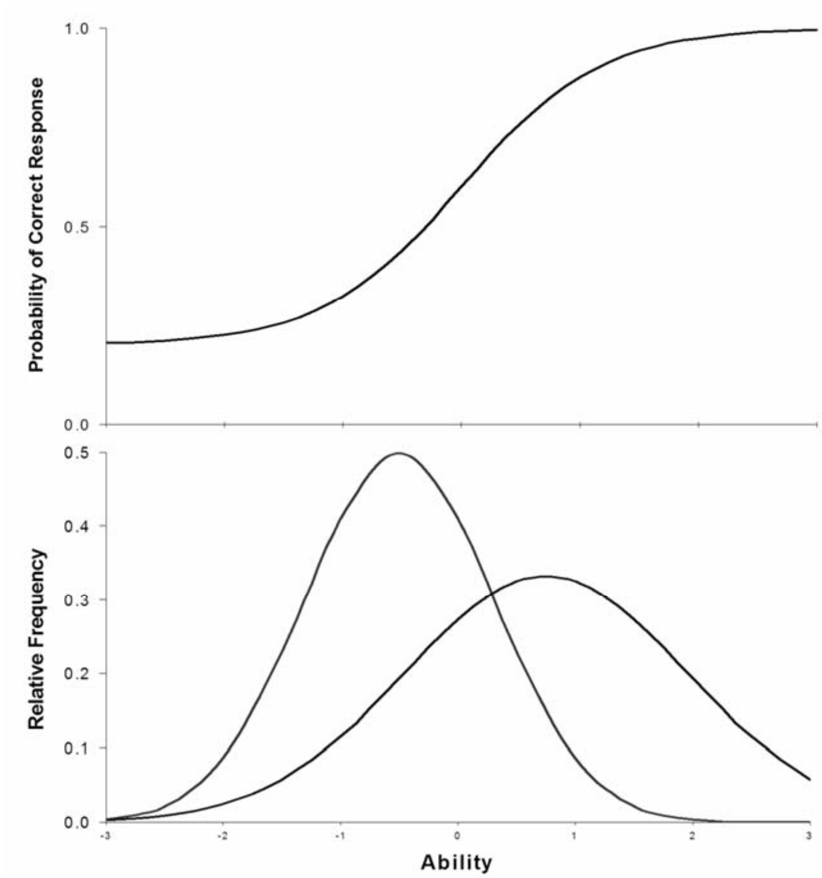
- This will become important soon

Nature of the Item Characteristic Curve

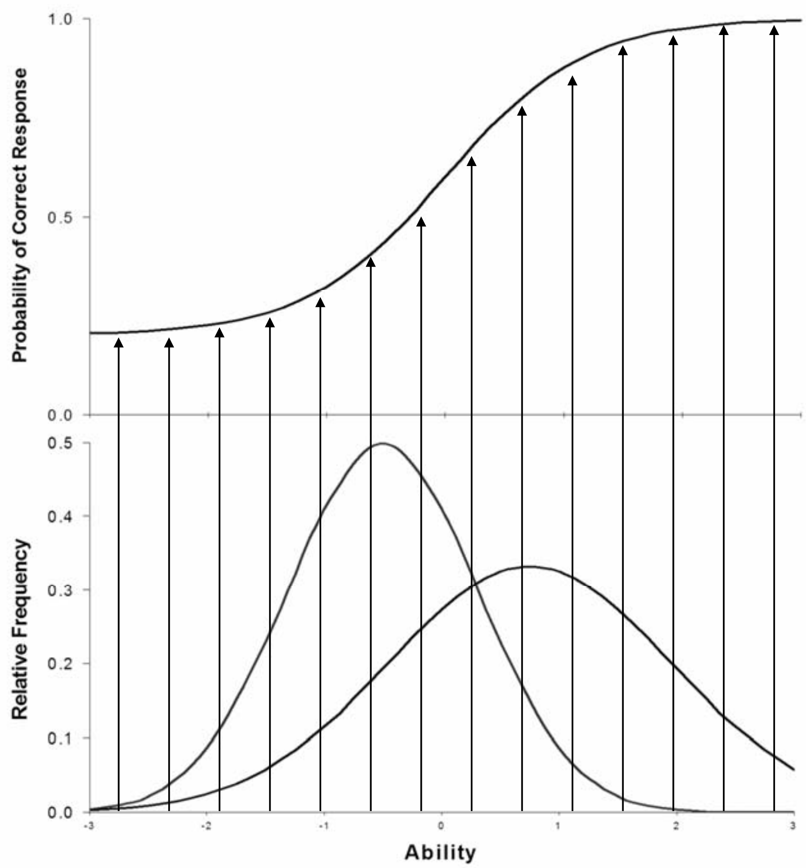
- For dichotomously scored test items (i.e., binary items scored "0" or "1") logistic functions are used to model the probability of "success" (i.e., a "1" vs. a "0")
- The logistic function specifies a monotonically increasing function, such that higher ability results in a higher probability of success
  - Appropriateness of this function depends on situation
  - Educational tests: more theta = higher chance of getting item correct – plausible
  - Opinions on Politics: more of some type of ideology may not give a higher chance of endorsing some political position
    - More on these types of models on Friday

Parameter Invariance

- IF THE IRT MODEL FITS…
  - Item parameters are invariant over samples of examinees from the population for whom the test is intended
  - Ability parameters are invariant over samples of test items from the population of items measuring the ability of interest



Two groups may have different distributions for the trait being measured, but the same model should fit



Frequencies may differ, but matched ability groups should have the same probability of success on the item

Parameter Invariance

- If this seems like a very strong assumption, you've made it before (perhaps without knowing it!)

- The assumption of parameter invariance is a cornerstone of linear regression
  - How else could we apply the model to individuals other than those used to estimate the model?

IRT MODELS

Model Identification in IRT (Setting the Scale)

- Before we begin, we must first decide on the anchoring method for our scale (our latent trait)
  - This means deciding on a mean and standard deviation for the latent trait

- The choice is arbitrary :: several popular methods are used
  - **Anchor by persons** :: Set a fixed mean and variance (such as mean = 0; SD = 1)
    - Done when explaining the variance of the latent trait is not important – rather, when providing a latent trait score is the focus
  - **Anchor by Items ::** Estimate either the mean or SD or both (typically the SD; done by "fixing" other model parameters)
    - Done when explaining the variance of the latent trait

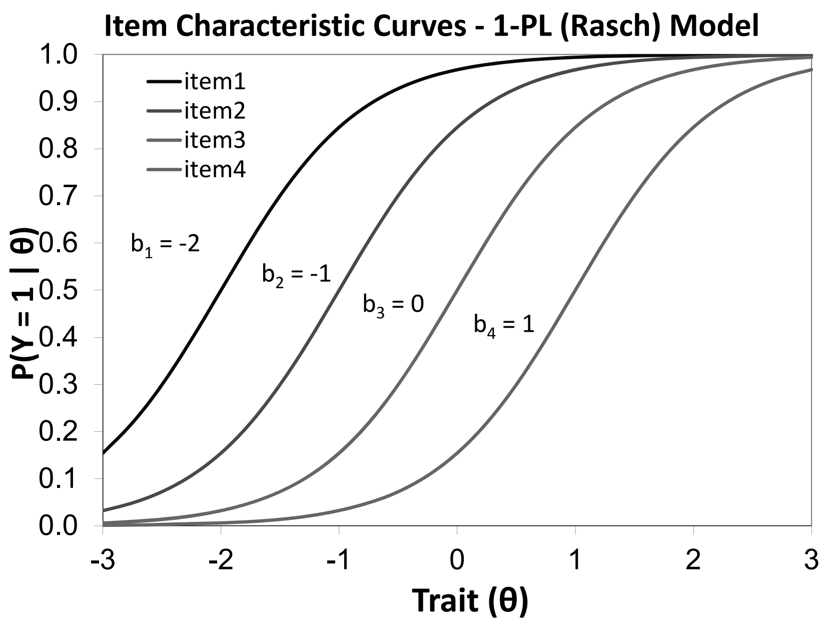
- We will focus on the first: we will fix our latent trait to have a mean of zero and a SD of one
  - Important: The numerical scale doesn't matter, all that matters is that persons and items are on the same scale

The One-Parameter Logistic Model (A.K.A. Rasch Model)

$$P(Y_{is} = 1|\theta_s) = \frac{\exp\big(1.7a(\theta_s - b_i)\big)}{1 + \exp\big(1.7a(\theta_s - b_i)\big)}$$

- θ_s is the **subject ability** (for subject s)
  - most likely latent trait score (Theta) for subject *s* given pattern of item responses
- b_i is the **item difficulty** (for item i)
- a is the common discrimination parameter
- 1.7 is a “scaling constant” which places the parameters of the logit onto a similar scale as the probit
  - Historical legacy which is slowly fadign away

1-PL (Rasch) Model Item Characteristic Curves



b_i = difficulty location on latent trait where p = .50

a = discrimination slope at p = .50, (at the point of inflection of curve)

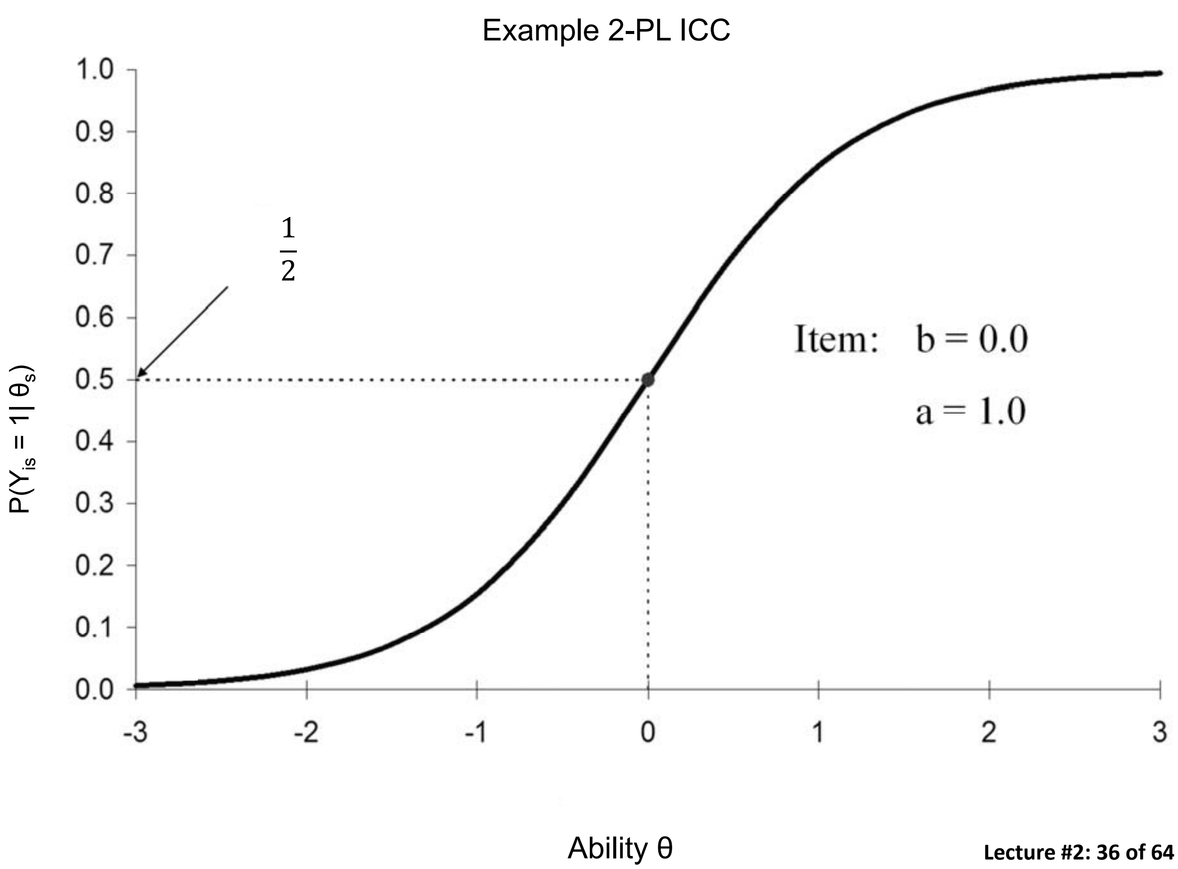
Note: **equal a’s** means curves will never cross :: this is called “Specific Objectivity”

The 2-Parameter Logistic Model

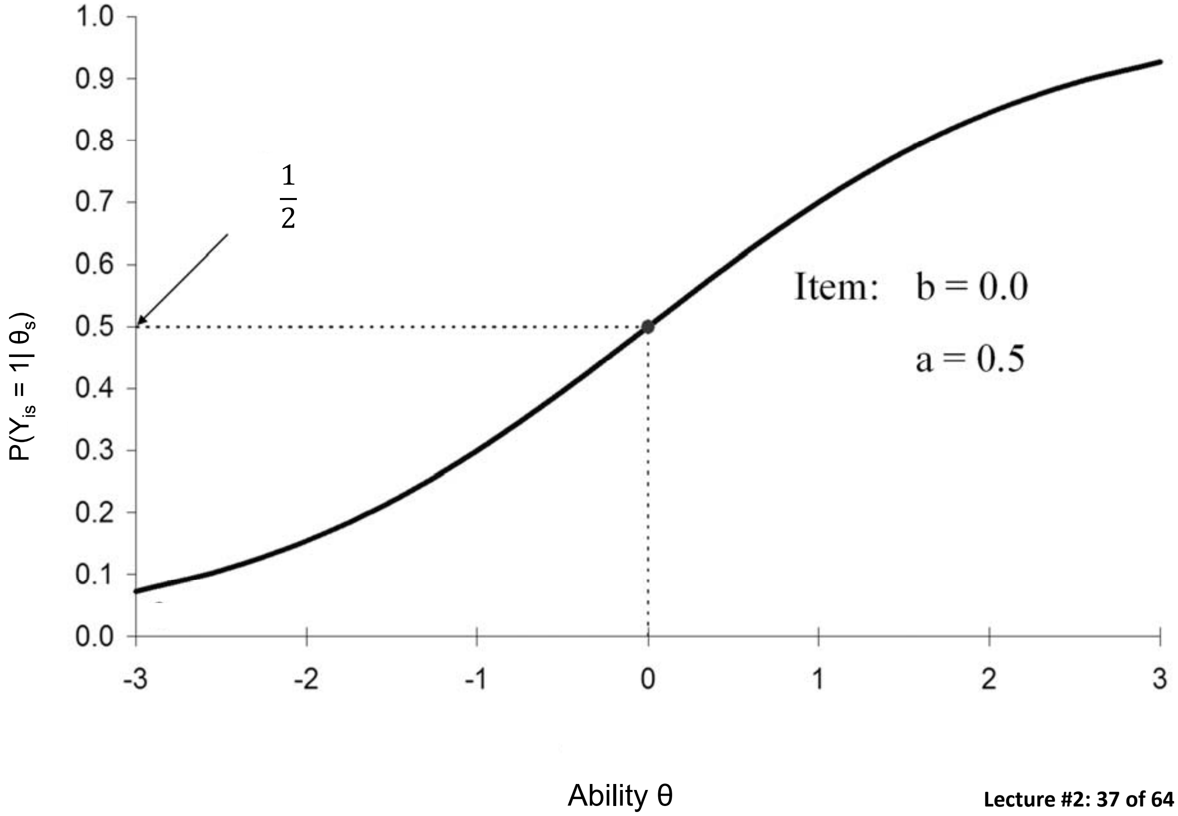
- The 1-PL (Rasch) model assumed each item had the same discrimination
  - This is unlikely to hold in most data

- The 2-PL model allows for each item to have it's own discrimination parameter:

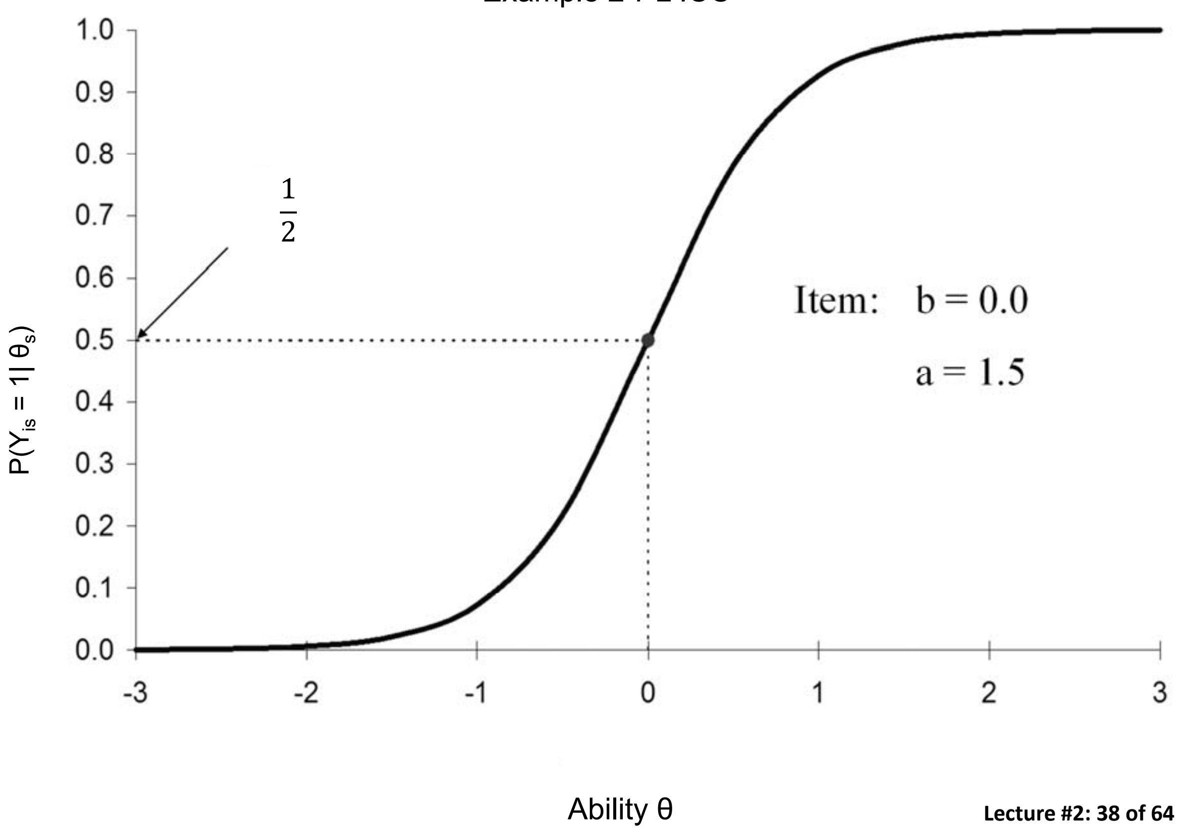
$$P(Y_{is} = 1|\theta_s) = \frac{\exp\left(1.7a_i(\theta_s - b_i)\right)}{1 + \exp\left(1.7a_i(\theta_s - b_i)\right)}$$



Example 2-PL ICC

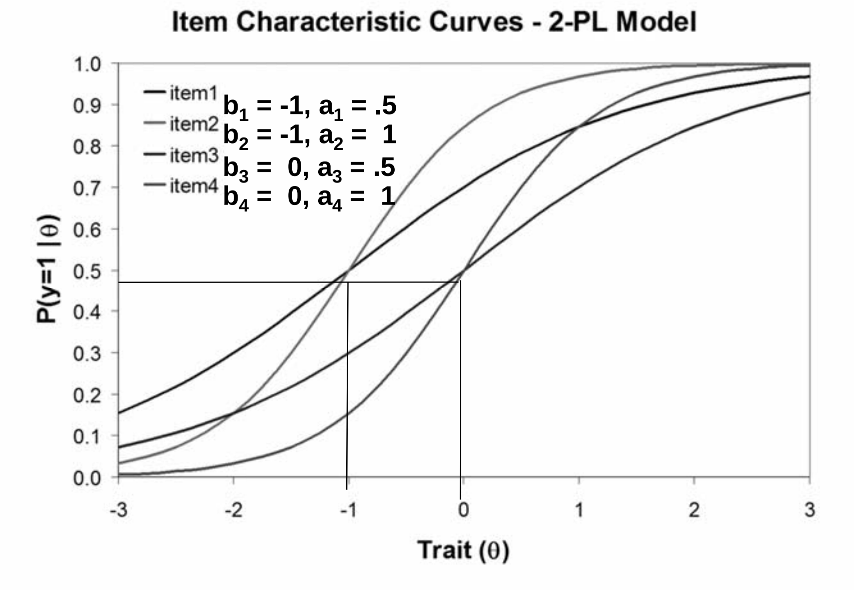


Example 2-PL ICC



2-PL Model Item Characteristic Curves

b_i = difficulty = location on latent trait where p = .50
a_i = discrimination slope at p = .50 (at the point of inflection of curve)



Note: **unequal a's** implies curves will cross. Violates Specific Objectivity

At Theta = -1:
Items 3 & 4 are 'harder' than 1 & 2 (lower prob of 1)

At Theta = +2:
Item 1 is now 'harder' than Item 4 (lower prob of 1)

"IRT" Modeling vs. "Rasch" Modeling

- According to most IRT people, a "Rasch" model is just an IRT model with discrimination $\mathbf{a}_i$ held equal across items
  - Rasch = 1-PL where difficulty is the only item parameter
  - Slope = discrimination a_i = strength of relation of item to latent trait
  - *"Items may not be equally 'good', so why not let their slopes vary?"*

- According to most Rasch people, the 2PL & rest of IRT is voo-doo
  - Rasch models have specific properties that are lost once you allow the item curves to cross (by using unequal a_i) :: "Specific Objectivity"
    - Under the Rasch model, persons are ordered the same in terms of predicted responses regardless of which item difficulty location you're looking at
    - Under the Rasch model, items are ordered the same in terms of predicted responses regardless of what level of person theta you're looking at
    - The a_i represents a person*item interaction :: the item curves cross, so the ordering of persons or items is no longer invariant, and this is "bad"
  - *"Items should not vary in discrimination if you know your construct."*

Which Model Fits Better? Relative Model Fit in IRT

- **Nested models** can be compared with -2LL difference tests
  - Step 1: Calculate -2*difference of LL_{old} and LL_{new}
  - Step 2: Calculate difference in df_{old} and df_{new} (given as "# free parms")
  - Compare $-2LL_{diff}$ on df = df_{diff} to χ^2 critical values (or excel CHIDIST)
  - Add 1 parameter? $-2LL_{diff} > 3.84$, add 2: $-2LL_{diff} > 5.99$…

- If **adding** a parameter, model fit gets **better** (LL up, -2LL down)
- If **removing** a parameter, model fit gets **worse** (LL down, -2LL up)
- AIC and BIC values (based off of -2LL) can be used to compare non-nested models (given same sample), smaller is better

- No easily obtainable trustable absolute global fit info available via ML for IRT
  - Stay tuned for why this is…

Local Model Fit under ML IRT

- IRT programs also provide "item fit" and "person fit" statistics (although not provided by Mplus)
  - Item fit: Predicted vs. observed ICCs – how well do they match?
    Or via inferential tests (Bock Chi-Square Index or BILOG version)
  - Person fit "Z" based on predicted vs. observed response patterns
  - Some would advocate removing items or persons who don't fit

- **Under ML in Mplus**: Local item fit available with **TECH10** output
  - **Univariate item fits**: How well did the model reproduce the observed response proportions? (Not likely to have problems here)
  - **Bivariate item fits**: Contingency tables for pairs of responses
    - Get χ^2 value for each pair of items that directly tests their remaining dependency after controlling for Theta(s); assess significance via χ^2 table
    - This approach is more likely to be useful than traditional 'item fit' measures because those use Theta estimates as known values
- *Stay tuned for an easier option for assessing local fit…*

Two Types of IRT Models: Logistic and Ogive

1. **Logistic**: $P(Y_{is} = 1|\theta_s) = \frac{\exp(1.7a_i(\theta_s - b_i))}{1+\exp(1.7a_i(\theta_s - b_i))}$

   Model predicts **logit** value that corresponds to prob(Y=1)

2. **Ogive**: $P(Y_{is} = 1|\theta_s) = \int_{-\infty}^{z_{is}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right) dt$
   $= \Phi(z_{is}) = \Phi(-b_i + a_i\theta_i)$

   Model predicts **z-score** for the to area to the left of prob(Y=1)

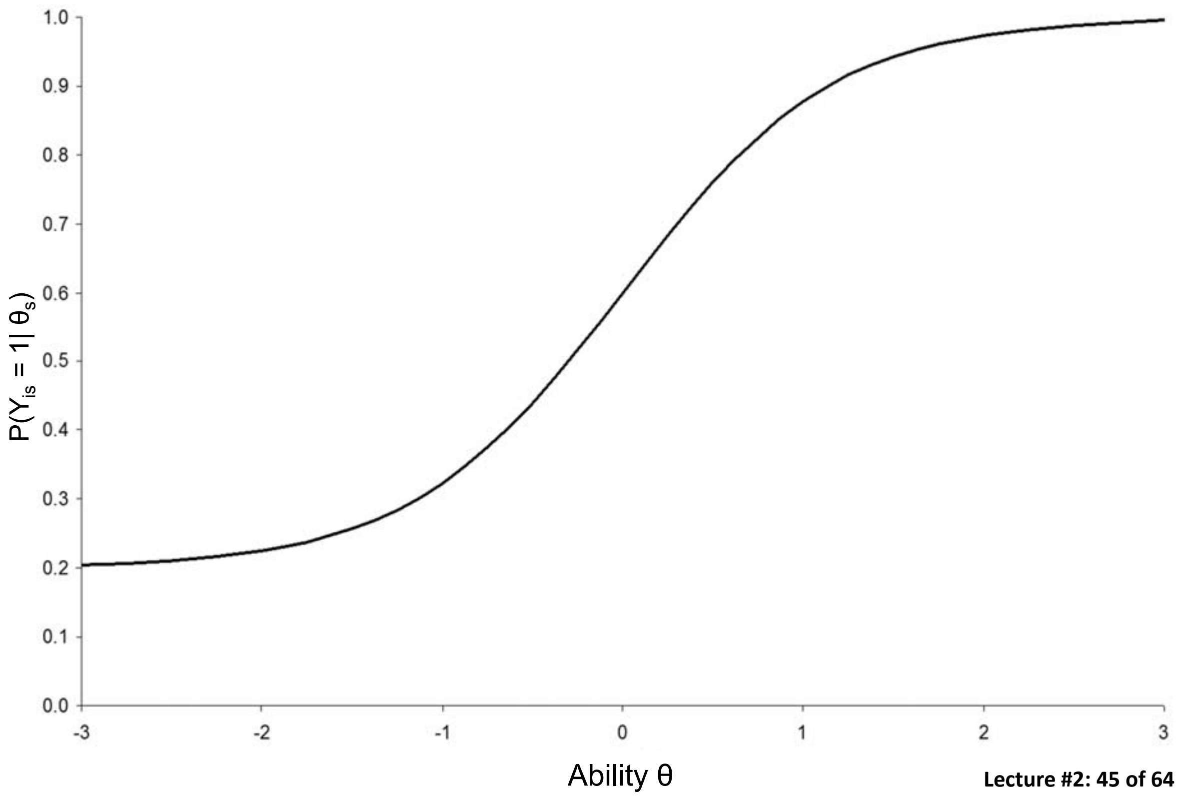
- This is the same distinction as "logit" vs. "probit"
  - **Logit scale = Probit scale*1.7**, so they predict the same curves
  - Probit came along first, but used to be harder to estimate, so logit was developed… and now logit is usually used instead

Additional IRT Models: 3-Parameter Logistic

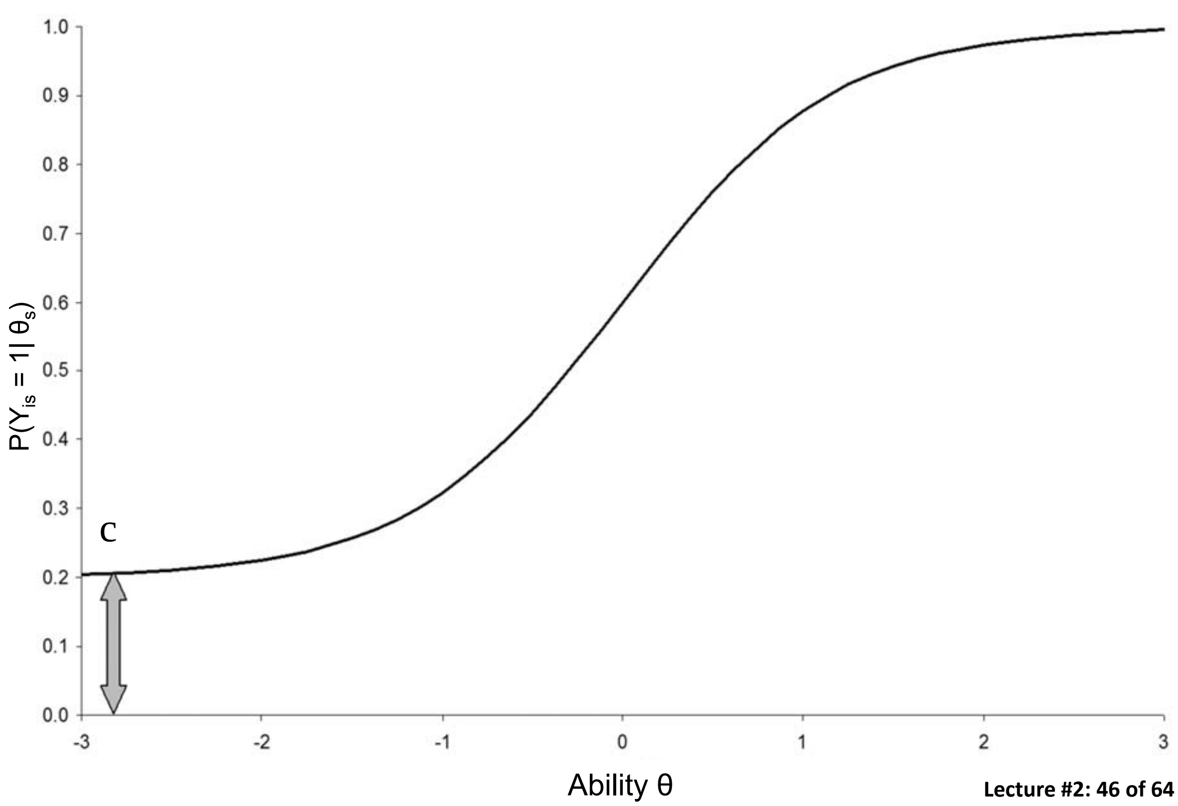
$$P(Y_{is} = 1|\theta_s) = c_i + (1 - c_i)\,\frac{\exp(1.7a_i(\theta_s - b_i))}{1 + \exp(1.7a_i(\theta_s - b_i))}$$

- **b_i** = item difficulty :: location
  - Higher values mean more difficult items (lower chance of a 1)
- **a_i** = item discrimination :: slope
  - Higher values = more discriminating items = better items
- **c_i** = item lower asymptote :: "**guessing**" (where $c_i > 0$)
  - Lower bound of probability independent of Theta
  - Can estimate a common **c** across items as an alternative
- Probability model starts at 'guessing', then depends on Theta and a_i, b_i
  - 3-PL model with c or c_i currently not available within Mplus

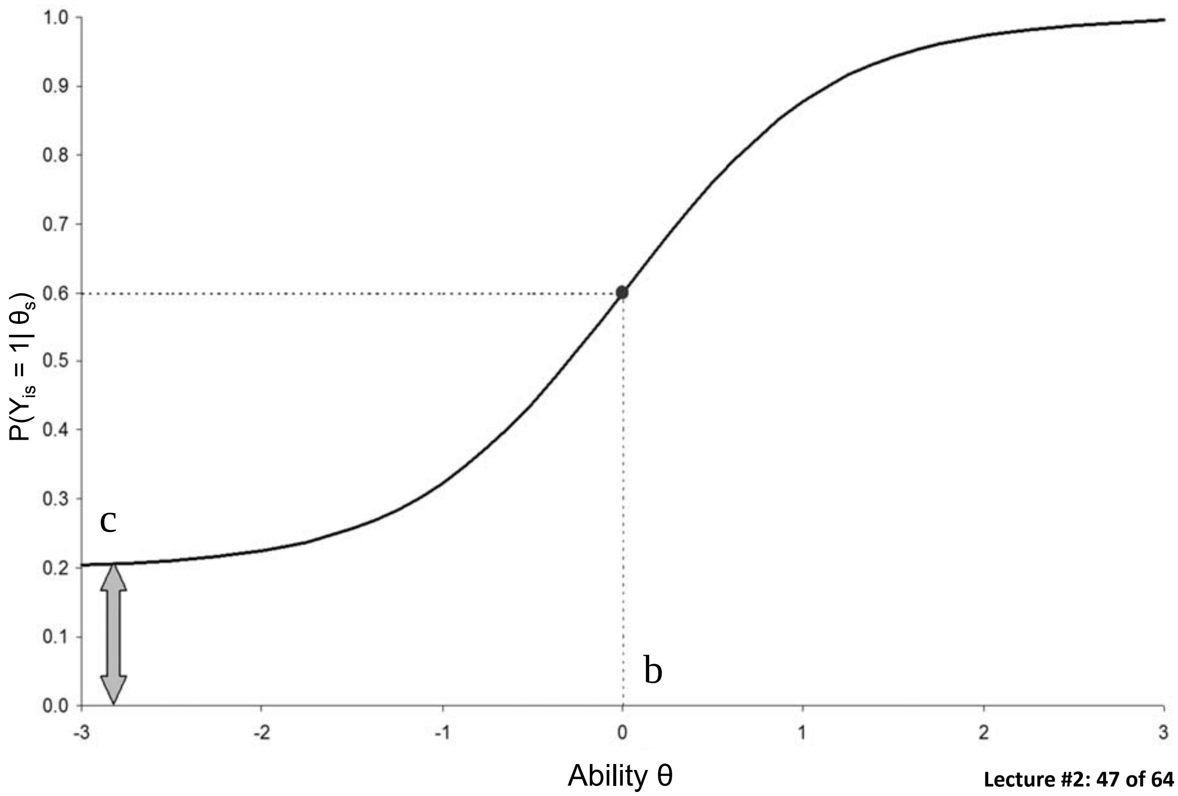
Example 3-PL ICC



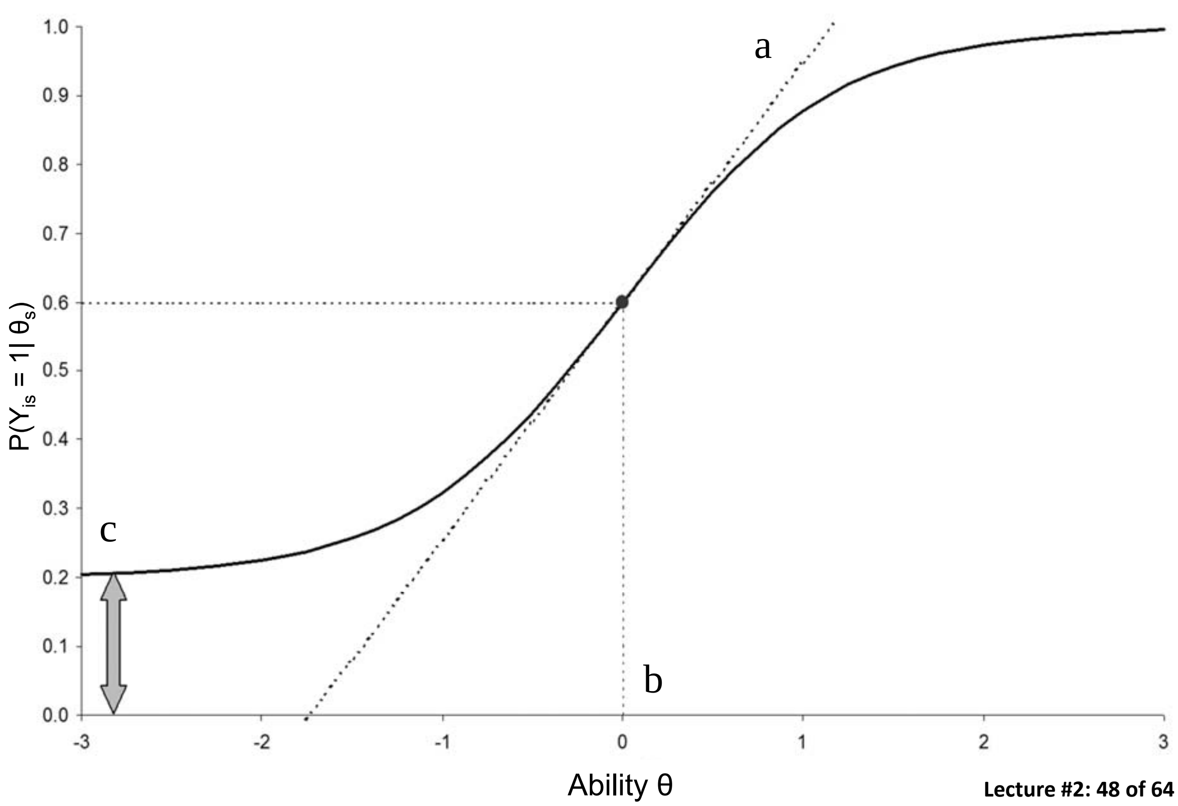
Example 3-PL ICC



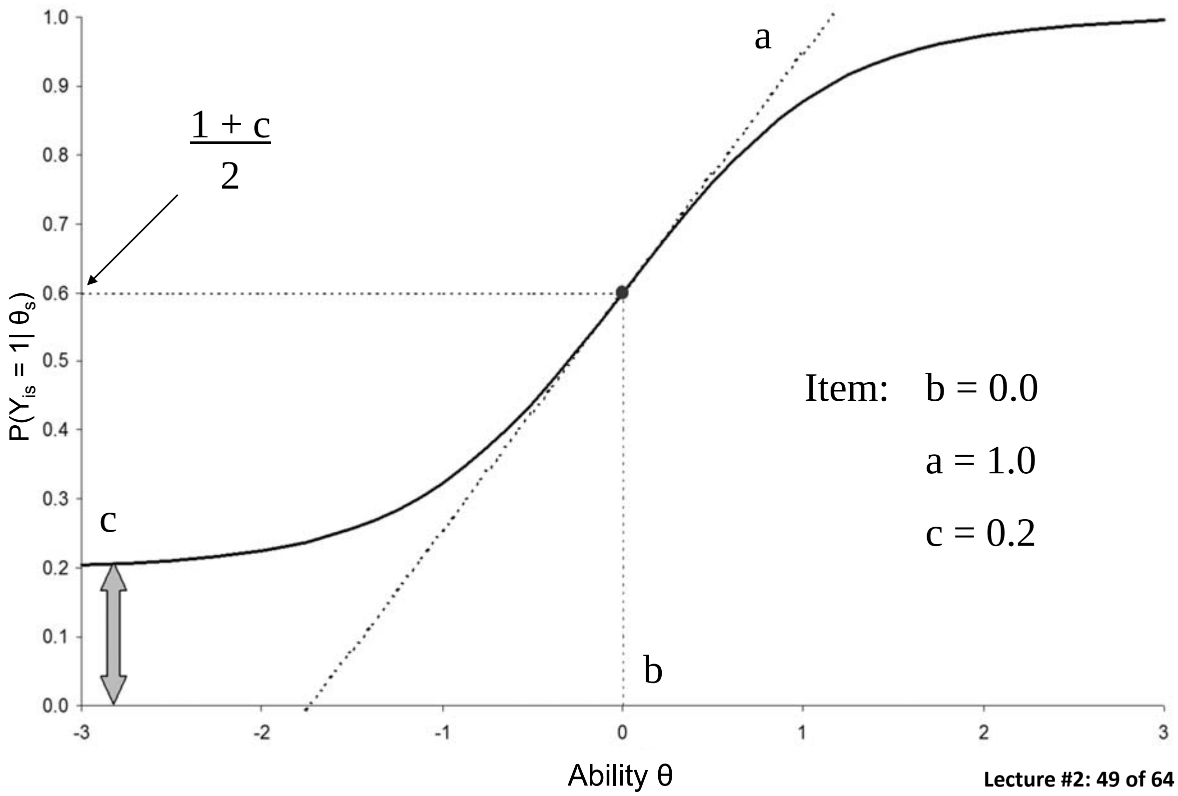
Example 3-PL ICC



Example 3-PL ICC

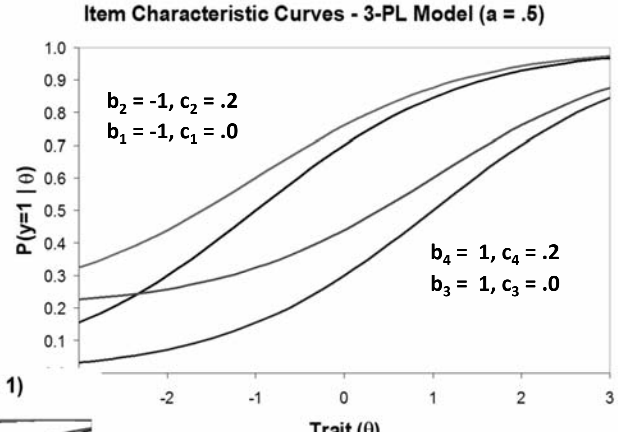


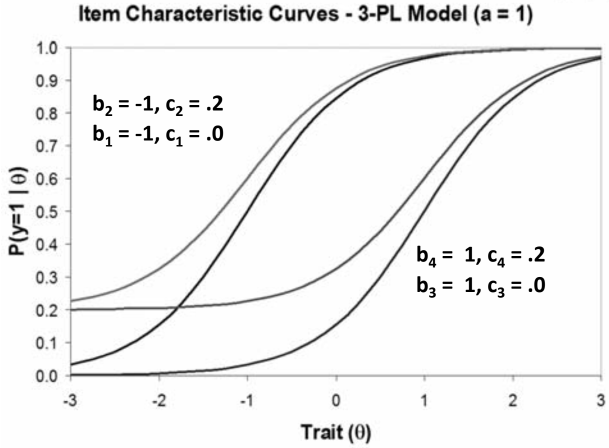
Example 3-PL ICC



Top: Items with lower discrimination (**a_i = .5**)

Below: Items with higher discrimination (**a_i = 1**)





Note that difficulty b_i values are no longer where p = .50

The expected probability at b_i is moved upwards by the lower asymptote c_i parameter

Yet Another One: The 4-Parameter Logistic Model

$$P(Y_{is} = 1|\theta_s) = c_i + (d_i - c_i)\,\frac{\exp(1.7a_i(\theta_s - b_i))}{1 + \exp(1.7a_i(\theta_s - b_i))}$$

- **b_i** = item difficulty :: location
- **a_i** = item discrimination :: slope
- **c_i** = item lower asymptote :: “guessing”
- **d_i** = item upper asymptote :: “**carelessness**” (so $d_i < 1$)
  - Maximum probability to be achieved independent of Theta
  - Could be carelessness or unwillingness to endorse no matter what
- Probability model starts at ‘guessing’, tops out at ‘carelessness’, then depends on Theta and a_i, b_i in between
  - 4-PL model with d or d_i currently not available within Mplus

IRT MODEL SPECIFICS AND PREDICTIONS

An Expected Score in IRT

- The probability of a correct response for a given ability level is equal to the expected score for subjects on that item

$$E(Y_{is}) = P(Y_{is} = 1|\theta_s) = \text{[IRT MODEL]}$$

- The relative frequency of correct answers for subjects of a given ability should be equal to the model predicted probability
  - This is sometimes used to assess the fit of a model

More on the Expected Score

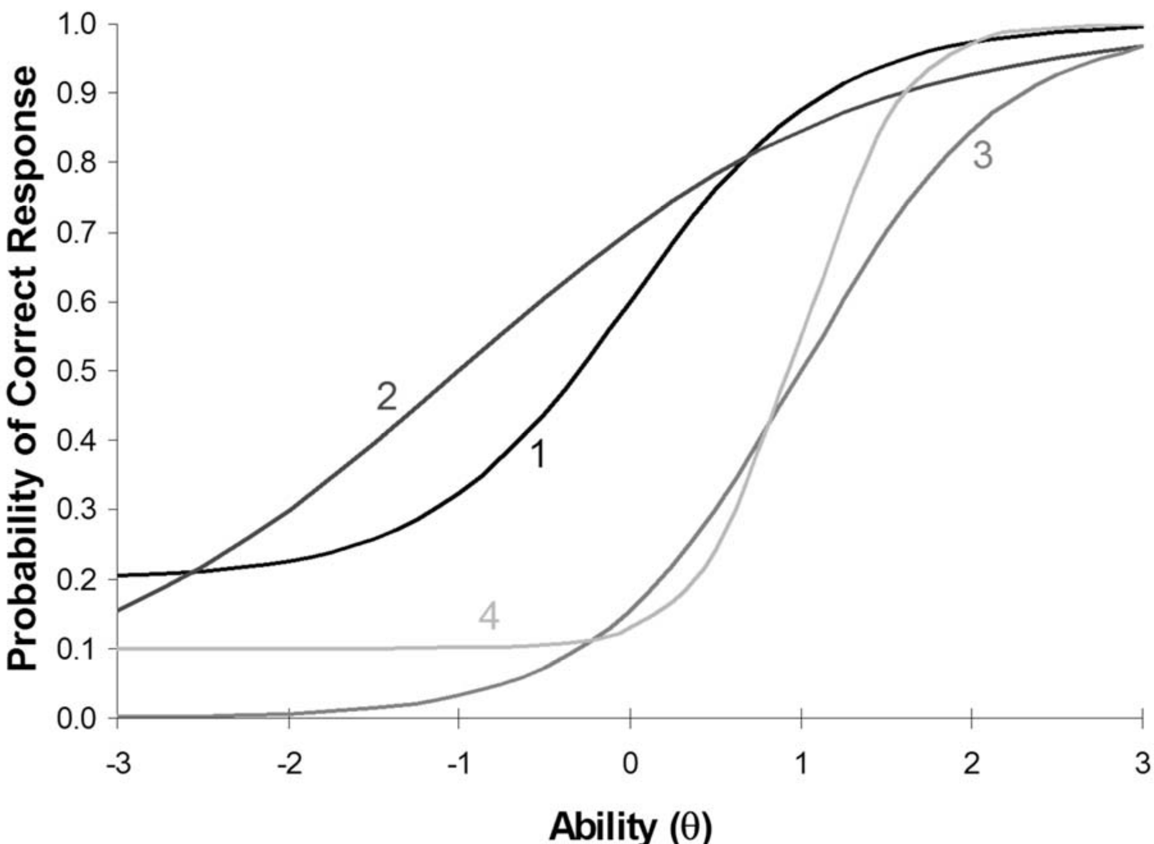
- If $P(Y_{is} = 1|\theta_s) = 0.80$ then 80% of the subjects with that theta should answer the item correctly
  - The remaining 20% should answer the item incorrectly

- Since dichotomous items are scored either right or wrong, from basic statistics:

$$E(Y_{is}) = (0.80 \text{ x } 1) + (0.20 \text{ x } 0) = 0.80$$

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



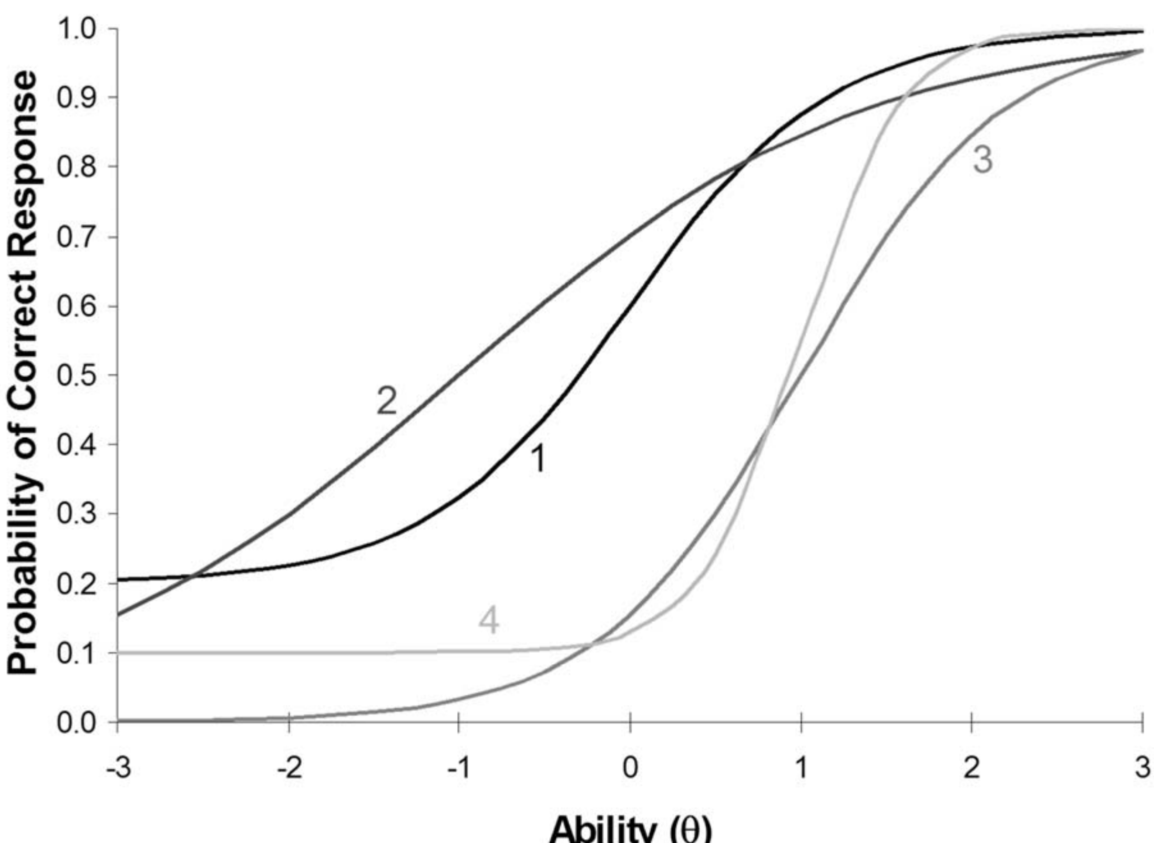
Test Characteristic Curve

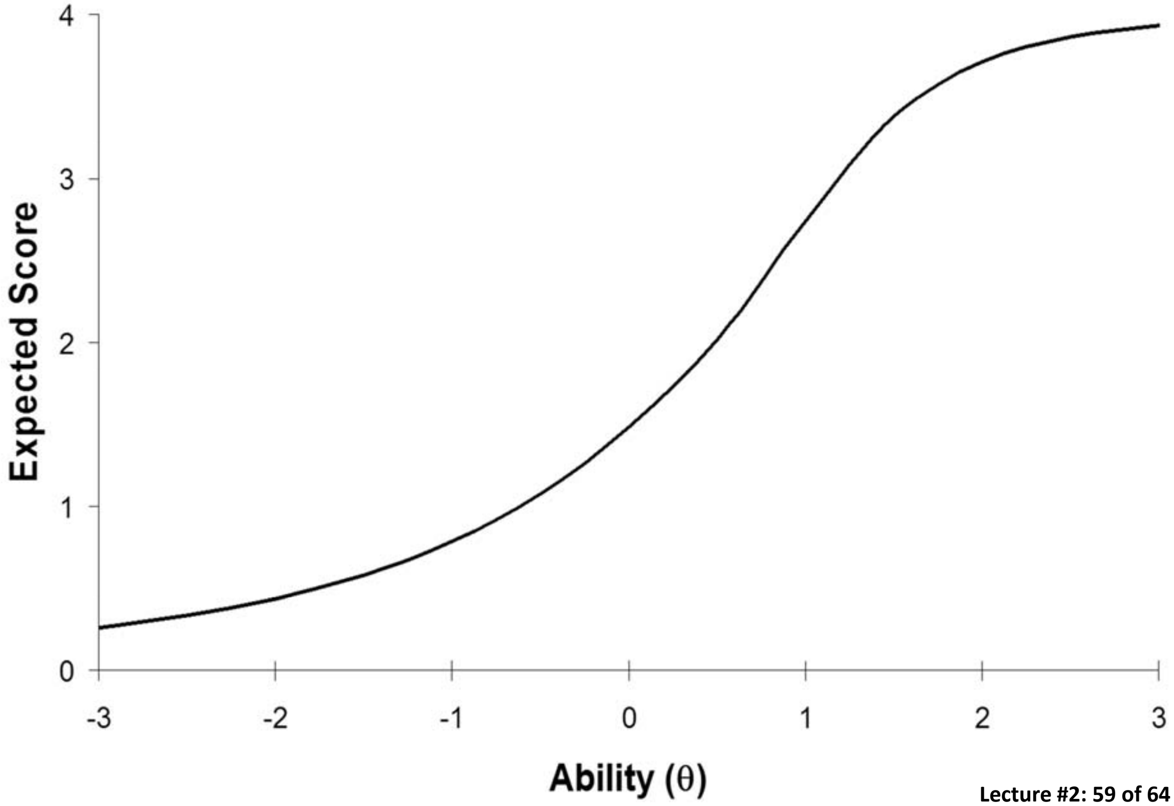
- A test characteristic curve (TCC) is created by summing each ICC across the ability continuum

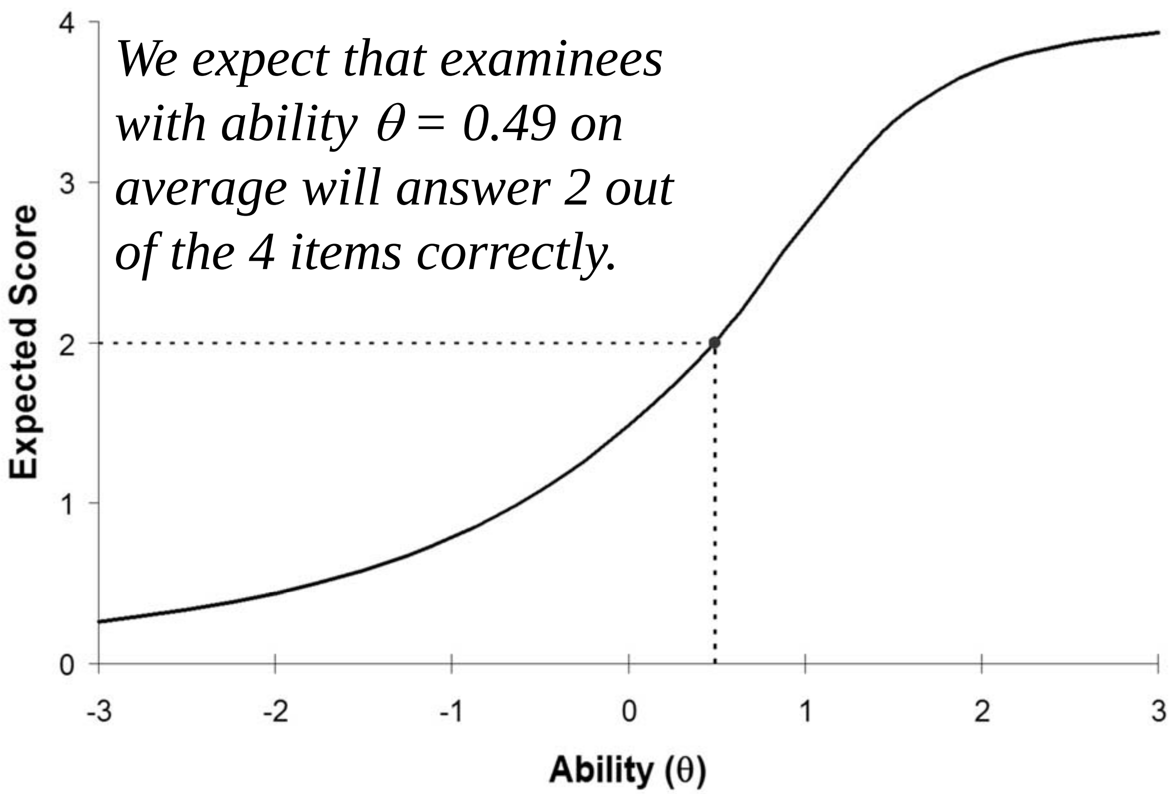
$$TCC(\theta_s) = \sum_{i=1}^{I} P(Y_{is} = 1|\theta_s)$$

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

- Since $P(Y_{is} = 1|\theta_s)$ 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




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)

WRAPPING UP

Lecture #2 Wrap Up

- IRT is a family of models that specify the relationship between the latent trait ("Theta") and a link-transformation of probability of Y
  - **Linear** relationship between Theta and **Logit** (Y=1) (or **probit** of y=1)
    :: **nonlinear** relationship between Theta and **Prob** (Y=1)
- The form of the relationship depends on:
  - At least the location on the latent trait (b_i)
  - Perhaps the strength of relationship may vary across items (a_i)
    - If not, its a "1-PL" or "Rasch model"
  - Also maybe lower and upper asymptotes (c_i and d_i)
- Ability is unidimensional; item responses are locally independent
- Item, ability parameters are estimated, assumed invariant, and model-data fit is assessed

The Big Picture

- If the model fits the data and the assumptions are met (IMPORTANT), IRT model fitting gives rise to a whole host of powerful procedures
  - Construct tests with known properties
  - Create banks of items on a common scale
  - Equate separate test forms reliably
  - Evaluate Differential Item Functioning
  - And many more…

Up Next

- More basics of IRT models, some review
- Model Specifications
- Scale Characteristics