Latent Trait Reliability

Lecture #7

ICPSR Item Response Theory Workshop

Lecture Overview

- Classical Notions of Reliability

- Reliability with IRT
  - Item and Test Information Functions
    - Concepts
    - Equations
    - Uses and Examples

Two Big Concerns about Scale Scores

- **Reliability:**
  - "Extent to which the instrument does what it is supposed to *with sufficient consistency* for its intended usage"
  - "Extent to which same results would be obtained from the instrument after repeated trials"
  - Operationalized in many ways…

- **Validity:**
  - "Extent to which the instrument measures *what it is supposed to* (i.e., it does what it is intended to do)" or "Validity for WHAT?"
  - Is measure of degree, and depends on USAGE or INFERENCES
    - Scales are not "valid" or "invalid" – validity is NOT a scale property
    - e.g., Test of intelligence: Measure IQ? Predict future income?

An Example From Practice

- From the *Graduate Record Examinations® Guide to the Use of Test Scores* (2010-2011; p. 20)
  - http://www.ets.org/s/gre/pdf/gre_guide.pdf

Table 6A: Conditional Standard Errors of Measurement at Selected Scores for General Test Measures*

Measure	200	250	300	350	400	450	500	550	600	650	700	750	800
Verbal	14	21	26	28	31	35	34	33	33	33	34	32	20
Quantitative	26	42	48	55	55	54	50	49	42	39	35	26	9

CLASSICAL NOTIONS OF RELIABILITY

Conceptualizing Reliability:

Y_{total} = True Score + error

- Wait a minute… if E(Y) = T…
  - This idea refers to a single person's data… if a test is reliable, then a given person should get pretty much the same score over repeated replications…(except for random "error" processes)
  - But we can't measure everybody a gazillion times…
  - So, we can conceptualize reliability as something that pertains to a sample of persons instead… by writing it in terms of variances
- $Var(Y) = Var(T) + Var(e) + 2Cov(T, e)$
  $= Var(T) + Var(e)$
- Reliability = Var(T) / Var(Y)
  - Proportion of variance due to "true score" out of total variance

How Only Two Scores Give Us a Reliability Coefficient in CTT

- $Y_1 = T + e_1$
- $Y_2 = T + e_2$

CTT assumptions to calculate reliability:

- Same true score (T) observed at both times
- e_1 and e_2 are uncorrelated with each other and T
- e_1 and e_2 have same variance
- Y_1 and Y_2 have same variance

$$r_{y1,y2} = \frac{\sigma_{y1,y2}}{\sigma_{y1}\sigma_{y2}} = \frac{\sigma_{t+e1,t+e2}}{\sigma_{y1}\sigma_{y2}} = \frac{\sigma_{t,t} + \sigma_{t,e1} + \sigma_{t,e2} + \sigma_{e1,e2}}{\sigma_{y1}\sigma_{y2}} = \frac{\sigma_t^2}{\sigma_y^2}$$

- Same as: Reliability of Y = Var(True) / Var(Y)
- We express unobservable true score variance in terms of the correlation between the two total scores and the variance of the total scores (assumed to be the same across tests)
- We now have an index of how much of the observed variance is "true" (if we believe all the assumptions)

Y = T + e, so how do we get Var(e)?

3 main ways of quantifying **reliability**:

1. Consistency of same test over time
   - Test-retest reliability
2. Consistency over alternative test forms
   - Alternative forms reliability
   - Split-half reliability
3. Consistency across items within a test
   - Internal consistency (alpha or KR-20)

Internal Consistency

- For quantitative items, this is Cronbach's Alpha…
  - Or 'Guttman-Cronbach alpha' (Guttman 1945 > Cronbach 1951)
  - Another reduced form of alpha for binary items: KR 20

- Alpha is described in multiple ways:
  - Is the mean of all possible split-half correlations
  - Is expected correlation with hypothetical alternative form of the same length
  - Is lower-bound estimate of reliability under assumption that all items are tau-equivalent (more about that later)
  - As an index of "internal consistency"
    - Some very much dislike this term (not a measure of "consistency")

Where Alpha Comes From

- The **sum of the item variances** is given by:
  - $Var(I_1) + Var(I_2) + Var(I_3)\ldots. + Var(I_k)$ → just the item variances

- The **variance of the sum of the items** is given by the sum of ALL the item variances and covariances:
  - $Var(I_1 + I_2 + I_3) = Var(I_1) + Var(I_2) + Var(I_3) \ldots + 2Cov(I_1,I_2) + 2Cov(I_1,I_3) + 2Cov(I_2,I_3) \ldots$
  - Where does the '2' come from?
    - Covariance matrix is symmetric
    - Sum the whole thing to get to the variance of the sum of the items

	I_1	I_2	I_3
I_1	σ_1^2	σ_{12}	σ_{13}
I_2	σ_{21}	σ_2^2	σ_{23}
I_3	σ_{31}	σ_{32}	σ_3^2

Cronbach's Alpha

Covariance Version:
k = # items

$$\text{alpha} = \frac{k}{k\text{-}1} \bullet \frac{\text{variance of total Y - sum of item variances}}{\text{variance of total Y}}$$

- Numerator reduces to just the covariance among items
  - ***Sum of the item variances…***
    - Var(X) + Var(Y) = Var(X) + Var(Y) → just the item variances
  - ***Variance of total Y (the sum of the items)…***
    - Var(X+Y) = Var(X) + Var(Y) **+ 2Cov(X,Y) → PLUS covariances**
  - So, if the items are related to each other, the variance of the total Y item sum should be bigger than the sum of the item variances
    - How much bigger depends on how much covariance among the items – the primary index of relationship

Cronbach's Alpha

- **Alpha** is a lower-bound estimate of reliability under assumption that **all items are tau-equivalent :: equally related to true score**

Correlation Version:
k = # items

$$\text{alpha} = \frac{k\bar{r}}{1 + \bar{r}\,(k\text{-}1)}$$

Where $\bar{r}$ is mean inter-item correlation

- You'll note alpha depends on two things (k and r), and thus there are 2 potential ways to make alpha bigger…
  - Get more items and/or increase the average inter-item correlation

- Potential problems:
  - But can you keep adding more items WITHOUT decreasing the average inter-item correlation???
  - Does not take into account spread of inter-item correlation, and thus **alpha does NOT assess dimensionality of the items**

Kuder Richardson (KR) 20: Alpha for Binary Items

- KR20 is actually the more general form of alpha
- From 'Equation 20' in 1937 paper:

$$\mathrm{KR}20 = \frac{k}{k-1}\left(\frac{\text{variance of total Y} - \text{sum of } pq \text{ over items}}{\text{variance of total Y}}\right)$$

k = # items
p = prop. passing
q = prop. failing

- Numerator again reduces to covariance among items…
  - **Sum of the item variances** (sum of pq) is just the item variances
  - **Variance of the sum of the items** has the covariance in it, too
  - So, if the items are related to each other, the variance of the total sum should be bigger than the sum of the item variances
    - How much bigger depends on how much covariance among the items – the primary index of relationship

How to Get Alpha UP

TABLE 1
Values of Cronbach's Alpha for Various Combinations of Different Number of Items and Different Average Interitem Correlations

Number of Items	Average Interitem Correlation					
	.0	.2	.4	.6	.8	1.0
2	.000	.333	.572	.750	.889	1.000
4	.000	.500	.727	.857	.941	1.000
6	.000	.600	.800	.900	.960	1.000
8	.000	.666	.842	.924	.970	1.000
10	.000	.714	.870	.938	.976	1.000

Ta-da! Alpha as Reliability… What could go wrong?

- Alpha does not index **dimensionality** → it does not index the extent to which items measure the same construct

TABLE 13.2. Interitem Correlation Matrices for Two Hypothetical Tests with the Same Coefficient Alpha Reliability of .81

	Test A with 10 items									
Variable	1	2	3	4	5	6	7	8	9	10
1.	—									
2.	.3	—								
3.	.3	.3	—							
4.	.3	.3	.3	—						
5.	.3	.3	.3	.3	—					
6.	.3	.3	.3	.3	.3	—				
7.	.3	.3	.3	.3	.3	.3	—			
8.	.3	.3	.3	.3	.3	.3	.3	—		
9.	.3	.3	.3	.3	.3	.3	.3	.3	—	
10.	.3	.3	.3	.3	.3	.3	.3	.3	.3	—

	Test B with 6 items					
Variable	1	2	3	4	5	6
1.	—					
2.	.6	—				
3.	.6	.6	—			
4.	.3	.3	.3	—		
5.	.3	.3	.3	.6	—	
6.	.3	.3	.3	.6	.6	—

- The *variability* across the inter-item correlations matters, too!
- We will use item-based models (CFA, IRT) to examine dimensionality

Another Problem with Reliability

- Note that the formula for reliability is basically the Pearson correlation
  - Pearson *r* standardizes each variable, so that differences in mean and variance between variables don't matter…
  - So Pearson correlation indexes *relative*, not *absolute* agreement
- But the reliability formula assumes that the mean and variance of the true and observed scores are the same…
  - What if this is not the case?
  - Pearson correlation won't pick this up!
  - A different kind of correlation is needed… **Intraclass correlation**
    - Note: There are LOTS of different versions of these… visit the McGraw & Wong (1996) paper for an overview

Problems with Reliability for Binary Items…

- In binary items, the variance is dependent on the mean
- If two items (X and Y) differ in p, such that $p_y > p_x$:
  - Maximum covariance: $Cov(X,Y) = p_x(1-p_y)$
  - **Maximum correlation will be smaller than -1 or 1:**

$$r_{x,y} = \sqrt{\frac{p_x(1-p_y)}{p_y(1-p_x)}}$$

  - For Example:

px	py		max r
0.1	0.2		0.67
0.1	0.5		0.33
0.1	0.8		0.17
0.5	0.6		0.82
0.5	0.7		0.65
0.5	0.9		0.33
0.6	0.7		0.80
0.6	0.8		0.61
0.6	0.9		0.41
0.7	0.8		0.76
0.7	0.9		0.51
0.8	0.9		0.67

Summary: Reliability in CTT

- Reliability is supposed to be about the consistency of an individual's score over replications… but it's not, really
- Instead, we get 2 scores per person (test-retest; alternate forms) or *k* items for person (alpha), and do:

- **$Y_{Total} = T + E$ or $Var(Y_{Total}) = Var(True) + Var(Error)$**
  - **True** score is an internal characteristic of the person
    - True score variance is assumed to *differ* across samples
  - **Error** is an external characteristic (test + environment)
    - Error variance is assumed to be the *same* across samples
  - ***Reliability is a characteristic of a sample, not of a test***

- Want to improve reliability? Examine the items…
  - Because individual items are not in the CTT measurement model, we have to make assumptions about them instead

RELIABILITY IN IRT

Reliability with IRT

- In IRT, we are focused on a latent variable but the unit of measure is the item, not the test
- Reliability (precision) is a desirable property for a test
- The more reliable a test is, the more precisely we can measure the construct
- For any scaling procedure (IRT or CTT), as reliability goes up, the standard error of measurement goes down

Reliability with IRT

- In CTT, reliability is a one-number summary of test precision, and there is a corresponding single standard error of measurement that is used for any test score

- In IRT, test precision is conceptualized as something called ***Information***, which is conditional on the trait level being measured
  - Some tests could measure certain trait levels very well but measure others poorly…

Reliability with IRT

- A further advantage of IRT with respect to evaluating reliability is that we can consider the amount of ***Information*** an item and/or a test provides

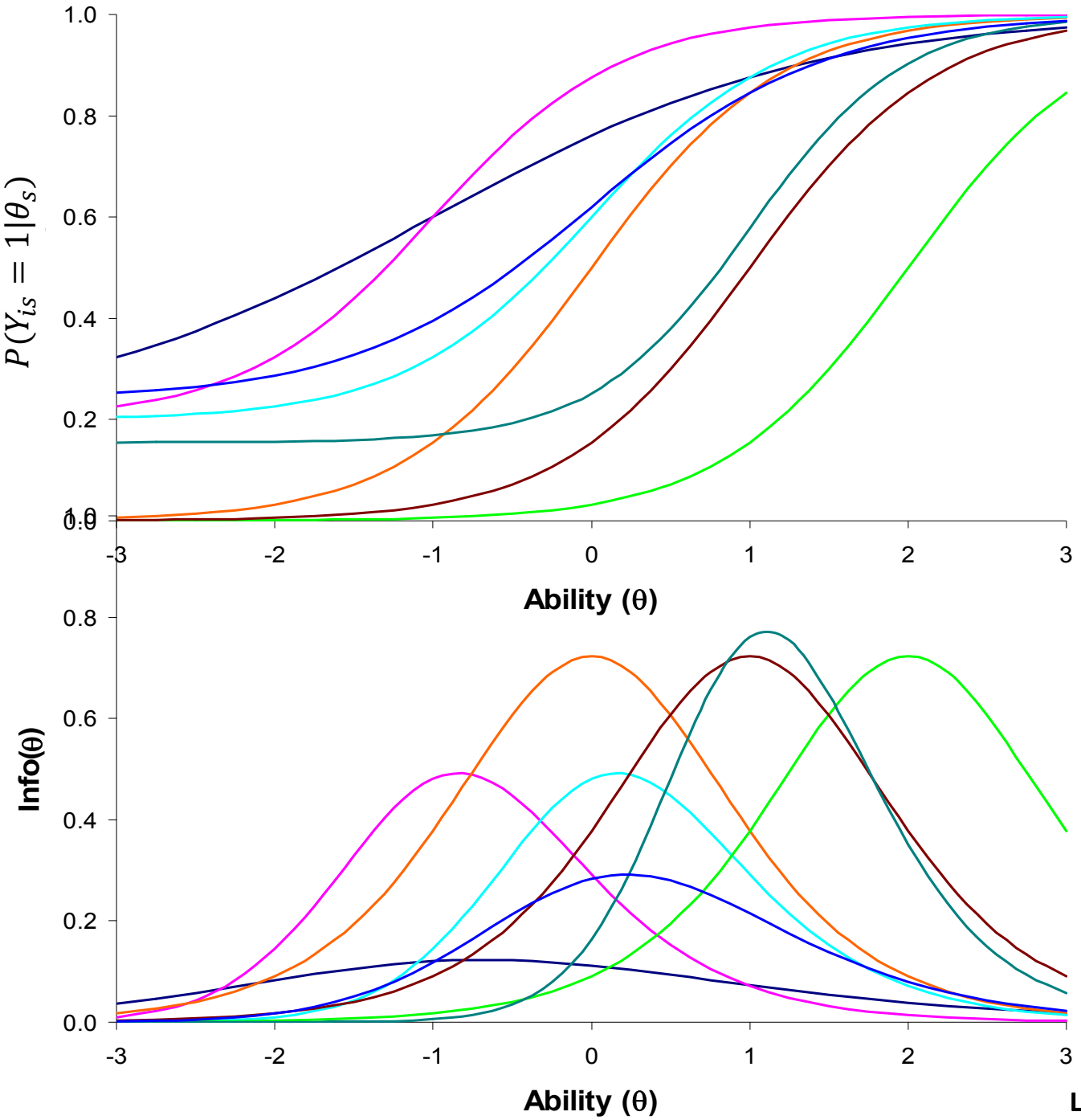
- In CTT, measures of item quality exist, but these are only indirectly related to what the reliability of the test will be
  - Item/Total correlation, for instance

Item Information Function

- "*Item Information*" indicates an item's usefulness for assessing ability

- By "usefulness" we basically mean how good an item is at distinguishing people with lower ability levels from those with higher ability levels

- Information :: Precision

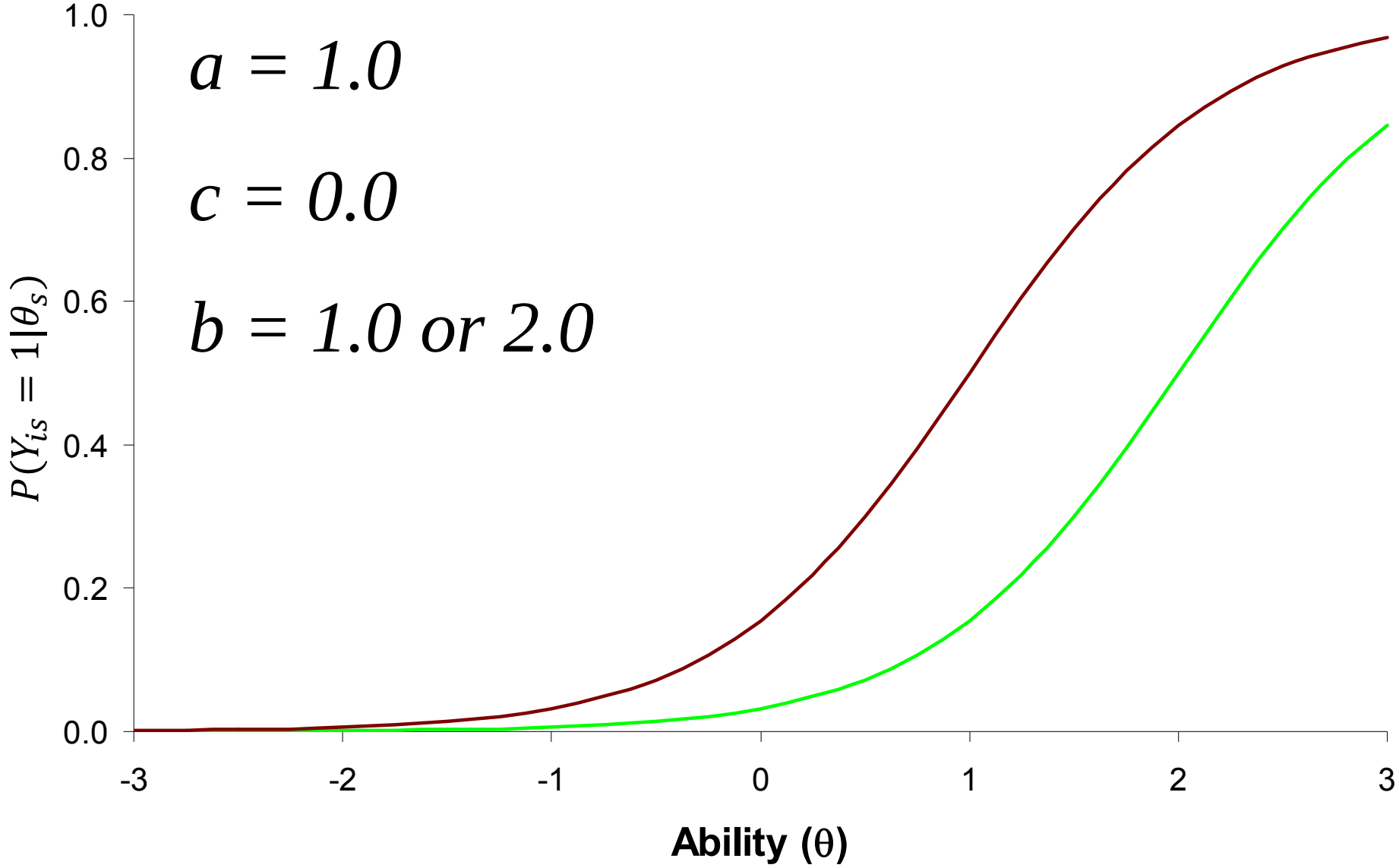

1.0
0.8
0.6
0.4
0.2
0.0
$P(Y_{is} = 1|\theta_s)$
-3
-2
-1
0
1
2
3
Ability (θ)
0.8
0.6
0.4
0.2
0.0
Info(θ)
-3
-2
-1
0
1
2
3
Ability (θ)

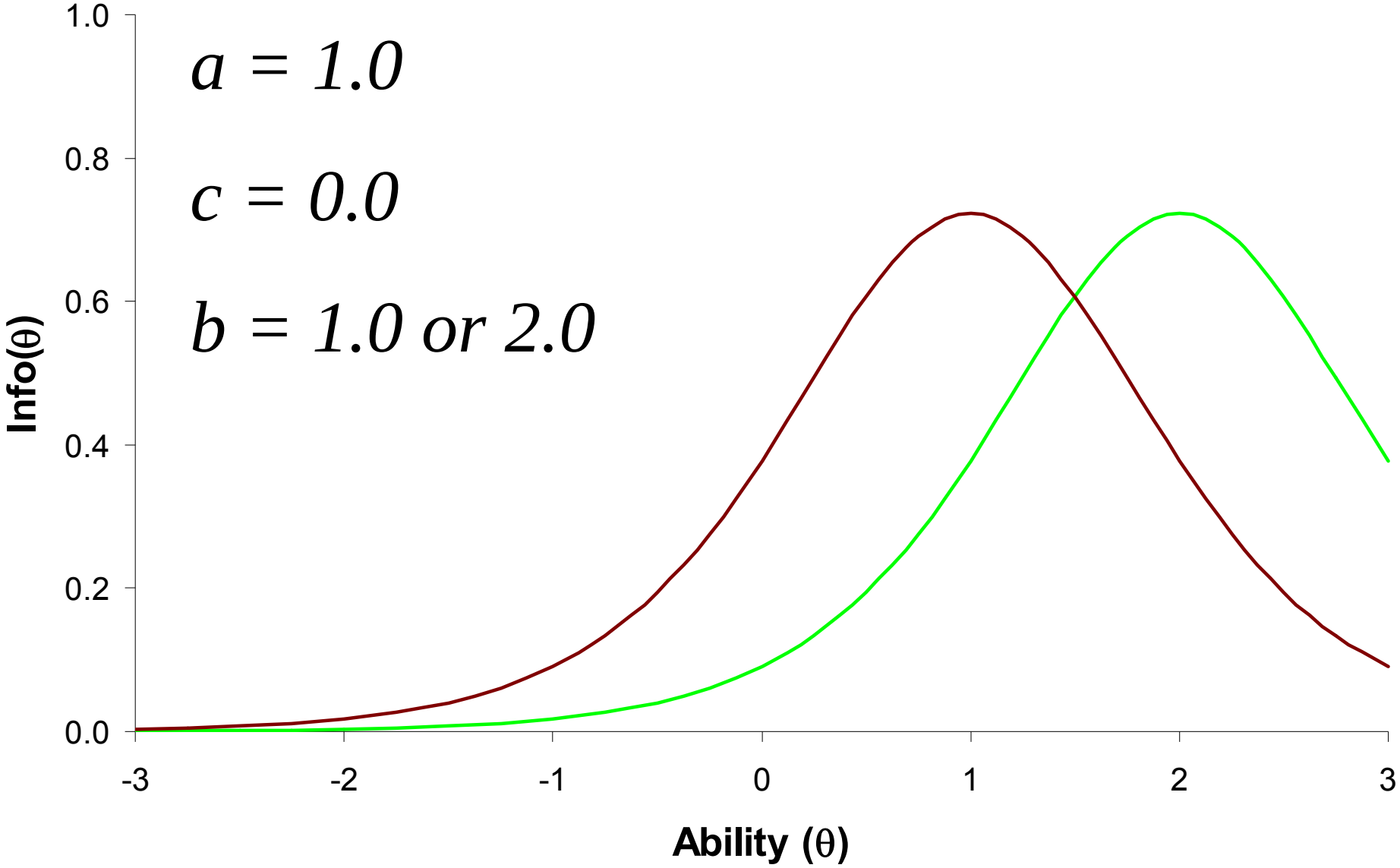
Item Information Function

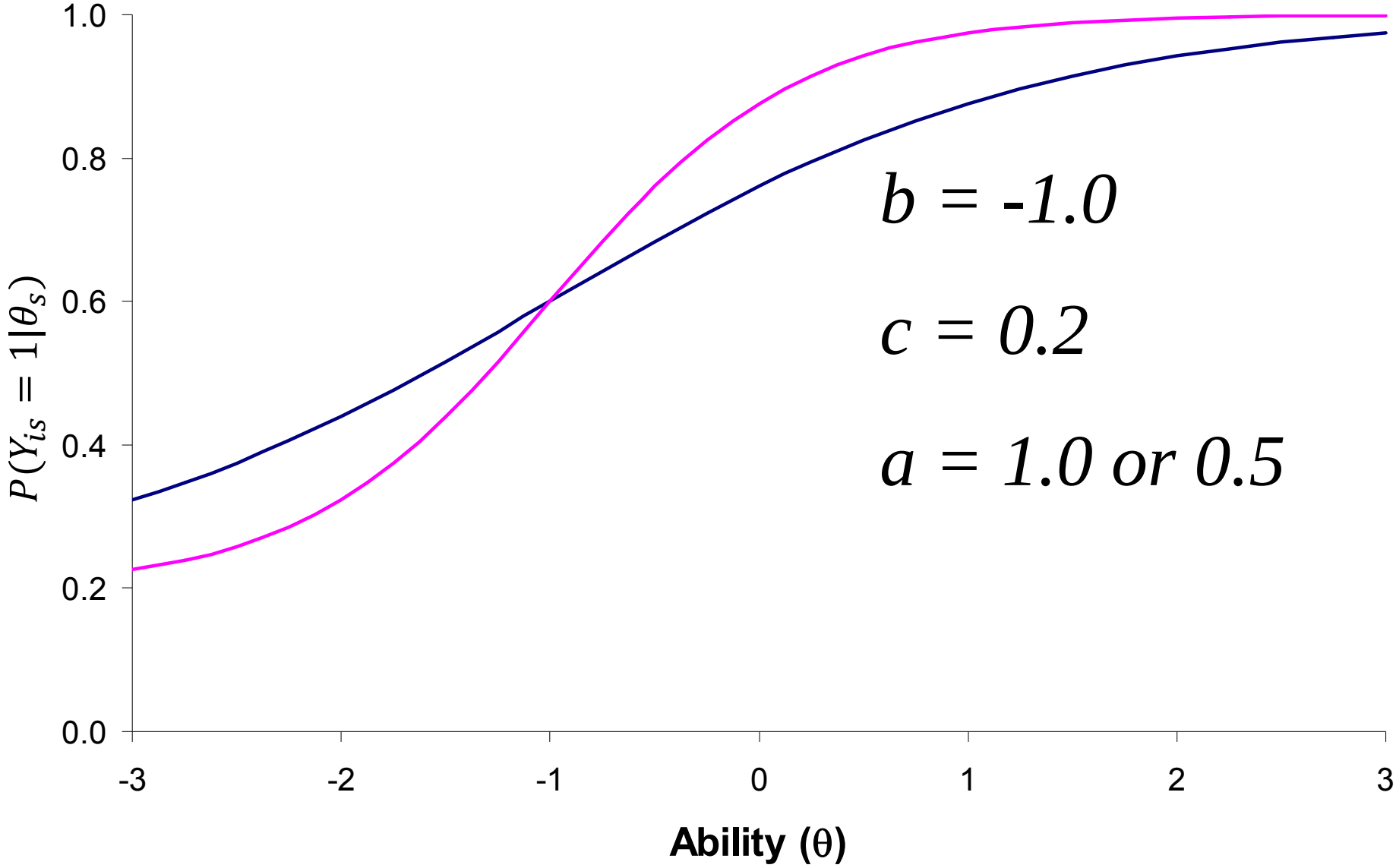
- Items are basically more informative where the slope of the ICC is steepest, which happens when…

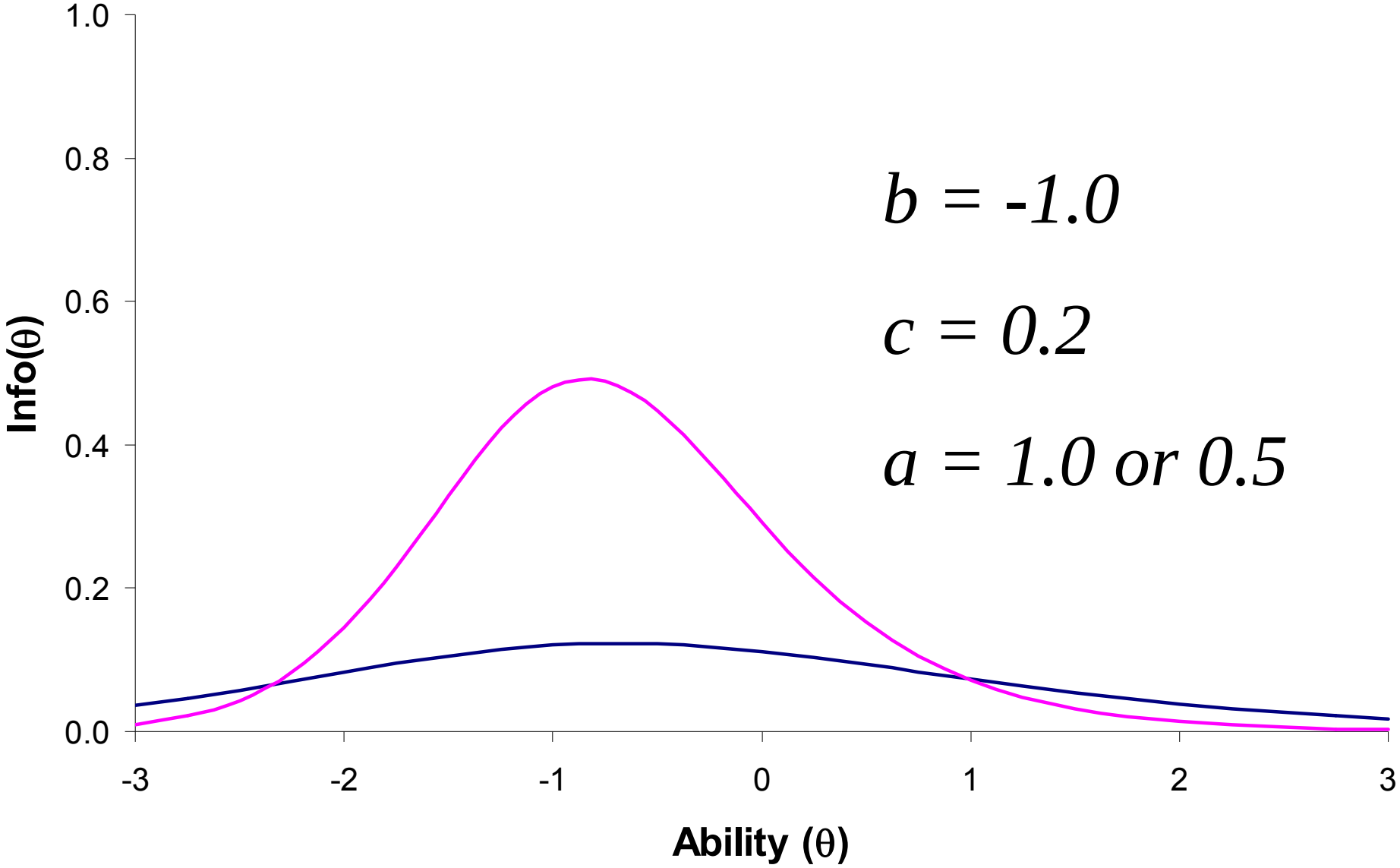
  b_i is relatively close to θ_i,
  a_i is relatively high, and
  c_i is relatively low

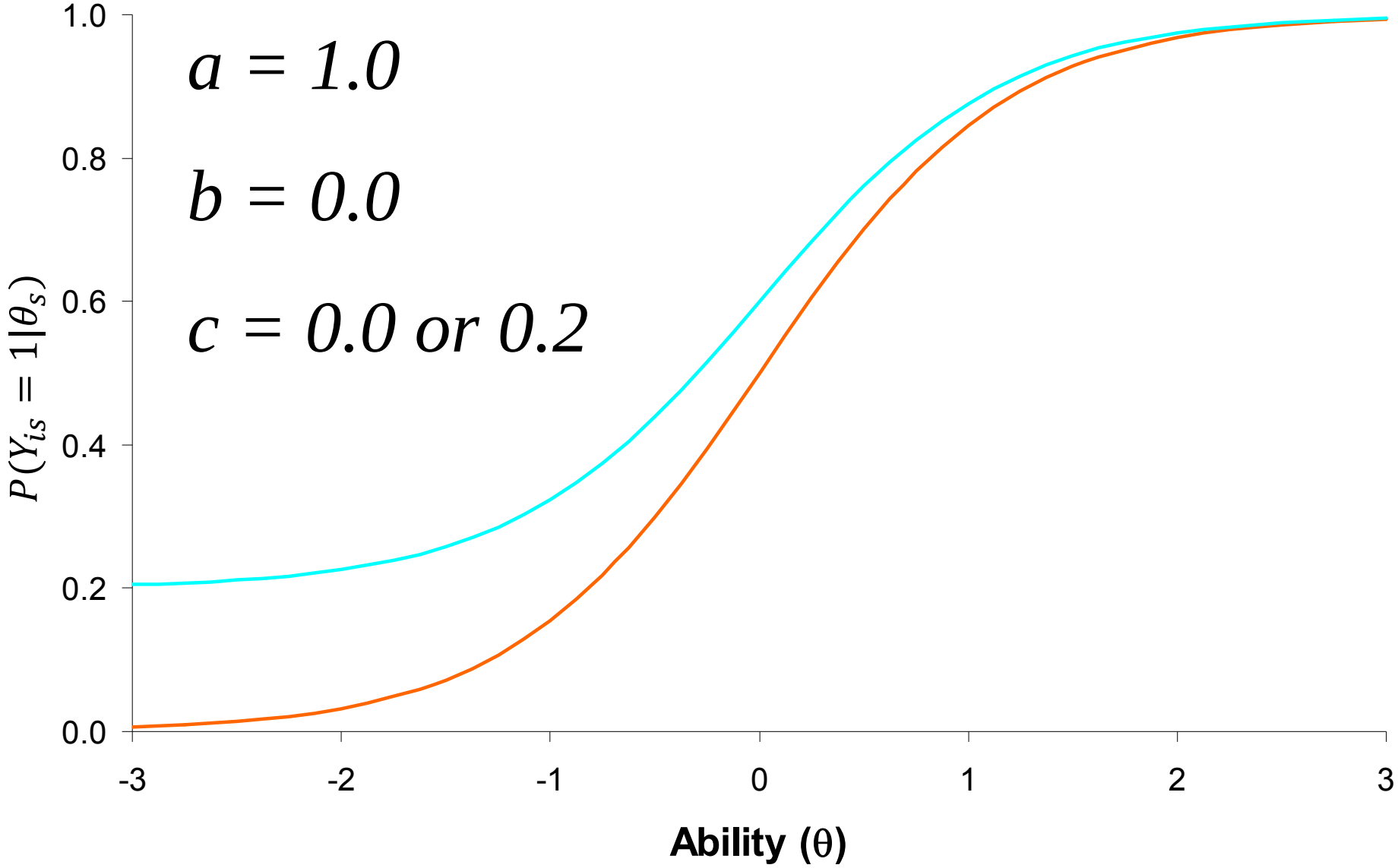
- If $c_i = 0$, an item provides its maximum information when $\theta_s = b_i$

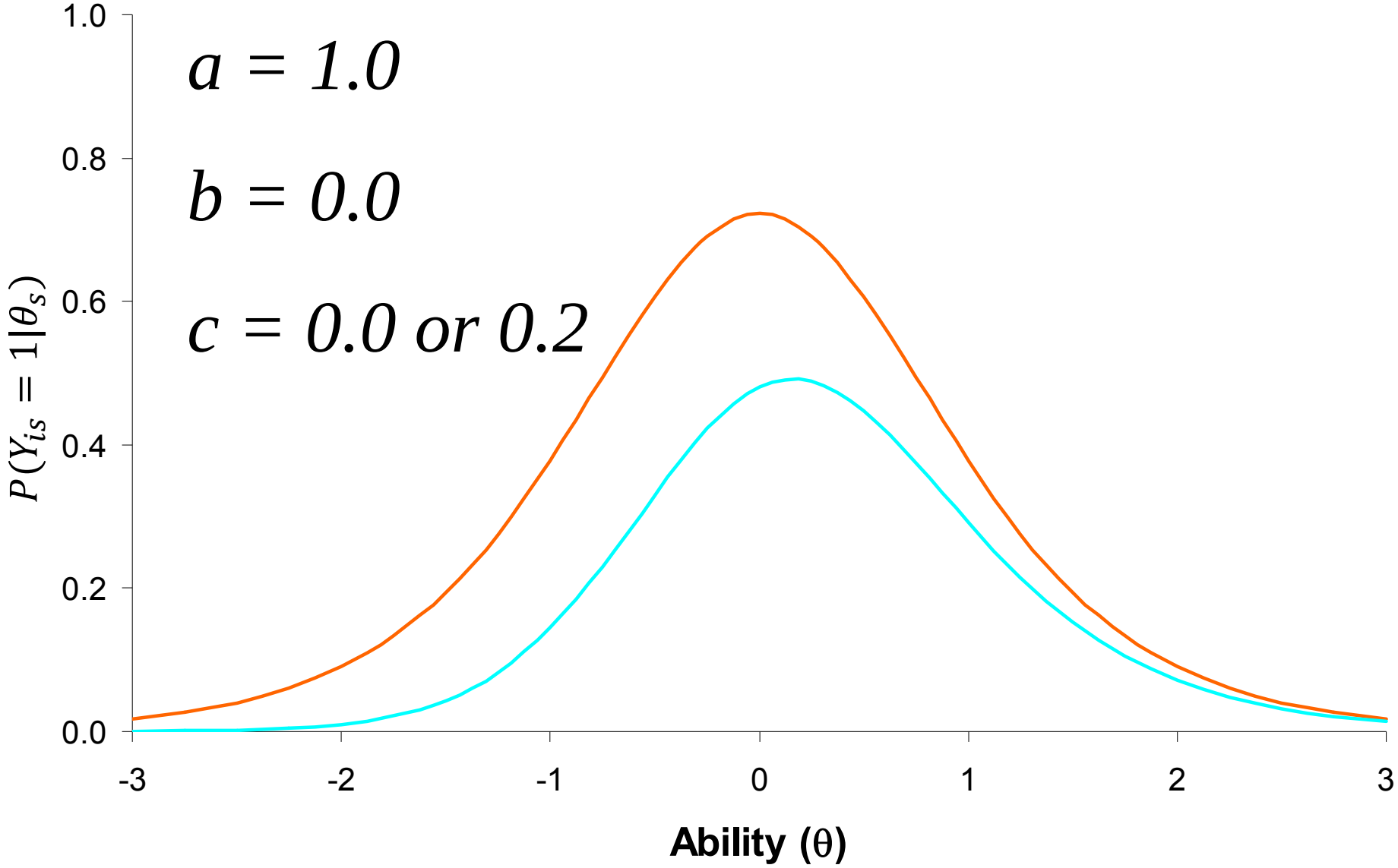

$a = 1.0$
$c = 0.0$
$b = 1.0$ or 2.0
$P(Y_{is} = 1|\theta_s)$
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
Ability (θ)


a = 1.0
c = 0.0
b = 1.0 or 2.0
Info(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
Ability (θ)


1.0
0.8
0.6
0.4
0.2
0.0
$P(Y_{is} = 1|\theta_s)$
-3
-2
-1
0
1
2
3
Ability (θ)
b = -1.0
c = 0.2
a = 1.0 or 0.5


1.0
0.8
0.6
0.4
0.2
0.0
Info(θ)
-3
-2
-1
0
1
2
3
Ability (θ)
b = -1.0
c = 0.2
a = 1.0 or 0.5

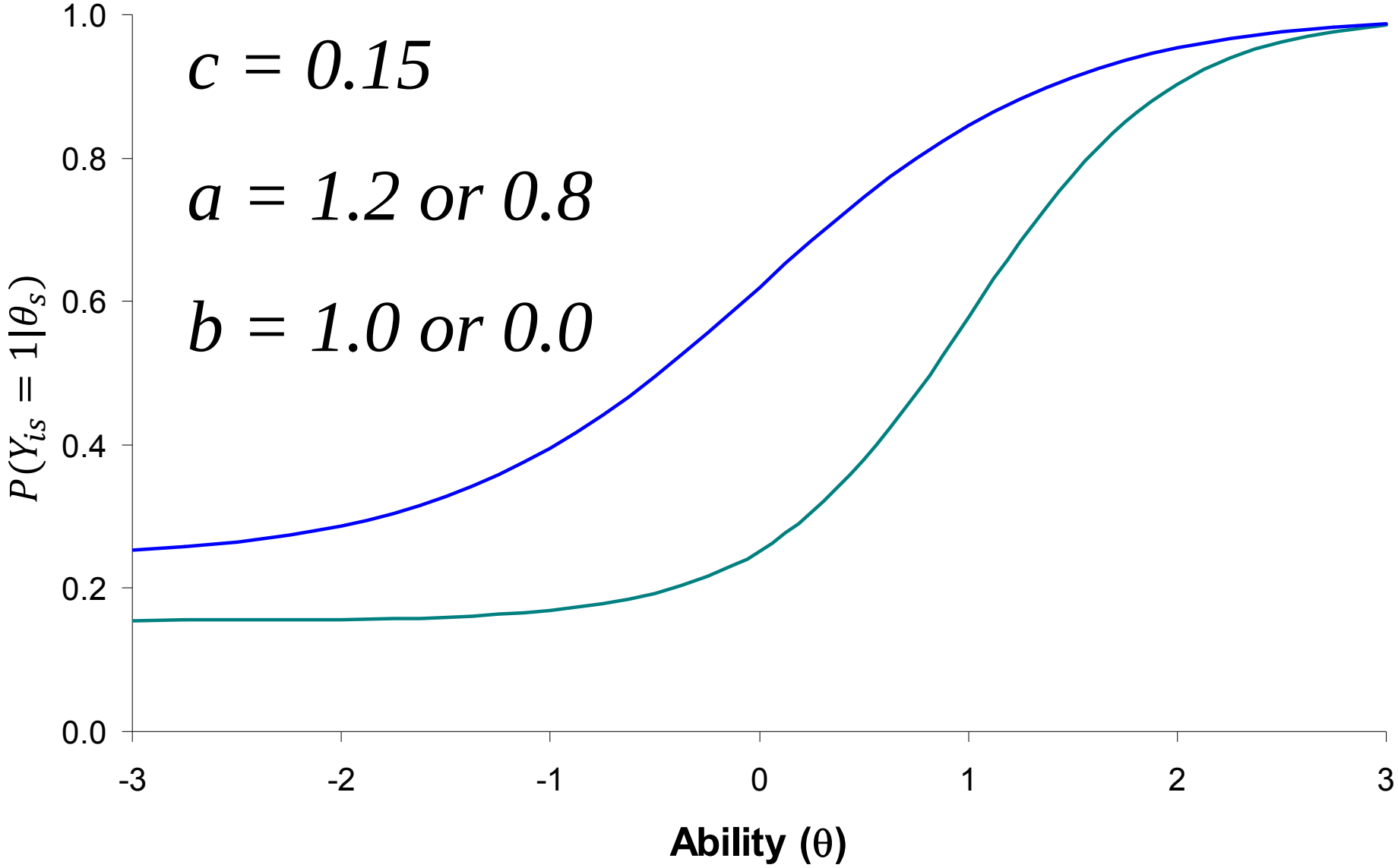

$a = 1.0$
$b = 0.0$
$c = 0.0$ or 0.2
1.0
0.8
0.6
0.4
0.2
0.0
$P(Y_{is} = 1|\theta_s)$
-3
-2
-1
0
1
2
3
Ability (θ)

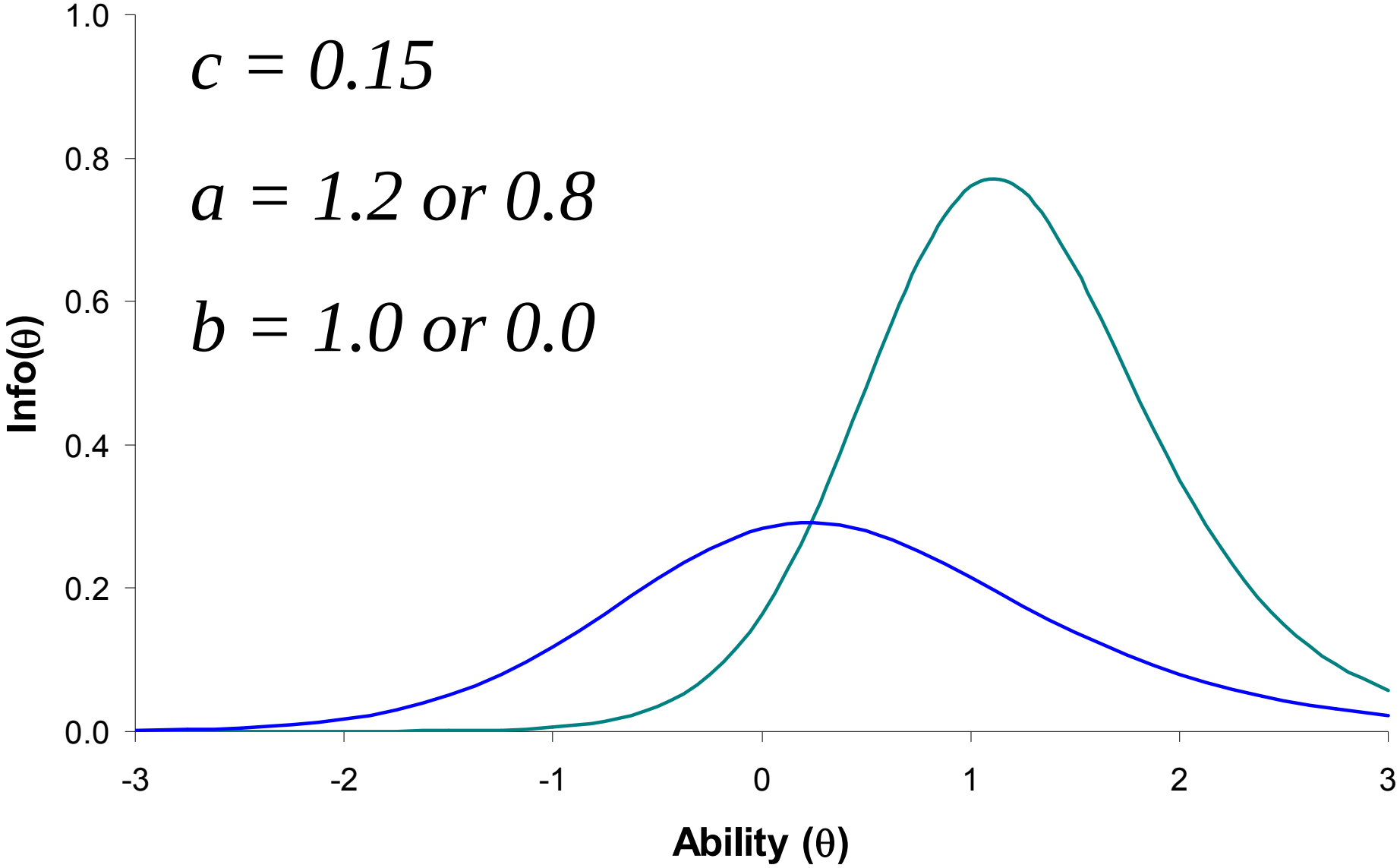

$a = 1.0$
$b = 0.0$
$c = 0.0$ or 0.2
$P(Y_{is} = 1|\theta_s)$
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
Ability (θ)

Item Information Function

- IMPORTANT: information is a ***function*** of θ, which means that an item could be very informative for some ability levels and relatively uninformative for others

- Example: difficult items are informative for higher ability levels, but don't tell us much about lower ability levels (because they mostly get all those items wrong!).


$c = 0.15$
$a = 1.2$ or 0.8
$b = 1.0$ or 0.0
$P(Y_{is} = 1|\theta_s)$
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
Ability (θ)


c = 0.15
a = 1.2 or 0.8
b = 1.0 or 0.0
Info(θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
Ability (θ)

Item Information Function for the 3-PL

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)(1-P_i(\theta))} =$$

$$\frac{1.7^2 a_i^2 (1 - c_i)}{\left[c_i + \exp(1.7a_i(\theta - b_i))\right]\left[1 + \exp\left(-1.7a_i(\theta - b_i)\right)\right]^2}$$

Notes on IIF

- The roles of a_i and c_i are:
  - as a_i increases, information increases
  - as c_i increases, information decreases

- As ability moves away from b_j (+ or -) the denominator increases, so information approaches zero

Maximum Information

$$\theta_{max} = b_i + \frac{1}{1.7a_i}\ln\left[.5\left(1+\sqrt{1+8c_i}\right)\right]$$

If $c_i = 0$, then Information is maximized at b_i

If $c_i > 0$, then Information is maximized at an ability level slightly greater than b_i

TEST (THETA) INFORMATION

Test Information Function

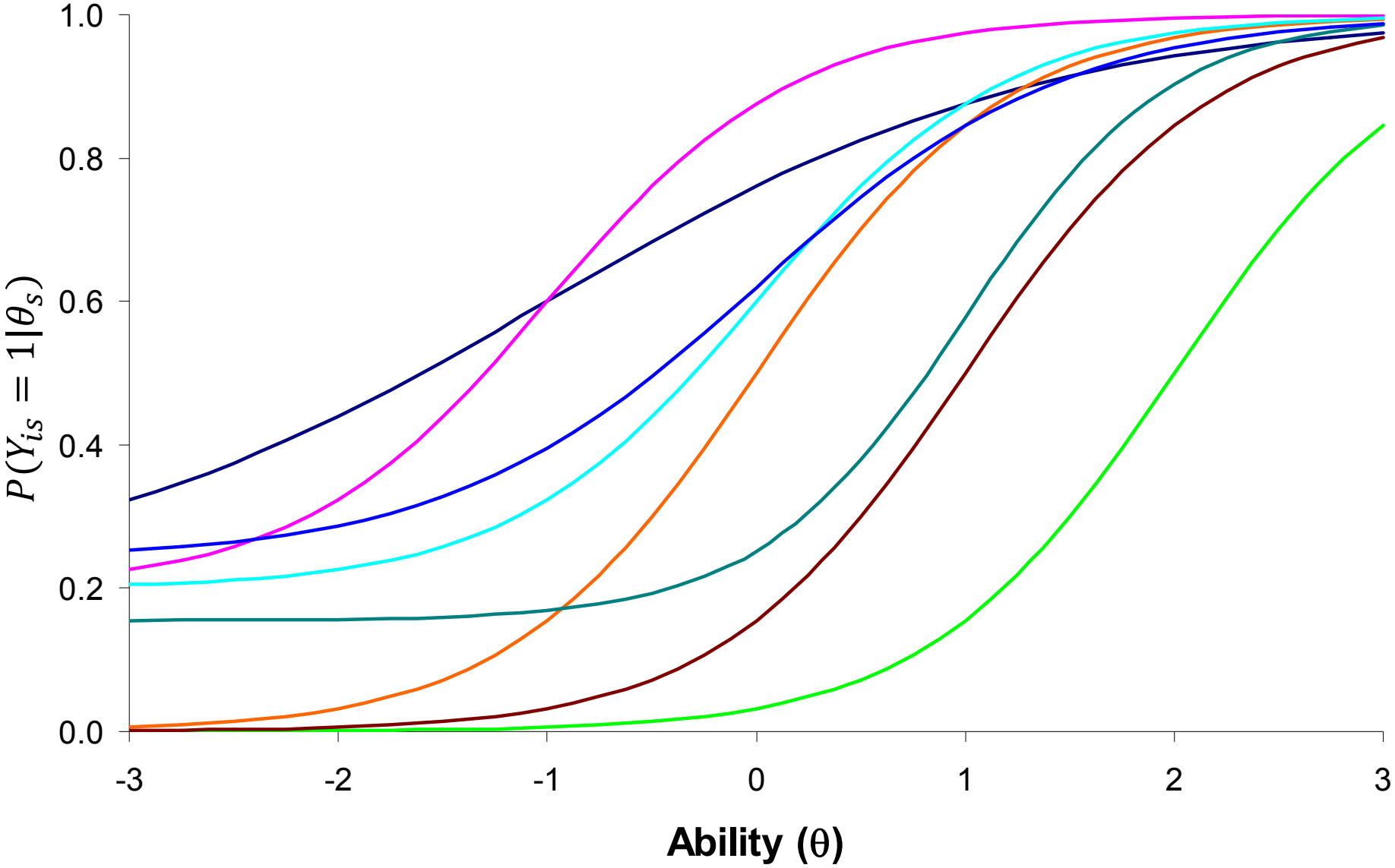
- Just like we add up ICCs to get a TCC, we add up IIFs to get a TIF
- Information will continue to increase as we add test items, therefore increasing precision
- All things equal, longer tests provide increased measurement precision

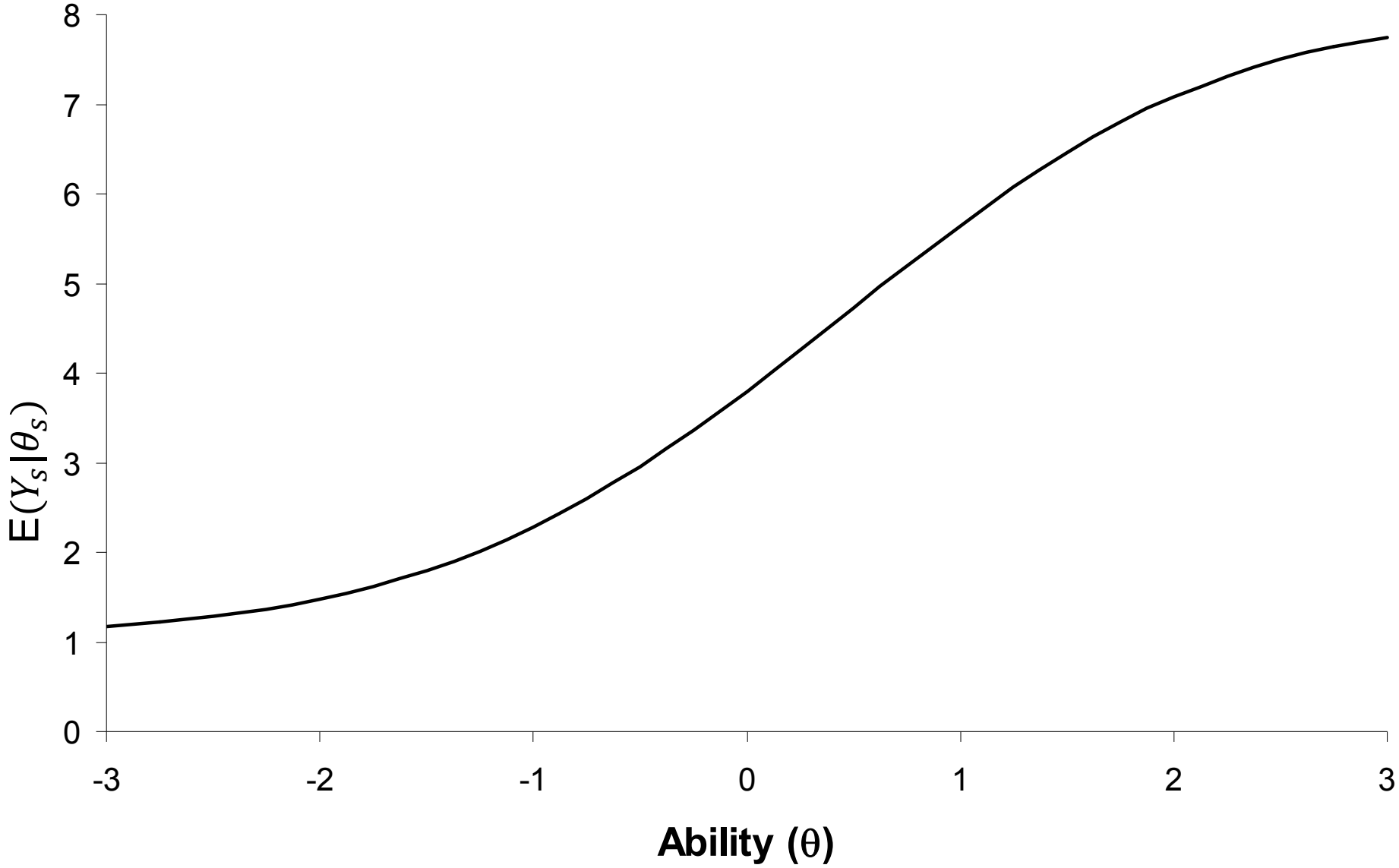
Test Information Function

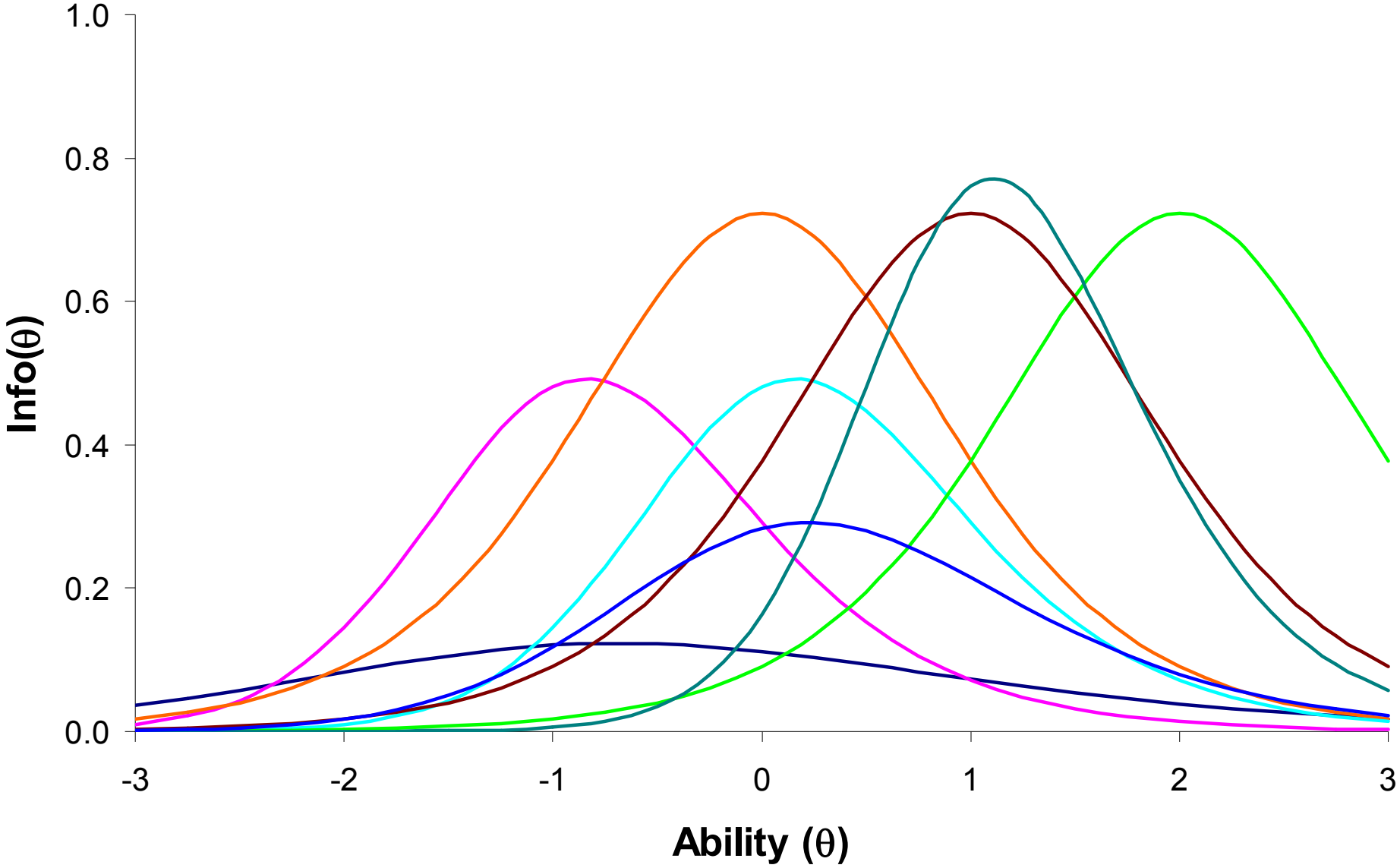
- Defined for a set of items at each point along the ability (θ) scale

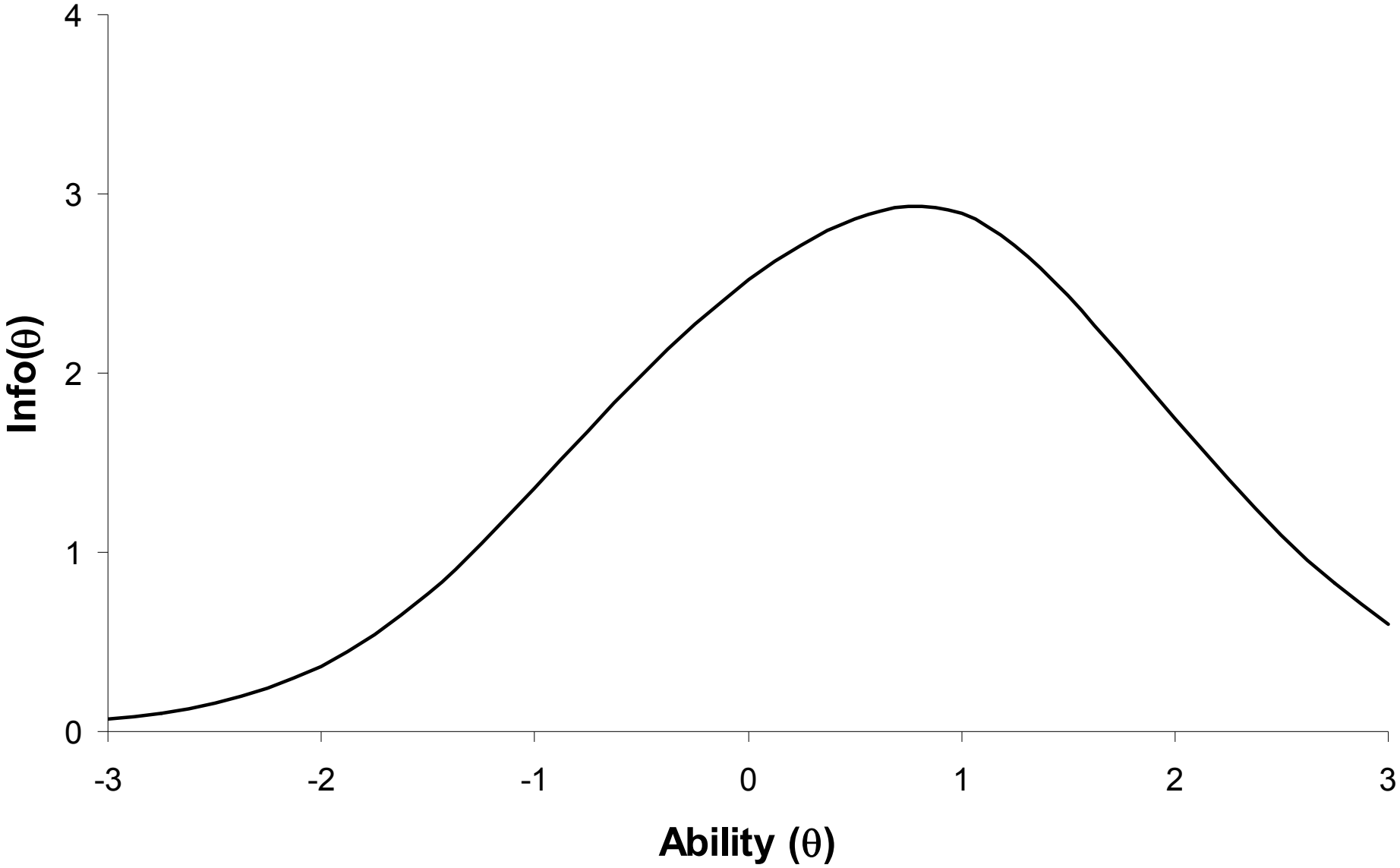
- Test information is influenced by the 'quality' and the number of test items:
  - *I* = total number of test items
  - *i* = item index
  - *I( )* = test information function

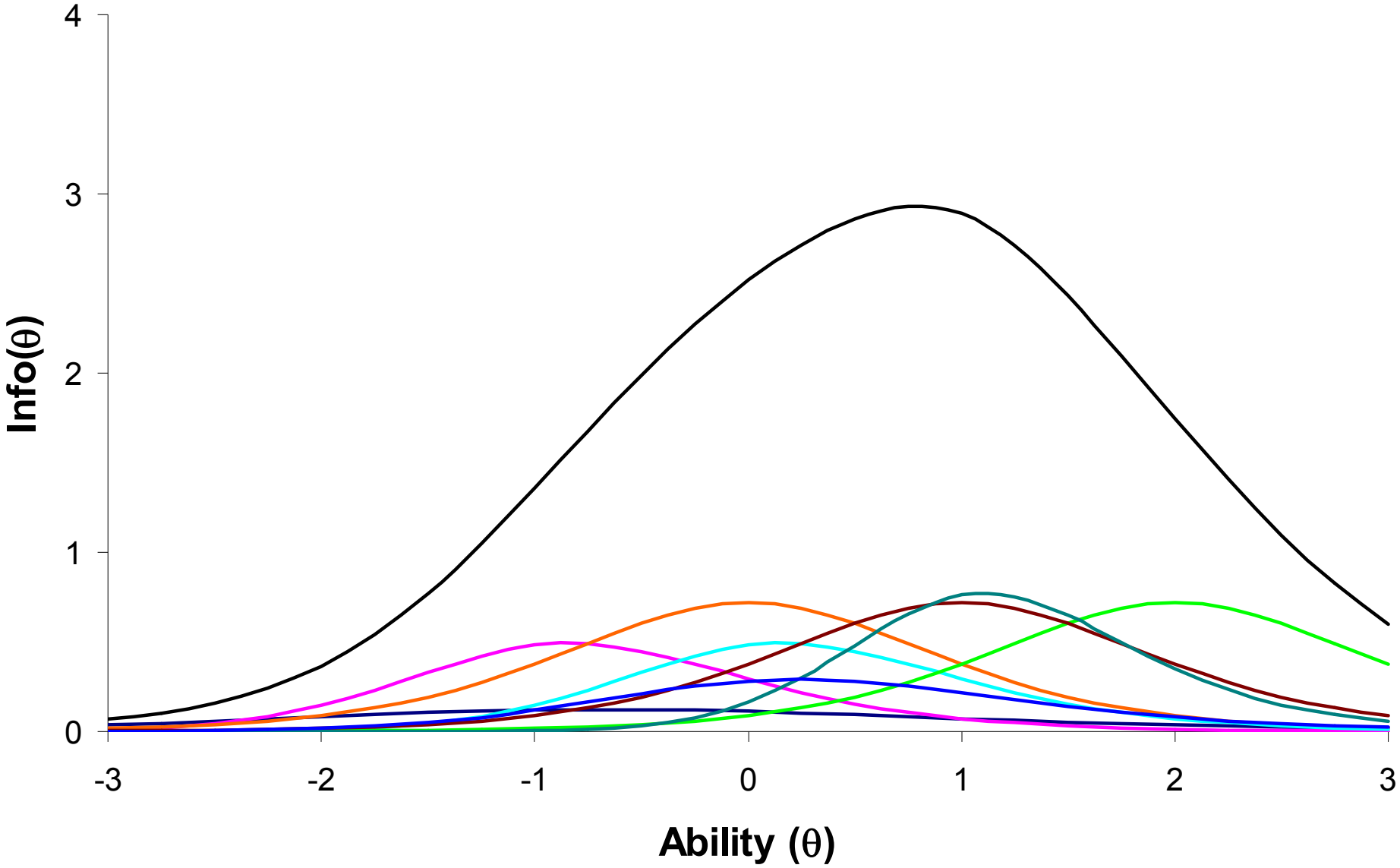
$$I(\theta) = \sum_{i=1}^{I} I_i(\theta)$$


1.0
0.8
0.6
0.4
0.2
0.0
$P(Y_{is} = 1|\theta_s)$
-3
-2
-1
0
1
2
3
Ability (θ)


8
7
6
5
4
3
2
1
0
$E(Y_S|\theta_S)$
-3
-2
-1
0
1
2
3
Ability (θ)


1.0
0.8
0.6
0.4
0.2
0.0
Info(θ)
-3
-2
-1
0
1
2
3
Ability (θ)


4
3
2
1
0
Info(θ)
-3
-2
-1
0
1
2
3
Ability (θ)


4
3
2
1
0
Info(θ)
-3
-2
-1
0
1
2
3
Ability (θ)

Conditional Error for ML Estimates

- Measurement precision and error are considered conditional on θ
- Standard error of an MLE is: $SE(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}}$
- The ***imprecision*** of ability estimation is therefore inversely related to the amount of ***Information*** with respect to ability that is available
- Since Information increases with the quality and number of items, the SE conversely decreases…which hopefully makes some sense!

Information vs. Reliability

- In terms of Reliability (for standard mean zero variance/one thetas):

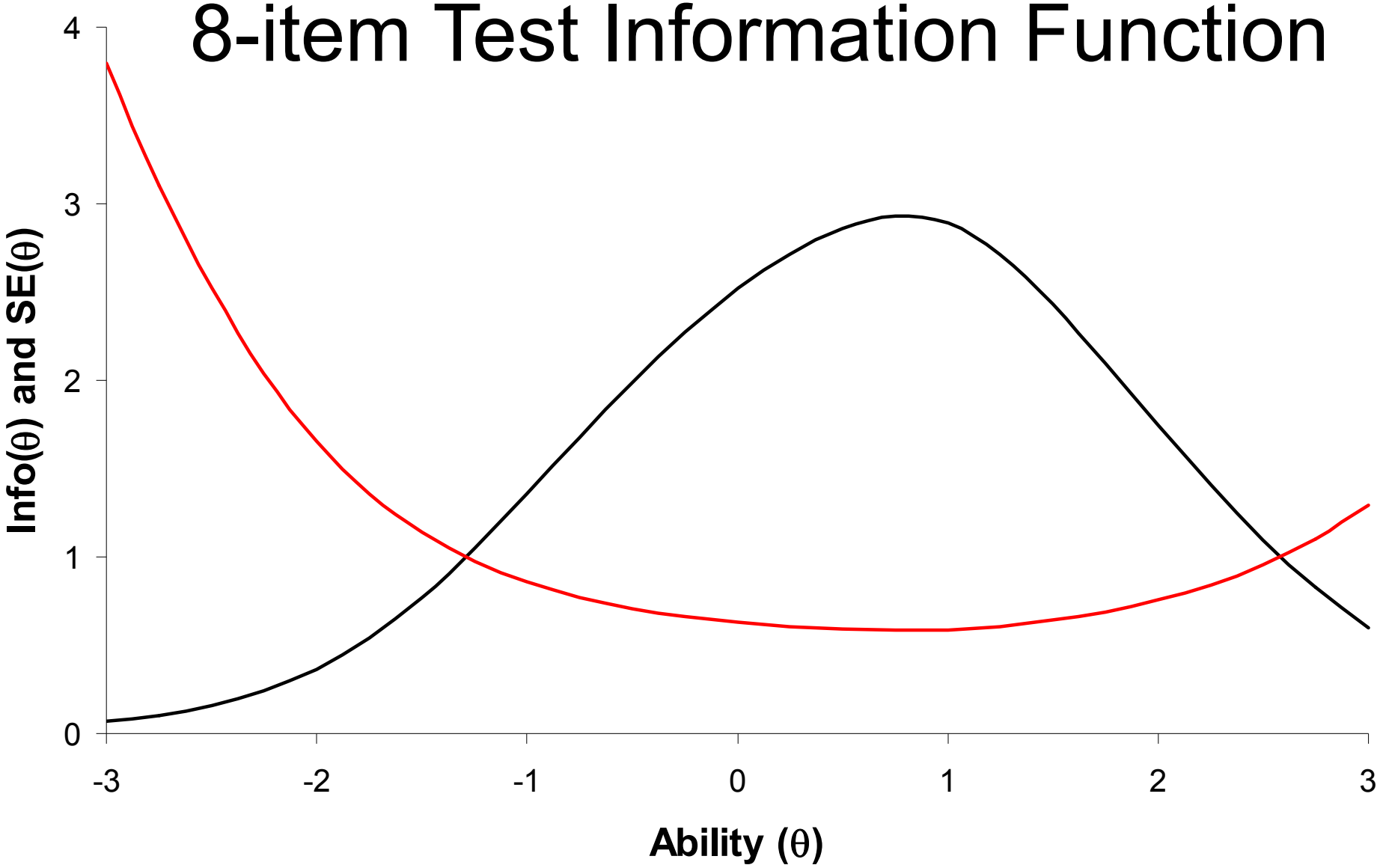
$$\text{Reliability} = \rho(\hat{\theta}) = \frac{I(\hat{\theta})}{I(\hat{\theta}) + 1}$$

- This comes from the classical definition of reliability (only with theta representing the "true score" of a person):

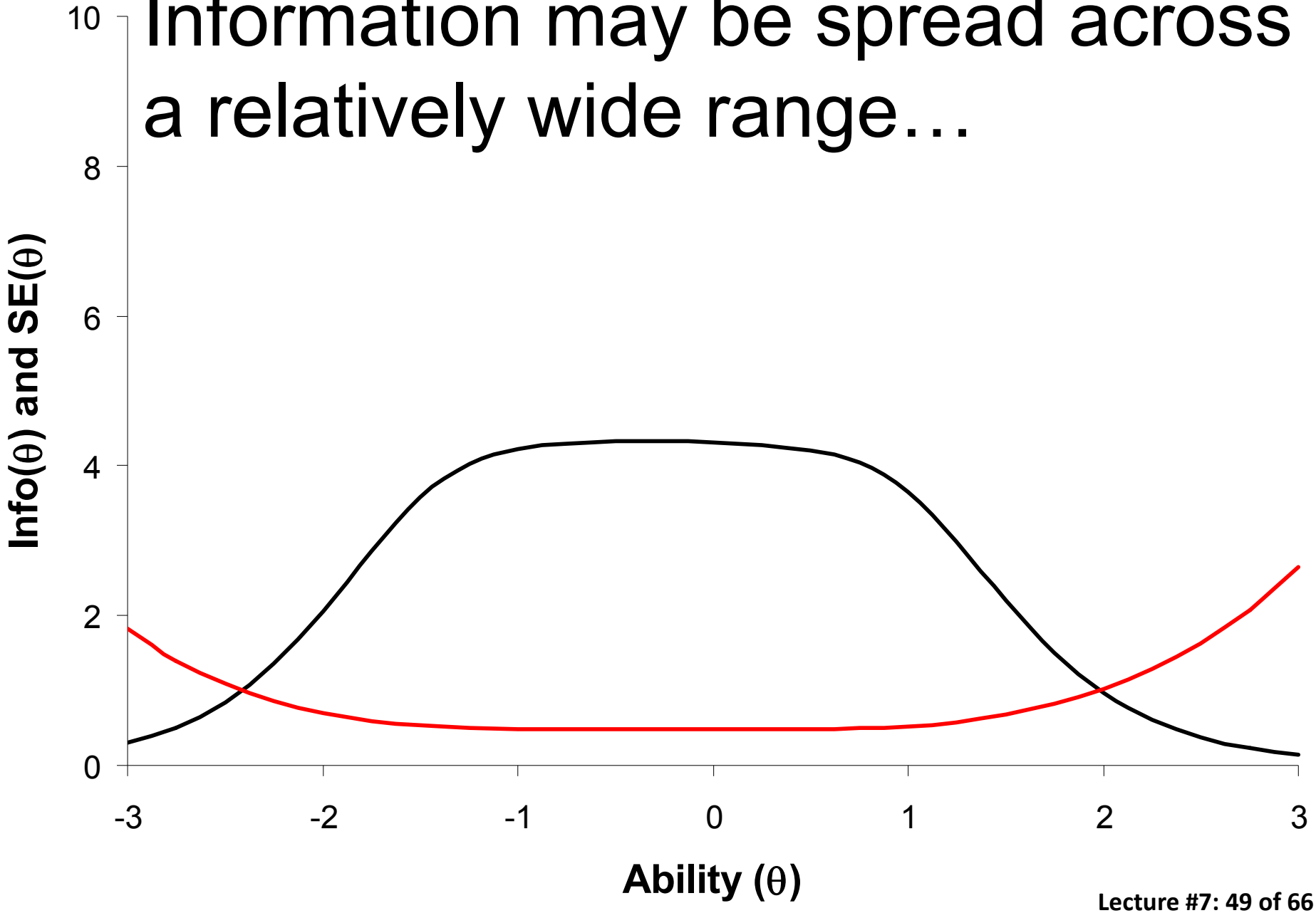
$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

- Here σ_T^2 is the variance of the estimate of theta (the true score here); σ_E^2 is the variance of error (the overall population variance for the true score)

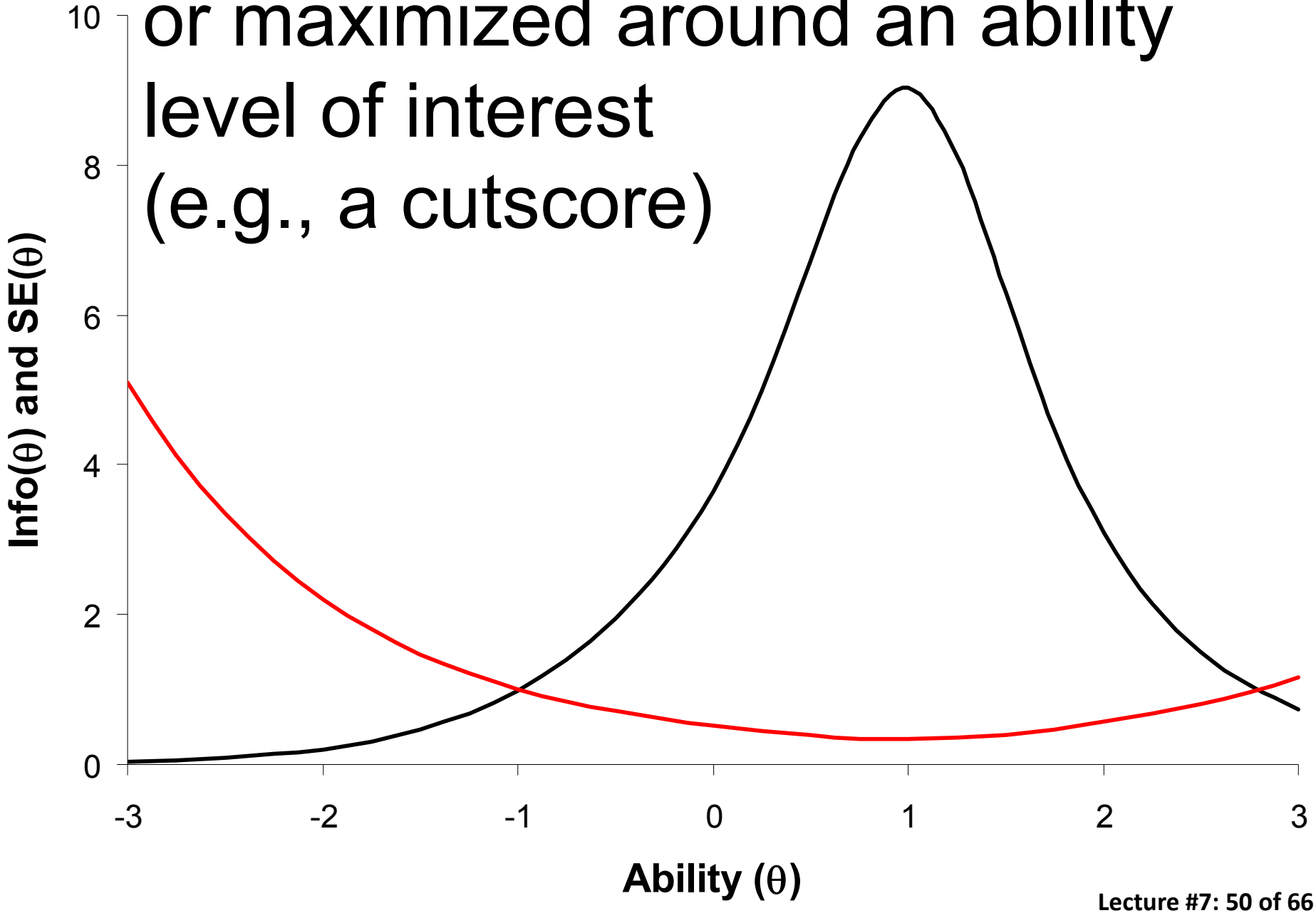
8-item Test Information Function



Information may be spread across a relatively wide range…



or maximized around an ability level of interest (e.g., a cutscore)



Info and SE Example

$$\text{At} \qquad \theta = 1.0, \qquad I(\theta = 1) = 9$$

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}} = \frac{1}{\sqrt{9}} = 0.3\overline{3}$$

$$\text{If } \hat{\theta} = 1.0, \qquad SE(\hat{\theta}) = 0.33$$

Info and SE Example

$$\text{At} \qquad \theta = 0.0, \qquad I(\theta = 0) = 3$$

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}} = \frac{1}{\sqrt{3}} = 0.58$$

$$\text{If } \hat{\theta} = 0.0, \qquad SE(\hat{\theta}) = 0.58$$

Info and SE Example

$$\text{At} \qquad \theta = -1.0, \qquad I(\theta = -1) = 1$$

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}} = \frac{1}{\sqrt{1}} = 1.0$$

$$\text{If } \hat{\theta} = -1.0, \qquad SE(\hat{\theta}) = 1.0$$

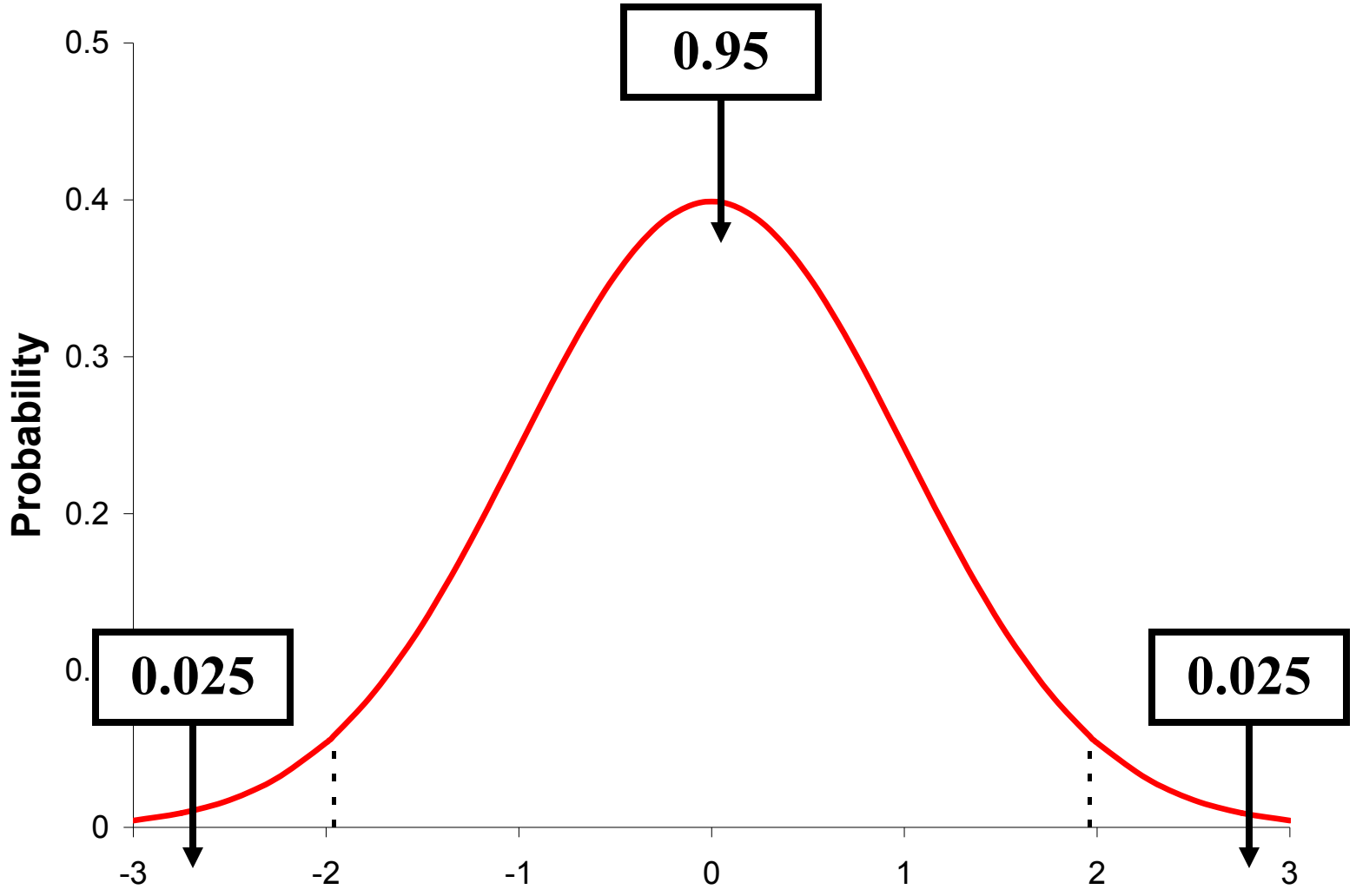
95% Confidence Interval

- Because MLEs are asymptotically normally distributed, we create a 95% confidence interval around a point estimate of ability by adding and subtracting 1.96 standard errors:

$$\text{Estimate} \pm 1.96 \text{ SE}$$

(recall critical values from a standard normal distribution)

Standard Normal Distribution



95% Confidence Interval

- For $\theta = 1$, SE=0.33 → 1.0 ± 0.65
  - 95% chance that examinee's true ability is in between 0.35 and 1.65
- For $\theta = 0$, SE=0.58 → 0.0 ± 1.14
  - 95% chance that examinee's true ability is in between -1.14 and 1.14
- For $\theta = -1$, SE=1.0 → -1.0 ± 1.96
  - 95% chance that examinee's true ability is in between -2.96 and 0.96

95% Confidence Interval

- As information increases…
  - SE decreases
  - CI becomes narrower
  - Increased trust in ability estimate
- As information decreases…
  - SE increases
  - CI becomes wider
  - Decreased trust in ability estimate

Notes on IIF and TIF

- Note that the contribution of $I_i(\theta)$ to $I(\theta)$ does not depend on the particular *combination* of test items
  - Each item contributes independently

- This is a very big advantage of IRT over CTT: reliability can be described conditionally (as information), and it does not depend on the particular set of items

Mini-CTT lesson

- In CTT, item discrimination (quality) is the item-total correlation
- This will depend on the item itself, but is also influenced by the other test items
- Adding items changes the total score, thus changing the correlation
- Therefore, it is difficult to anticipate the reliability of a test when creating a form from a bank of previously piloted items, unless those items all appeared together

CTT versus IRT

- In IRT, item quality is ***Information***, which is affected by a_i, b_i, c_i, and θ
- An item's information function will be independent of the other items on the test, as will its contribution to the TIF
- Adding more and/or better items will increase TIF, but won't impact any IIF
- It is ***easy*** to anticipate the reliability of a test when creating a form from a bank of previously piloted items

Excel Spreadsheet Demo

- Show Excel Spreadsheet containing eight items, their ICCs, TCC, IIFs, TIF and SE

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

Uses of Item and Test Information Functions

1) Providing conditional SE of trait
2) Building a test to meet desired statistical specifications
3) Revising an existing test
4) Comparing tests

WRAPPING UP

Concluding Remarks

- Reliability in IRT is built upon item information

- Item information tells us
  - Where items are best at measuring a latent trait

- Item information becomes test information when aggregated across items of a test
  - Which can be used for building a test of a given target
  - Is used in many applications

A Return to the Example From Practice

- From the *Graduate Record Examinations® Guide to the Use of Test Scores* (2010-2011; p. 20)
  - http://www.ets.org/s/gre/pdf/gre_guide.pdf

Table 6A: Conditional Standard Errors of Measurement at Selected Scores for General Test Measures*

Measure	200	250	300	350	400	450	500	550	600	650	700	750	800
Verbal	14	21	26	28	31	35	34	33	33	33	34	32	20
Quantitative	26	42	48	55	55	54	50	49	42	39	35	26	9

Up Next…

- Putting item and test information to good use:
  - Test Development