Test Development with IRT

Lecture #8

ICPSR Item Response Theory Workshop

Lecture Overview

- A discussion of scale building:
  - From the basics of psychometrics…
  - …to scale building using a model-based approach (relying on IRT)

SCALE DEVELOPMENT

Practical Problems in Measurement

- To demonstrate the types of issues we will discuss related to test development and evaluation, consider the following two examples of measurement:
  1. A teacher wishing to evaluate student knowledge of math
  2. A psychologist wishing to measure depression

- Note the common denominator here is not topic, but rather than each person is trying to assess a **latent trait**
  - These concerns apply any time you are trying to do that, regardless of what the trait is
  - The key issue that is often overlooked is the nature of the trait
    - Is it a scale? Is it measurable? Does experience say it functions the way you think it does?

Example #1 – The Math Teacher

- A teacher constructs 20 pass/fail items for a math test that covers algebra and geometry, administers the test, and adds up the number of correct items to use as the math score for each student

- In doing so, the teacher wonders…
  - Should there be one score or two scores for math ability?
    - One score for geometry items AND one score for algebra items?
    - If so, what about items that require both algebra and geometry?
  - If one score is sufficient…
    - How accurate is that single score as a measure of math ability?
    - How accurate would two scores be?
  - Are 20 items sufficient to give a reasonably accurate determination of each student's knowledge?
    - Should more be used? Could fewer have been used?

Questions about Questions…

- Are all items good measures of math ability or are some items better than others? Are there other ways of getting the right answer besides ability?
- If different items had been used, would they have measured the same thing?
  - Equally well? Can two tests be made (with different items) so that the scores are interchangeable? Could a computer be used to administer the test adaptively?
- Are students who have low scores measured as accurately as students scoring highly or in the middle?
  - Test floor? Test ceiling?
- Are the items free from bias when given to students of different cultural backgrounds? In different languages?
  - Could some students have irrelevant problems with certain items because of differences in their background and experience?
  - How would we be able to know?

Example #2 – The Psychologist

- A clinical psychologist writes a set of items to measure depression, with 5 options ranging from “rarely” to “almost always” such as:
  - “I have lots of energy.”
  - “I sometimes feel sad.”
  - “I think about ending my life.”
  - “I cry.”

- The psychologist may have similar questions about measurement…
  - Dimensionality of traits to be measured?
  - Overall accuracy and efficiency of measurement?
  - Item quality, exchangeability, and bias?
  - Reliability across trait levels?
  - Do positively and negatively worded items measure same trait?
  - Are all ‘almost always’ responses created equal?

A Non-Exhaustive List of Potential Worries in Test Construction…

- Dimensionality of traits and items:
  - How many traits are you measuring?
- Overall test accuracy vs. efficiency
  - Do you need to add or remove items?
  - Add or remove response options?
  - Just any items? Or targeted items?
- Reliability across trait levels
  - Avoid ceiling and floor effects
  - Customize test for specific measurement purposes
- Bias and generalizability across populations:
  Does your test 'work' for different groups?
  - Sufficiently unbiased?
  - Sufficiently sensitive for groups with different ability levels?

Defining Constructs
(adapted from *Constructing Measures*, Wilson, 2005)

- Purpose of measurement:
  - Provide a reasonable and consistent way to summarize the responses that people make to express their abilities, attitudes, etc. through tests, questionnaires, or other types of scales
- Classical definition of measurement:
  - "process of assigning numbers to attributes"
  - But important steps precede and follow this part!
- All measurement begins with a *construct*, or unobserved (latent) trait, ability, or attribute that is the focus of study
  - i.e., the 'true score' in CTT, 'factor' in CFA, or 'theta' in IRT

Defining Constructs, continued

- The models we'll utilize each assume the construct to be a *unidimensional* and *continuous* latent variable
  - Wilson (2005) calls this a 'construct map'
  - If not strictly unidimensional, try to think of sub-constructs that would be unidimensional, and focus efforts on each one of those
  - Qualitative distinctions (benchmarks) are ok as a means of *description*, but should be continuous in between those points

- Constructs made up of categorical latent 'types' instead? You might even need another kind of measurement model:
  - Diagnostic Classification Models (e.g., Rupp, Templin, & Henson, 2010)
    - Goal is measurement of discrete attributes or skills, not traits
    - Useful when classification is the goal of measurement
    - Potential for use when reliability of multidimensional traits is low

Construct Maps should include…

- Coherent, substantive definition of the construct
- An underlying continuum that can be manifested 2 ways:
  - *Ordering of persons to be measured (low to high)*
    - Could include descriptive labels for 'types of people'
    - Could include other characteristics (e.g., age, disease state)
  - *Ordering of item responses (low to high)*
    - Behaviors (e.g., 'sits quietly'…. 'kicks and screams on the floor')
    - Item options ('no problems', 'some problems', 'many problems')
  - Key idea: Responses have to *orderable*

- Some examples of construct maps…

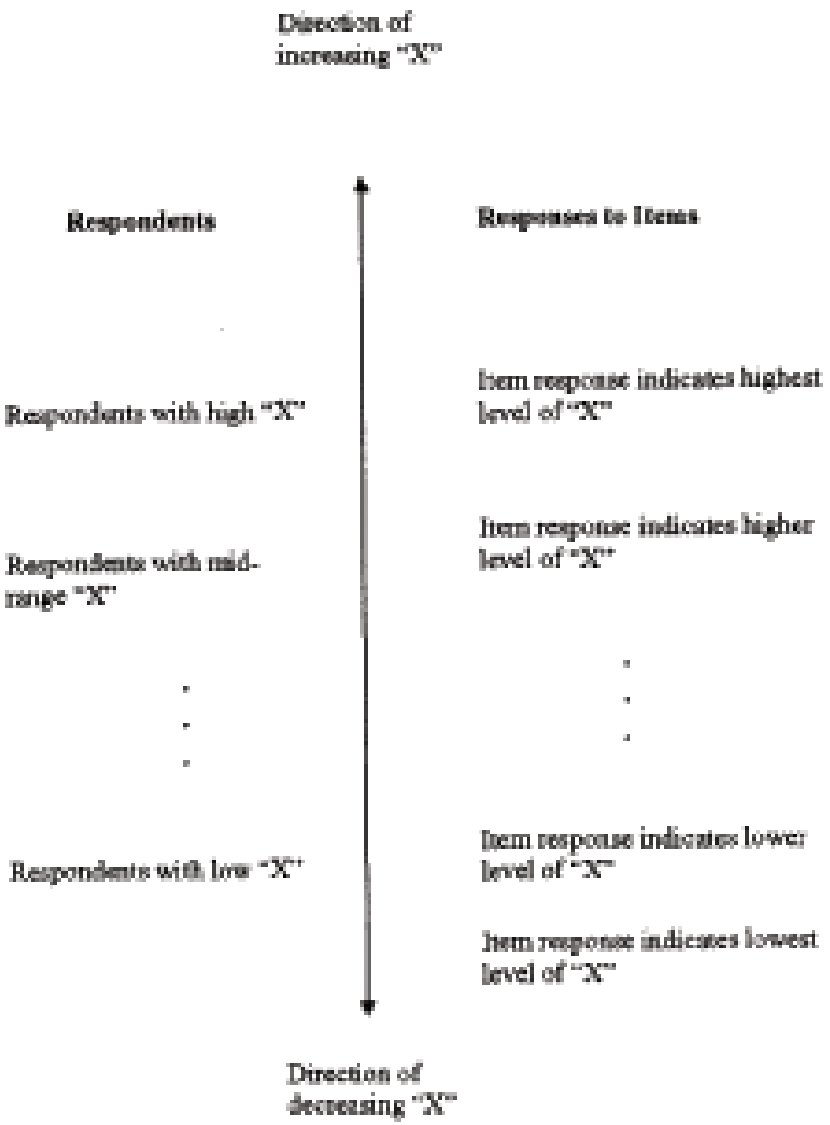


FIG. 2.1 A generic construct map in construct "X."

From Wilson (2005)

Template for a Construct Map

Left = PERSONS

qualities

characteristics

Right = ITEMS

responses

behaviors

Direction of increasing speech sound development for *girls*

Respondents	Responses to Items
9 ½ yrs.	All speech sounds are accurate
9 yr. olds	spr, thr, skr, str
8 yr. olds	r-, -er, pr, br, tr, dr, gr, kr, fr
7 yr, olds	-ng, s, z, th, sp, st, sk, sp, sm, sn, sw, sl, spl, skw
6 yr. olds	sh, ch, j, th, -l
5 ½ yr. olds	-f, v, pl, bl, kl, gl, fl
5 yr. olds	l-
4 yr. olds	y-, t, tw, kw
3 ½ yr. olds	n, g, k, f-
3 yr. olds	m, h, w, p, b, d
1 yr. olds	No accurate speech sounds

Direction of increasing speech sound development for *boys*

Respondents	Responses to Items
9 ½ yr. olds	All speech sounds are accurate
9 yr. olds	spr, thr, skr, str
8 yr. olds	th, \r-, -er, pr, br, tr, dr, gr, kr, fr
7 yr, olds	-ng, s, z, th, sp, st, sk, sp, sm, sn, sw, sl, spl, skw, -l, j, ch, sh
6 yr. olds	l-, pl, bl, kl, gl, fl
5 ½ yr. olds	-f, v, tw, kw
5 yr. olds	y-
4 yr. olds	g
3 ½ yr. olds	t, k, d, f-
3 yr. olds	m, h, n, w, p, b, d
1 yr. olds	No accurate speech sounds

Construct Map for Standardized Interviewing

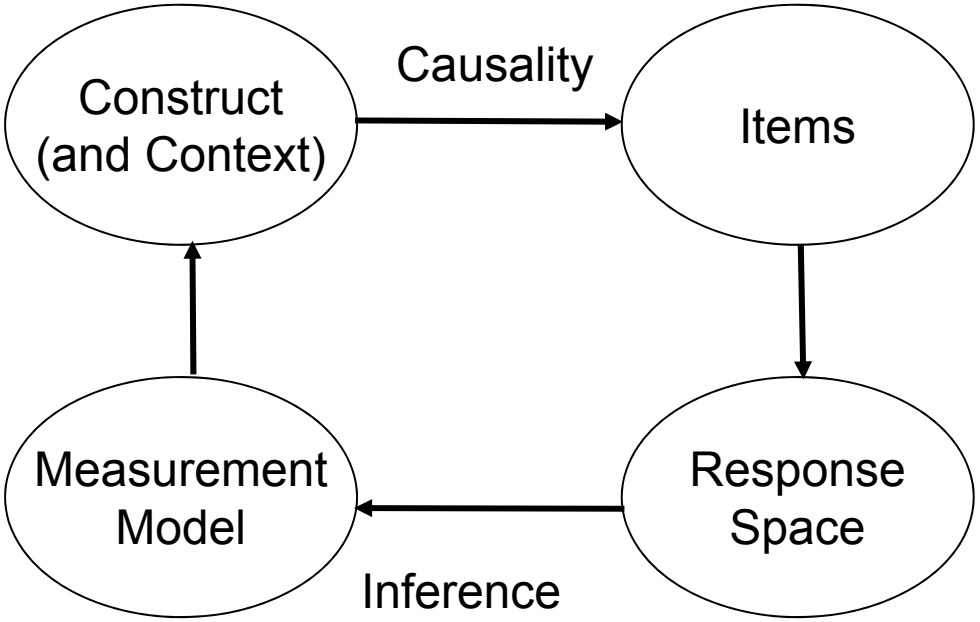
Types of people	Item response options
ATA-certified SLLs specifically trained to work with surveys	Can translate survey questions, maintaining standardization of question wording
SLLs who are certified by the American Translators Association (ATA)	Can translate documents from second language into first language
SLLs who have studied both languages and have studied translation theory	Can revise translated documents
SLLs with at least 5 years of language study	Can write in the first and second language
SLLs with at least 3 years of language study	Can speak in the first and second language
SLLs with at least 1 year of language study	Can read in the first and second language
An individual with at least 10 years of educ	Can write in at least one language
Any literate individual	Can read in at least one language
Anyone over the age of two who has not been raised in isolation	Can speak at least one language

SLI = Second Language Learners

Instrument Construction

- Once your construct is mapped in terms of ordering of persons and responses, next is instrument construction
- Instrument :: Measurement method through which observable responses or behaviors in the real world are related to a construct that exists only as part of a theory
- 4 components of instrument construction:
  - Construct (and Context)
  - Item Generation
  - Response (Outcome) Space
  - Measurement Model

4 Instrument Building Blocks



Direction of causality: The construct determines which items are relevant (to represent the construct), the content of the items then causes a response, and *the response format then directs which measurement model to use*.

We then use the measurement model to make inferences about people's standing on the latent construct (trait as measured in a given context).

Construct and Context

- Instruments should be secondary – they are created:
  - For the purpose of measuring a pre-existing latent **construct**
  - Within a specific **context** in which that measurement is needed

- Instruments should be seen as **logical arguments**:
  - Can the results be used to make the intended decision regarding a person's level of a construct in that context?
  - Build instrument purposively with this in mind, but pay attention to information gathered after-the-fact as to how well it is working

- Instruments are created from items, which have 2 parts:
  - **Construct** component: Location on the construct map?
    - Want to include both hard and easy items to measure full range
  - **Descriptive** component: Other relevant item characteristics
    - Language? Context? Method of administration? Reporter/rater?

Steps to Item Design

- Do your homework:
  - Literature review
    - What's been done before…And what's wrong with it?
  - Ask relevant people (participants, professionals):
    - What should we be focusing on? How should we ask the questions?

- Design the instrument:
  - Item design (construct and descriptive components)
  - Response format (location on 'openness' continuum)

- Get feedback from participants:
  - 'Think aloud' while solving problems
  - Exit interview

(Good) Item Generation

- Ideally, items are *realizations* of existing constructs
  - Hmm…How do I measure this construct? (write item 1, 2, 3…)
  - In reality, this is an iterative process…
- Items should be unambiguous
  - Cover a single concept (no 'ands') with a clear referent
- Items should be simple to process
  - Short, common vocabulary
  - Negatives can be harder to process – and research has suggested negatively-worded (reverse-coded) items to be less discriminating
- Good items should span the full range of construct…
  but without going too narrow or too broad

Actual (Not so Good) Items…

- *How important to you is it that…*
  - My family members have good relationships with extended family members (grandparents, in-laws, etc.).
  - My family is physically healthy.

- Assess the quality of the relationship that you have with your children?

  ___excellent ___very good ___good ___fair ___poor

- To what extent did others make it difficult for you to engage in various activities before your imprisonment?

  ____ 1. never _____2. rarely _____3. often _____4. most of the time

Response (Outcome) Space

- **Outcome space = response format** :: varies in flexibility
  - Most flexible: Open-ended response
    - e.g., essay, performance
    - Less work at beginning; more work at the end
  - Least flexible: Fixed format
    - e.g., multiple choice or likert scales
    - More work at beginning; less work at the end

- Ideally, instrument development **would start by seeking open-ended responses**, from which representative fixed format options would be created that are:
  - Research-based, well-defined, and context-specific
  - Finite and exhaustive (orderable responses; include n/a)

Specificity of Response Space

Response options can be item-specific to maximize their utility:

Do you feel confident in explaining your religious beliefs to others?

_____ Not at all confident
_____ Mostly not confident
_____ Confident
_____ Very confident
_____ Totally confident

How often do you explain your religious beliefs to others?

_____ Never
_____ Once a year
_____ Every couple months
_____ Couple times a month
_____ Once a week,
_____ Couple times a week
_____ Everyday

How good are you at explaining your religious beliefs?

_____ I have no idea how to explain my beliefs
_____ I struggle a lot in explaining my beliefs
_____ I struggle a little in explaining my beliefs
_____ I am pretty good at explaining my beliefs
_____ I am very good at explaining my beliefs
_____ I am extremely good at explaining my beliefs

Item response formats DO NOT all have to be the same if you are using a latent trait model – you can and should customize them to be most informative for the question at hand.

Specificity of Response Space

Versus something like this:

- Sometimes I feel caught between wanting to buy things to make me look better in some way to others, when I really should be spending more money in ways that have more spiritual meaning.

_____ Strongly Disagree
_____ Disagree
_____ Somewhat Disagree
_____ Neither
_____ Somewhat Agree
_____ Agree
_____ Strongly Agree

Another instance of what not to do: unlabeled options:

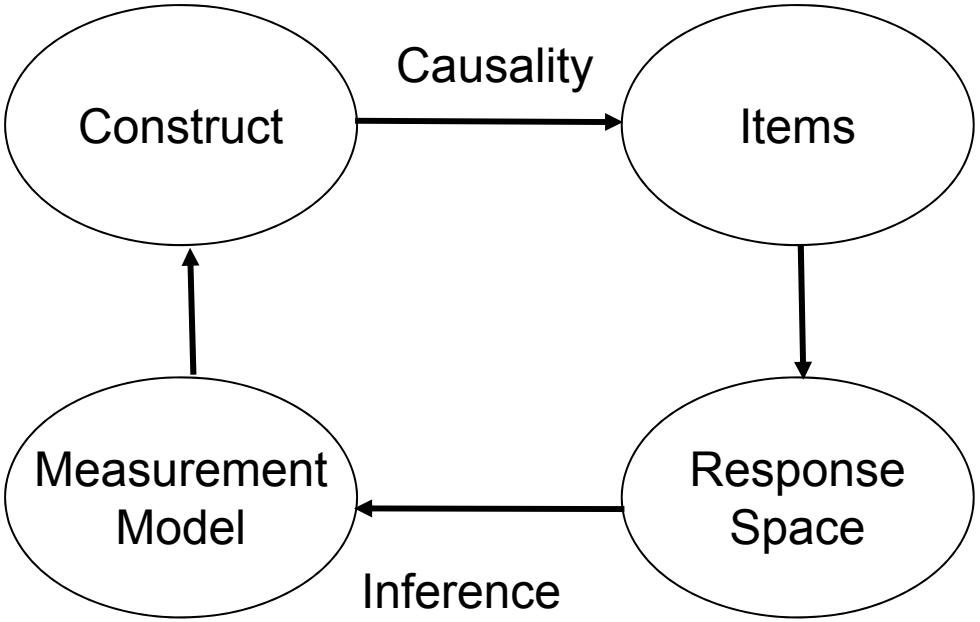
1. "Never"
2. …
3. …
4. …
5. "Always"

Item-Level Measurement Models

- Type of response format will generally lend itself to an appropriate measurement model
  - Dichotomous (binary) item? (yes/no, MC :: correct/not)
    - Logistic/probit model (IRT)
    - Normal approximation (CFA) won't work very well
  - Polytomous (quantitative) item? A few IRT options…
    - Graded response model
    - Partial credit model
    - Normal approximation (CFA) *may* not be too bad…
  - Unordered categorical item? Only one IRT option:
    - Nominal model (way hard to estimate)
  - No clear measurement model for many other types of item choices (i.e., forced choice, rankings)

4 Instrument Building Blocks

- Process of Inference:
  - Relate responses to construct via measurement model
  - In other words, *translate scores to locations on construct map*



Note that causality does NOT go through the measurement model – items would be caused by the construct regardless of response format, and thus regardless of the choice of measurement model.

TEST CONSTRUCTION WITH IRT

Uses of Item and Test Information Functions

Test construction in IRT centers on test and item information functions, used for:

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

Conditional SE in IRT

- As previously stated, the precision (reliability) and imprecision (error) of a test scaled with IRT is conditional on theta

- Tests may be better or worse for measuring certain trait levels

Test Development

- From a pool of previously piloted test items, IRT makes it relatively easy to switch items in and out and determine what the resulting information function will be

- This tells the test maker what the conditional standard errors will be, too

Test Development

- Another benefit to test development is that multiple forms may be built to the same statistical specifications
- This process is often referred to as “Pre-equating”
- Building strictly parallel forms is always difficult, but these procedures can help

Test Revision

- Likewise, test items may be removed from previously existing forms (e.g, to create a “short form” of a test)
- Test items may also need to be added if the previous form is found to be unreliable
- Estimating the new reliability of the test is straightforward with IRT

Test Information Function

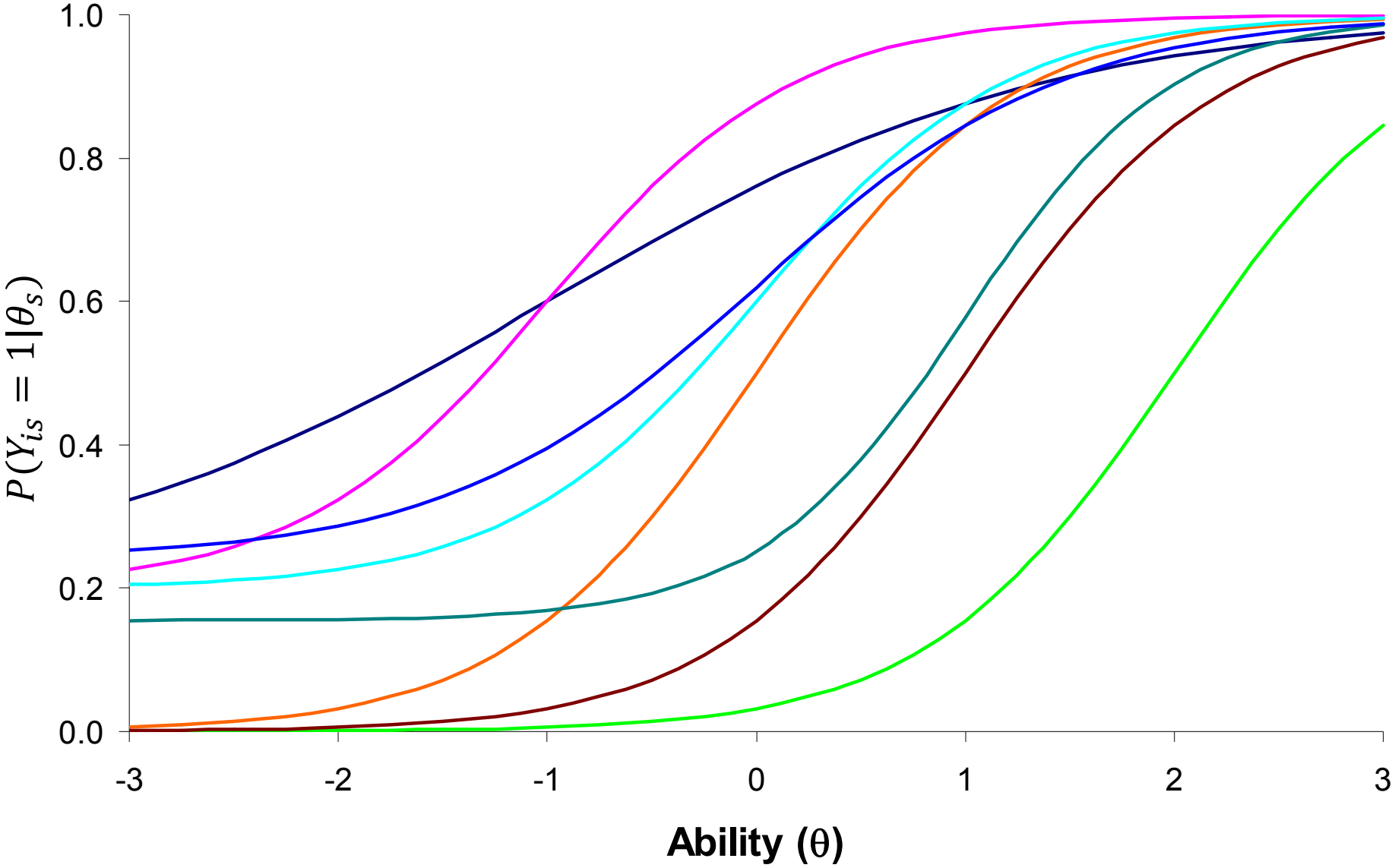
- Just like we add up ICCs to get a TCC, we add up IIFs to get a TIF
- Information will continue to increase as we add test items, therefore increasing precision
- All things equal, longer tests provide increased measurement precision

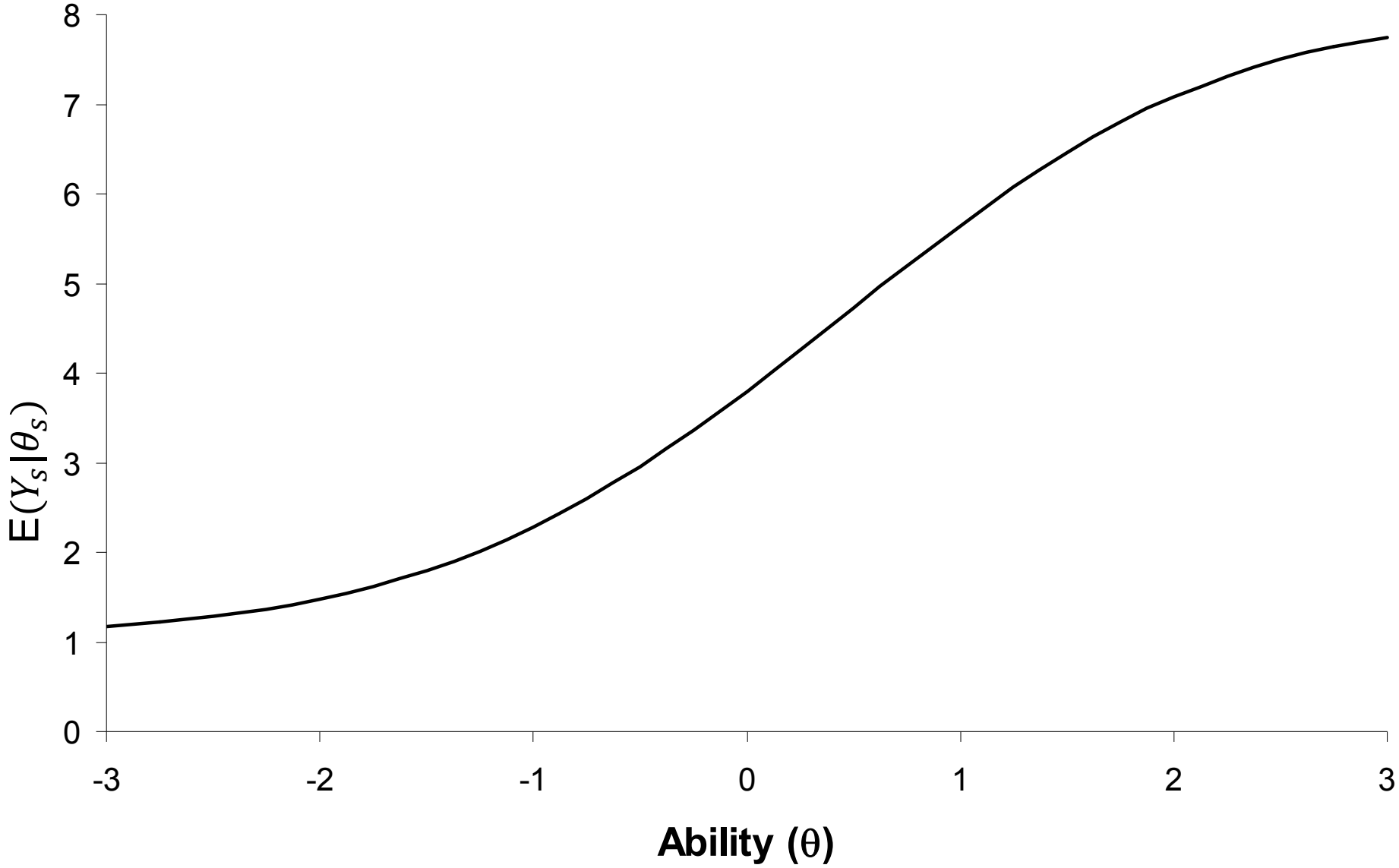
Test Information Function

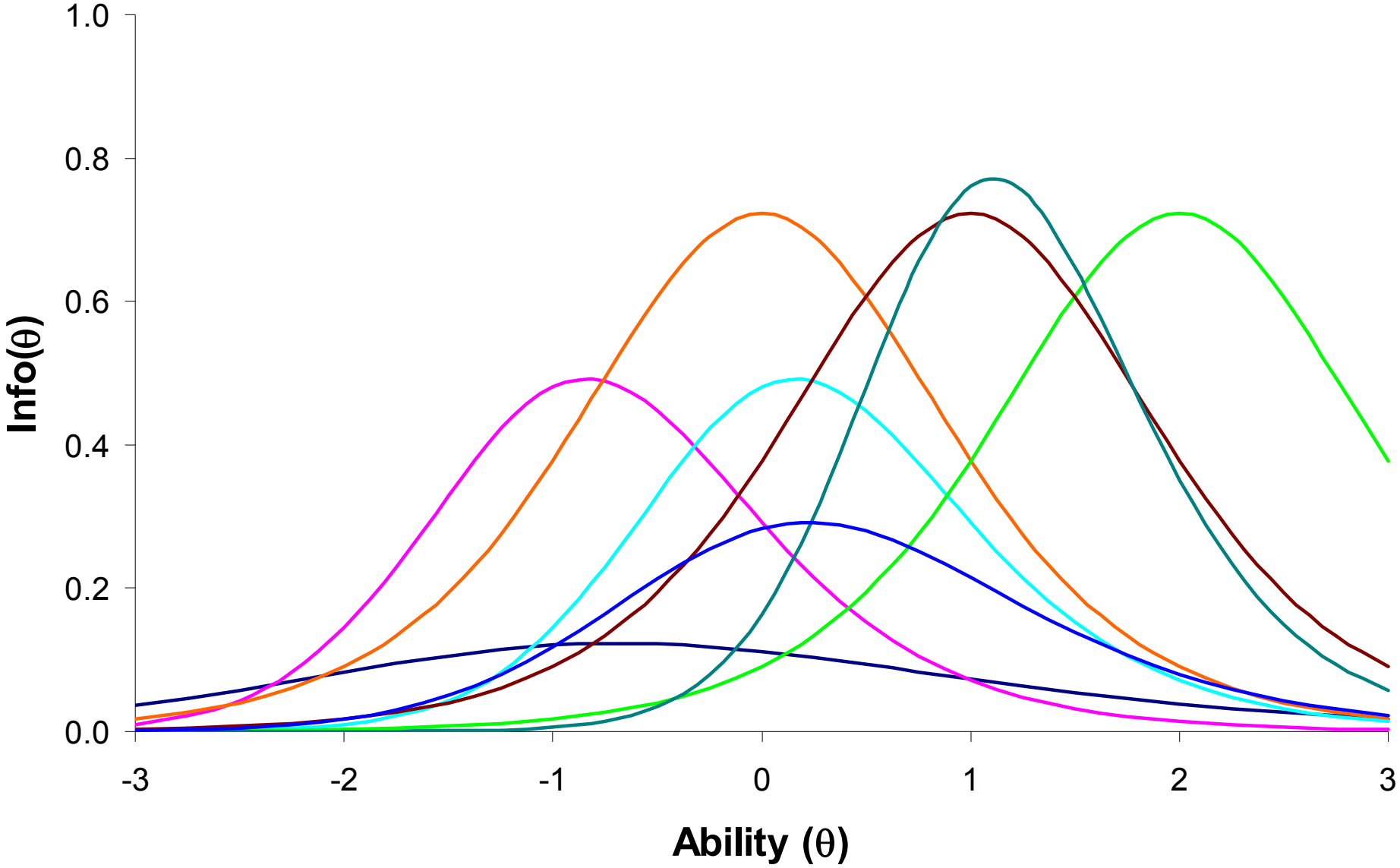
- Defined for a set of items at each point along the ability (θ) scale

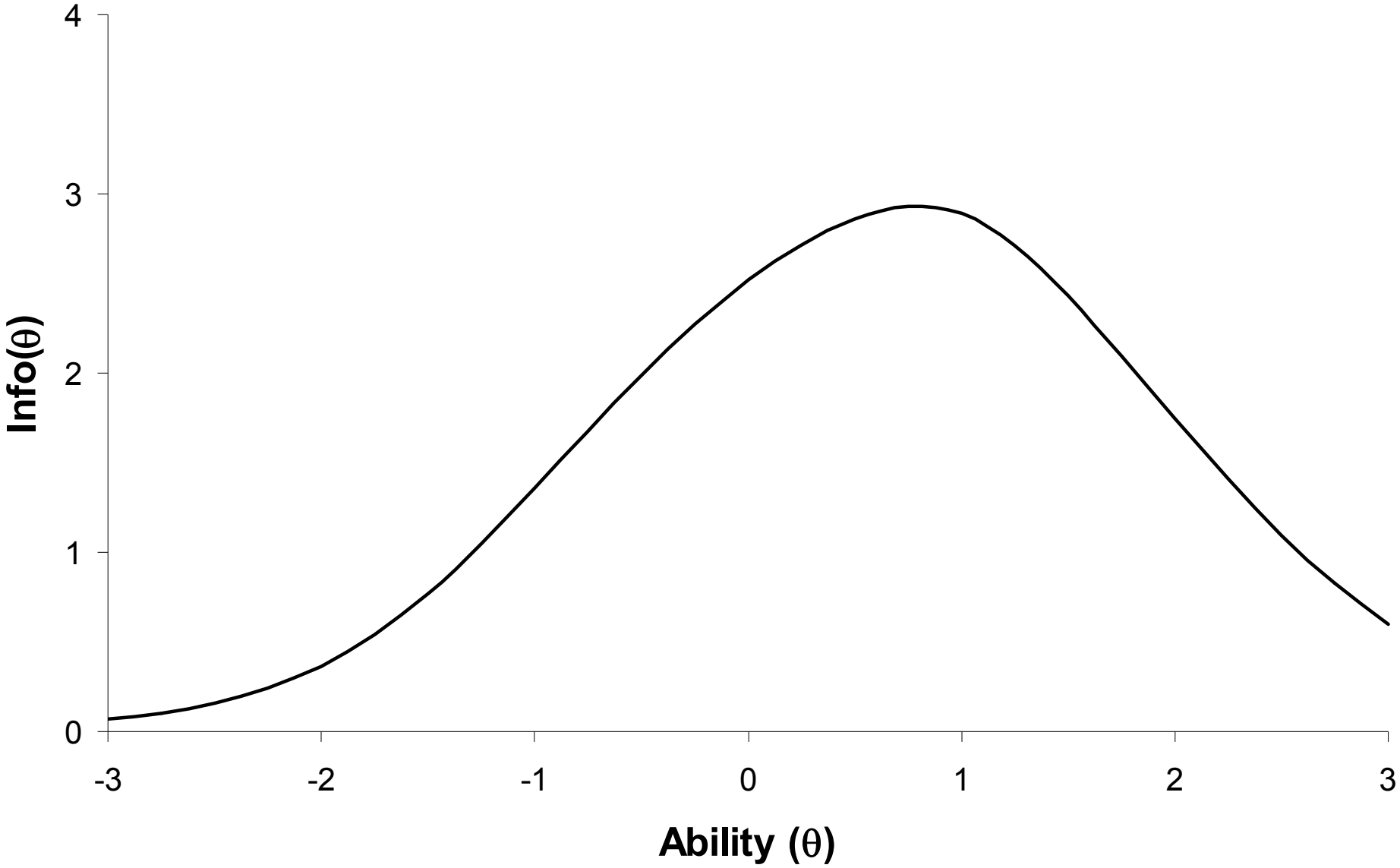
- Test information is influenced by the 'quality' and the number of test items:
  - *I* = total number of test items
  - *i* = item index
  - *I( )* = test information function

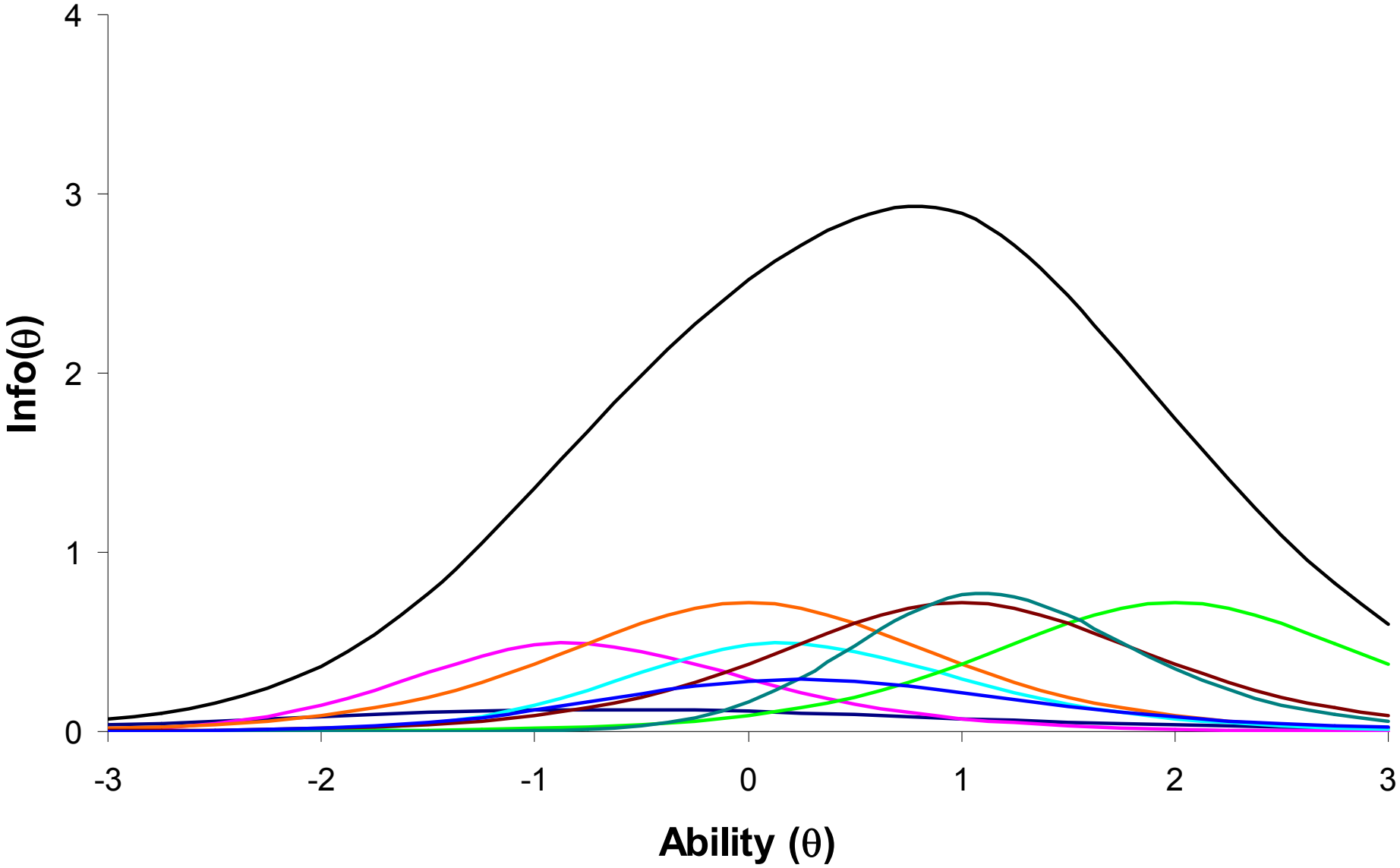
$$I(\theta) = \sum_{i=1}^{I} I_i(\theta)$$


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):

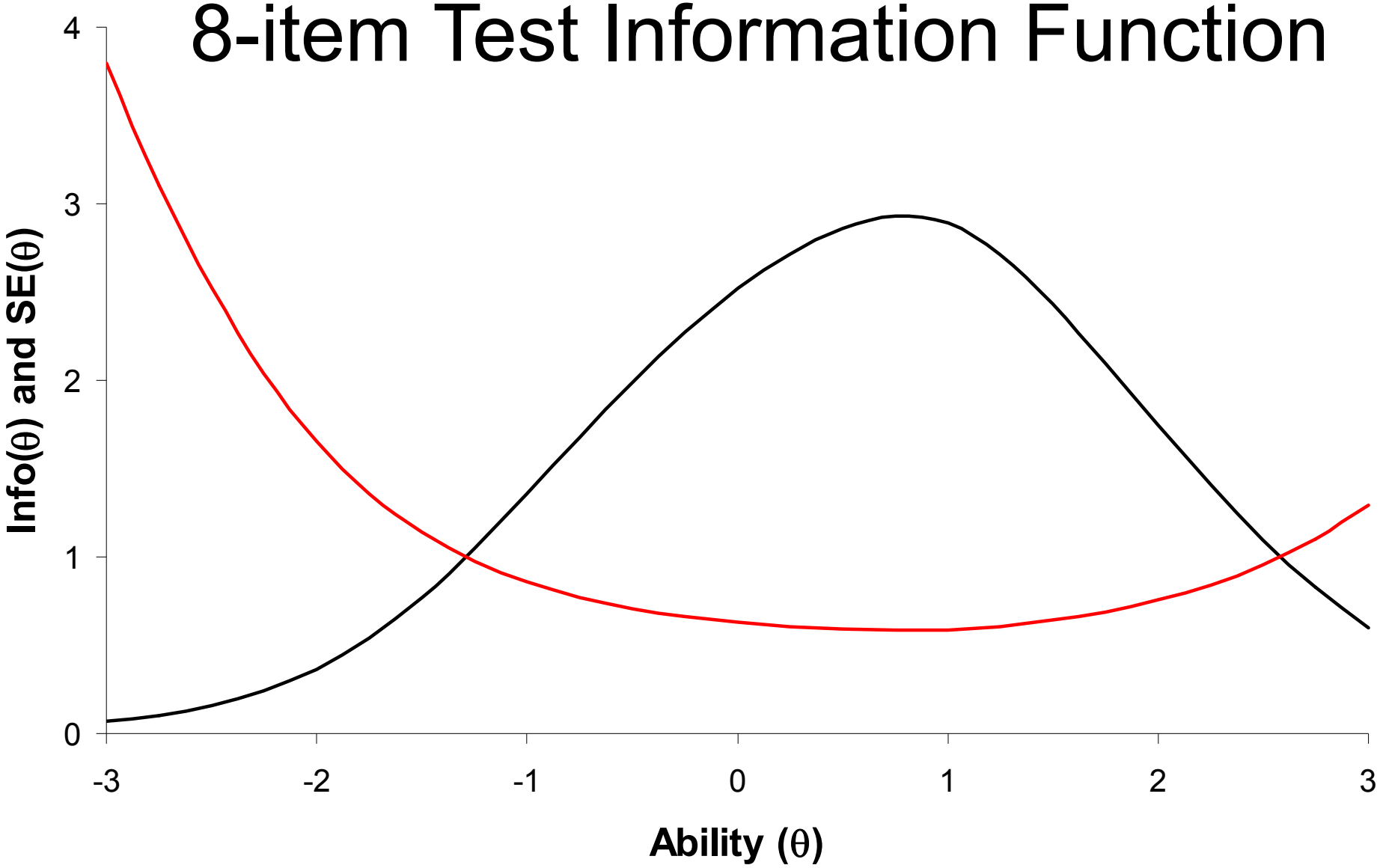
$$\text{Reliability} = \rho(\hat{\theta}) = \frac{I(\hat{\theta})}{I(\hat{\theta}) + 1}$$

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

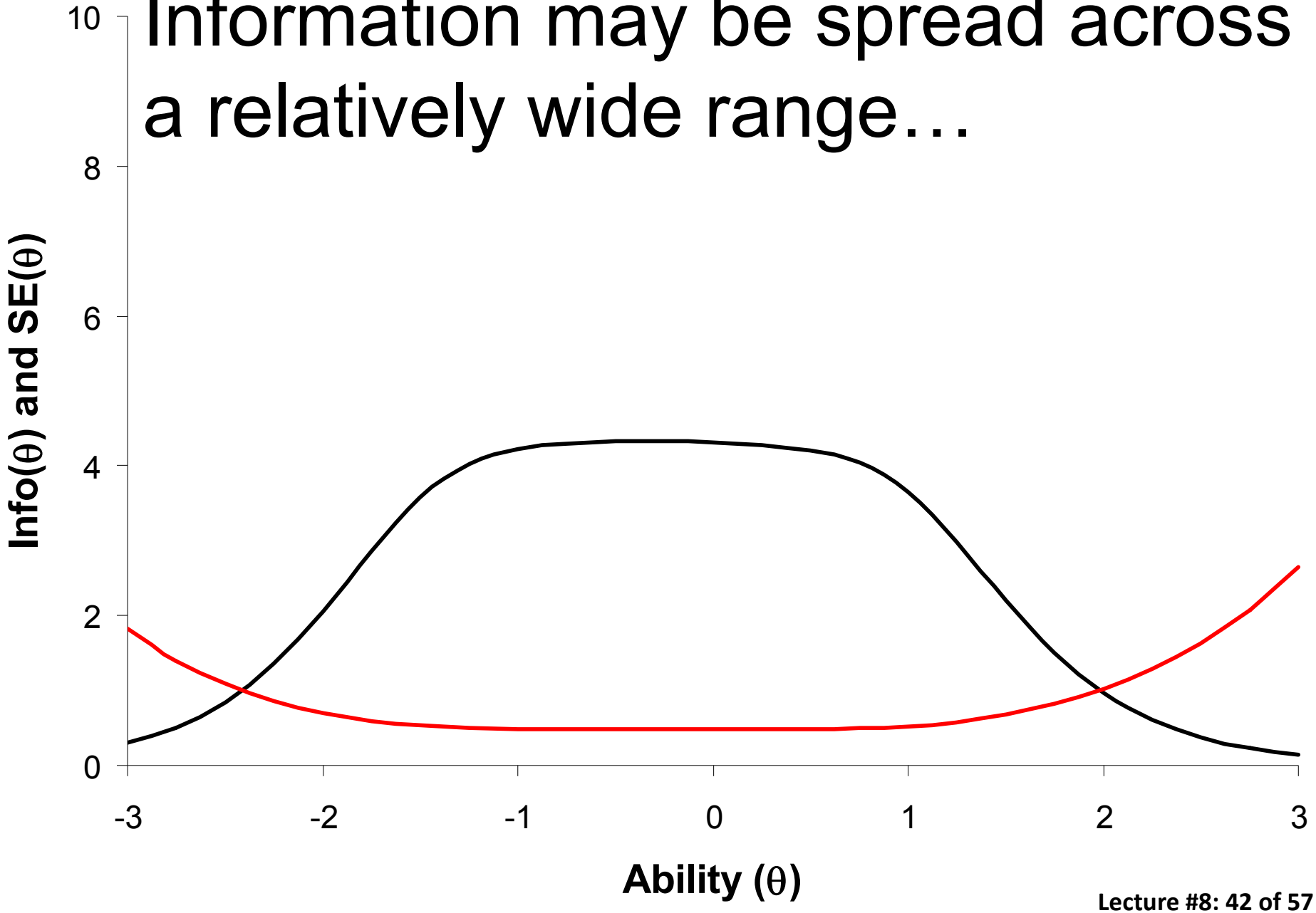
$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

- 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)

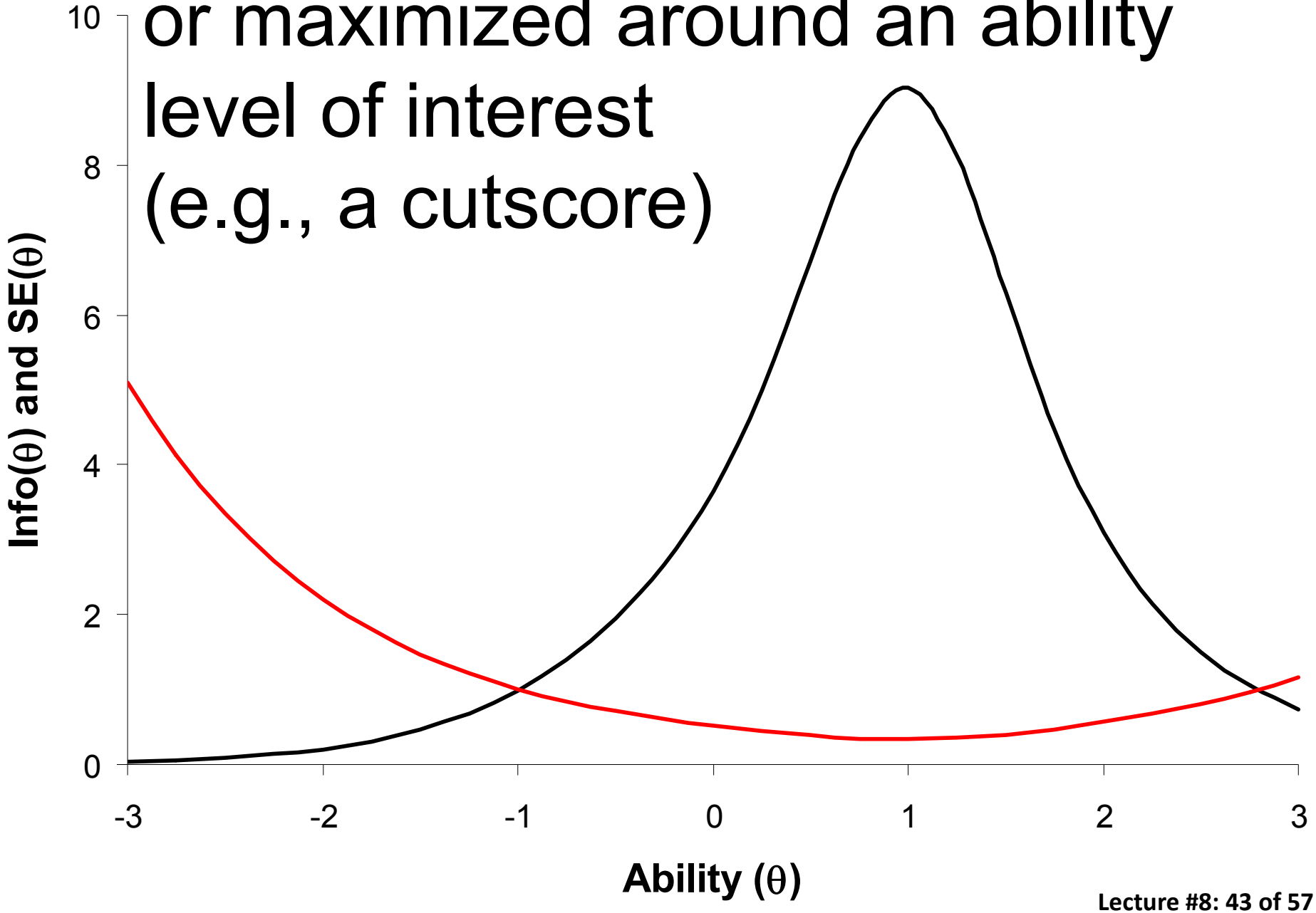
8-item Test Information Function



Information may be spread across a relatively wide range…



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



Test Revision

- In CTT, such test revisions require the assumption that the deleted or added items are of comparable statistical quality to those already on the test
  - Spearman-Brown prophecy formula
  - This may or may not be true!

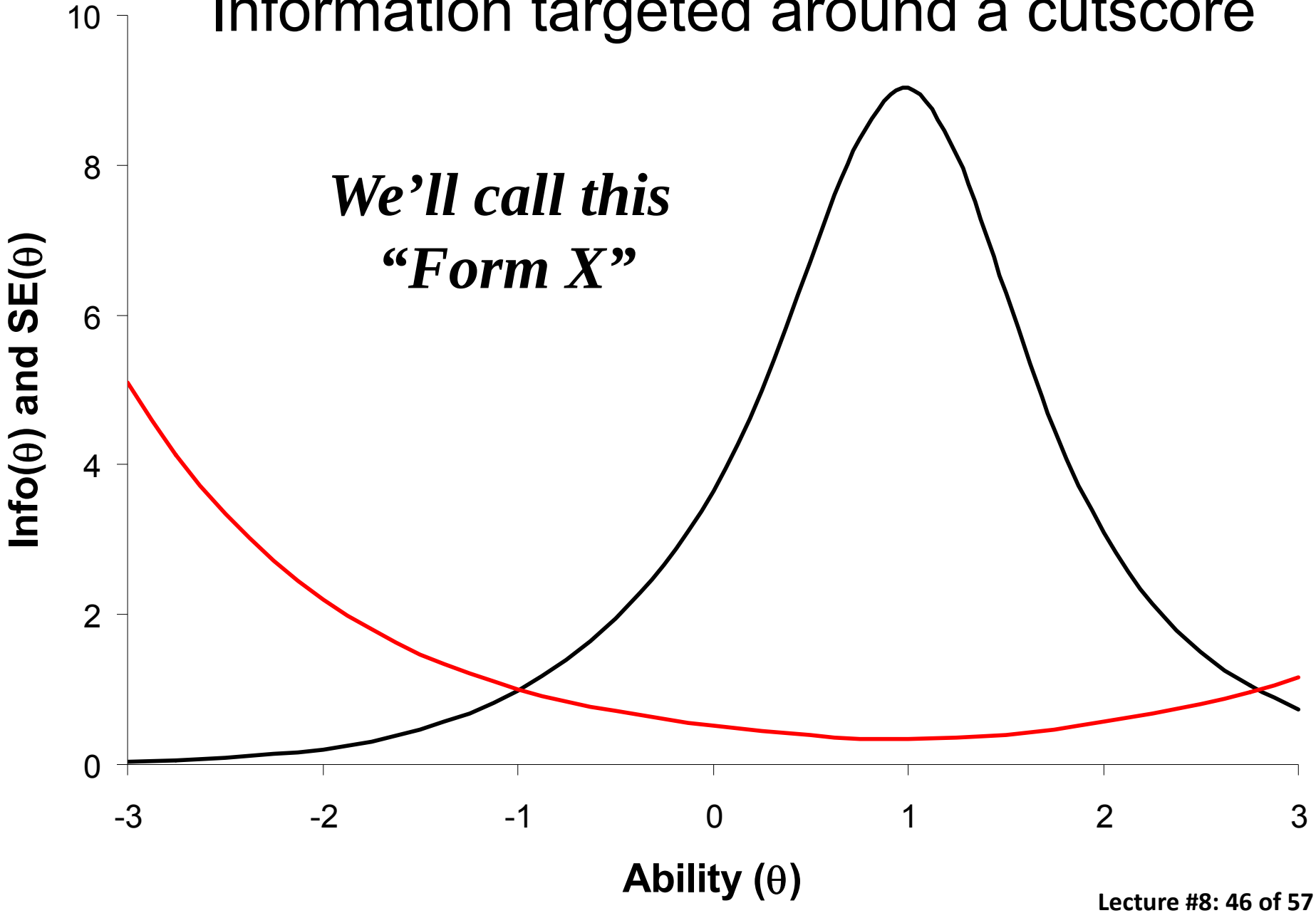
Comparing Tests

- When comparing the reliability (i.e., precision) of two test forms, its useful to determine the ratio of their information with respect to theta

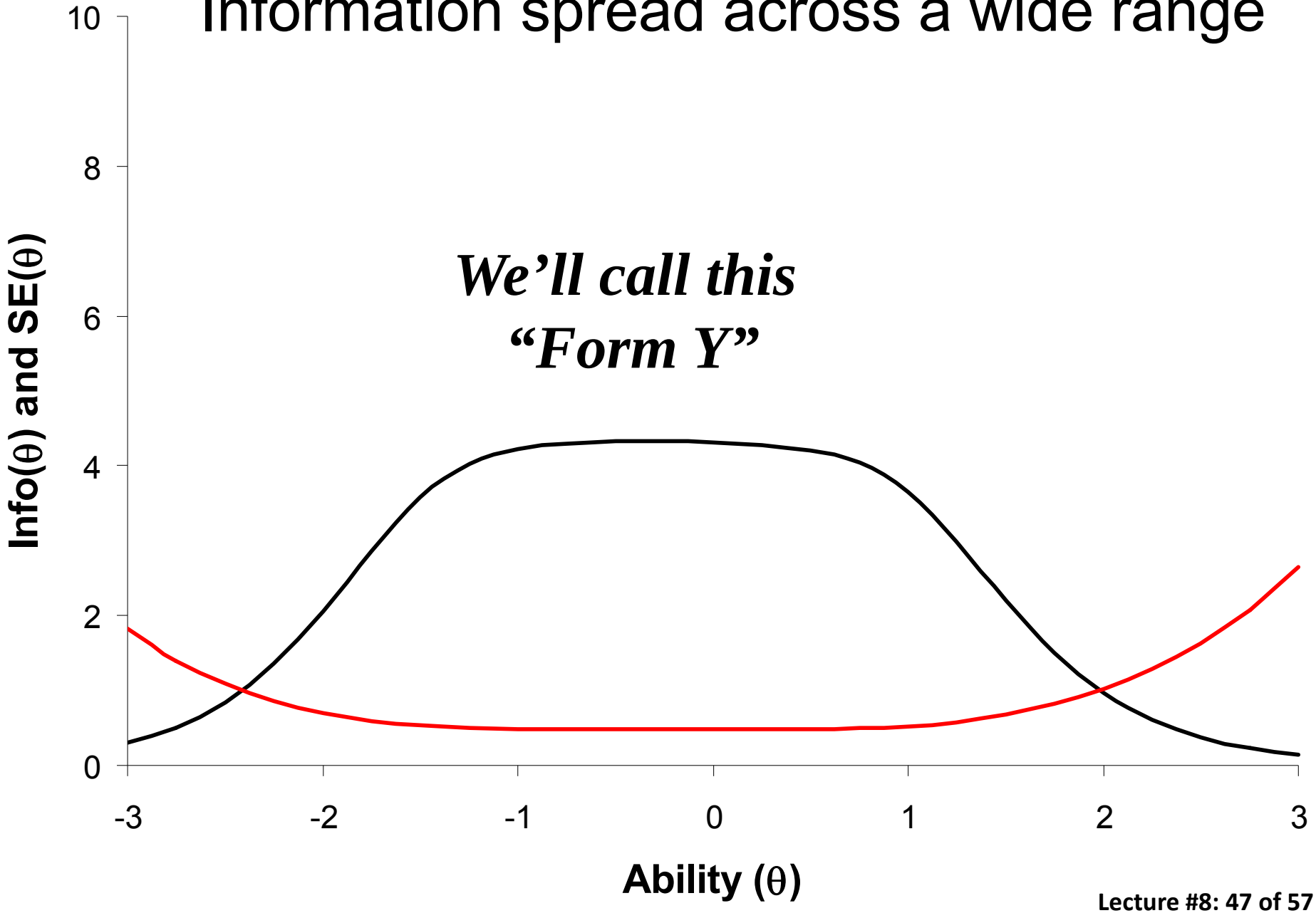
- This ratio is known as the relative efficiency of a test: RE(theta)

- Consider two previous example TIFs

Information targeted around a cutscore



Information spread across a wide range



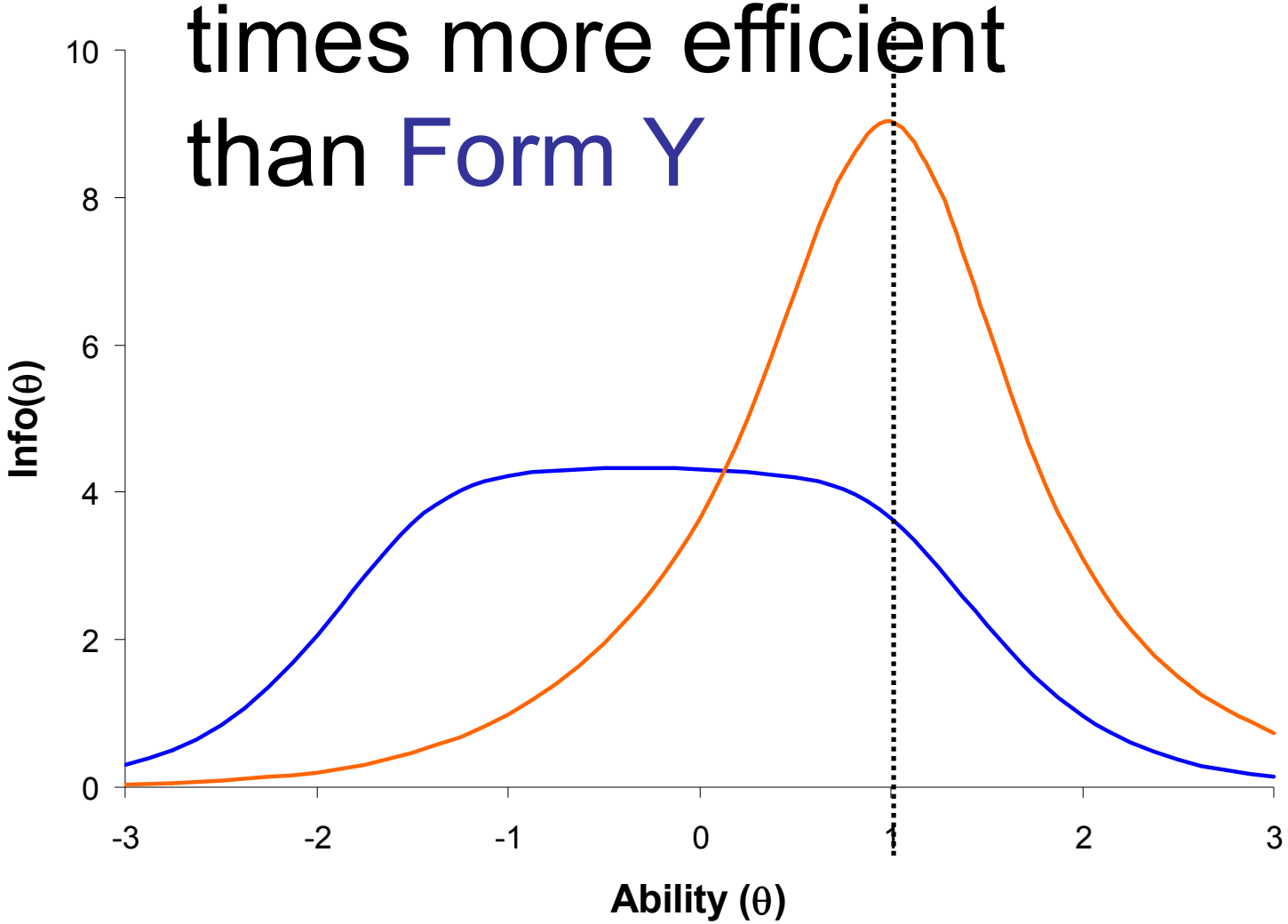
$$RE(\theta) = \frac{I_X(\theta)}{I_Y(\theta)} \rightarrow \frac{\text{info for form X at } \theta}{\text{info for form Y at } \theta}$$

Suppose at θ=1 $\rightarrow I_X(\theta) = 9.0$

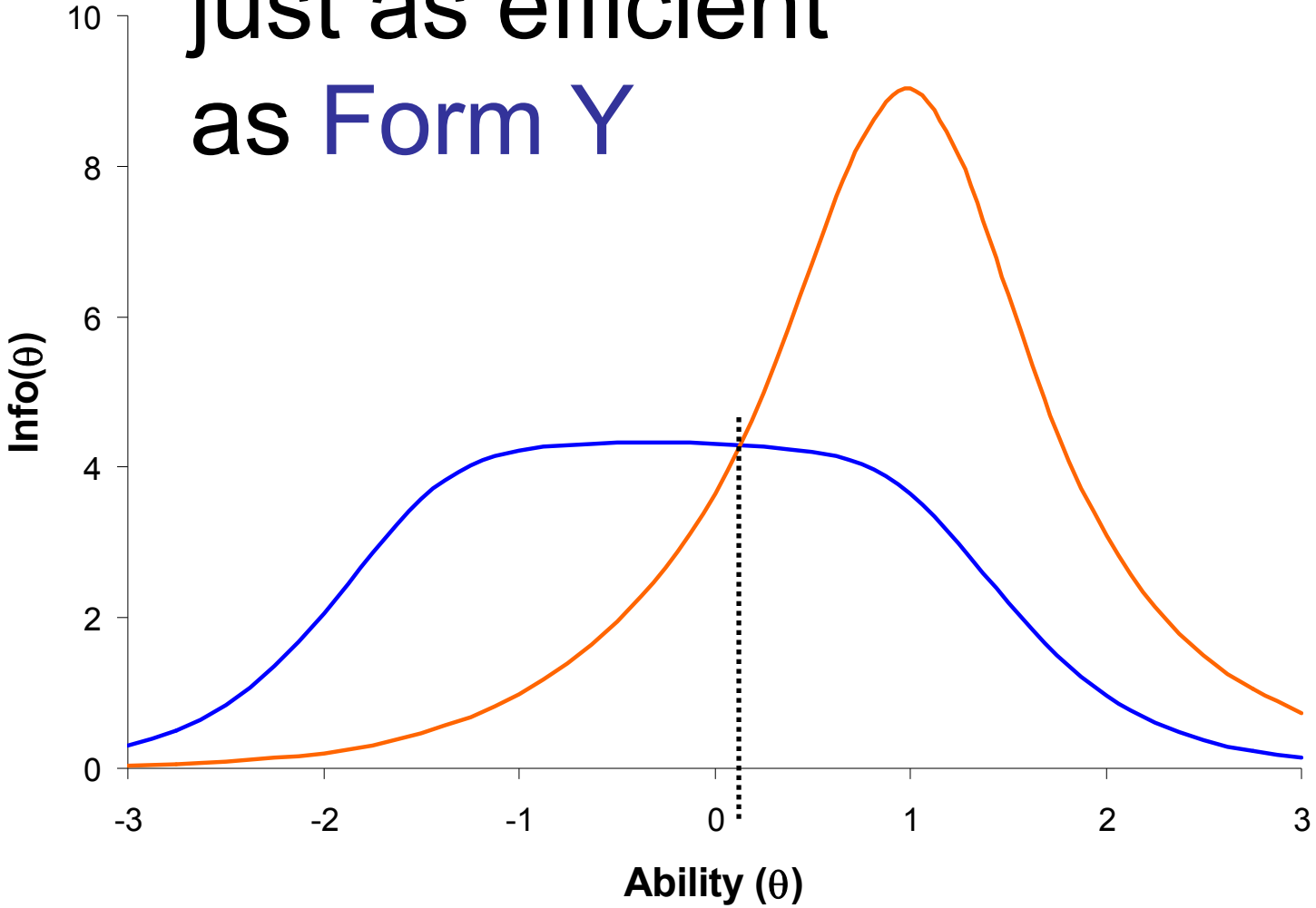
θ=1 $\rightarrow I_Y(\theta) = 3.6$

Then, $RE(\theta = 1) = \frac{9}{3.6} = 2.5$

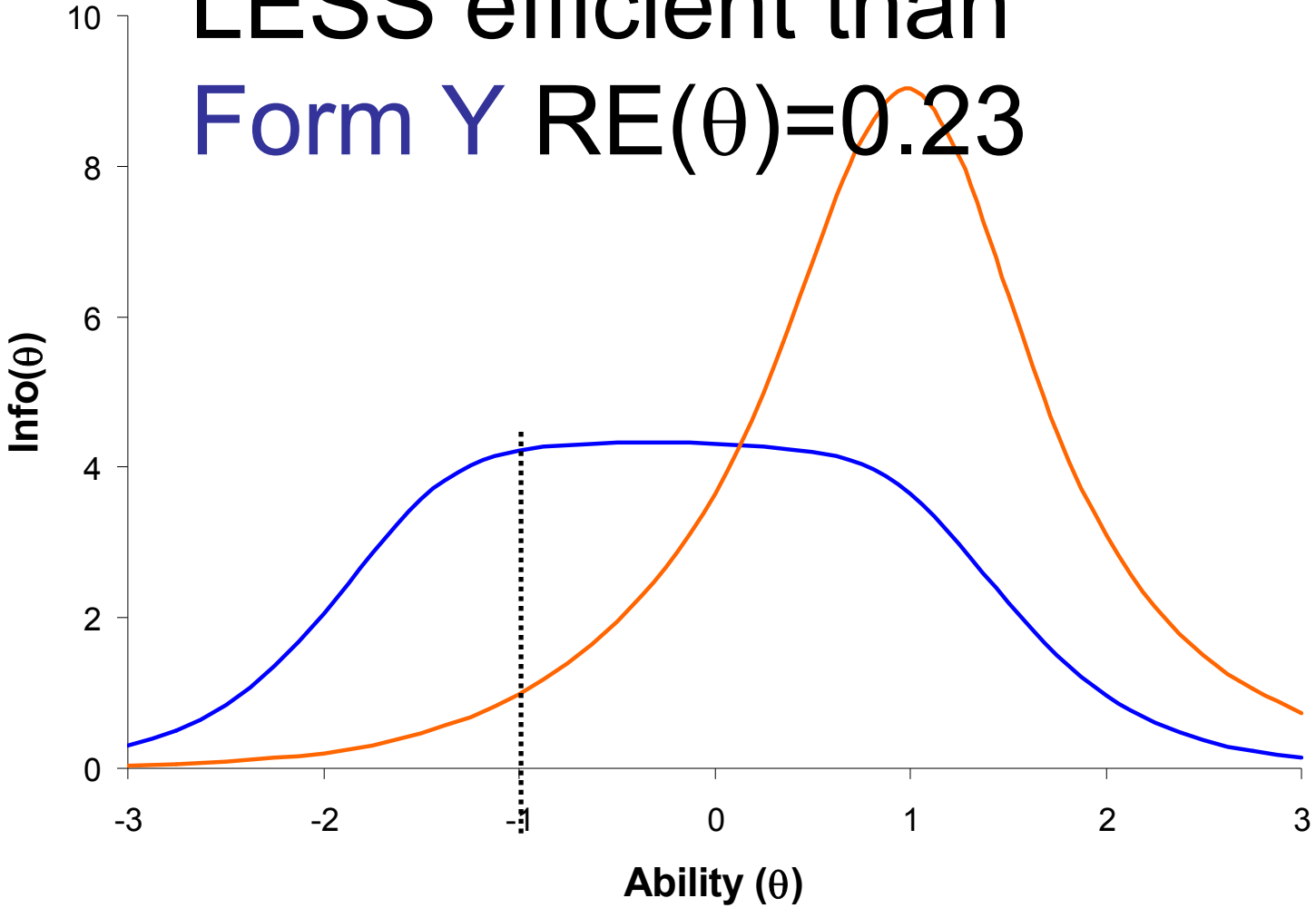
In the region θ = 1, Form X is 2.5 times more efficient than Form Y



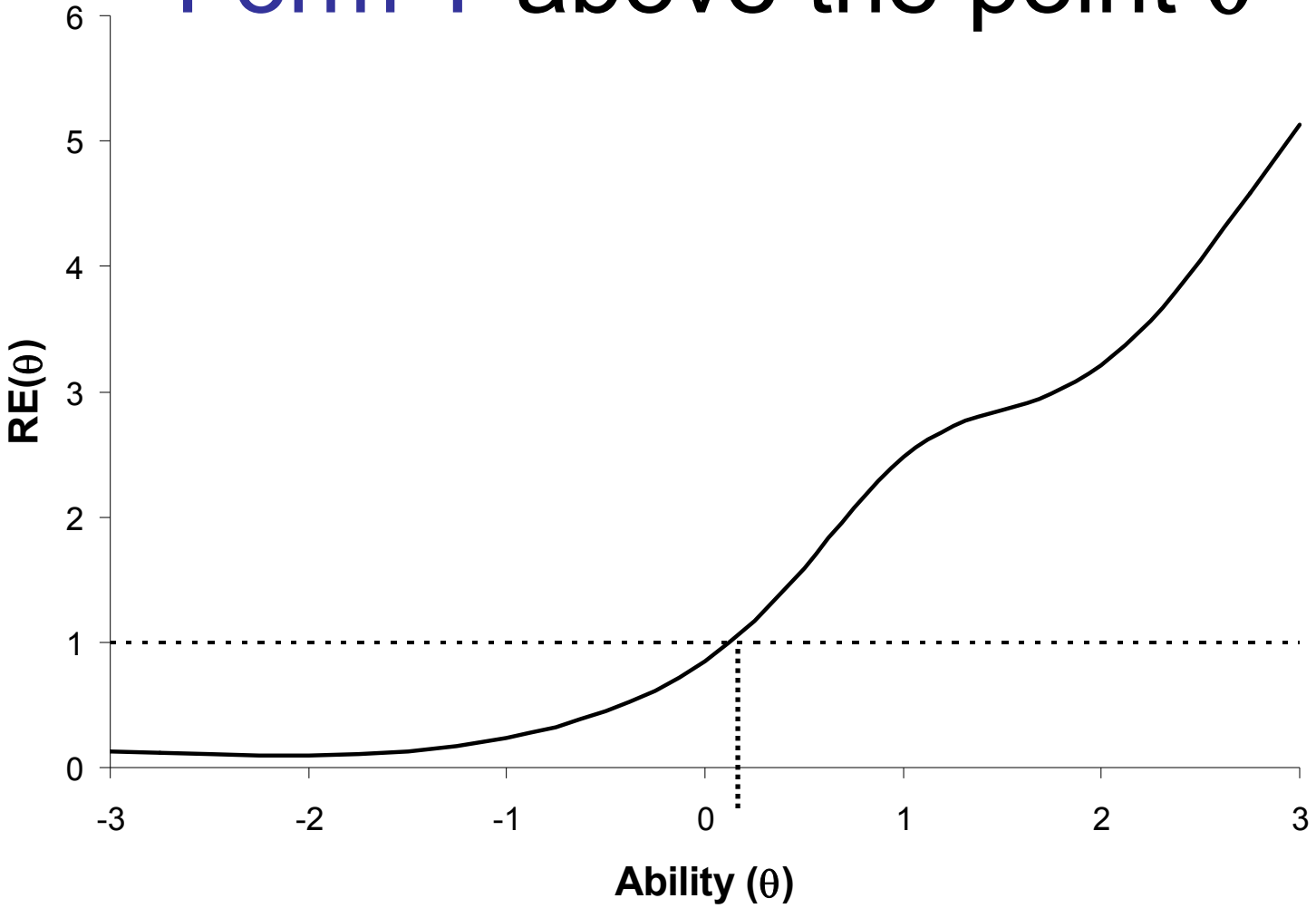
In the region θ ≈ 0.10, Form X is just as efficient as Form Y



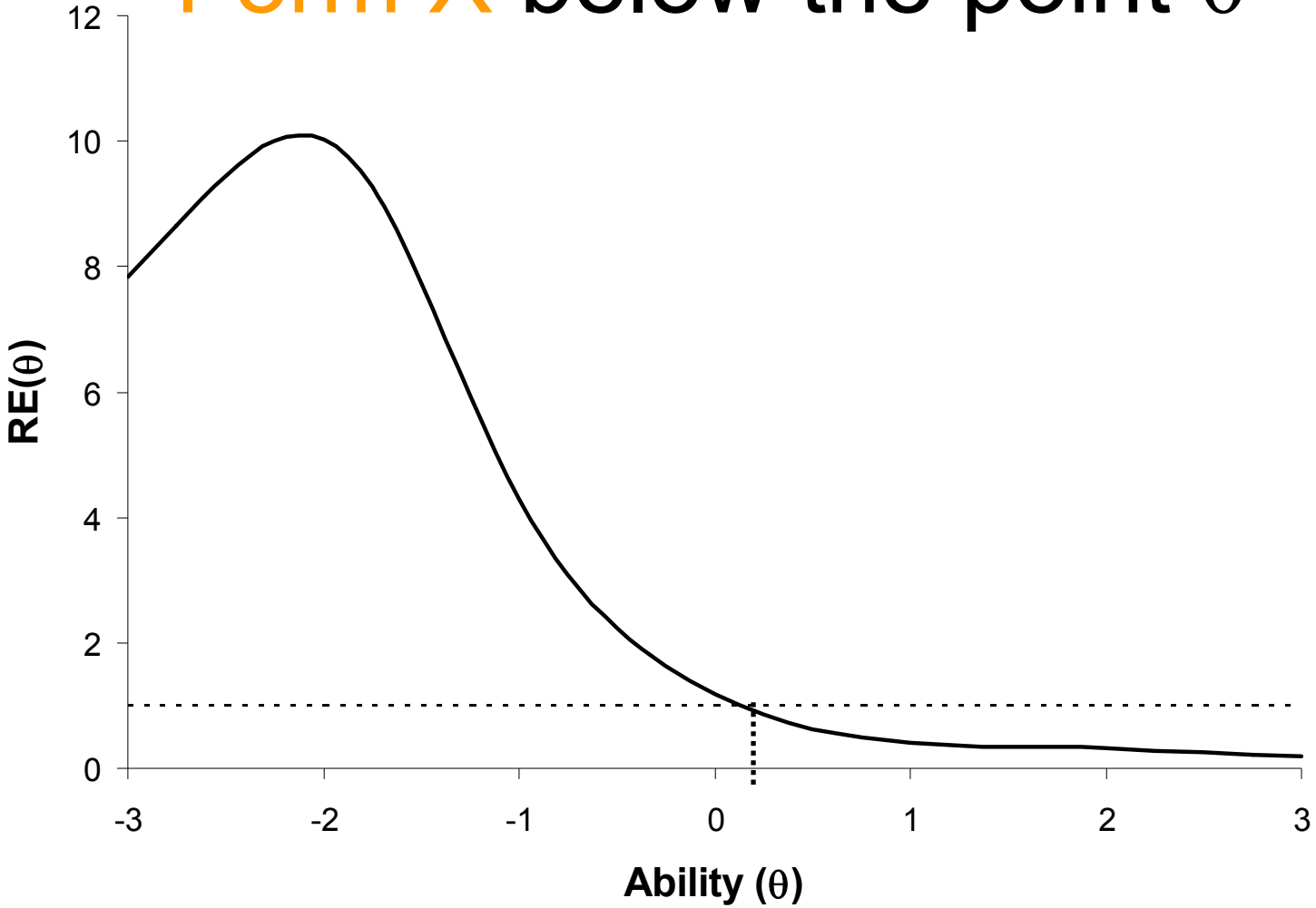
In the region θ = -1, Form X is LESS efficient than Form Y RE(θ)=0.23



Form X is more efficient than Form Y above the point $\theta \approx 0.1$



Form Y is more efficient than Form X below the point $\theta \approx 0.1$



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

CONCLUDING REMARKS

Wrapping Up…

- Instruments are created to measure pre-existing latent constructs: latent traits within desired contexts
  - Item construction is part art, part science
  - Seek as much info as possible before and after about your items
- Response options should be carefully considered:
  - Start with open-ended responses
  - Come up with flexible but fixed response categories eventually
- Measurement models provide basis for inference back to a person's position on the latent construct:
  - Specific model chosen on the basis of response format
  - The ones we'll use assume continuous underlying latent variable on which BOTH persons and items can be ordered
- Constructing tests for use with IRT is a process by which the theory of the latent trait interacts with the statistical model
  - Case in point: most end-of-grade tests in state assessments

Up Next…

- Using test information to create adaptive tests
  - Testing on the fly…