Assessment 101

Michigan Schools Testing Conference

Pre-Session

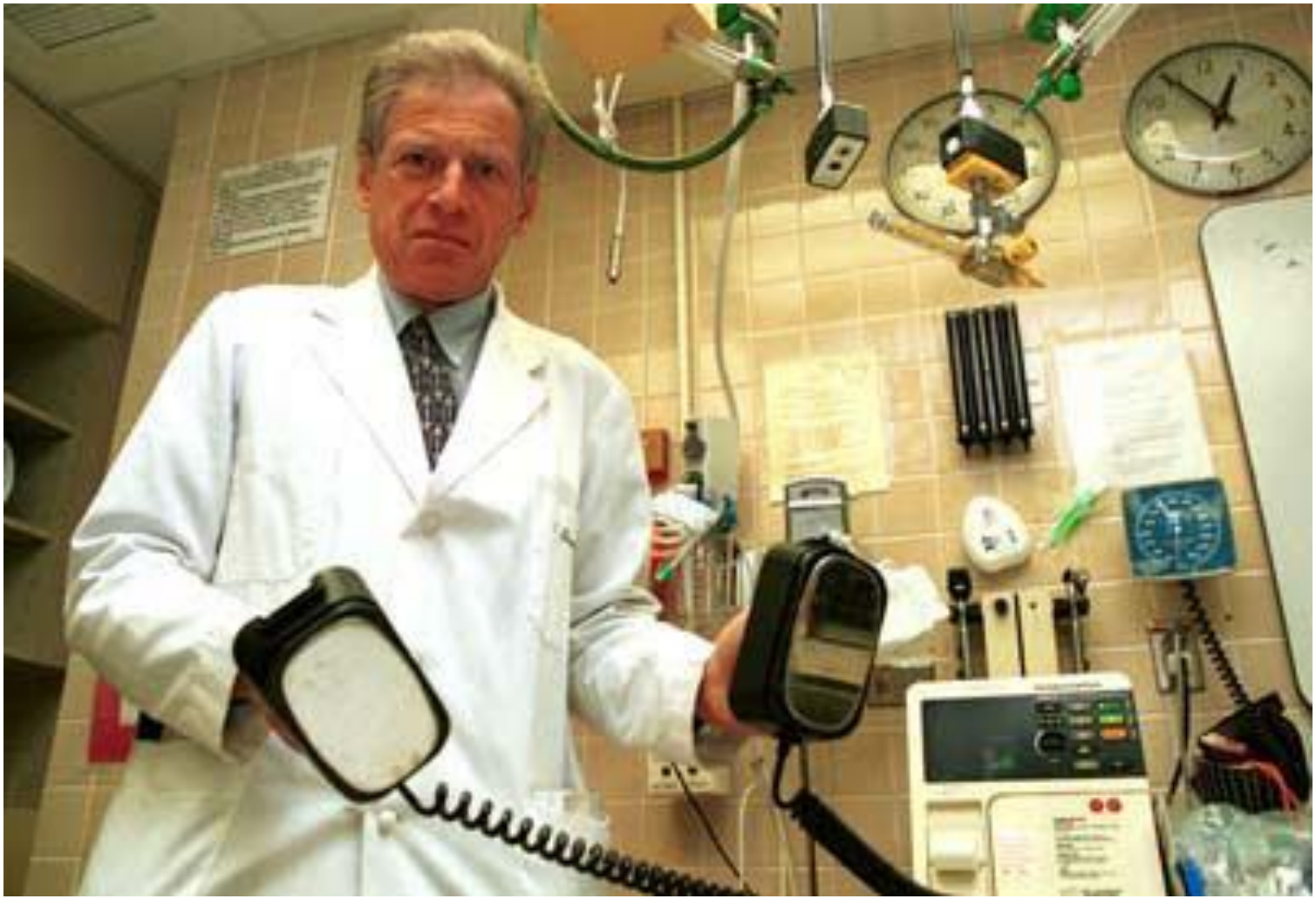
February 18, 2015

Let's start slow with some basics…

- It is obvious that

$$I(\Theta, y) = \frac{\left[\sum_{i=1}^{n} w_i P_i^{'}\right]^2}{\sum_{i=1}^{n} w_i P_i Q_i}$$

JUST KIDDING!!!! ☺



Today's Agenda

Foundational Topics

- Reliability
- Validity
- Frames of Reference
- Types of Scores
- Classical Test Theory
- Item Response Theory
- Standard Setting

Practical Applications

- Balanced Assessment
  - Formative Practices
  - Interim Assessments
  - Summative Assessments
  - Graphical Representation
- Depth of Knowledge
- Target-Method Match
- Domain Sampling
- Blueprinting

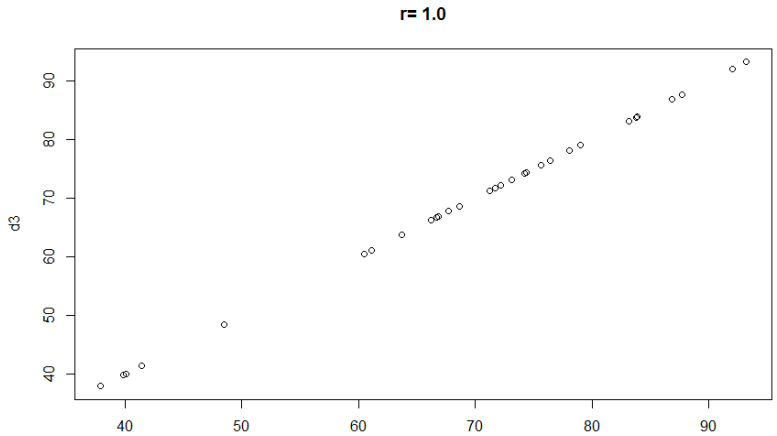
Reliability

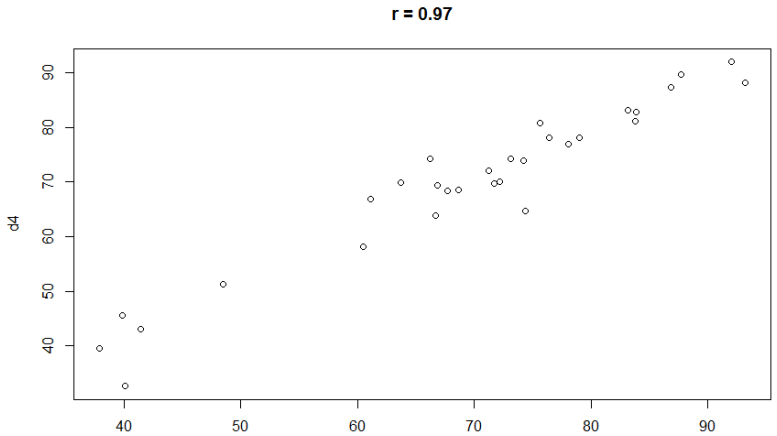
- How consistent are the scores we get from our tests?
- Consistency is the first step toward having good measurement.

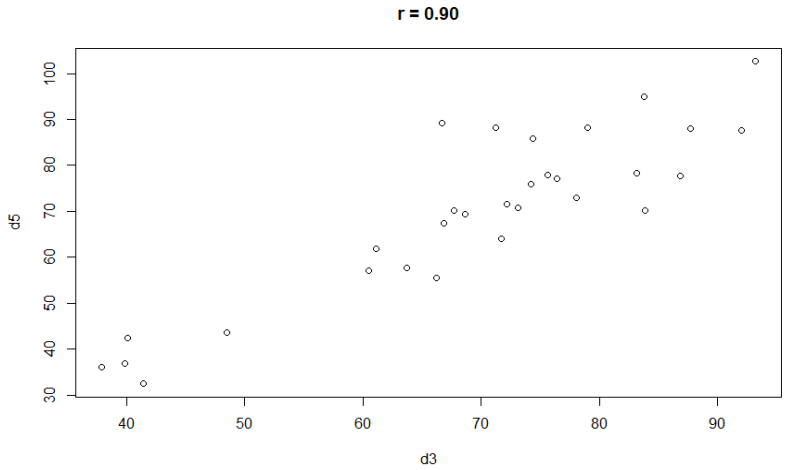
Reliability

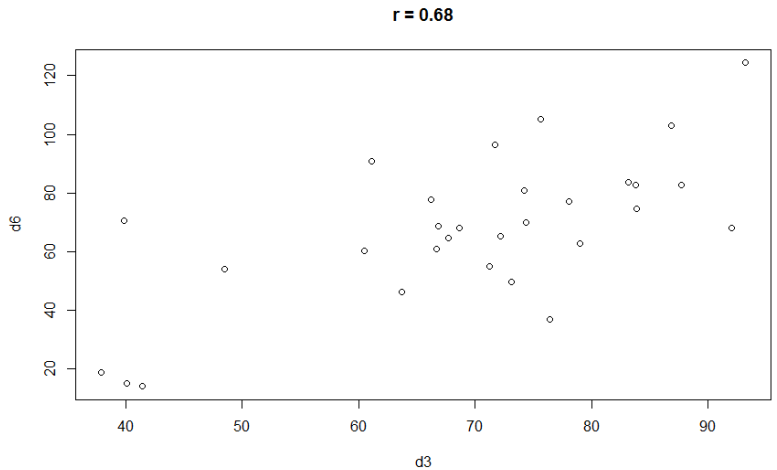
- We assess the strength of (a linear) relationship between two variables using the Pearson product-moment correlation coefficient.
- Everyone calls it Pearson's r
- Ranges from -1 to 1
  - -1 or 1 show perfect correlation
  - 0 shows no correlation
- Only captures a linear relationship

Reliability

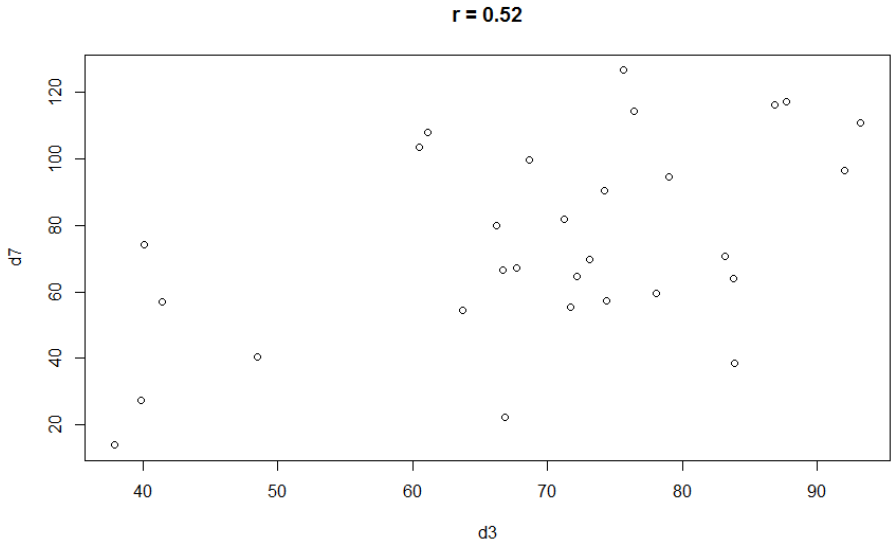


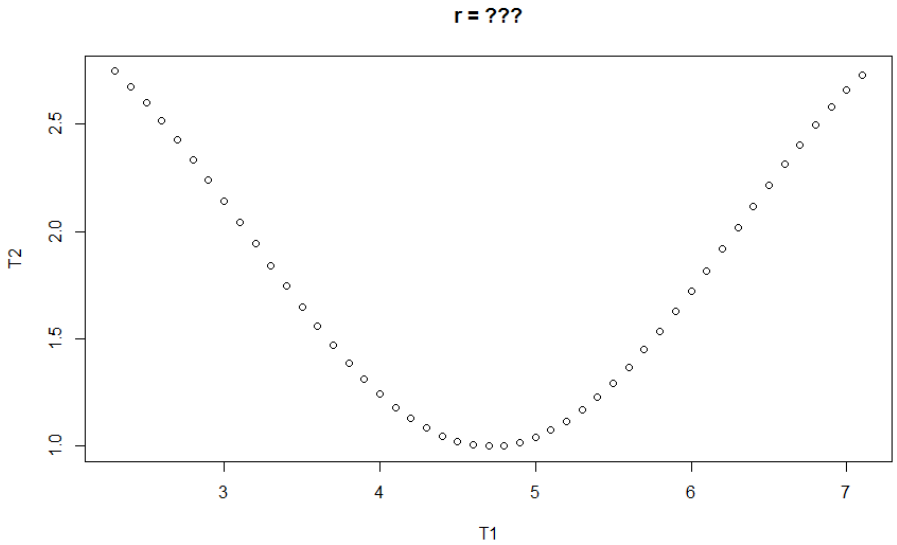






Reliability





Reliability

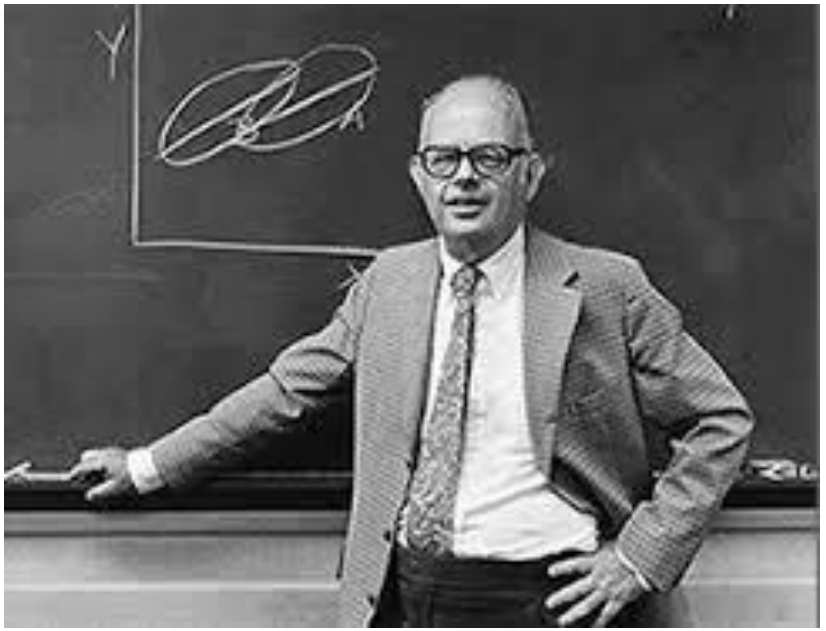
- We use Pearson's r to assess the relationship between two sets of scores and to determine the reliability of a test's scores.
- Where do we get two sets of scores from?
  - Give the test twice (Test-retest)
    - How much time between tests?
    - What if students learn more stuff between tests?
    - What if students forget some stuff between tests?

Reliability

- Rather than give the same test twice, what if we split the test in half to get two tests?
  - This is “split-half” reliability
  - What if students are more tired for the second half of the test?
  - We could split even-odd items…but that has problems as well.

Reliability

- Fortunately, we have a good statistic for this: Cronbach's Alpha
- Alpha gives us the average value of all possible split-halves of a test.
- We interpret Alpha the same way as Pearson' r



Reliability

- If we have true parallel forms of a test, we can correlate those two scores to get “alternate forms” reliability
  - Having two strictly parallel forms requires A LOT of psychometric development
  - Probably not possible/efficient for classroom assessments

Reliability

- What about performance tasks or observations?
- For these types of items, we look at inter-rater reliability
  - How correlated are the scores of two raters on the same set of tasks?
  - This is important so that a person's score depends on what they do, not who observes it!

Reliability

- In general, all other things being equal, longer tests are more reliable than shorter tests
- Tests that have higher stakes associated with them need to have higher reliability.
  - MEAP reliabilities are typically greater than 0.90
  - For our classroom assessments, 0.80 or greater is a good target.

Take a few minutes to reflect…

WHAT'S SO IMPORTANT ABOUT RELIABILITY?

Reliability is primarily a statistical computation…

Validity is much more….


LAW &
ORDER

Validity

- We have to gather and present evidence to establish the validity of a test for a stated purpose.
- Note: Tests aren’t valid or invalid. The inferences or uses of a test are what need to be evaluated for validity
- If you remember nothing else from today, remember that!

Validity

- So, what types of evidence do we gather?
- There are different “flavors” of validity
  - Face Validity
    - Sure, it kinda looks like a biology test.
  - Content Validity
    - People who know biology well think it looks like a biology test.

Validity

- Concurrent Validity
  - Results on this test “agree” with results on other, well established tests of biology.
- Criterion Validity
  - Results of this test “agree” with other ways to assess one’s knowledge of biology.
- Predictive Validity
  - Results on this test do a good job of predicting results on other tests (or things)

Validity

- Construct Validity
  - A construct is something we hypothesize exists but can't see directly. This is also referred to as a latent variable or latent trait
  - Most of the things we use tests for in education are really assessing constructs.
  - Construct validity subsumes all the other types of validity we've talked about.
    - You choose which flavor(s) of validity evidence you will gather based on the purpose of your test.

With some help from James Popham,

LET'S PRACTICE LOOKING AT VALIDITY

Let’s take a few minutes to reflect…

WHAT’S SO IMPORTANT ABOUT VALIDITY?

Different ways to look at test scores…

FRAMES OF REFERENCE

Frames of Reference

- There are different contexts in which to look at test scores:
  - Norm-Referenced
    - Student performance is compared to other test takers
  - Criterion-Referenced
    - Student performance is compared to pre-specified content
  - Standards-Referenced
    - Student performance is compared to a set of content standards and associated performance level descriptors

Looking at norm-referencing by

AUDIENCE PARTICIPATION!


YOU MUST BE
THIS TALL TO RIDE


48"
48"

Frame of Reference

- Standards-Referenced is very similar to criterion referenced
  - Rather than the criterion being any old content, the criteria are the content standards
  - May also be scored into performance levels based on performance standards
    - Yes….two different types of standards!
    - No…that doesn't add ANY confusion! ☺

Let’s take a few minutes to reflect…

WHAT DO WE NEED TO KNOW ABOUT FRAMES OF REFERENCE?

Converting student responses to numbers…

TEST SCORES

Test Scores

- Much of the analysis and reporting we do with test scores requires numbers
- The conversion of test responses to numbers is referred to as scoring
- The first, and most simple, score is a raw score
  - Presents the number of points earned on a test
  - May equal the number of items answered correctly
  - Is largely uninterpretable(!)

Test Scores

- There are two-types of test items
  - Dichotomously scored items
    - Right/wrong (two categories)
    - Scored 1 for a correct answer and 0 for a wrong answer
    - Could have the score multiplied for larger weight
  - Polytomously scored items
    - More than two categories
    - Scored 0 for no response up to the number of categories
    - Categories are ordered…higher scores indicate more/better information in the response

Test Scores

- The points assigned to the dichotomously scored and polytomously scored items are combined to provide the raw score.

- For example: “I got a ‘5’ on my last test.”

- Now, we can think about how to interpret the raw score.

Test Scores

- Percent Correct
  - Probably the most common score scale
  - Percent correct = raw score / total points possible
- Percentile Rank
  - Gives the percentage of raw scores that are equal to or less than a given student's raw score
  - The percentile may reference to all test takers who took the test at the same time as the student or to a previous group of test takers (norm group)

Test Scores

- z-scores (standard scores)
  - Expresses a data points distance from its mean in standard deviation units
  - This “centers” the distribution around the mean of the set of scores
  - It also removes the original scale and replaces it with standard deviation units.
  - Allows comparison of distributions that are on different scales.

Calculating a z-score

A z-score is defined as

$$z = \frac{x - \mu}{\sigma}$$

- Where:
  - x is an element of the data set
  - μ is the mean (average) of the distribution
  - σ is the standard deviation of the distribution

Calculating a z-score

- Find a data set
- Calculate the mean (μ) of that data set
  - The mean is the sum of the elements divided by the number of elements or

$$\mu = \frac{\sum x}{n}$$

Calculating a z-score

- Calculate the standard deviation (σ) of the data set
  - The standard deviation is a measure of how "spread out", or variable, the data are.
  - We calculate the standard deviation by

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

Or we can just ask Excel to calculate it! ☺ =stdevp()

Let's Practice…

$$z = \frac{x - \mu}{\sigma}$$

- A school's mean 5th grade MEAP science scale score is 526.8 with a standard deviation of 24.8.
- Calculate the z-scores for the following students

- Student A: 502
- Student B: 566
- Student C: 553
- Student D: 508
- Student E: 527
- Student F: 585

- A: -1, B: 1.6, C: 1.1 D: -.8, E: 0, F: 2.3

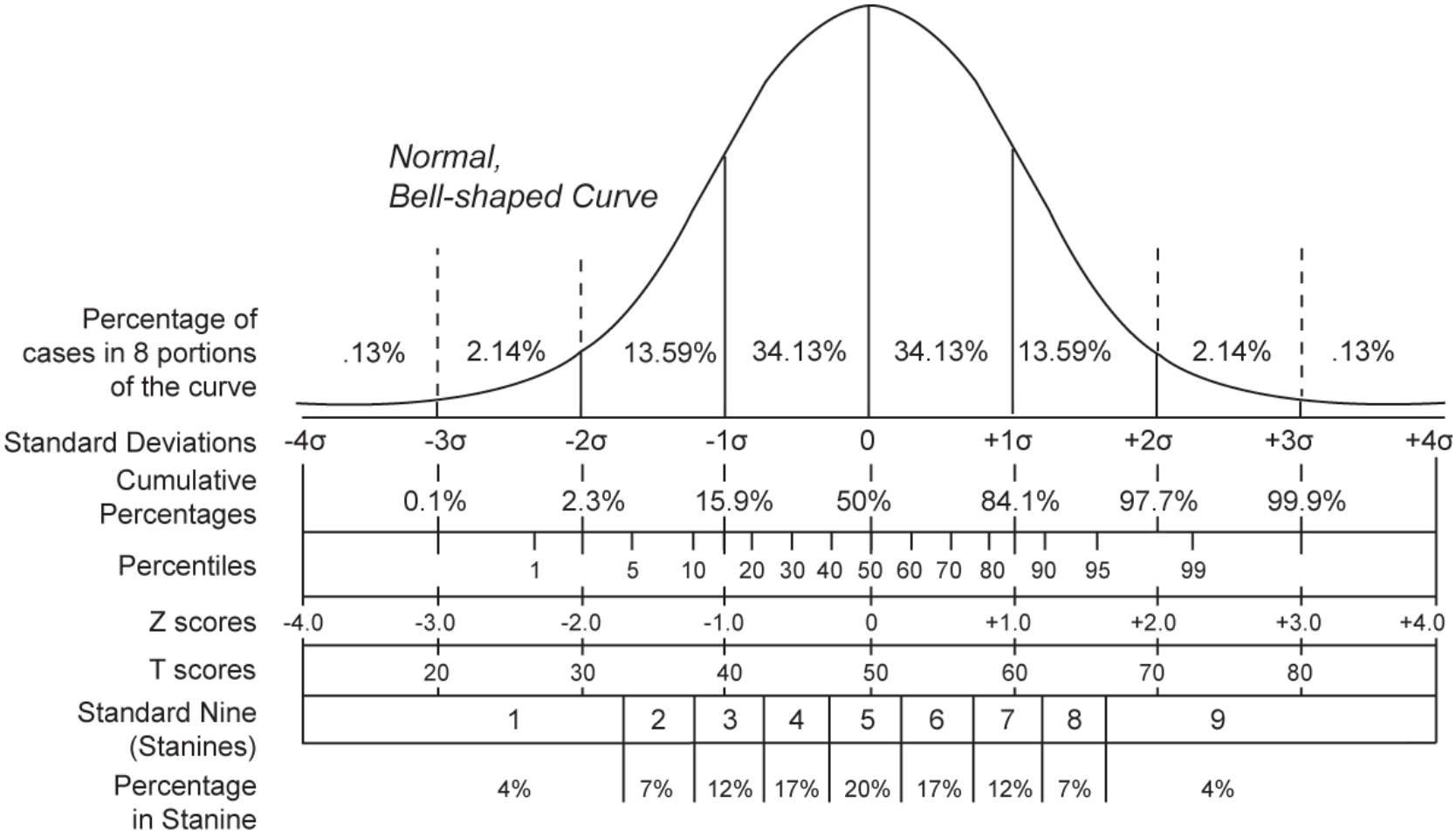
That wasn’t so bad…

LET’S FIRE UP EXCEL AND DO ANOTHER EXERCISE…

Test Scores

- There are a few other scales that are worth mentioning
  - Grade Equivalents
    - Very prone to misinterpretation!
  - Lexile
    - A very useful scale to measure reading ability in people and reading difficulty in text
  - There are others…

Many of these score types are related



Frames of Reference and Score Types

- Different frames of reference need different types of scores:
  - Norm-referenced
    - Percentile rank
    - Standard score (z-score) and transformations
    - Developmental Scales and Grade Equivalents
  - Criterion-referenced
    - Percent correct
  - Standards-referenced
    - Performance level

There's one additional scale that's worth mentioning, θ

Also known as the ability scale. This scale in important in item response theory and we'll talk more about it later.

This is a good place to take a break!

Let’s come back in 15 minutes…

We put a lot of effort into test scores

WHAT DO TEACHERS NEED TO KNOW ABOUT TEST SCORES?

Classical Test Theory

Sounds impressive, huh? ☺

Classical Test Theory

- O = T + E
  - where

    O = the observed score of a student

    T = the “true” score of a student

    E = some random error

- If you feel that it doesn’t quite live up to its name, that’s OK! ☺

Classical Test Theory

- In CTT, we look at
  - Test reliability coefficients…remember those?
  - Item difficulties
    - The percentage of students answering an item correctly (p-value) or the distribution of point values for a polytomously scored item
  - Item discriminations
    - How well does each item distinguish between high performing students (on the test) and low performing students (on the test).

Item Discrimination

- Ranges from -1 to 1
- “1” represents perfect discrimination, “0” represents no discrimination
- Negative values indicate a problem
  - Students who did well on the test had more difficulty on the item than students who didn’t do well on the test
  - Often indicates a mis-keyed answer.

Item Discrimination

- There are a few ways to calculate D
  - Order the test papers from high to low
  - Separate into two groups: high performing and low performing
  - Calculate the percentage of correct responses to the item in both groups. This gives you p_h and p_l
  - Discrimination (D) is given by: $D = p_h - p_l$

Item Discrimination

- Let's practice
  - Class of 30 divided into two groups of 15
  - Item one was answered correctly by 15 of the students in the top group and by 5 students in the bottom group
  - What is the items discrimination, D?
  - D= 1.0 - .33 = .67
  - How do we interpret this value?

Item Discrimination

- Ebel and Frisbie (1991) suggest the following interpretations of discrimination values:

.40 and above	Very good
.30 to .39	Reasonably good
.20 to .29	Marginal, needs improvement
.19 and below	Poor, reject or re-write

These are good guidelines, but interpreting discrimination is a little more complex.

Discrimination and difficulty are related

Item Discrimination

- Most testing software will calculate discrimination differently than how we just did.
- They will calculate a point-biserial correlation
- It is interpreted in the same way
- Makes the nomenclature a bit tricky
- D, pbis, p_{bis} , r_p, Disc, are all ways I've seen it denoted in software…it's all discrimination

Standard Error of Measurement

- The standard error of measurement (SEM) gives an estimate of the precision with which we've measured students
- Smaller values indicate more precision, larger values indicate less.
- We have to have a formula, so here it is!

$$SEM = s_x \sqrt{1 - r}$$

Where s_x is the standard deviation of the observed test scores and r is the reliability of the test

Standard Error of Measurement

- By adding and subtracting the SEM (or 2*SEM) from an observed score, we get a confidence interval.
  - The observed score is our “best guess” as to what the true score is…but we’re pretty certain that it’s not exact.
  - We are more confident that the true score lies within the confidence interval

Standard Error of Measurement

- Here's an example:
  - A student took the 5th grade MEAP test and earned a scale score of 646. The standard error of measurement is 8.
    - The 68% confidence interval is found by
    - $646 \pm 8 = 638 < T < 654$
    - The 95% confidence interval is found by
    - $646 \pm 2*8 = 630 < T < 662$

Standard Error of Measurement

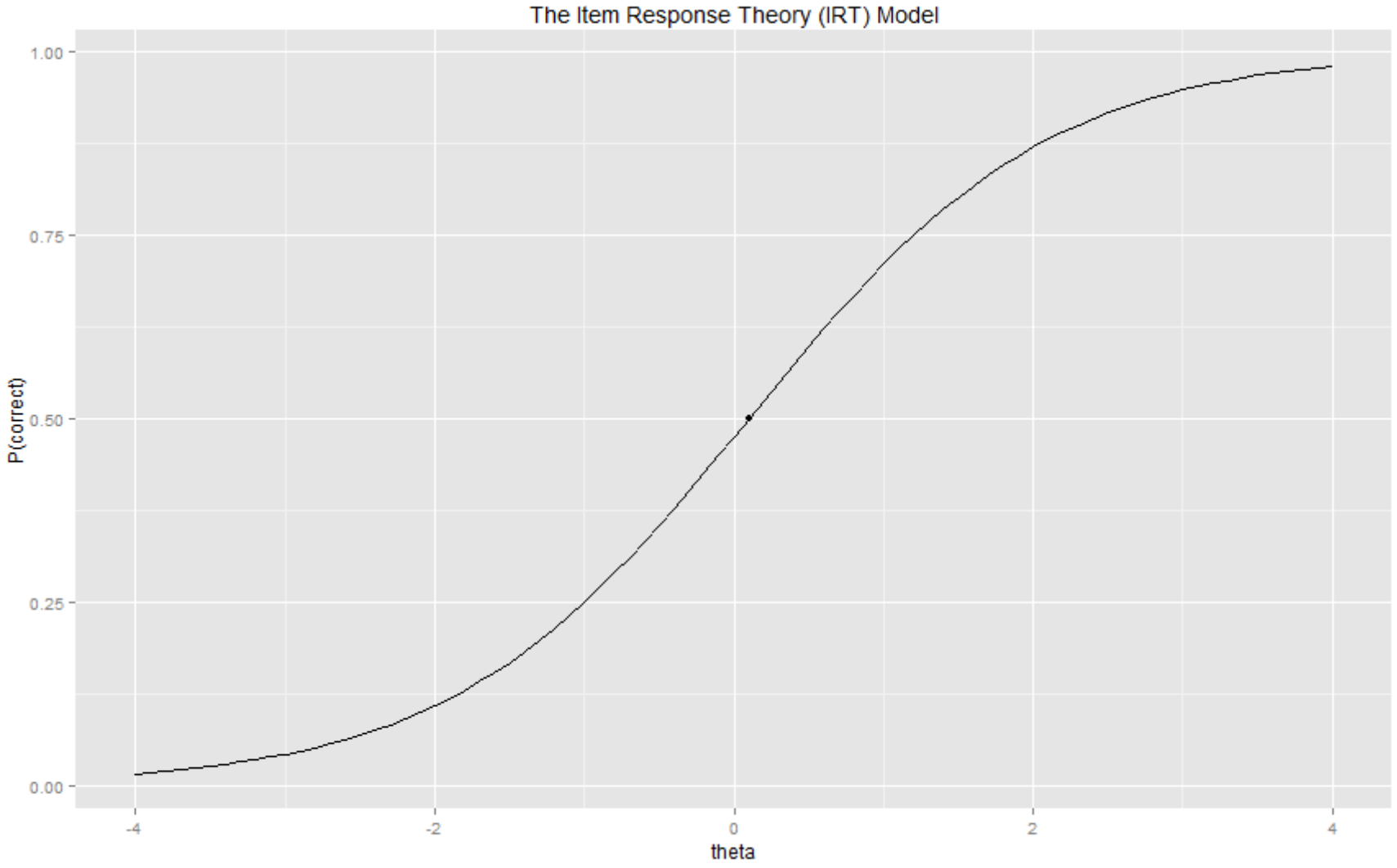
- The state uses the SEM and a confidence interval for accountability calculations
- If a student is not proficient, but the proficient cut score falls within the 68% confidence interval, the student is deemed “provisionally proficient” and counts as proficient in the calculations.
- A student is Level 3 with a SS = 334, SEM = 7 and the cut score is 336. This student is provisionally proficient.

Whew… Take a few moments, write down the important points about

CLASSICAL TEST THEORY

ITEM RESPONSE THEORY

Item Response Theory



Item Response Theory

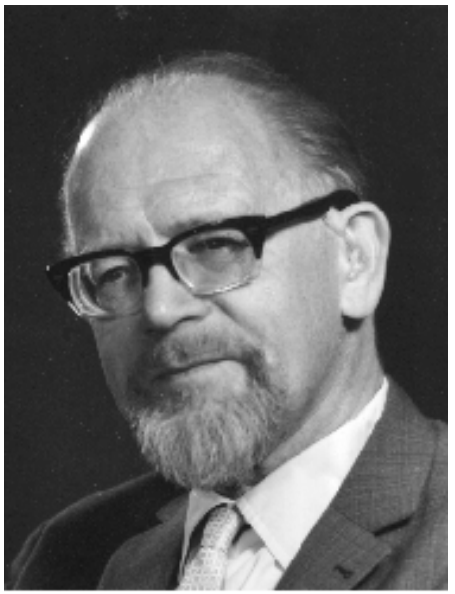
- As ability increases, the probability of correctly answering a question increases
  - This gives us the (logistic) ogive shape in the graph
- The point on the theta scale where p =.5 is the items difficulty (D, d, or δ)
- The slope of the curve is the discrimination of the item
- The curve approaches its maximum and minimum asymptotically

Item Response Theory

- Comes in a variety of flavors
- We'll talk about three today
  - 1-parameter logistic and Rasch Models
    - VERY similar but have some philosophical differences
    - We'll consider them together
  - 2-parameter logistic model
  - 3-parameter logistic model
- The difference between all of these models is the number of parameters that are allowed to change

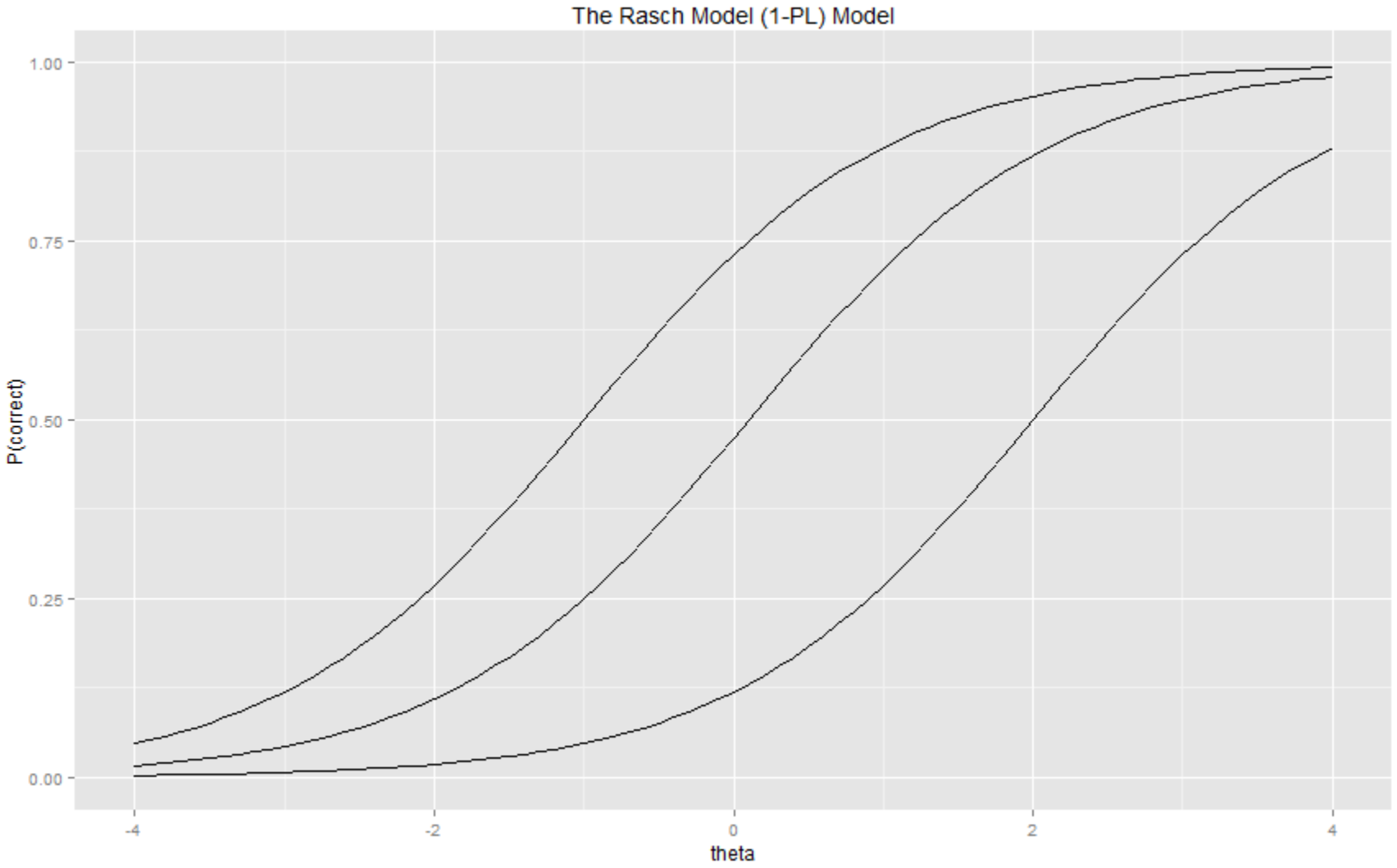
Item Response Theory

- The Rasch Model (1-PL model)
  - In the Rasch model, only the item difficulties are allowed to vary
  - The shapes of the curves (discriminations)are the same
  - 1-PL is scaled slightly differently than the Rasch model



Georg Rasch (1901-1980)

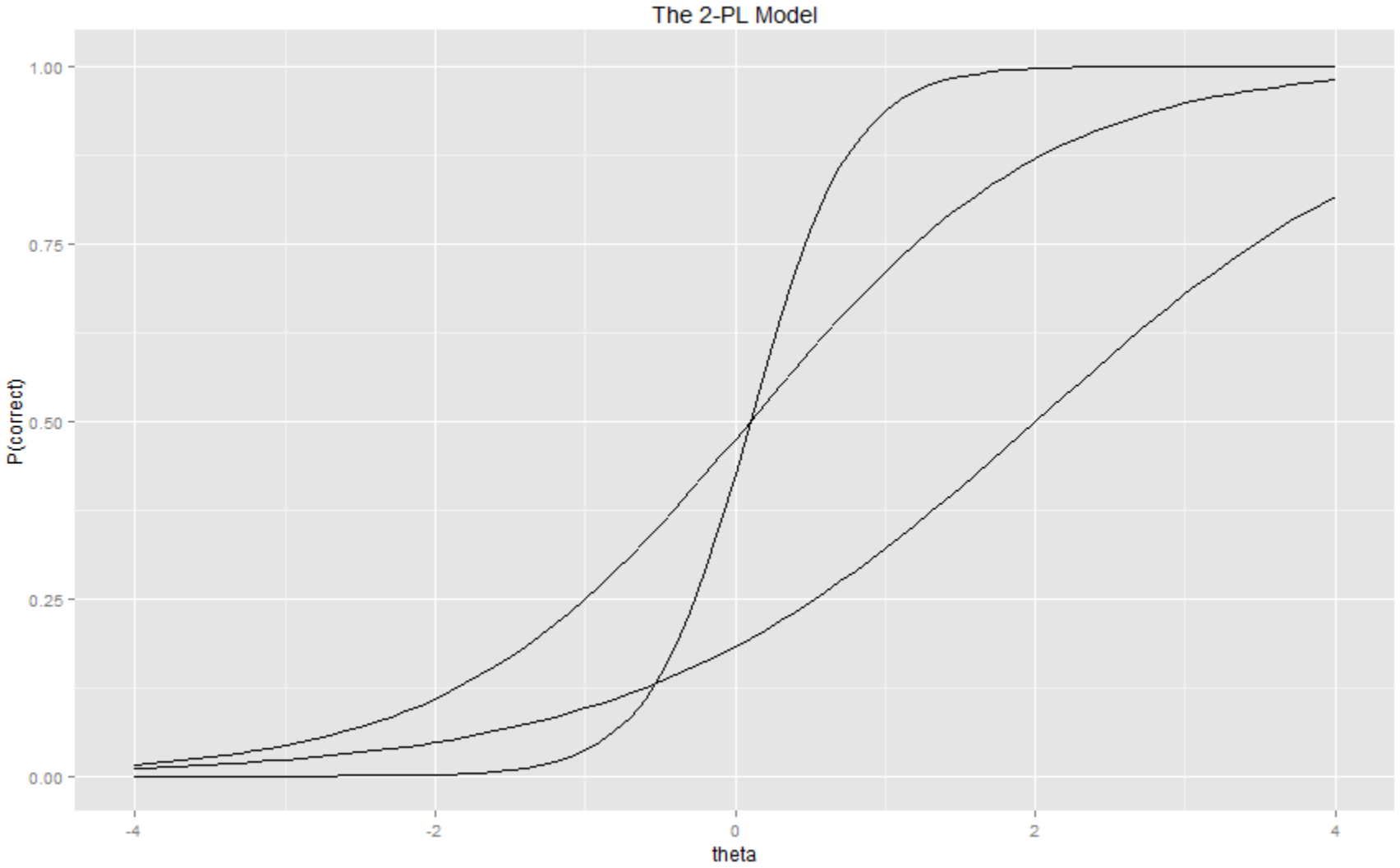
The Rasch Model



The 2-Parameter Logistic (2-PL) Model

- In addition to the difficulties varying, the discriminations are allowed to vary
  - The difficulties continue to be the location of the curve
  - The discriminations are the slopes of the curves

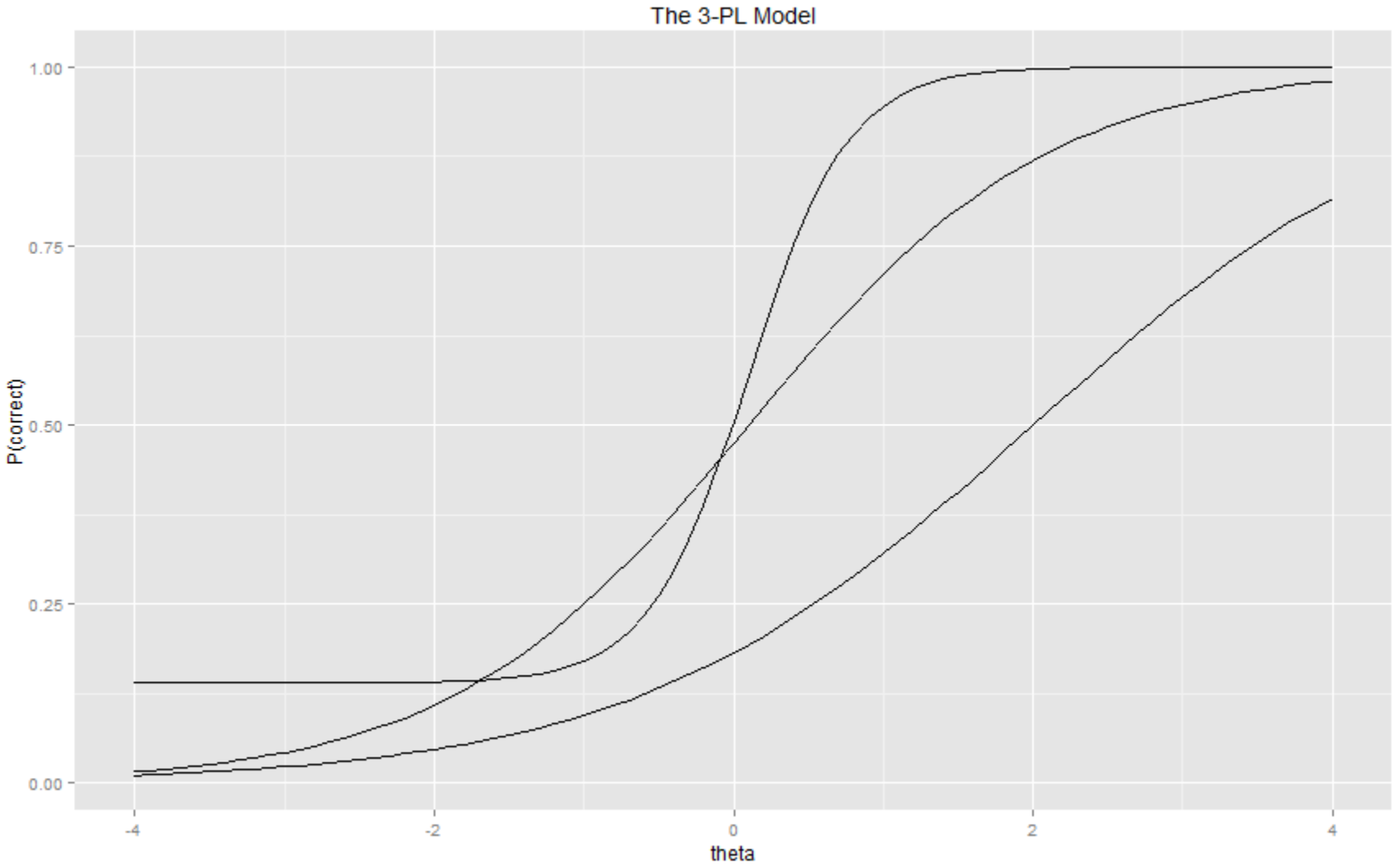
The 2-Parameter Logistic (2-PL) Model



The 3-Parameter Logistic (3-PL) Model

- In addition to the difficulty and discrimination parameters changing, the lower-asymptote can also vary from 0
  - This is referred to as “pseudo-chance”
  - It will typically be a value less than a random guess probability
    - This is why we don’t call it a guessing parameter
    - It is lower than chance because item writers create distractors that will be chosen by people who don’t know the content

The 3-Parameter Logistic Model



Item Response Theory

- A few final notes on IRT
  - Note that we can place item difficulty and person ability on the same scale
  - A person's ability (location estimate) is based on not only how many questions are answered correctly but also which items are answered correctly for the 2-PL and 3-PL models
  - Rather than a SEM for the whole test, IRT gives us a standard error of estimate that is different at different points along the scale
  - IRT is at the heart of computer adaptive testing

IRT Appendix

- Let's look at the formulas for the three IRT models
- Rasch: 1-PL:

$$P(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad P(\theta) = \frac{e^{1.7(\theta - b_i)}}{1 + e^{1.7(\theta - b_i)}}$$

Note that there are only two variables in this equation: θ and b_i (the i is a subscript denoting a particular item)

IRT Appendix

- 2-Parameter Logistic

$$P(\theta) = \frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}}$$

The 1.7 is a scaling constant

There are three variables in this equation:

θ, b_i, and a_i (discrimination)

IRT Appendix

- 3-Parameter Logistic

$$P(\theta) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}}$$

There are four variables in this equation:

θ, b_i, a_i, and c_i ("pseudo-chance")

That’s probably enough…

WHAT WILL YOU TAKE AWAY ABOUT ITEM RESPONSE THEORY?

STANDARD SETTING

Standard Setting

Definition (Gregory Cizek):

> The task of deriving levels of performance on educational or professional assessments, by which decisions or classifications of persons (and corresponding inferences) will be made.

Standard Setting

- There are LOTS of different standard setting techniques
  - We'll talk about two that you could use at your school
  - We'll talk a little about how the "college and career ready" standards were set on MEAP/MME
- Standard setting is basically a policy (or political) decision
  - There is no right answer or "true" standard

Angoff's Procedure

- William Angoff worked as a distinguished research scientist at ETS.
- He described the procedure in a footnote to something else he was writing
- He was actually uncomfortable with the method and attributed it to someone else
- It has a number of variations that have been used

A (modified) Angoff Procedure

- Picture in your mind the minimally competent student
- For each question, determine how many…out of 100…minimally competent students would get that question correct
- After you've done this for each item, add up your numbers and divide by 100. This is your cut score
- Compile the cut scores of all the standard setters and take the median value

The Bookmark Method

- Reorder the items on the test from easiest to hardest and distribute them to the panel
- Think about a minimally competent student
- Start at item 1 and decide if the minimally competent student would get that right. If so go on to the the next item and repeat
- When you get to an item you think the MCS wouldn’t get right, put your bookmark there.
- Take all panelists bookmark locations and take the median

MEAP and MME

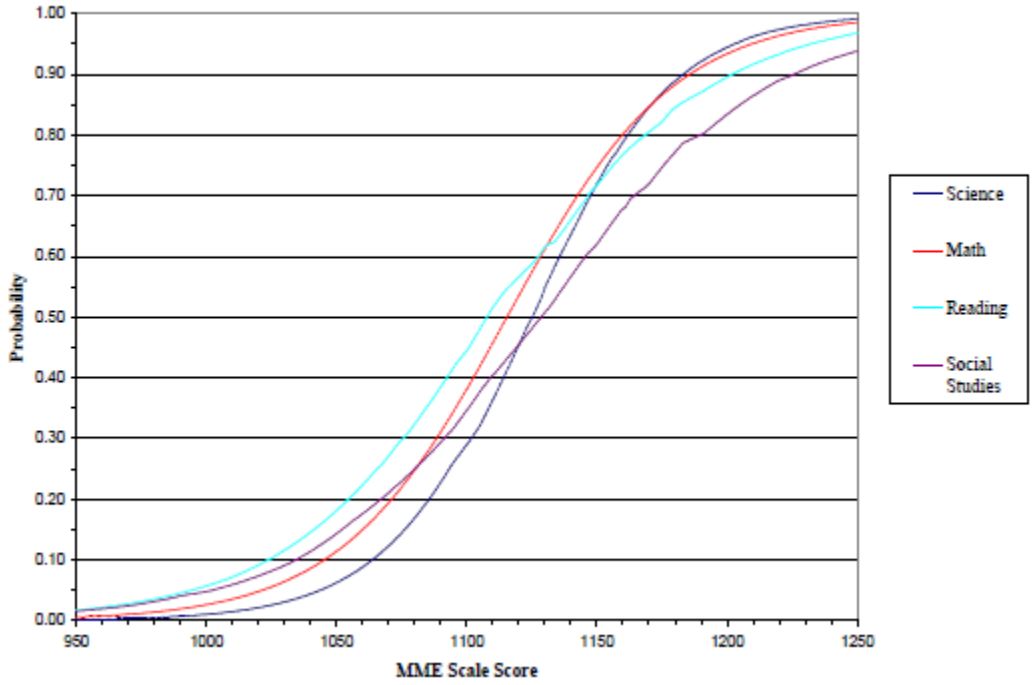
COLLEGE AND CAREER READY CUTSCORES

The “New” Cutscores

- New college and career ready standards were developed during the Fall of 2011
- Attempts to link student test performance to college course grades
- Links MME scores to college grades
- Links MEAP scores to MME scores
- Uses a number of different methodologies
- What follows is a simplified overview.

The “New” Cutscores

Figure 1. *Probability of a B or Higher in College Courses by MME Subject Area Test*



- Cut scores based on probability of earning a “B” or higher:
- Partially Proficient (33%)
- Proficient (50%)
- Advanced (67%)

The “New” Cutscores

- Once the link from MME to college grades was made, links were made between MEAP scores and MME scores
- A number of different methodologies were used (don’t worry about the details of these today)
  - Signal Detection Theory
  - Logistic Regression
  - Equipercentile Cohort Matching

The “New” Cutscores

Links in Tying Cut Scores on MME and MEAP to College Grades

Cut score		
Content area	**Grade**	**Links created**
Mathematics and Reading	3	#1. Grade 11 MME to College Grades #2. Grade 7 MEAP to Grade 11 MME #3. Grade 3 MEAP to Grade 7 MEAP
	4	#1. Grade 11 MME to College Grades #2. Grade 7 MEAP to Grade 11 MME #3. Grade 4 MEAP to Grade 7 MEAP
	5	#1. Grade 11 MME to College Grades #2. Grade 7 MEAP to Grade 11 MME #3. Grade 5 MEAP to Grade 7 MEAP
	6	#1. Grade 11 MME to College Grades #2. Grade 7 MEAP to Grade 11 MME #3. Grade 6 MEAP to Grade 7 MEAP
	7	#1. Grade 11 MME to College Grades #2. Grade 7 MEAP to Grade 11 MME
	8	#1. Grade 11 MME to College Grades #2. Grade 8 MEAP to Grade 11 MME
	11	#1. Grade 11 MME to College Grades
Science and Social Studies	5/6	#1. Grade 11 MME to College Grades #2. Grade 8/9 MEAP to Grade 11 MME #3. Grade 5/6 MEAP to Grade 8/9 MEAP
	8/9	#1. Grade 11 MME to College Grades #2. Grade 8/9 MEAP to Grade 11 MME
	11	#1. Grade 11 MME to College Grades

Looking At Your Assessment System: A Graphical Perspective of Balance

Let's give credit…

- Much of this presentation is based upon:
- Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. Educational measurement: Issues and practice. 28(3) pp. 5-13.
- The entire issue is devoted to thinking about formative and interim assessments.

A Balanced Assessment System meets the *legitimate* needs of all stakeholders.

- Students & Parents
- Teachers
- School Administrators
- District policy makers
- State policy makers

What information do we need?

- Students:
  - How am I progressing toward mastery?
- Teachers
  - How are my 25 (40, 150) students progressing toward mastery
- Administrators
  - Are the tools that my teachers have providing acceptable student achievement?
- State
  - What schools need more support/recognition?

Each of these information needs is important…

…but they require different types of data and tests that are built to provide that data.

We’re really talking about validity across our system.

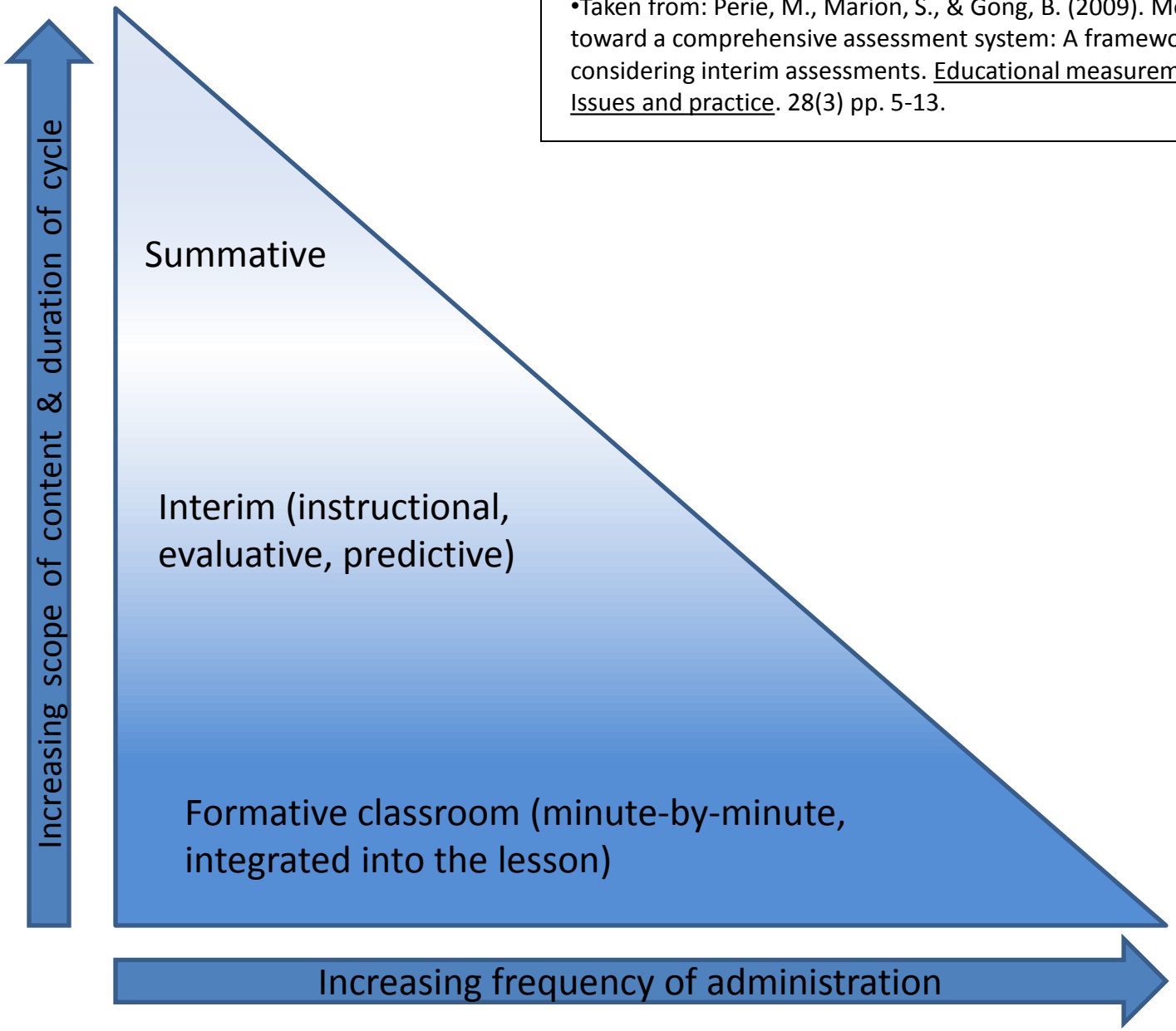
It all comes down to being very clear about why you’re giving a certain test and what you need from it.

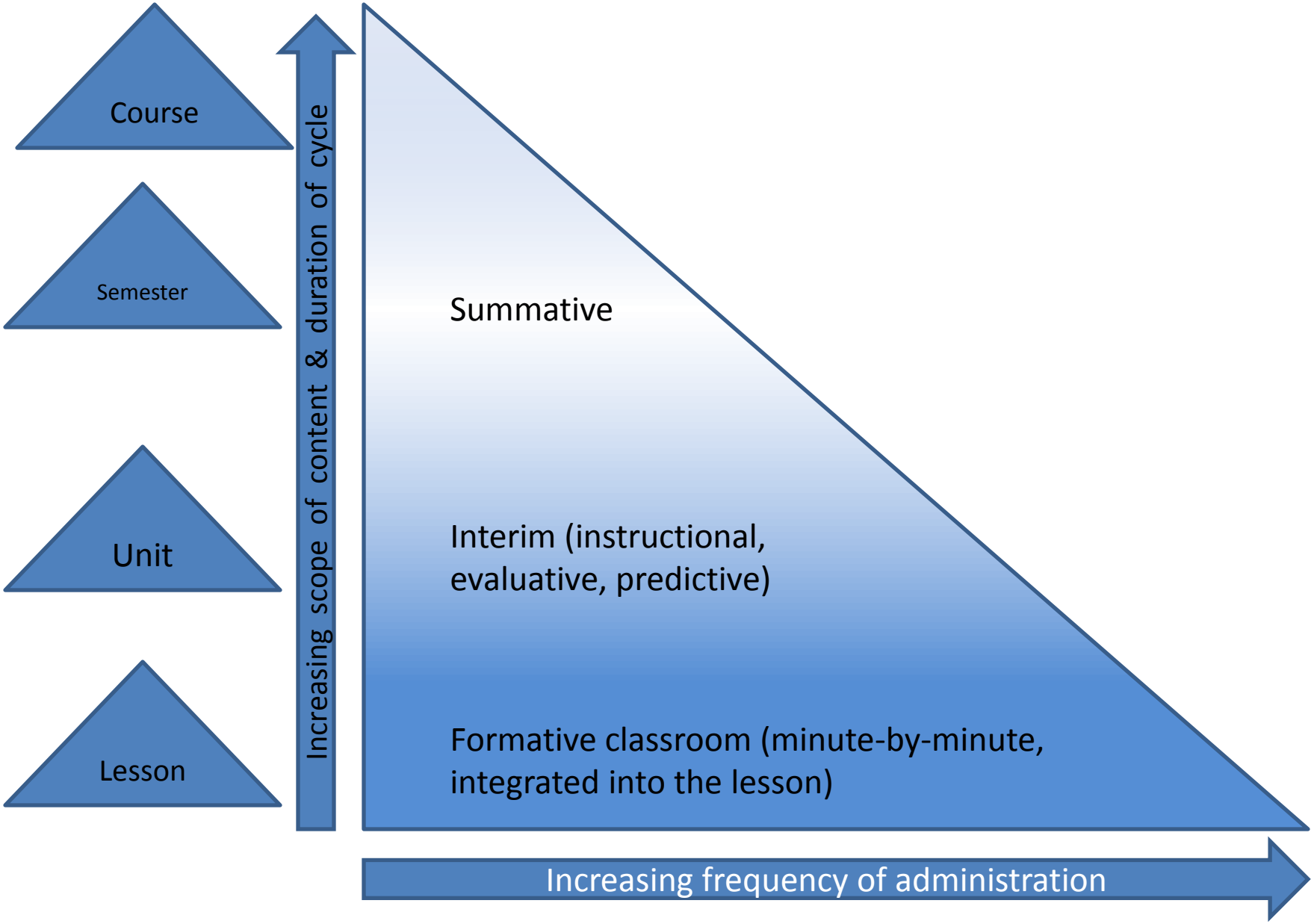
All we've got is language, really.

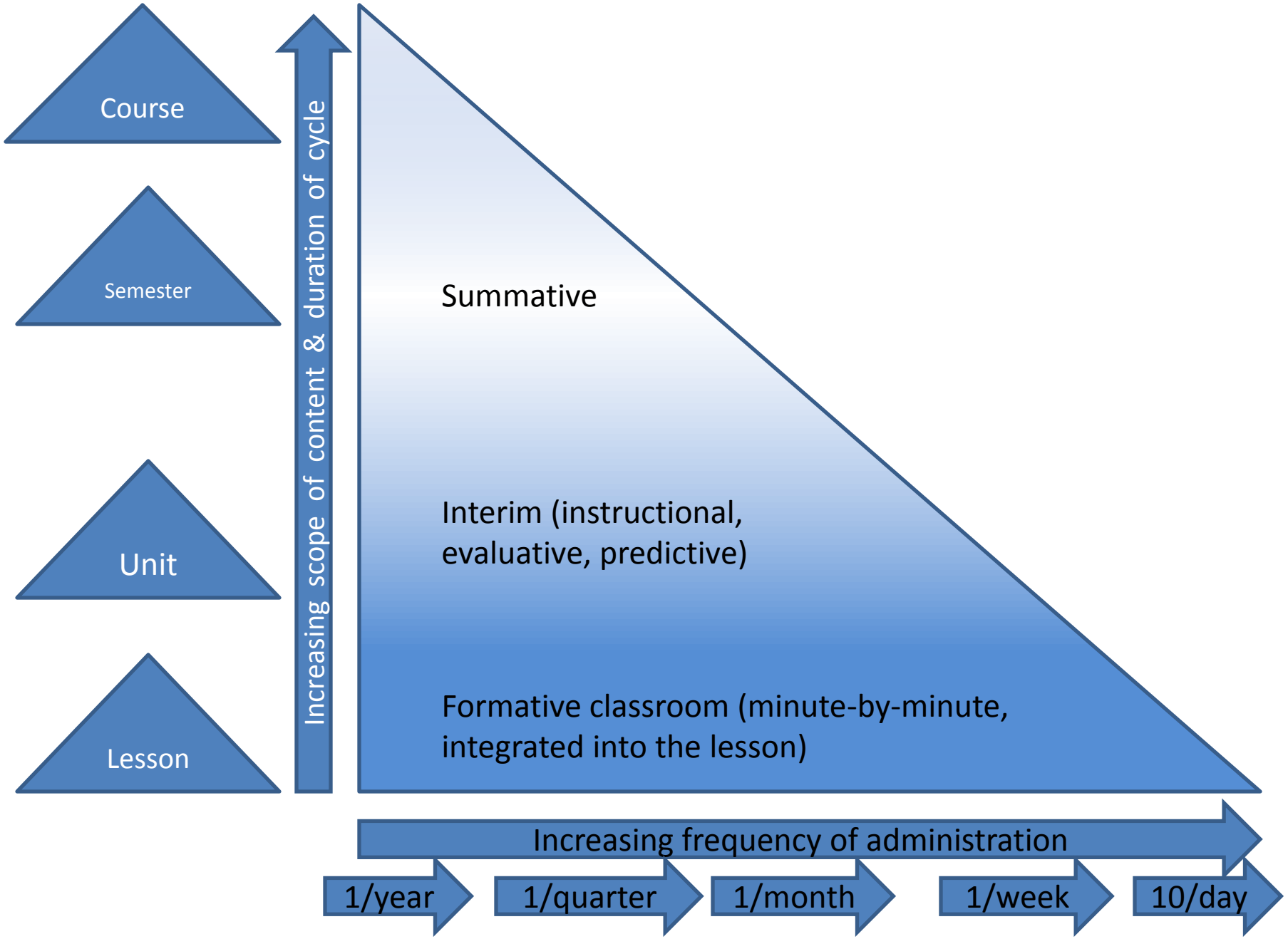
- Language that caused "significant" discussion as we thought about comprehensive assessment systems:
  - "Assessment"
  - "Balanced Assessment"
  - "Cognitive Assessment"
  - "Common Assessment"
  - "interim assessment"
  - "Curriculum"

Perhaps a picture will help…

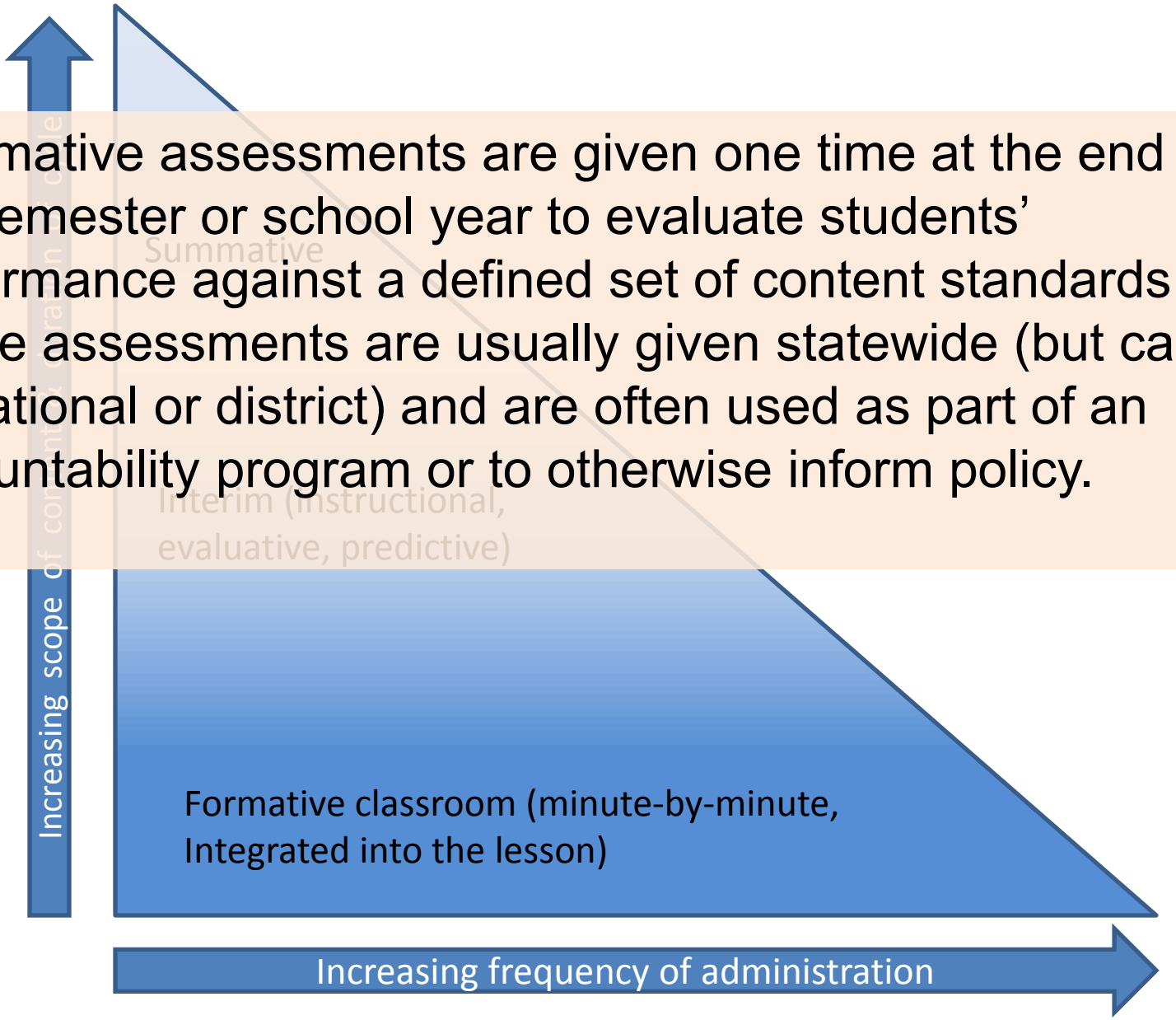
•Taken from: Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. Educational measurement: Issues and practice. 28(3) pp. 5-13.

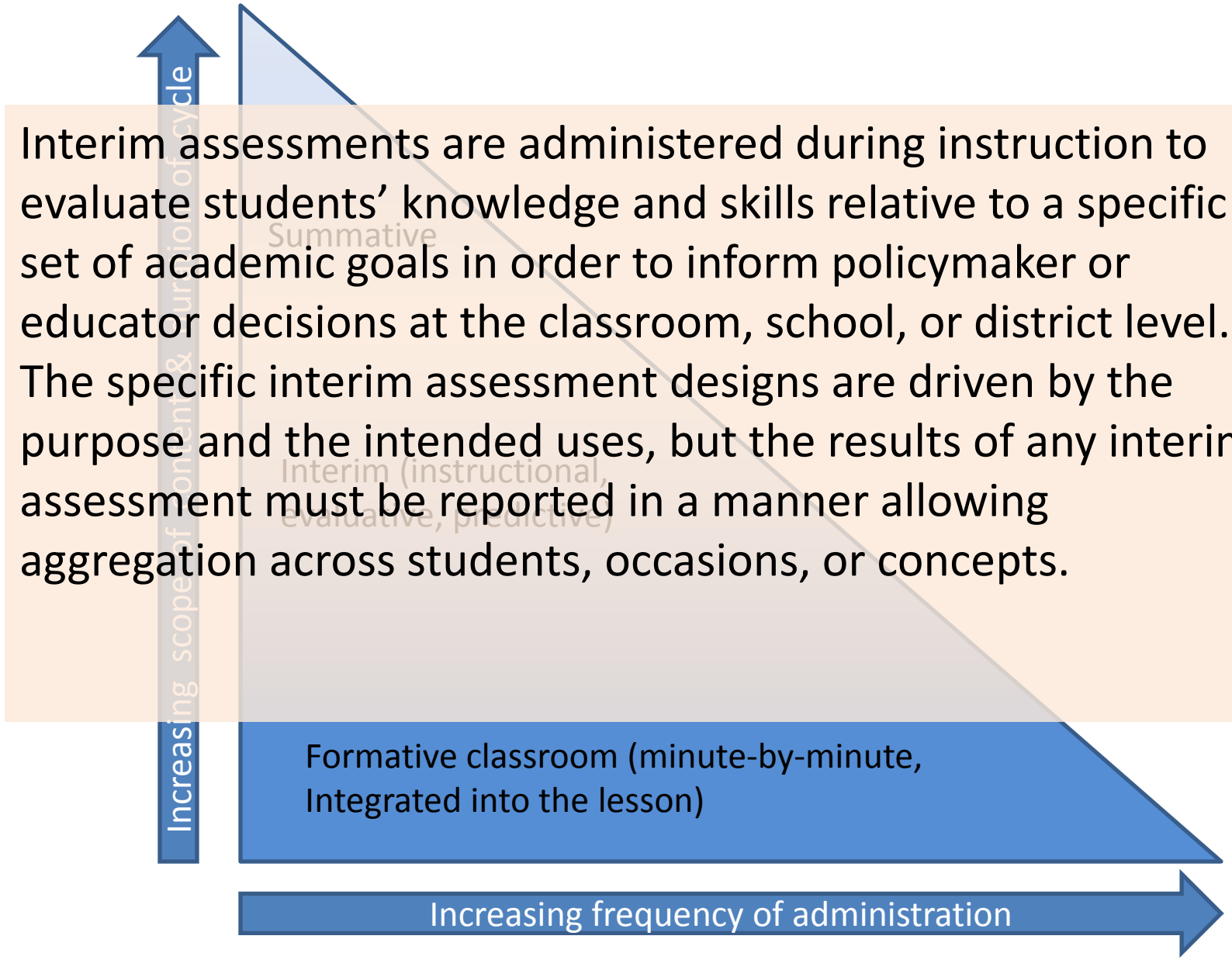



Course
Semester
Unit
Lesson
Increasing scope of content & duration of cycle
Summative
Interim (instructional, evaluative, predictive)
Formative classroom (minute-by-minute, integrated into the lesson)
Increasing frequency of administration


Course
Semester
Unit
Lesson
Increasing scope of content & duration of cycle
Summative
Interim (instructional, evaluative, predictive)
Formative classroom (minute-by-minute, integrated into the lesson)
Increasing frequency of administration
1/year
1/quarter
1/month
1/week
10/day

Summative assessments are given one time at the end of the semester or school year to evaluate students' performance against a defined set of content standards. These assessments are usually given statewide (but can be national or district) and are often used as part of an accountability program or to otherwise inform policy.

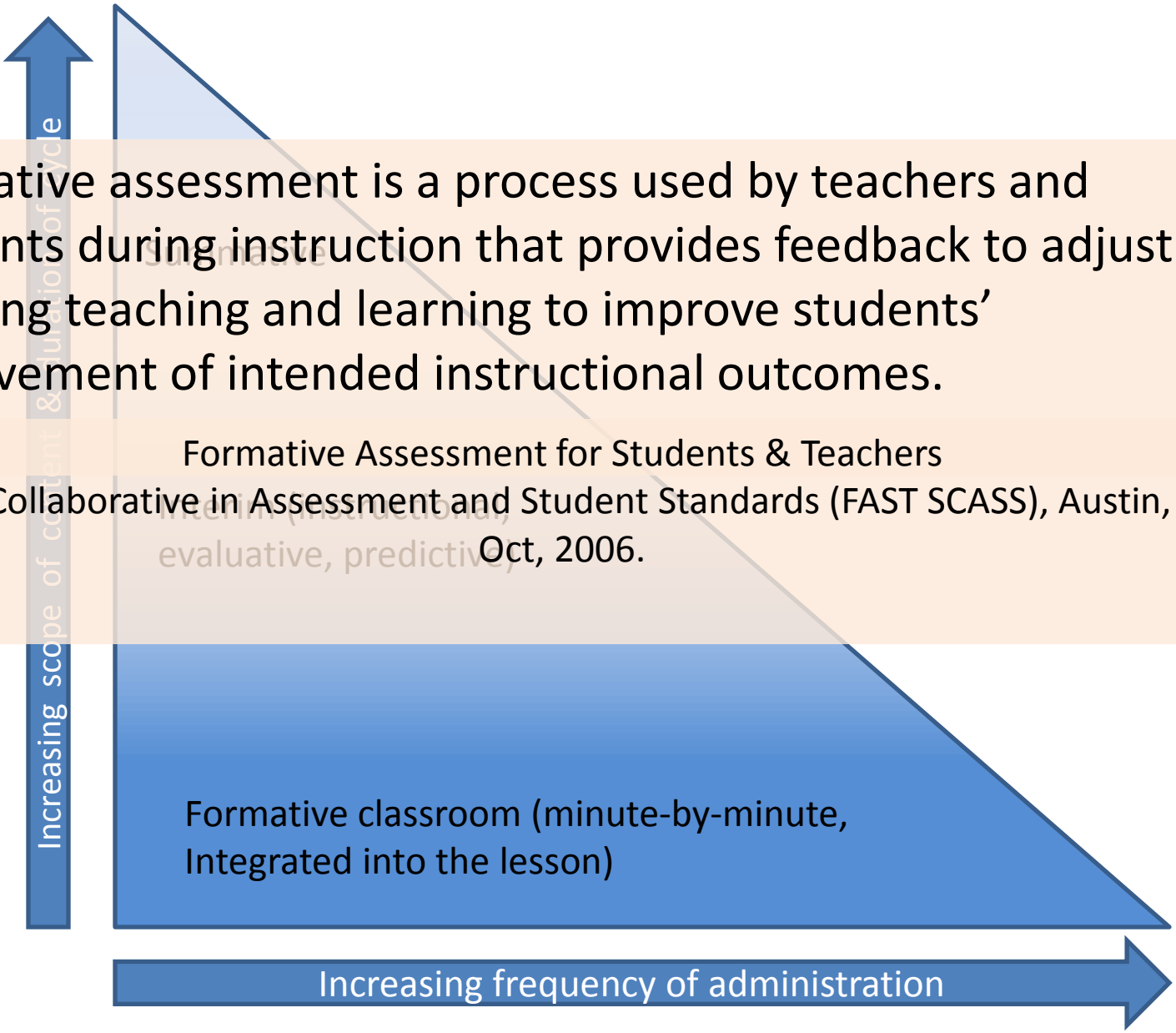




Interim assessments are administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level. The specific interim assessment designs are driven by the purpose and the intended uses, but the results of any interim assessment must be reported in a manner allowing aggregation across students, occasions, or concepts.

Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes.

Formative Assessment for Students & Teachers
State Collaborative in Assessment and Student Standards (FAST SCASS), Austin, TX, Oct, 2006.



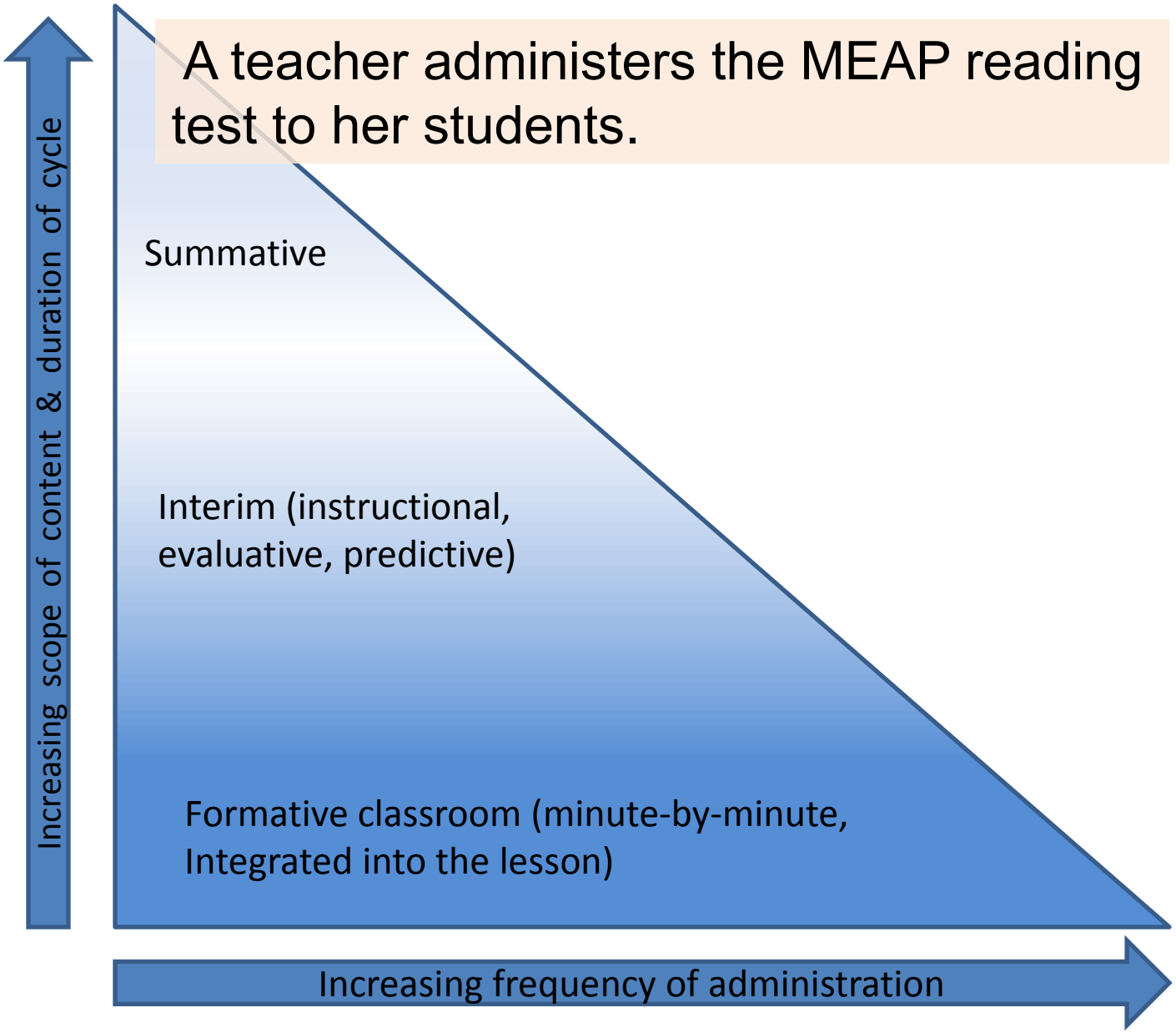
Formative assessment is a planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their ongoing instructional procedures or by students to adjust their current learning-tactics.

James Popham, MSTC, Feb. 24, 2009

An assessment is formative to the extent that information from the assessment is used, during the instructional segment in which the assessment occurred, to adjust instruction with the intent of better meeting the needs of the students assessed.

From Perie, 2007

1. **At your table, place the assessment scenarios on the triangle.**
2. **Share what your table did with another table.**



DEPTH OF KNOWLEDGE

Depth of Knowledge

Recall
Skills and Concepts
Strategic Thinking
Extended Thinking

Depth of Knowledge (DOK)

- A taxonomy that can be used to classify test items and content expectations
  - Much like Bloom or UbD.
- Developed by Norm Webb at the University of Wisconsin.
  - http://facstaff.wcer.wisc.edu/normw/
- This is taxonomy that Michigan uses in developing the MEAP and MME. SMARTER Balanced also uses DOK.

Recall

The recall of information (fact, definition, or term), or performing a simple procedure, or applying a simple algorithm or formula.

Requires only a rote response, a well-known formula, or following a well defined procedure that typically involves only **one** step.

Recall

(verbs that might be used)

- ✓ Identify
- ✓ Select
- ✓ Name
- ✓ Describe
- ✓ Define
- ✓ Locate
- ✓ Label
- ✓ Match
- ✓ Give an Example
- ✓ Cite
- ✓ Recall
- ✓ State

Recall (Science)

What information can always be obtained from a topographic map?

a. types of wildlife

b. elevation*

c. temperature

d. types of rocks

Skills and Concepts

Items require students to make some decisions and typically involve more than one step.

Students use information or conceptual knowledge when selecting the response.

Skills and Concepts

- ✓ Restate
- ✓ Change
- ✓ Solve
- ✓ Illustrate
- ✓ Confirm
- ✓ Extend
- ✓ Summarize
- ✓ Predict
- ✓ Classify
- ✓ Choose
- ✓ Convert
- ✓ Discuss
- ✓ Estimate
- ✓ Explain
- ✓ Generalize
- ✓ Construct
- ✓ Determine
- ✓ Use

Skills and Concepts (Science)

Which of the following observations would provide the best evidence to support that Eris is NOT a star?

A. It has mass, density, and a circumference that can be estimated.

B. Its light, which has been observed, is reflected from the Sun.

C. It is found in the Milky Way Galaxy, which includes our solar system

D. Its brightness can be used to help calculate its size and temperature.

Skills and Concepts (Soc. St.)

Snowfall and Temperatures in Lansing, Michigan, 2004

Month	Snowfall (inches)	Average Monthly Temperature (°F)
January	28.3	16.4
February	7.2	23.3
March	2.9	38.8
April	0.5	48.3
May	Trace	58.2
June	0	65.0
July	0	69.0
August	0	65.1
September	0	64.7
October	0	51.3
November	5.5	40.4
December	13.9	28.2

- What is the relationship shown in the chart?

- A. There were two months with two inches of snow in 2004.
- B. The coldest three months have the most snowfall.
- C. There were 6 months without snow in 2004.
- D. The warmest five days are in the fall months.

Strategic Thinking

Items at this level require planning, using evidence, and complex and abstract reasoning. (often explain their thinking)

Students are asked to draw conclusions, cite evidence, develop logical arguments, solve complex problems, explain concepts, and justify their responses

Strategic Thinking

- ✓ Compare
- ✓ Contrast
- ✓ Summarize
- ✓ Construct
- ✓ Organize
- ✓ invent

- ✓ Differentiate
- ✓ Analyze
- ✓ What if . . .
- ✓ Critique
- ✓ Assess
- ✓ Create

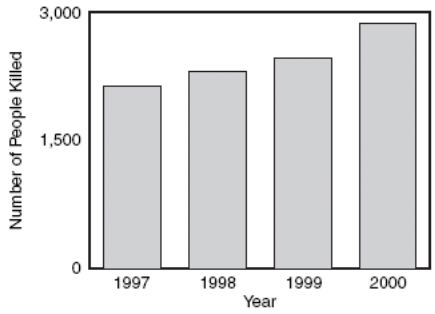
Strategic Thinking (Soc. St.)

DATA SECTION

Part A

The National Highway Traffic Safety Administration determined that the number of people involved in fatal motorcycle accidents increased 30% between 1997 and 2000.

Fatal Motorcycle Accidents in the United States, 1997-2000



Source: National Highway Traffic Safety Administration

DATA SECTION (continued)

Part B

Currently each state has its own helmet law. This table shows the helmet laws in selected states and the percent of traffic deaths that are motorcycle riders.

Deaths of Motorcycle Riders in Traffic Accidents, 2003

Selected State	Helmet Law	Percent of Total Traffic Deaths
Alabama	Required for all riders	5.1
California	Required for all riders	9.1
Connecticut	Required for riders under 18	9.5
Indiana	Required for riders under 18	9.7
Iowa	No helmet use required	11.6
Wisconsin	No helmet use required	12.1

Source: National Highway Traffic Safety Administration

- Should the United States government require all motorcyclists to wear helmets?
- Include:
  - clear and supported position
  - Core democratic value
  - Supporting knowledge from history, geography, civics, or economics
  - Supporting information from the data section.

Strategic Thinking (Math)

- If a and b are real numbers such that

$$0 < a < 1 < b$$

which of the following must be true of the value ab?

A. $0 < ab < a$
B. $0 < ab < 1$
C. $a < ab < 1$
D. $a < ab < b$
E. $b < ab$

Extended Thinking

Items require complex reasoning, experimental design, and planning usually over extended periods of time.

Students are asked to make connections *within* or *among* content areas.

Many on-demand instruments will NOT include any items at this level.

Extended Thinking

- Based on provided data from a complex experiment that is novel to you, deduce the fundamental relationship between several controlled variables

- Conduct an investigation, from specifying a problem to designing
  and carrying out an experiment,
  to analyzing its data and forming
  and justifying conclusions.

Practicing with DOK

Target Method Match

- Rick Stiggins talks about matching assessment methods with assessment targets.

CLASSROOM ASSESSMENT MATRIX		METHOD				
		WRITTEN RESPONSE (Paper & Pencil or Computer)			**PERFORMANCE/ OBSERVATION**	**INTERACTIVE/ CONVERSATION**
		Selected Response	*Short-Response*	*Extended-Response*		
		True/False, multiple-choice, and matching	Fill-in-the-blank and short answer	Essays, research reports and lab reports	Public performances, investigations	Oral exams, interviews, discussion groups
TARGET	**KNOWLEDGE MASTERY**	Can sample mastery of knowledge elements.	Can sample mastery of knowledge elements and suggest understanding of relationships.	Can tap understanding of relationships among elements of knowledge.	Not recommended.	Allows the examiner to explore mastery selectively, but in depth, as responses to questions are evaluated.
	REASONING PROFICIENCY	Can assess understanding of basic patterns of reasoning.	Brief descriptions of simple problem solutions can provide a shallow window into reasoning proficiency.	Longer descriptions of complex problem solutions can provide a deeper window into reasoning proficiency.	Can infer reasoning proficiency from direct observation of student problem solving behaviors.	Can infer reasoning proficiency more deeply by asking student to "think aloud" or through focused, probing follow-up questions.
	SKILLS	Can assess mastery of knowledge prerequisite to the ability to create quality products–but cannot assess the quality of the products themselves.		Can assess skill in writing directly, but otherwise limited to prerequisite knowledge.	Can directly observe and evaluate skills as they are being performed.	Can assess skill in oral communication directly, can also assess mastery of prerequisite knowledge.
	ABILITY TO CREATE PRODUCTS	Can assess mastery of knowledge prerequisite to the ability to create quality products–but cannot assess the quality of the products themselves.		Can assess ability to create written product directly, but otherwise limited to prerequisite knowledge.	Can assess directly (a) proficiency in carrying out steps in product development, and (b) attributes of the product itself.	Can probe knowledge of procedures and attributes of quality—but not product quality itself.
	DISPOSITIONS	Surveys/questionnaires can tap student feelings and attitudes.	Open response items can capture additional information not included in a fixed survey.	Open-ended questions can elicit deep responses about feelings and attitudes.	Dispositions can be inferred from behavior and products.	Feelings and attitudes can be explored and probed in depth.

Learning Target Assessment Worksheet

Learning Target: _____

Depth of Knowledge: Recall Skills and Concepts Strategic Thinking Extended Thinking

Pre-requisite Knowledge/Skills: _____

Evidence needed to infer mastery:

Types of questions to be asked on the common assessment:

How many questions of each DOK level should be asked?

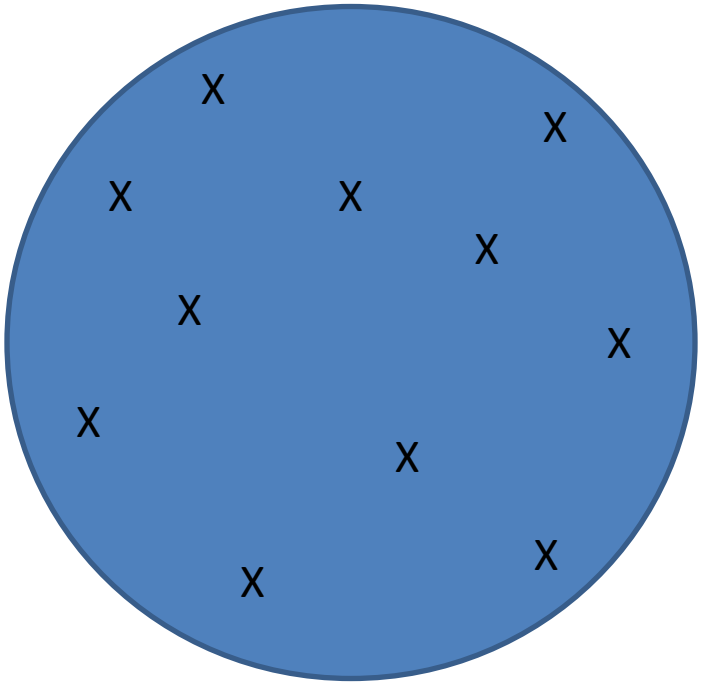
Recall: ____ Skills and Concepts: ____ Strategic Thinking: ____ Extended Thinking: ____

What Items should we use?

- It is rare that we can ask every possible question related to a content area
- We have to sample from the domain to choose test questions.
- How we sample should reflect the purpose of our test

Sampling

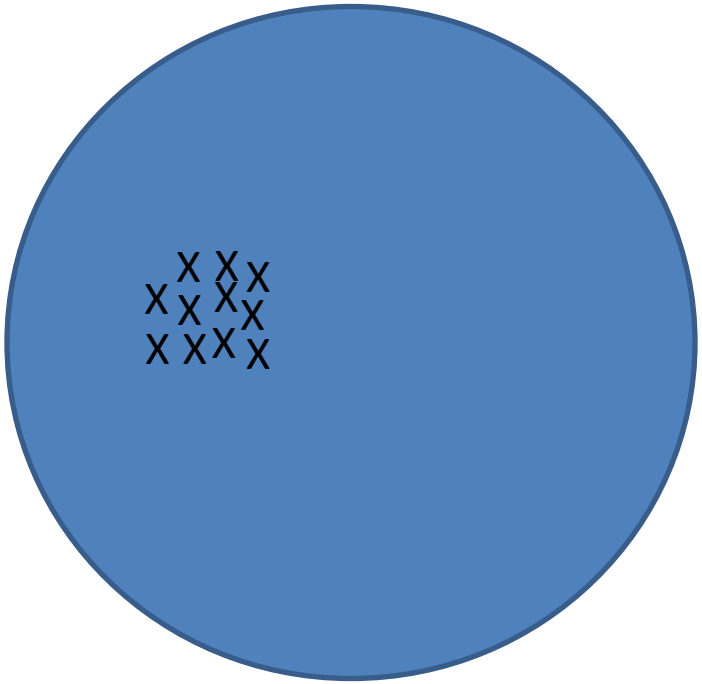
- Consider a test on Addition



- 1 + 2 = ___
- 4 + 4 = ___
- 5 + 9 = ___
- 153 + 34 = ___
- 1.3 + 6.0 = ___
- -1 + 5 = ___
- ½ + ¾ = ___
- 2x + 4x = ___

Sampling

- Consider another test on Addition



- 5 + 9 = ___
- 17 + 5 = ___
- 15 + 16 = ___
- 37 + 81 = ___
- 1.3 + 6.8 = ___
- 124 + 9 = ___
- 357 + 864 = ___
- 18x + 4x = ___

What are appropriate and inappropriate uses/interpretations of those two tests?

How does sampling impact the use of our test scores?

The Test Blueprint…

- summarizes the content and format of the test.
- is typically laid out as a grid:
  - The rows are the learning objectives
  - The columns are the level of cognitive complexity
  - The cells list the types and numbers of items.
- has margins that can be used to total points.

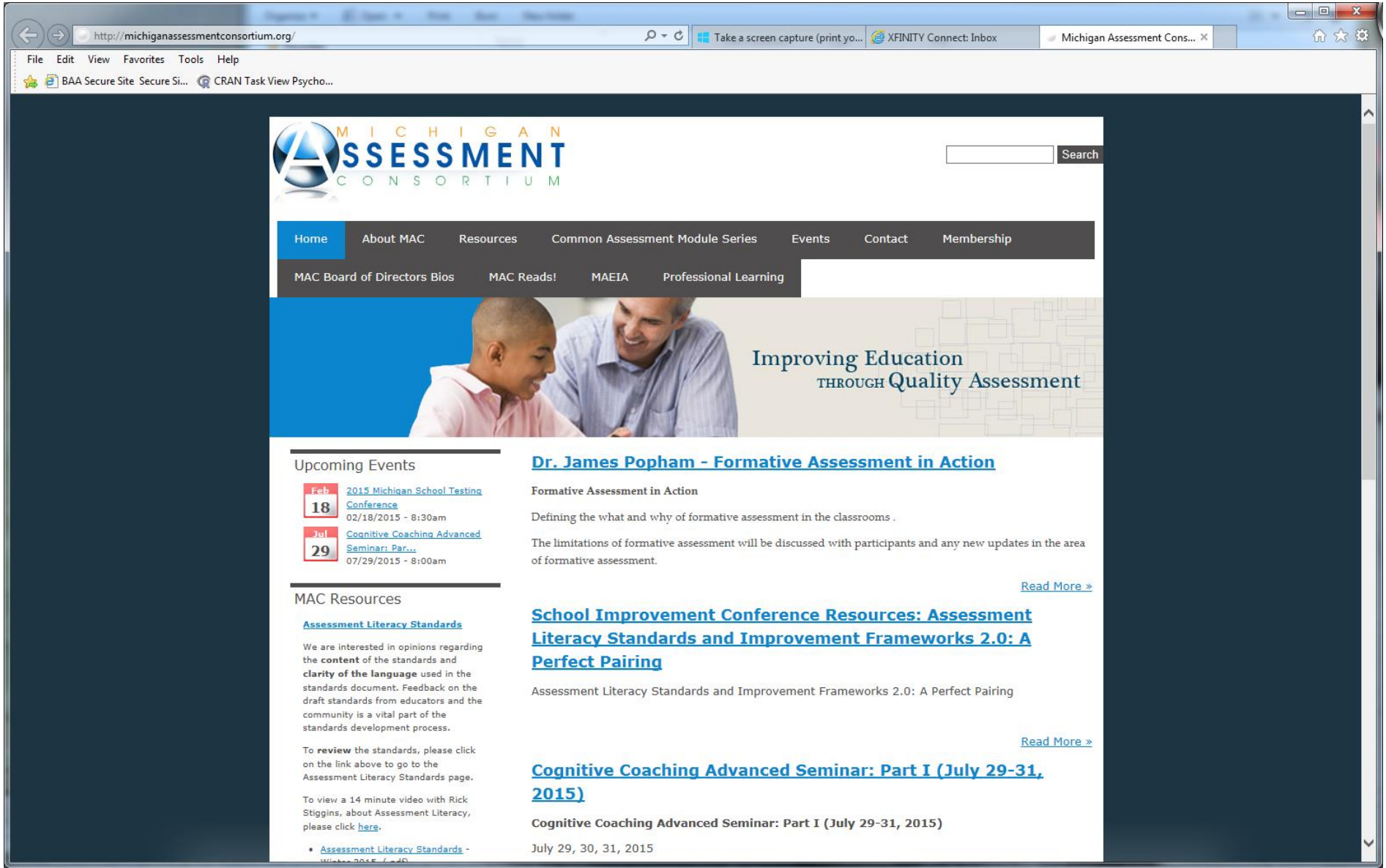
	Recall	Crit. Thinking	Products
Assessment Types	10 m/c 1pt each	1 essay 5 points	
Question Types	10 m/c 1 pt each		5 short answer (1pt)
Test Blueprints		1 essay 5 points	1 blueprint 20 points

AND FINALLY…

The Michigan Assessment Consortium

- A statewide association that is dedicated to improving assessment practice.
- Makes resources available to educators in Michigan
- Has developed a set of assessment literacy standards for educators, students, and policy makers
- Has a snazzy web page:
- www.michiganassessmentconsortium.org

Home of the MAC



That wasn't too painful…right? ☺

- Many thanks!
- Jim Gullen: Testing and Assessment Consultant, Macomb ISD
- jgullen@misd.net
- 586.228.3459