Tolerating Approximate Answers about Student Learning

Derek Briggs

University of Colorado Boulder

May 24, 2018

Invited Lecture

Oxford University

Overview

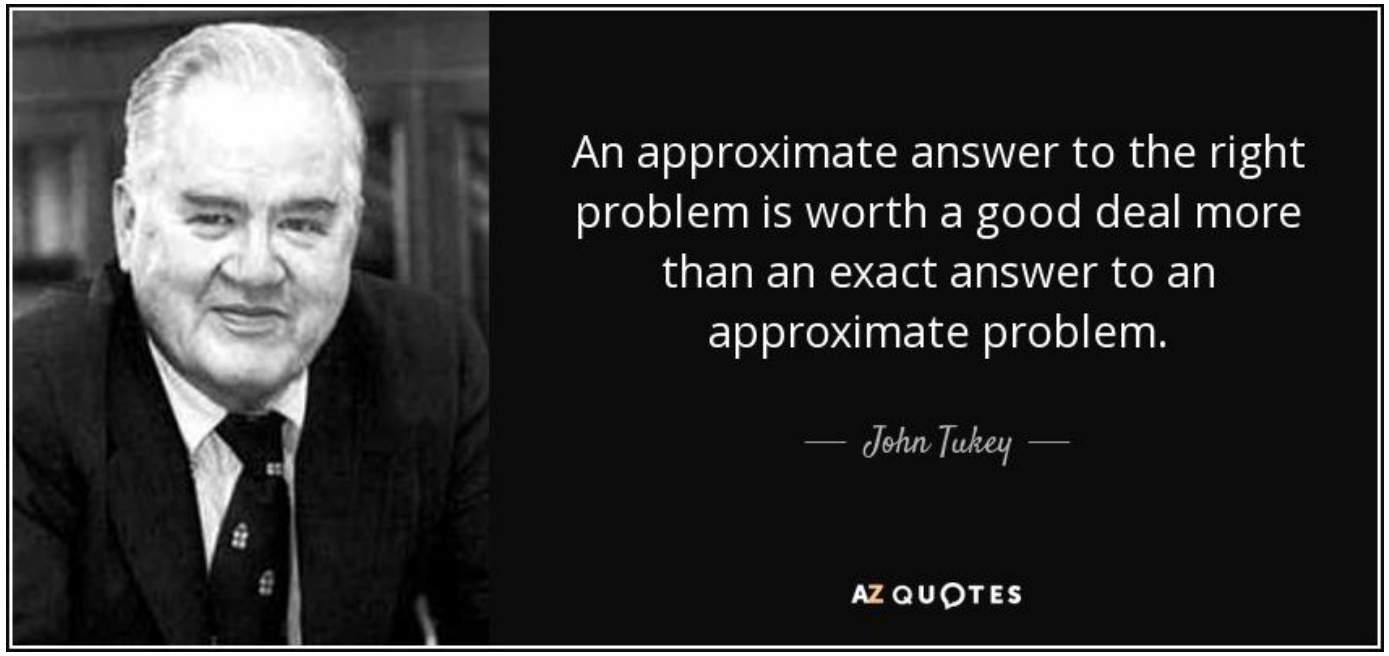
- Two Aphorisms and an Identity Crisis
- Defining the Right Problem, Asking More Relevant Questions
- Insights about Student Learning
  - Requires a Theory of How Students Learn (Learning Progressions)
  - Tests as Experiments (Waiting to Happen)
  - The Rasch Model and the Person-Item Tradeoff
  - Making the Scale Meaningful through Reference Units
- Tolerance for Approximate Answers

Two Aphorisms and an Identity Crisis

John Tukey (1915-2000)

- Home-schooled, masters degree in chemistry from Brown University in 1937
- Completed PhD in mathematics at Princeton 2 years later.
- World War II interrupted academic career, went to work in Fire Control Research Office
- Full professor at age 35, founding chairman of Princeton statistics department in 1965.
- Published the monograph “The Future of Data Analysis” in 1962.

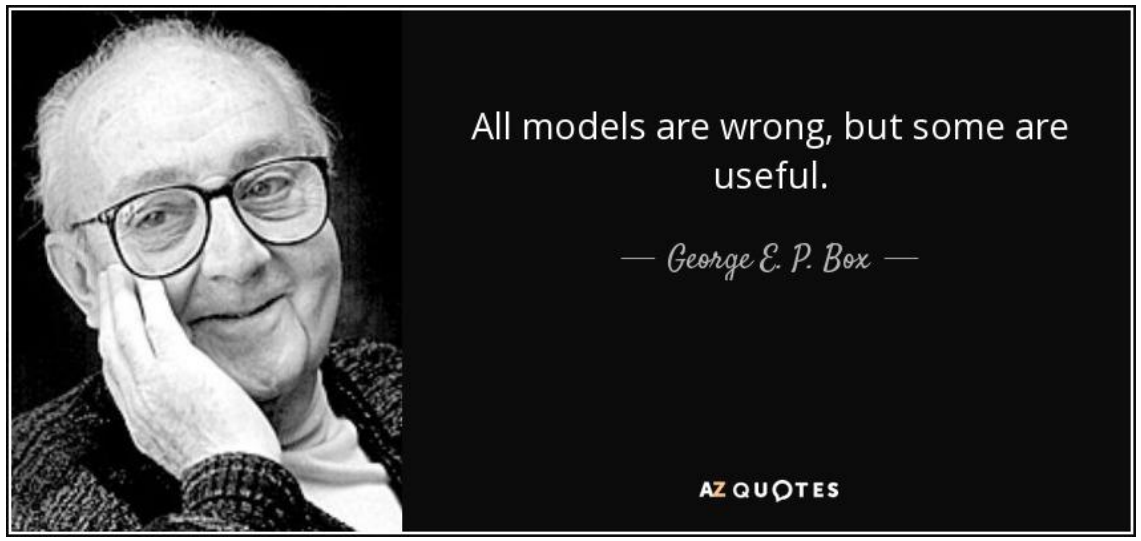
Tukey's Aphorism



A deconstruction. When research questions are posed

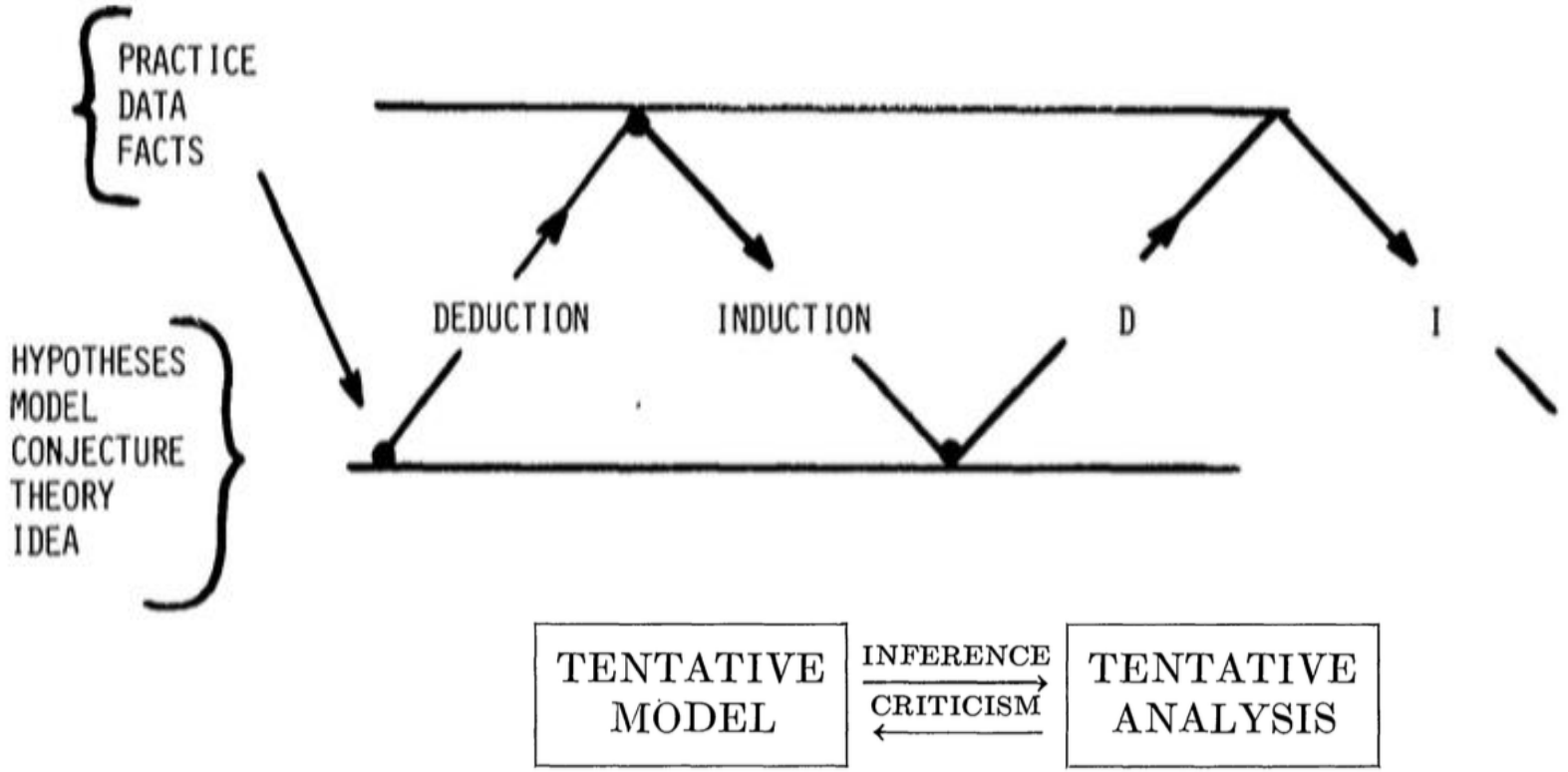
1. Some questions are better (more practically relevant) than others.
2. Some answers are more precise than others.
3. We can get very precise answers at the cost of asking less practically relevant questions.
4. When this happens, Cost > Benefit
5. **Therefore, It is better to live with approximate answers to practically relevant questions.**

George Box



- Idea originated as part of 1974 address to ASA in honor of R.A. Fisher
- Over 1100 citations in Google Scholar (as of 5/20/18)
- A love-hate relationship.

Utility of “Wrong” Models in Science



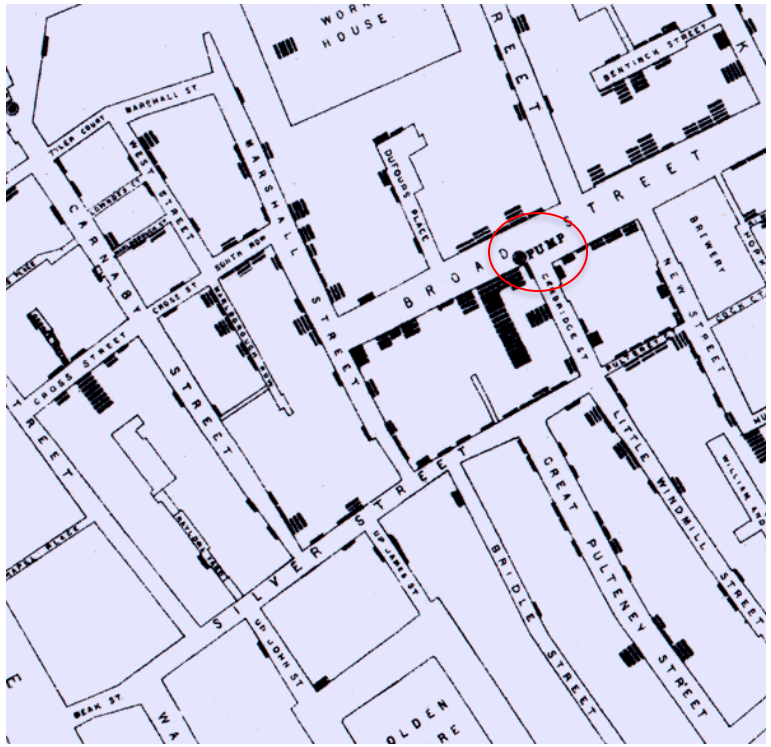
Iterations between model and practice are critical

Some Models are Wrong and Harmful

Cholera Outbreaks in London during early to mid 19th century

Miasma Theory

Transmission via Drinking Water



John Snow's Famous Map of Golden Square in St. James District, 1854

Some Possible Parallels with Psychometrics

1. Development of theory for theory's sake ("mathematistry")
2. Overly reliant on confirmatory analyses (model is King)
3. Theories are not typically validated by practice (no iterations)

Just as Tukey worried about the lack of separation between mathematics and statistics, I worry about the same lack of separation between statistics and psychometrics.

Psychometrics as a "Tool-Making Enterprise"?

Psychometric activity involves a sequence of three conceptual stages:

(i) engineering (defining a problem to be solved),

(ii) art (that involves lies and stealing), and

(iii) design of a solution

Thissen, 2001, p. 476

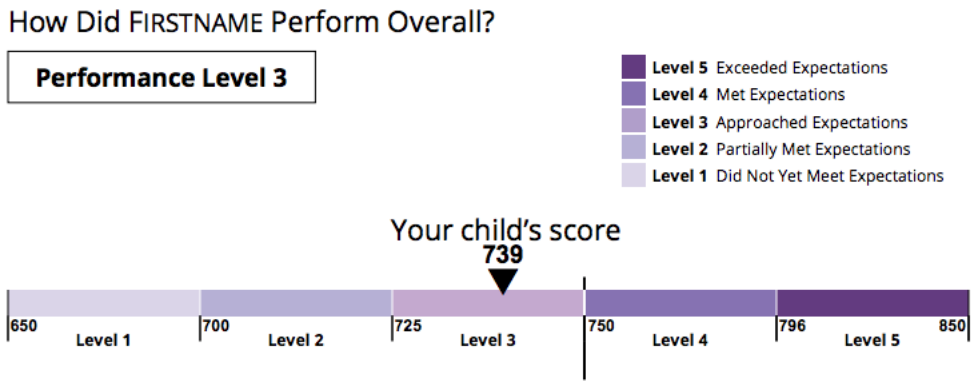
Questions about Test Scores that Psychometrics Answers with "Precision"

- How reliable is this test score?
- What is the standard error of measurement?
- At what location on the score scale does the test provide the most information about student ability?
- What is the equating function that adjusts unique test forms for differences in difficulty?
- Is there any evidence that test items are biased?

A Question for which Psychometrics Provides Surprisingly Few Insights

How <u>much</u> are students learning?

An Illustration: PARCC Assessment in the US



What scale should we interpret? The one from 650 to 850 or the one from 1 to 5? How do the two relate? The distance from 739 to 750 is 11 points—what does this mean?

School A: Mean Score of 755
School B: Mean Score of 740

Is this difference significant?

Conventional Large-Scale Tests Convey Information about Attainment Status, not Growth

THESIS: To answer questions about student learning…

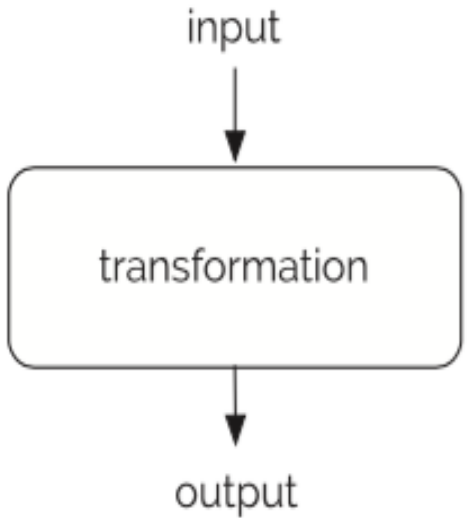
We will need

1. test scores on multiple occasions (longitudinal design)
2. to link these scores onto a common scale (vertical scale)
3. to define a meaningful unit for the scale (criterion-referencing), and
4. to interpret score changes over time (growth model).

But most importantly, to evaluate the validity of 1-4 we will need to iterate between a theory of student learning (learning progression) and a process of test item design (engineering).

Insights about Student Learning by Making Magnitudes Interpretable

Psychometrics and the Structure of Measurement



Source: Maul, Mari, Torres-Irribara & Wilson (2018)

- INPUT: What are we prepared to assume about the attribute to be measured? Is the attribute quantitative or qualitative?
- TRANSFORMATION: What evidence do we have that the instrument we have developed is sensitive to variability in the attribute?
- OUTPUT: What are the scale properties of the numeric scores that result?

Learning Progressions (LPs)

- Empirically grounded and testable hypotheses about how students' understanding of core concepts within a subject domain grows and become more sophisticated over time with appropriate instruction (Corcoran, Mosher, & Rogat, 2009)
- As theories of learning, can range from very complex sociocultural and sociocognitive theories to relatively crude speculations about sequences of instruction.

Maths in the US
"Common Core of State Standards"

Math Content Domains Associated With Grades 3 to 8 in the CCSS

	Grade in Which CCSS Includes Domain					
Content Standards by Domain	*3*	*4*	*5*	*6*	*7*	*8*
Operations & Algebraic Thinking	X	X	X			
Number & Operations in Base 10	X	X	X			
Number & Operations—Fractions	X	X	X			
Measurement & Data	X	X	X			
Geometry	X	X	X	X	X	X
Ratios & Proportional Relationships			X	X		
The Number System				X	X	X
Expressions & Equations				X	X	X
Functions						X
Statistics & Probability				X	X	X

Progression for FRACTIONS (Grades 3-5)

- develop understanding of fractions as numbers (grade 3),
- extend understanding of fraction equivalence and ordering (grade 4),
- build fractions from unit fractions (grade 4),
- understand decimal notation for fractions, and compare decimal fractions (grade 4),
- use equivalent fractions as a strategy to add and subtract fractions (grade 5),
- apply and extend previous understandings of multiplication and division (grade 5).

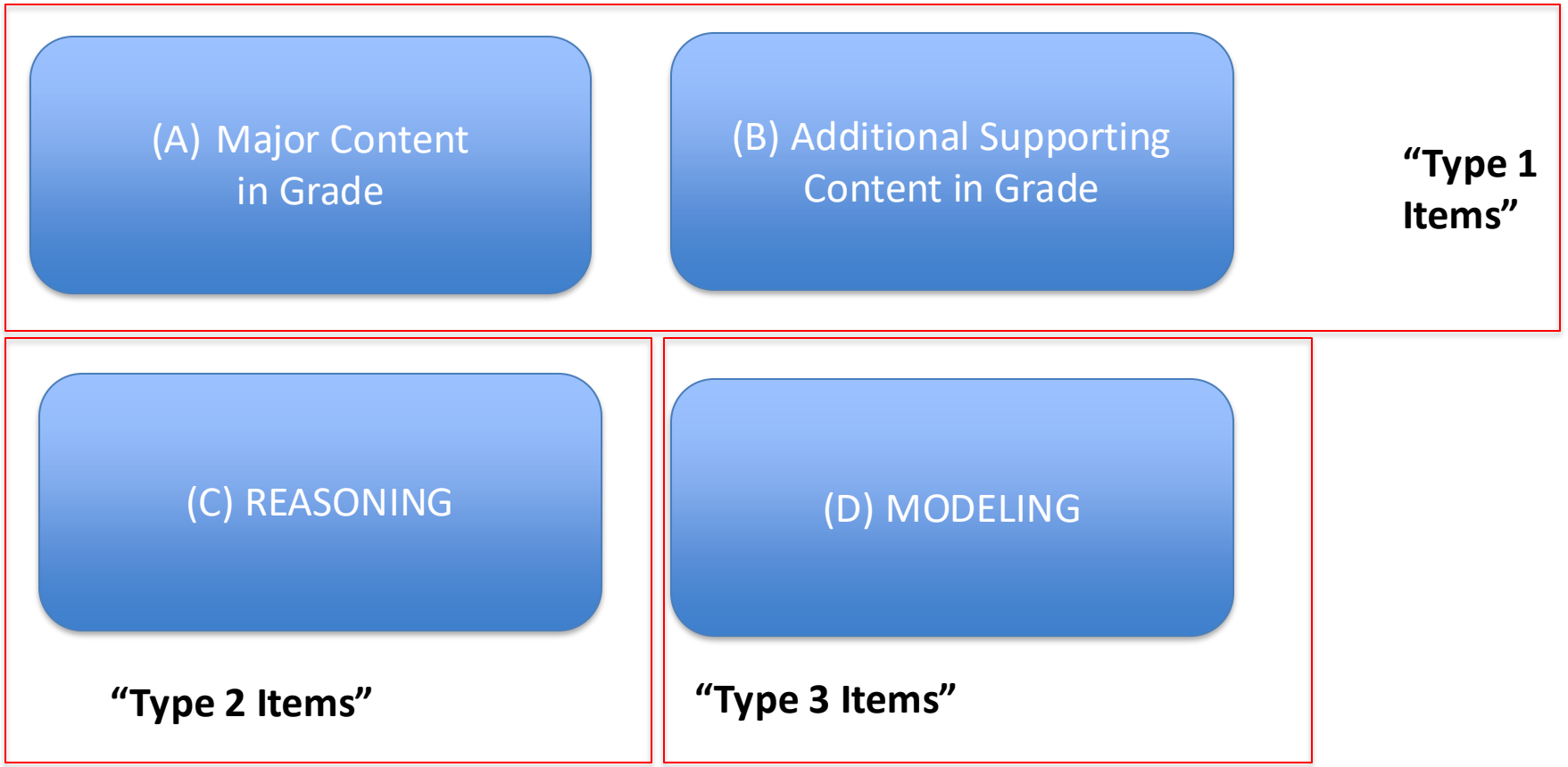
http://www.corestandards.org/Math/Content/NF/

A PARCC Within Grade LP for Fractions

	Grade 3 Math : Sub-Claim A The student solves problems involving the Major Content for the grade/course with connections to the Standards for Mathematical Practice.			
	Level 5: Exceeds Expectations	**Level 4: Meets Expectations**	**Level 3: Approaches Expectations**	**Level 2: Partially Meets Expectations**
Fractions as Numbers 3.NF.1 3.NF.2 3.NF.A.Int.1	Understands 1/*b* is equal to one whole that is partitioned into *b* equal parts – limiting the denominators to 2, **3**, 4, **6** and 8. Represents 1/*b* on a number line diagram by partitioning the number line between 0-1 into *b* equal parts recognizing that *b* is the total number of parts. Demonstrates the understanding of the quantity *a*/*b* by marking off *a* parts of 1/*b* from 0 on the number line and **states that the endpoint locates the number *a*/*b*.** **Applies the concepts of 1/*b* and *a*/*b* in real-world situations.** **Describes the number line that best fits the context.**	Understands 1/*b* is equal to one whole that is partitioned into *b* equal parts – limiting the denominators to 2, 4 and **8**. Represents 1/*b* on a number line diagram by partitioning the number line between 0-1 into *b* equal parts recognizing that *b* is the total number of parts. **Demonstrates the understanding of the quantity *a*/*b* by marking off *a* parts of 1/*b* from 0 on the number line.**	Understands 1/*b* is equal to one whole that is partitioned into *b* equal parts – limiting the denominators to 2 and 4. **Represents 1/*b* on a number line diagram by partitioning the number line between 0-1 into *b* equal parts recognizing that *b* is the total number of parts.** **Represents fractions in the form *a*/*b* using a visual model.**	Understands 1/*b* is equal to one whole that is partitioned into *b* equal parts – limiting the denominators to 2 and 4. Identifies 1/*b* on a number line diagram when partitioned between 0 and 1 into *b* equal parts.

Takes a grade-specific domain cluster (develop understanding of fractions as numbers), breaks it into two different components (one of these, fractions as numbers shown here) and integrates them with one or more of the standards for mathematical practices (problem solving).

PARCC Item Design for Maths Tests



A Test as an Experiment Waiting to Happen

- The learning progressions that influenced the design of the PARCC test items are nascent theories.
- But they come with large amounts of data, are great candidates for exploratory data analyses, and could also be a basis for experimentation.
- For example, say a distinction between levels hinges upon the presence of a visual model as scaffold
  - We can ask, what is the effect of the presence/absence of this scaffold on item performance?

Creating a Vertical Scale

Test	Grade					
	3	4	5	6	7	
1	X					
2	X		X			
3		X	X	X		
4		X	X	X	X	
5				X	X	

"Now, of course, the tests used changed from one occasion to another, but none the less our aim was to evaluate the progress of each pupil. Thus it became an urgent problem whether it would be possible ascertain the levels of attainment of a pupil independently of which tests were used and also independently of age, school group and time of school year."

Rasch, 1960

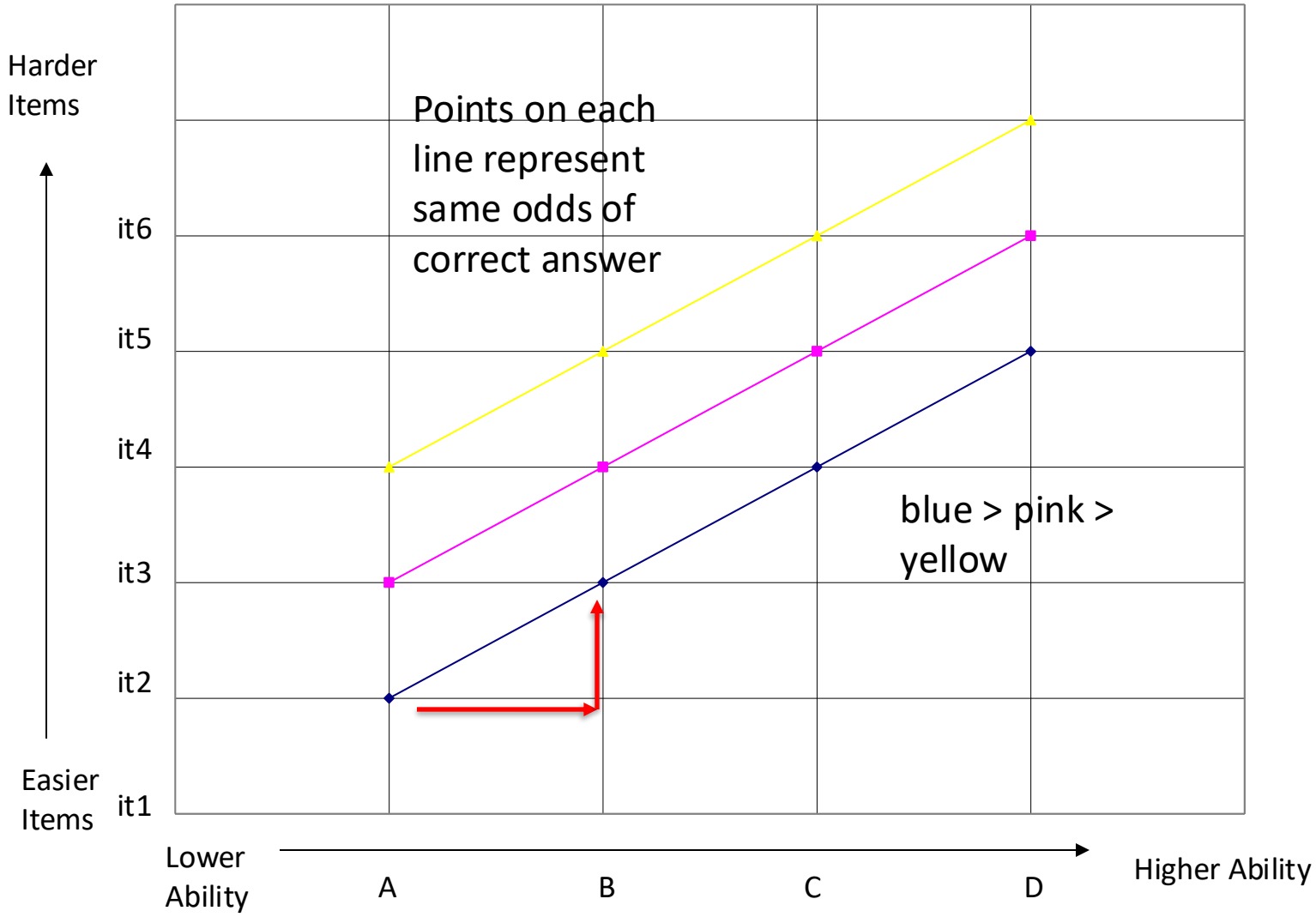
Rasch's Model for Dichotomously Scored Items

$$P(X_{pi}=1|\theta_p,b_i)=\frac{\exp(\theta_p-b_i)}{1+\exp(\theta_p-b_i)}$$

$$P=f(\theta,\delta)$$

$$\log\left[\frac{P(X_{pi}=1)}{P(X_{pi}=0)}\right]=\theta_p-b_i$$

Interval of Vertical Scale Defined by Person-Item Tradeoffs



If student B > A, it must be the case that I could give her a harder item (i.e., it3) that would "explain" the advantage.

Conjoint Additivity: Luce & Tukey (1964)

- Luce & Tukey establish a mathematical basis through which it would be possible to establish whether psychological attributes can be measured on an interval scale.
- Often seized upon by proponents of the Rasch Model to justify its use to establish an interval scale
- Interestingly, this was not something Tukey followed up on in his later work.

Conjoint Additivity (in a nutshell)

	Test (A)
Instruction (X)	(X, A)

Conjoint Additivity (in a nutshell)

	Test (A)
Instruction (X)	(X, A)
More Instruction (Y)	(Y, A)

Is (Y, A) > (X, A)?

Conjoint Additivity (in a nutshell)

	Test (A)	Test with Harder Items (B)
Instruction (X)	(X, A)	
More Instruction (Y)	(Y, A)	(Y, B)

Is (Y, B) = (X, A)?

The key idea: if we really understand the construct of measurement, we should understand how to manipulate the test to be harder or easier as a tradeoff against an additional instruction.

Learning Progression Basis for a Reference Unit

Build fractions from unit fractions

[GRADE 4]

- CCSS.MATH.CONTENT.4.NF.B.3.C
  Add and subtract mixed numbers with like denominators

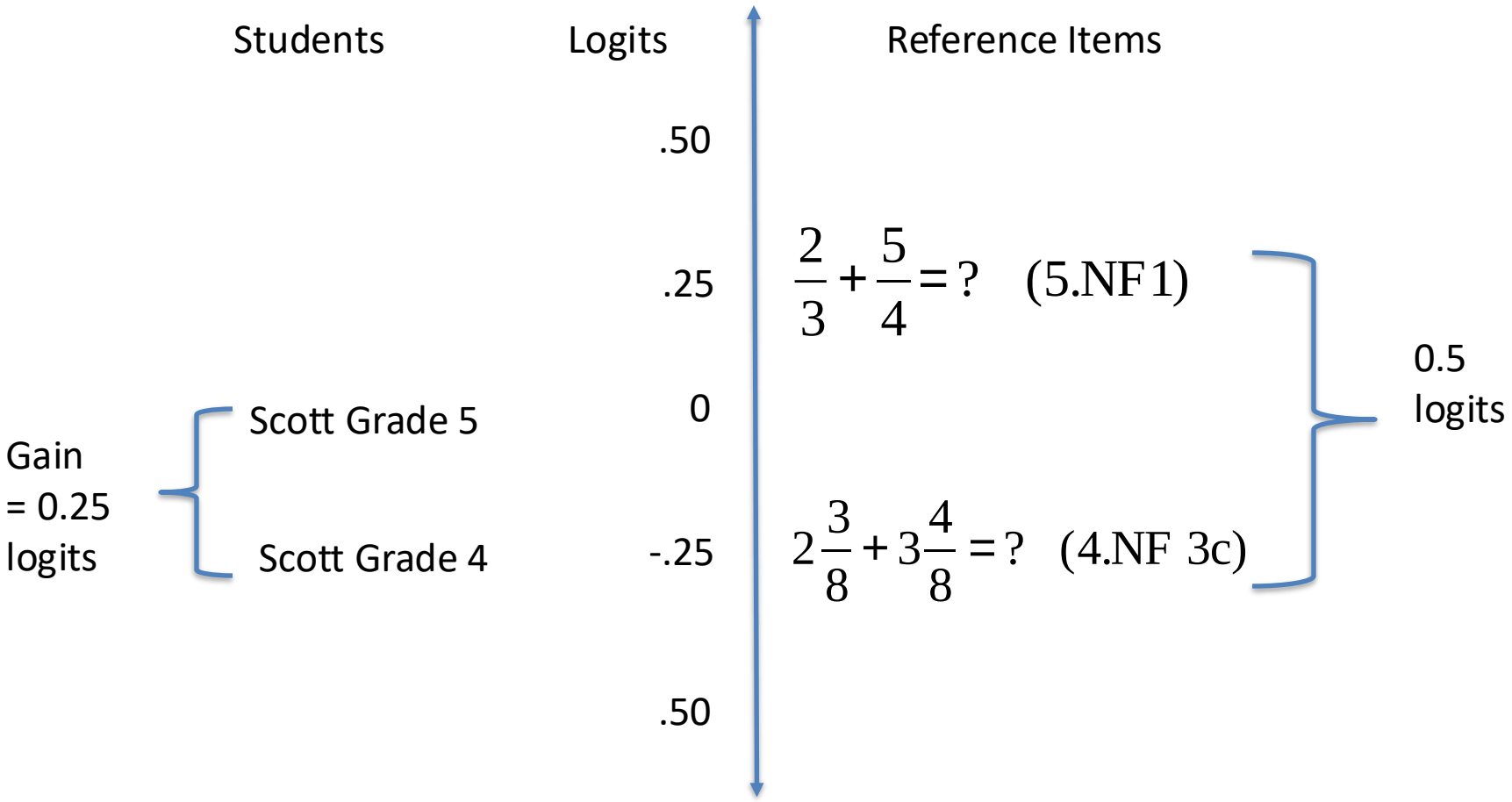
$$2\frac{3}{8}+3\frac{4}{8}=?\quad(4.NF\ 3c)$$

Use equivalent fractions as a strategy to add and subtract fractions. [GRADE 5]

- CCSS.MATH.CONTENT.5.NF.A.1
  Add and subtract fractions with unlike denominators (including mixed numbers)

$$\frac{2}{3}+\frac{5}{4}=?\quad(5.NF1)$$

Interpreting Magnitude Relative to Reference Unit



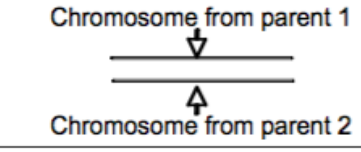
Another Example with Real Data

Learning Goal 4: "Describe the phenomenon of linkage and how it affects assortment of alleles during meiosis"

Bloom Level 2: Understand

13. A man is a carrier for Wilson's disease (Aa) and Rotor syndrome (Rr). Assume the genes involved in these two disorders are both on chromosome 13 (a non-sex chromosome). Below are possible representations of his genotype (labeled #1, #2, and #3). Which of them could be correct?

The horizontal parallel lines represent a pair of chromosome 13s.

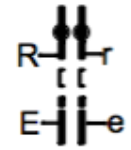


a) #1 only
b) #2 only
c) #3 only
d) #2 and #3 only
e) #1, #2 and #3

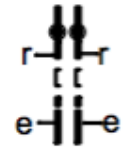
Difficulty = -1.074

Bloom Level 3: Apply

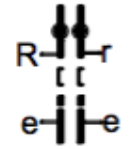
24. Two different genes are located on the same chromosomal pair in rabbits. A particular female rabbit is heterozygous for alleles of both these genes, with the alleles arranged as shown in the diagram to the right. Scientists know that the two genes are on the same chromosome, but do not know their exact position, as indicated by the dashed line.



Suppose this female mates with a male rabbit in which the same chromosome pair looks like this:



How likely is it that this pair of rabbits would have offspring with a chromosome pair that looks like this:

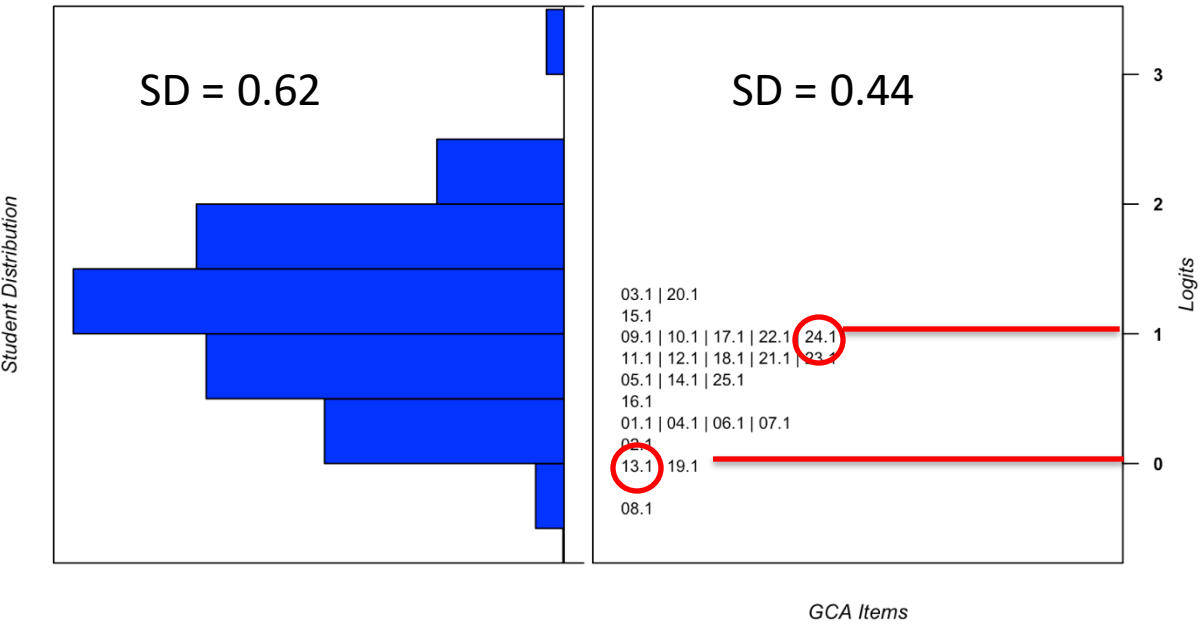


Difficulty = 0.608

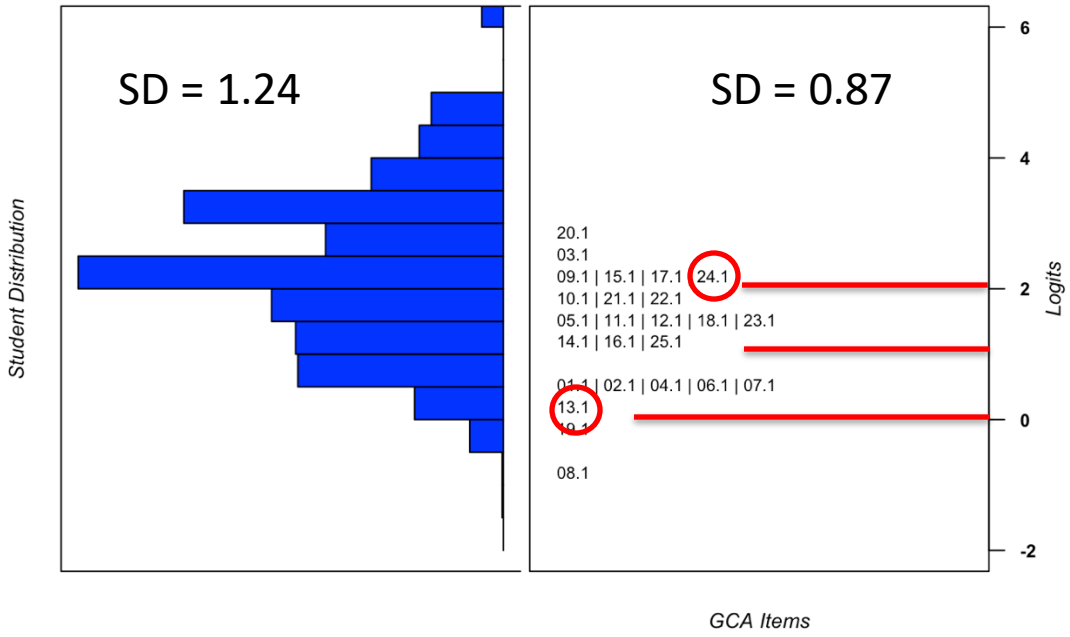
a) Not likely, because the R and e alleles are not on the same chromosome in either parent.
b) Very likely, because the random assortment of chromosomes during cell division to make sperm or eggs allows for the mixing of all alleles.
c) More likely if the two genes are very close together on the chromosome.
d) More likely if the two genes are not very close together on the chromosome.

Changing Reference Unit of Scale

1 unit = difference in location between items 13 and 24.



1 unit = ½ difference in location between items 13 and 24.



University of Colorado Boulder

Item Maps in NAEP

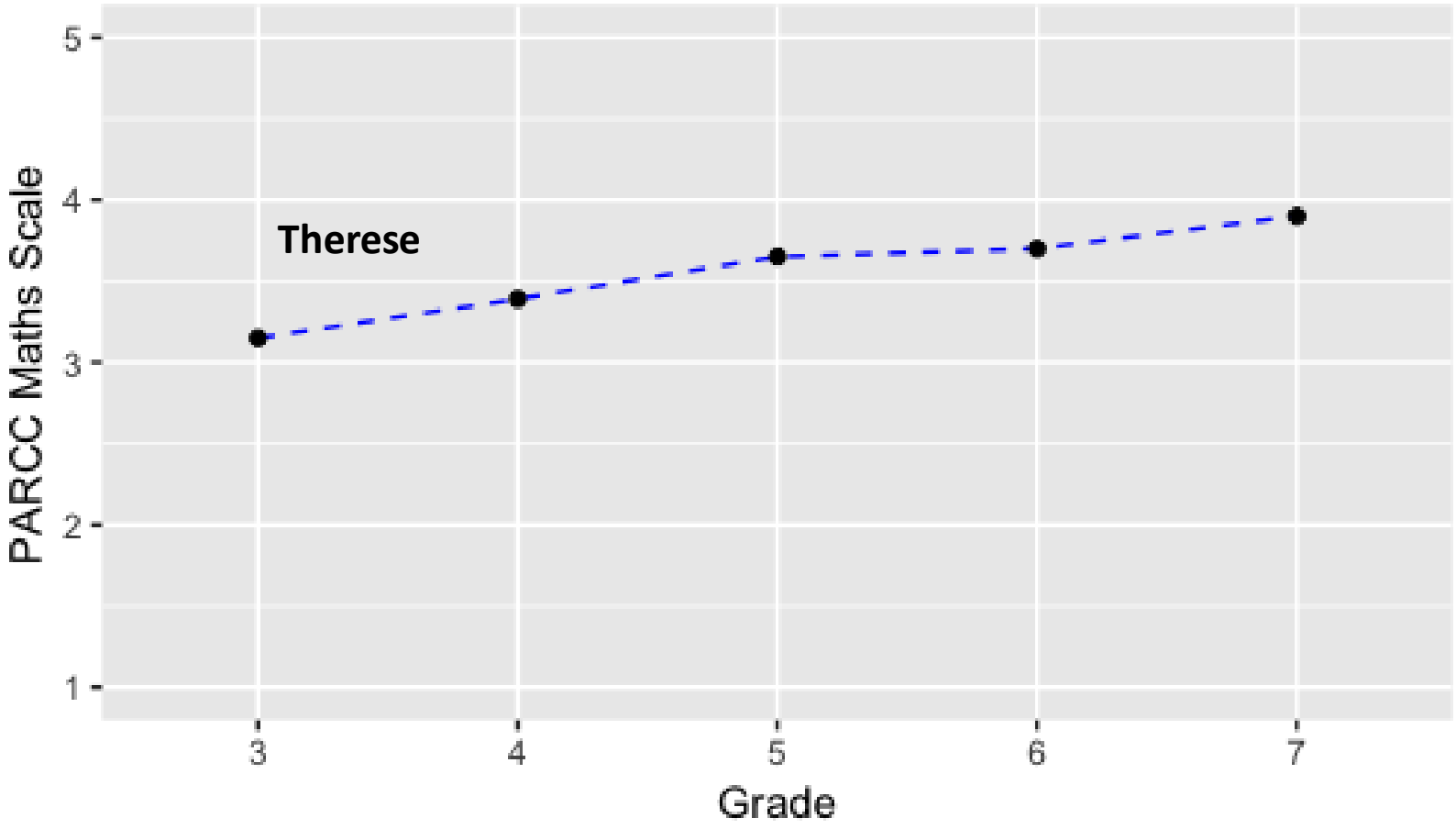
https://www.nationsreportcard.gov/itemmaps/?subj=MAT&grade=4&year=2017&jurisdiction=NT&variable=TOTAL

Tolerance for Approximate Answers

Why Are Answers about Student Learning only "Approximate?"

- It is certainly true that estimates of student growth based on differences between measures over time will have considerable uncertainty.
- This is definitely a concern, but that isn't what Tukey meant by an "approximate" answer in this context
- Learning is complicated!
- The approximation in this context comes from assuming that a construct relevant to inferences about learning is measurable
- Magnitudes can be defined, but are they meaningful?

Depicting Growth on a Vertical Scale: An Ideal Case



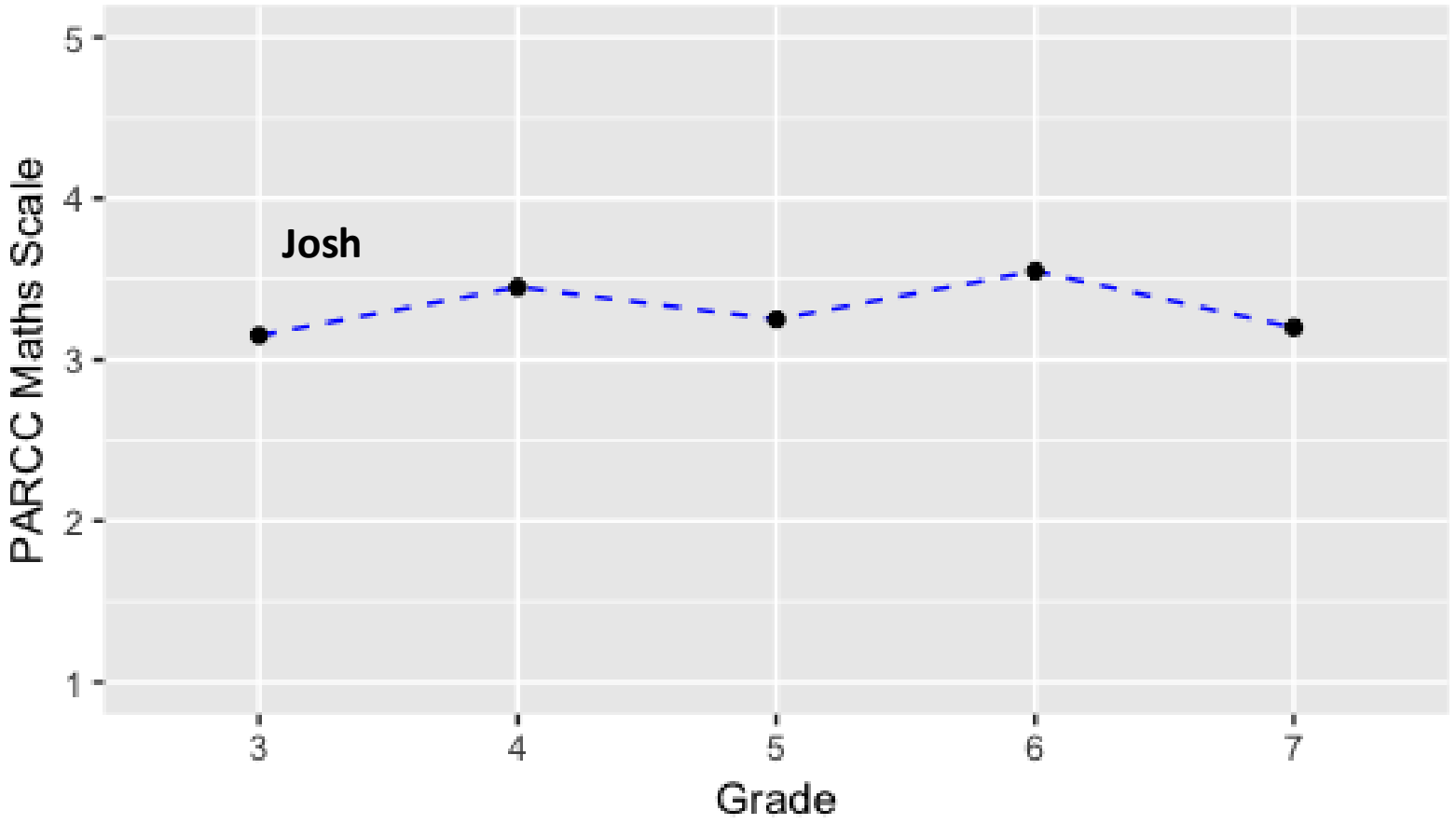
A reference unit on this scale defined to be the average distance between items written to differentiate levels 3 and 4 of PARCC LPs.

Grade 3 = 3.15
Grade 4 = 3.39
Grade 5 = 3.65
Grade 6 = 3.70
Grade 7 = 3.90

Change from Grade 3 to 7 = .75

➔ 75% of the distance between level 3 and 4 understanding

Growth on a Vertical Scale: Less Ideal



Grade 3 = 3.15
Grade 4 = 3.45
Grade 5 = 3.25
Grade 6 = 3.55
Grade 7 = 3.20

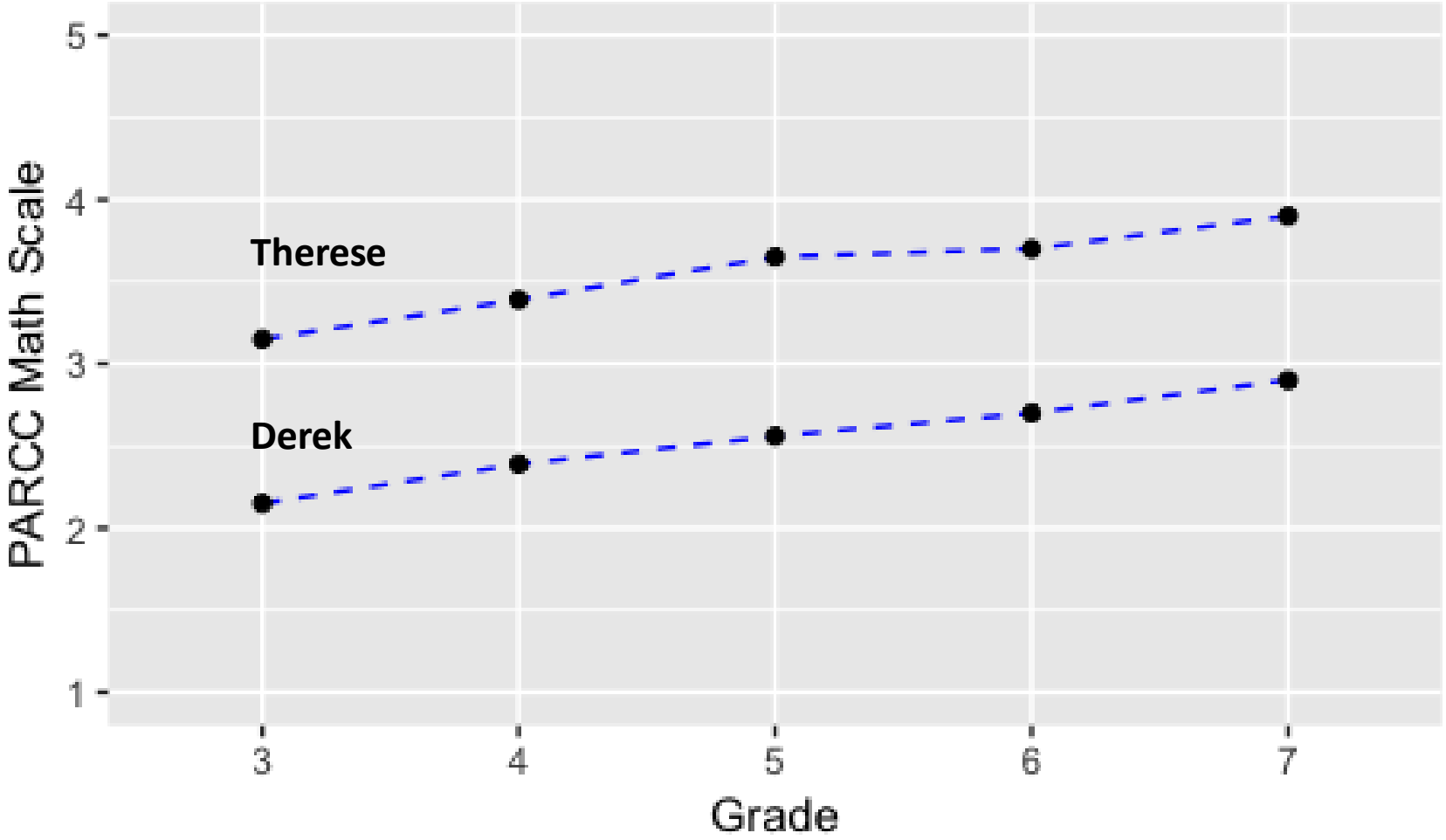
Negative Growth??

Change from Grade 3 to 7 = .05

➔ 5% of the distance between level 3 and 4 understanding

Equivocal interpretations…

Two Hypothetical Student Growth Trajectories



The changes from grade to grade are identical except for the starting point.

Can we conclude both students are learning equally?

An Ordinal Alternative

1 = Meeting Within Grade Performance Expectation
0 = NOT Meeting the Expectation

Students	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Therese	1	1	1	1	1
Josh	1	1	0	0	0
Derek	0	0	0	0	0

Back to Tukey

Most key questions in our world sooner or later demand answers to “by how much?” rather than merely to “in which direction”…in doing so we are asserting a belief in quantitative knowledge.

--preface to Tukey’s 1977 textbook *Exploratory Data Analysis*

But we need to be careful about the dangers…

The developmental [vertical] scale score is like a ruler that measures growth in reading and math from year to year. Just like height in inches, the student's scores in reading and math are expected to increase each year.

--Newsletter sent to the public from a state board of education

Briggs (2010; 2013)

Recall THESIS: To answer questions about student learning…

We will need

1. test scores on multiple occasions (longitudinal design)
2. to link these scores onto a common scale (vertical scale)
3. to define a meaningful unit for the scale (criterion-referencing), and
4. to interpret score changes over time (growth model)

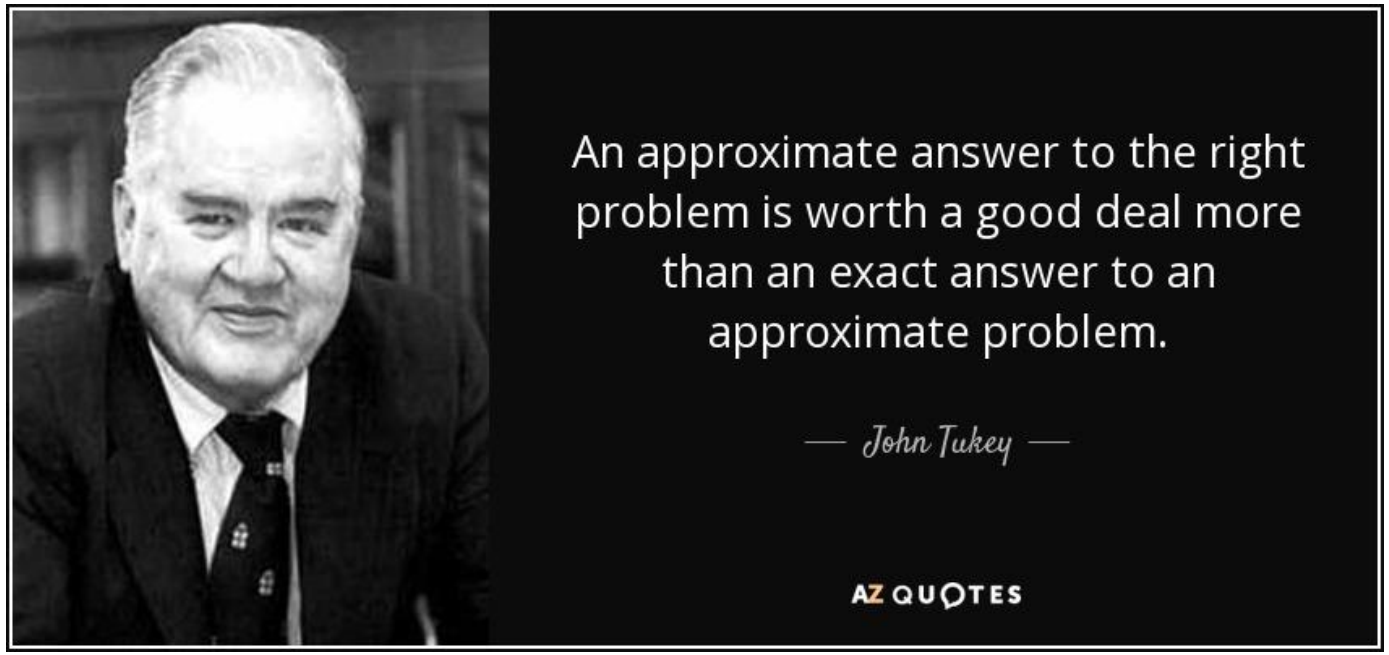
To evaluate the validity of 1-4 we will need to iterate between a theory of student learning (learning progression) and a process of test item design (engineering).

To TOLERATE approximate answers

We need to have an empirical basis for evaluating the validity of inferences along a putatively quantitative score score.

1. Person and item fit 
2. Item difficulty modeling 
3. Invariance of Item Distances 
4. Experimental Designs 

Conclusion


An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.
— John Tukey —
AZ QUOTES