The Use of Content-Referencing to Evaluate the Magnitude of Student Growth

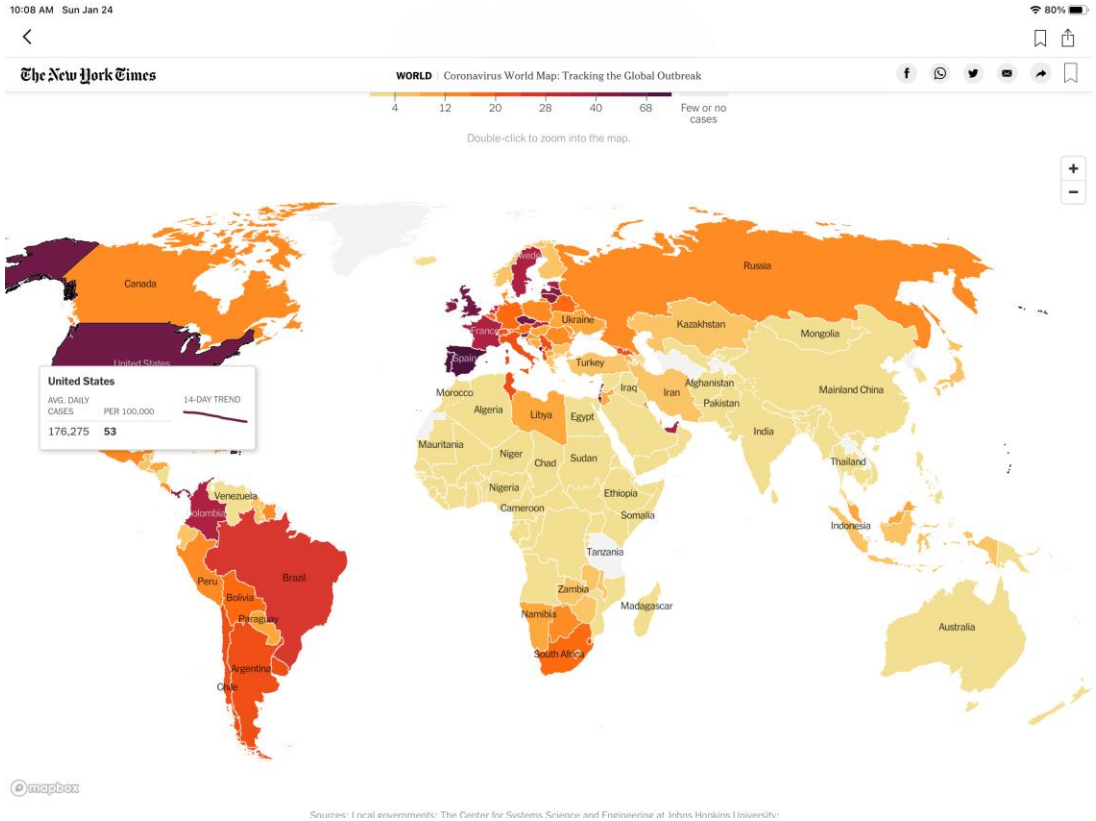
Derek Briggs

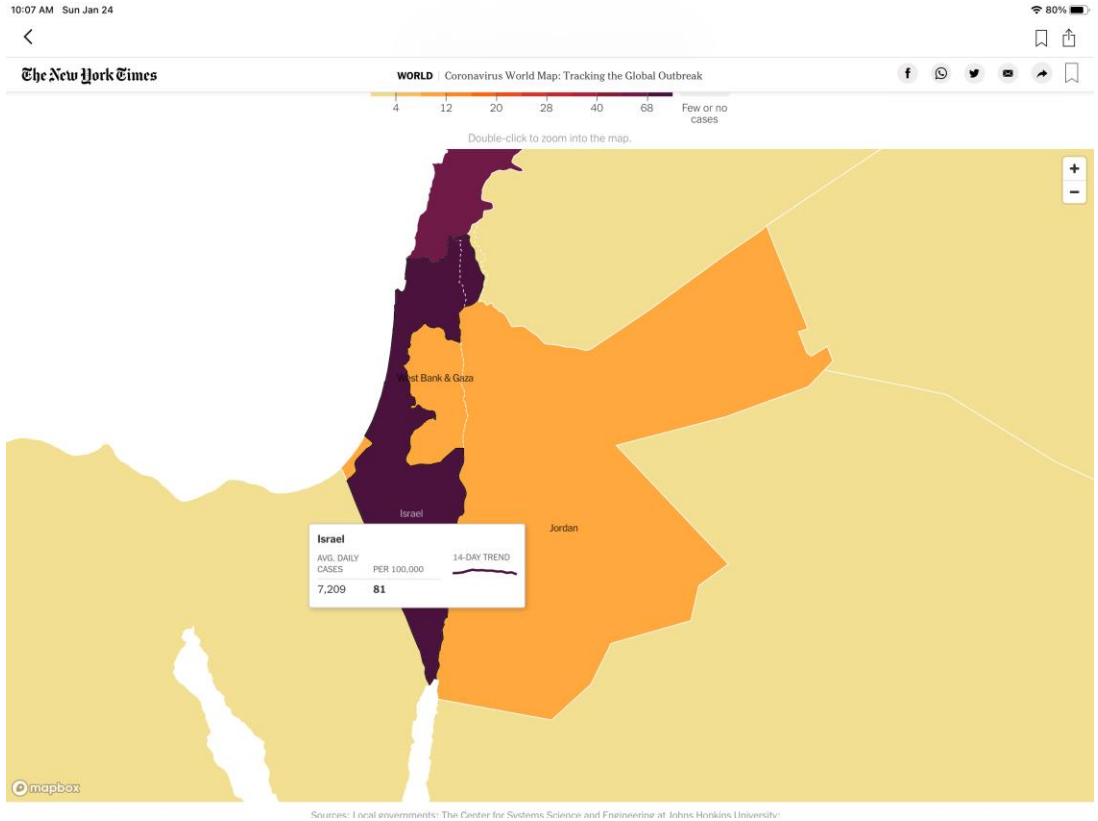
University of Colorado Boulder

January 27, 2021

Keynote Address at the 17th Conference of the Israeli Psychometric Association

The Elephant in the Room: COVID-19



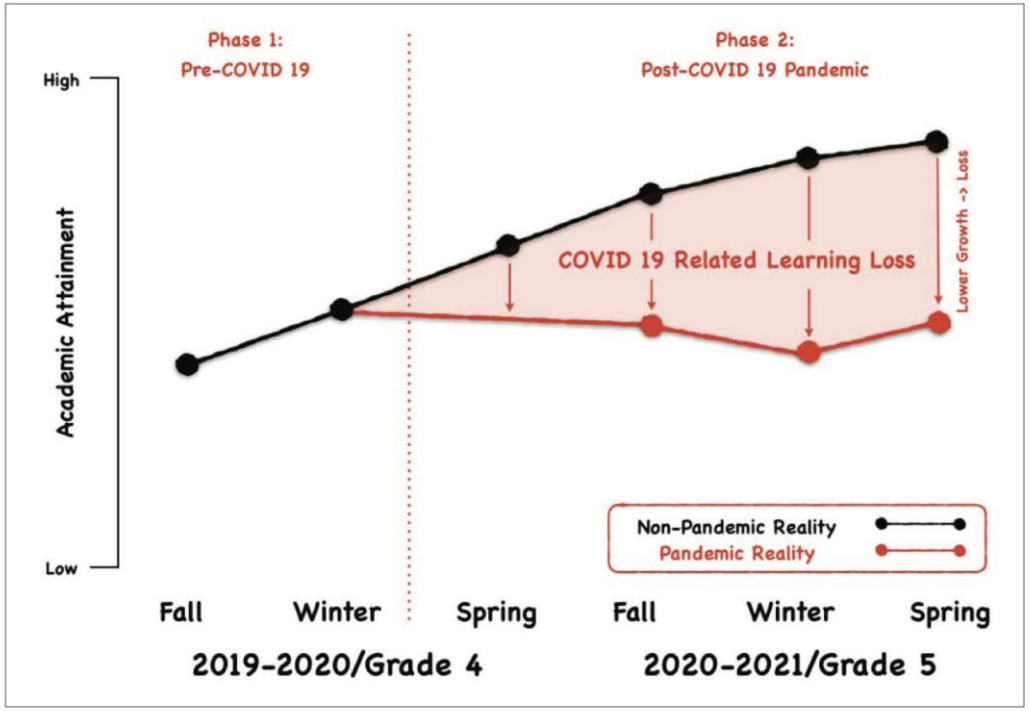


The Impact on Education

- It is hard to imagine a theoretical model of student learning in which one would predict that COVID has had a positive effect.
- A simple way to frame this is that most children have had less or lower quality contact time with teachers in school settings
- But even more disastrous is the reduced contact with peers and local community—opportunities for learning outside the school have likely been reduced.
- Do we need to quantify the magnitude of learning loss?
  - From an immediate policy perspective, no.
  - From a long-term research perspective, yes.

A Recent Visualization of Learning Loss

Figure 1: Visual depiction of learning loss based upon academic attainment measures (e.g., interim assessments) given in the fall, winter, and spring of each year.



https://www.nciea.org/sites/default/files/inline-files/CFA_LearningLossRecoveryGrowthStatewideTesting.pdf

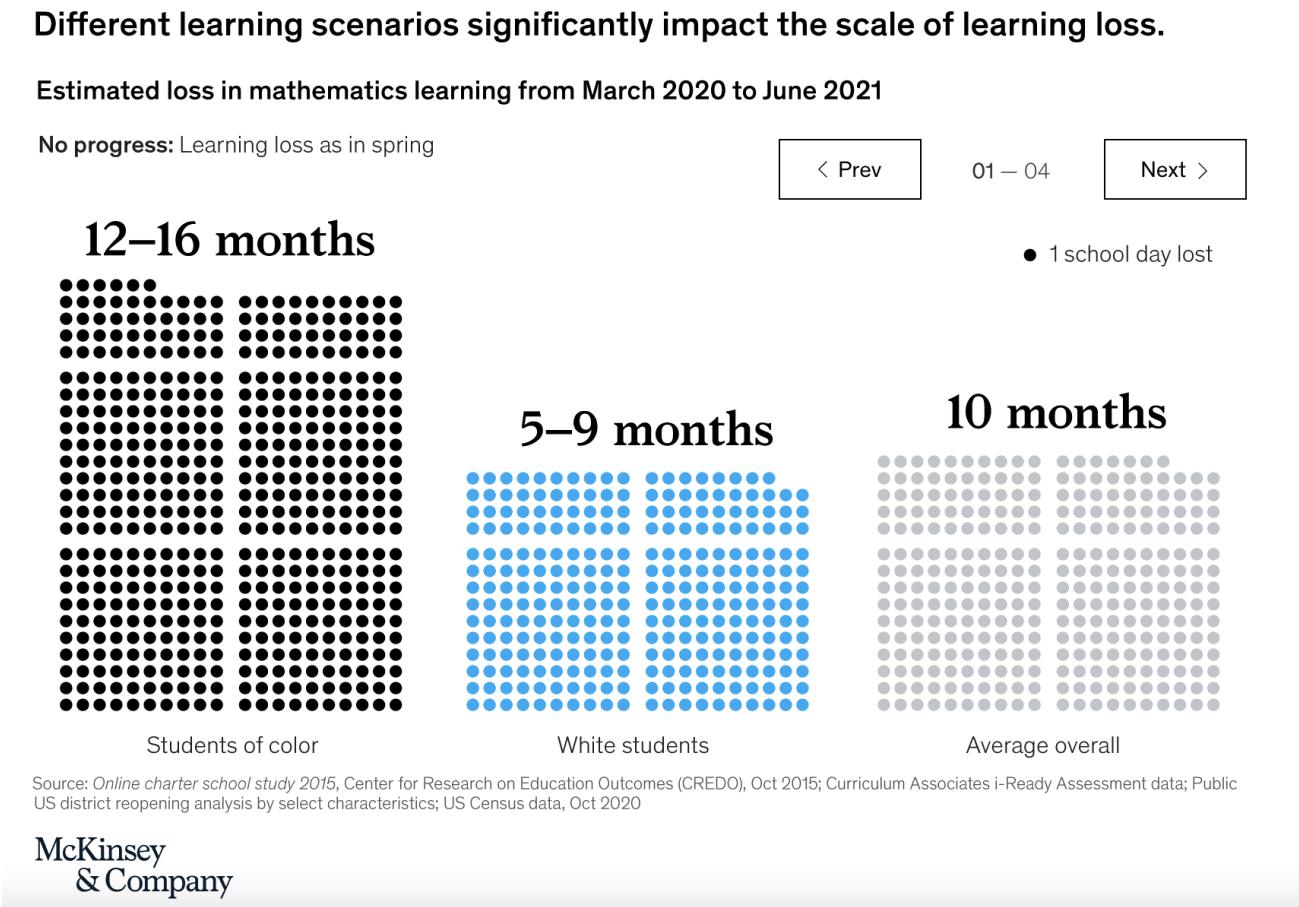
The Key Psychometric Issues

- What attribute (i.e., construct) should we attempt to measure, when and how often?
- Can the attribute being measured on a common (vertical) scale?

Even if these issues have been addressed…

- A decision must be made in regard to the measurement unit in which learning loss is conveyed.
- The two most common candidates: SD units (“effect size”), temporal units (“months of learning”).

A Recent Report on Learning Loss in the US (written by McKinsey & Company)



Source: https://www.mckinsey.com/industries/public-and-social-sector/our-insights/covid-19-and-learning-loss-disparities-grow-and-students-need-help

Thesis of this Talk

- Neither SD units nor transformations into temporal units provide a good interpretation of magnitude.
- SD units are too opaque, temporal units suggest misleading precision.
- I suggest an approach that can be used to establish “content-referenced” units of measurement.
- This unit can also be given a spatiotemporal representation that is a prerequisite for a cognitively meaningful interpretation of magnitude.

Some History

Forthcoming book (by this summer)

Historical and Conceptual Foundations of Measurement in the Human Sciences

Alfred Binet (1857-1911)

- Universal public education had only been established in France as of 1881.
- Prior to this, children with (real or perceived) cognitive disabilities were sorted out of schooling.
- In 1904, government appoints a commission to "investigate the state of the mentally subnormal in France."
- Binet is a member of commission, comes to appreciate need for a standard approach to decide which children require special services.
- Conventional French labels for subnormal/abnormal/retarded children at the time: Idiots, Imbeciles, Debiles (Americanized to "Morons" thanks to Henry Goddard)

Binet's Methods

- The goal was to establish a series of "tests" that could be ordered in difficulty.
- Hierarchy based on age when 80-90% normal children could answer correctly.
- Time-consuming process of trial and error.
- Each series of tests involved one on one interviews with children.
- 1908: Introduces scaling approach to establish a "mental level"
  - Find age where child has passed all but one item associated with that age
  - Next give child a year of credit for every additional 5 items answered correctly (irrespective of difficulty of the item)

The Binet-Simon Age Scales

1905 Age Scale

Tests	0-2	3	5	7	9	11	>
1	B	LC	S	M	S		LA
2	B	LC	M	LC	S		MI
3	B	LC	S	M	LA		MI
4	B		LC	M		LA	LA
5	B			M		LA	
6	B			M			
7				LC			

1908 Age Scale

Tests	0-2	3	4	5	6	7	8	9	10	11	12	13
1	B	LC	A	S	M	S	A	A	A	LA	M	MI
2	B	LC	LC	S	A	N	N	A	A	LA	LA	MI
3	B	M	M	M	LC	A	A	LA	LA	LA	M	LA
4	B	M	S	N	M	S	N	A	LA	LA	LA	
5	B	A		MI	A	N	LC	N	LA	LA	LA	
6	B				A	LC	A	S				
7						N						
8						A						

1911 Age Scale

Tests	0-2	3	4	5	6	7	8	9	10	11	12	13	14	15	Adults
1	B	LC	A	S	A	S	LC	A	S		S			M	MI
2	B	LC	LC	S	LC	LC	N	LA	MI		LA			LA	MI
3	B	M	M	M	S	M	S	A	LA		LA			M	A
4	B	M	S	N	N	N	A	A	LA		LA			LA	LA
5	B	A		MI	A	A	M	LA	LA		LA			LA	LA
6	B														

Note: Each test required one or more of four abilities: comprehension, direction, invention and censure. Distinguishing requirements of tests beyond these abilities were B = basic communication, LC = understanding of concrete language, LA = understanding of abstract language and reasoning, M = memory, S = sensory discrimination, MI = mental imagery, N = numeracy, A = acquired cultural knowledge

Intent was for Diagnostic Use with Individuals, not for Group Comparisons

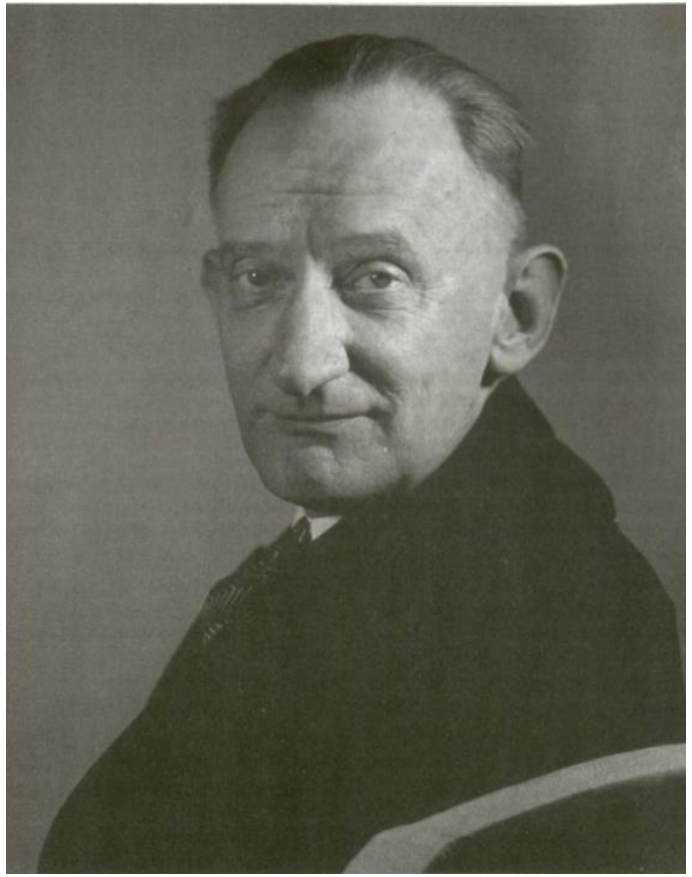
In spite of the system of annotation which we have devised, we think it is the duty of the experimenter to *judge*, *weigh* and *examine* the replies. Our method is not an automatic weighing machine like those in railway stations, which register automatically the weight of a person, without his intervention or assistance…The results of our examination have no value if deprived of all comment; they need to be interpreted

(Binet & Simon, 1908/1916, 222, 239)

Widespread Popularity of Binet-Simon Scale

- “Perhaps no device pertaining to education has ever risen to such sudden prominence in public interest throughout the world as the Binet-Simon measuring scale of intelligence” (Bell, 1912, 102).
- By 1914, a bibliography of literature related to the Binet-Simon scale included 254 citations.
- Used in Canada, England, Australia, New Zealand, South Africa, Germany, Switzerland, Italy, Russia, China, Japan and Turkey.
- Popularized in US by Goddard and Termin
- Led to test use for controversial group comparisons and interpretations

L. L. Thurstone (1887-1955)



(1925) A method of scaling psychological and educational tests. *The Journal of Educational Psychology*, 16(7), 433-451.

(1926) The mental age concept. *Psychological Review*, 33(4), 268-278.

(1927) The unit of measurement in educational scales. *The Journal of Educational Psychology*, 18(8), 505-523.

(1928) The absolute zero in intelligence measurement. *Psychological Review*, 35(3), 175-197.

Thurstone's Method of Absolute Scaling

- Units of a raw test score have an equivocal meaning because items can (and do) differ in difficulty
- Goal: The scale must be independent of the unit selected for the raw scores and from the shape of the distribution of raw scores (Thurstone, 1927, 519)

The whole study of intelligence measurement can hardly have two more fundamental difficulties than the lack of a unit of measurement and the lack of an origin from which to measure.

(Thurstone, 1928, 176)

Requirements for Absolute Scaling

- Applies when a test of ability is given to students at different ages
- Tests have items that are scored as right or wrong.
- Students in adjacent ages take overlapping/common items

Core Assumption: The underlying attribute being measured by each test can be conceptualized as a continuous quantitative variable with a normal distribution.

Finding a “Common Baseline”

Define the unit of measurement by the SD

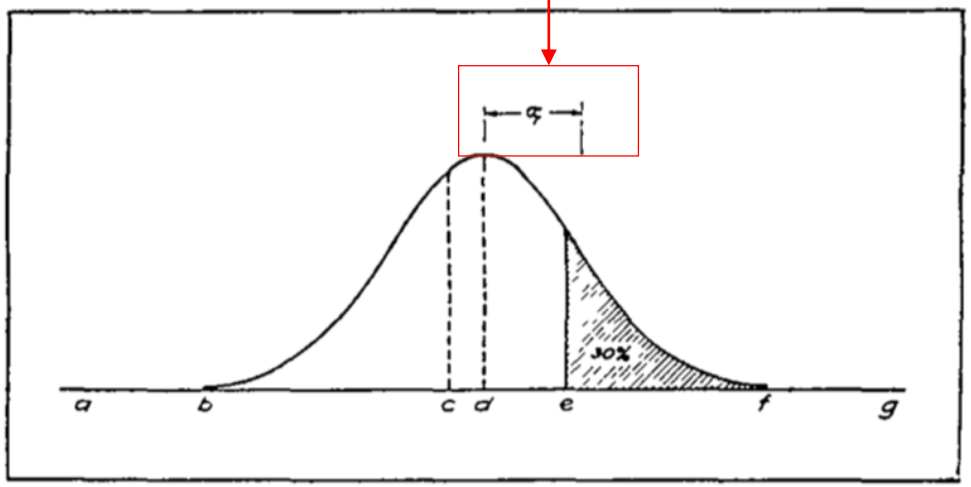


Fig. 1.

The letters represent test items that range from very easy (a, b) to very hard (f, g) to answer correctly in an absolute sense.

For any two distributions of student ages, need to find the locations on a common scale, and need to establish a common unit of measurement

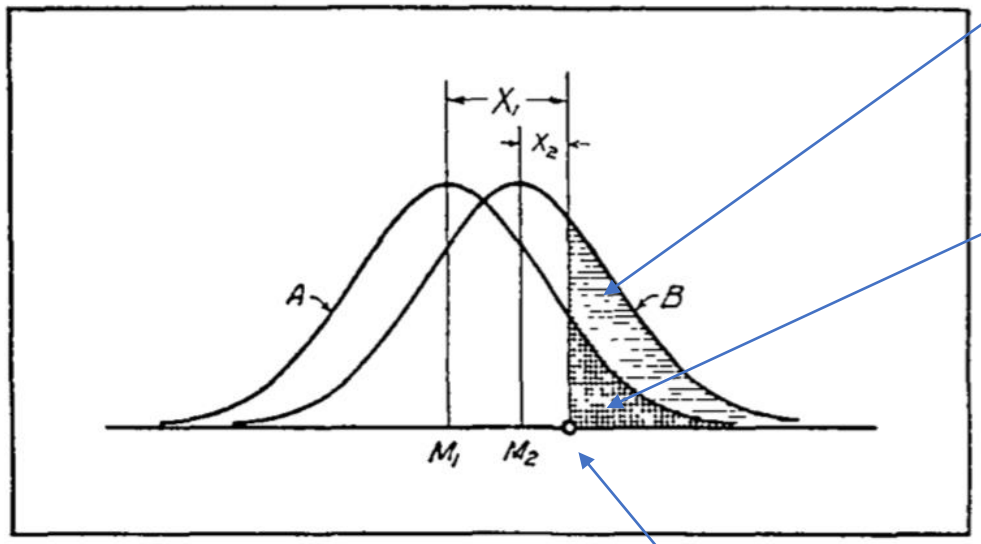


Fig. 2.

Older age group

Younger age group

A specific test question answered by both groups

Result: Depiction of Growth on an Absolute Scale

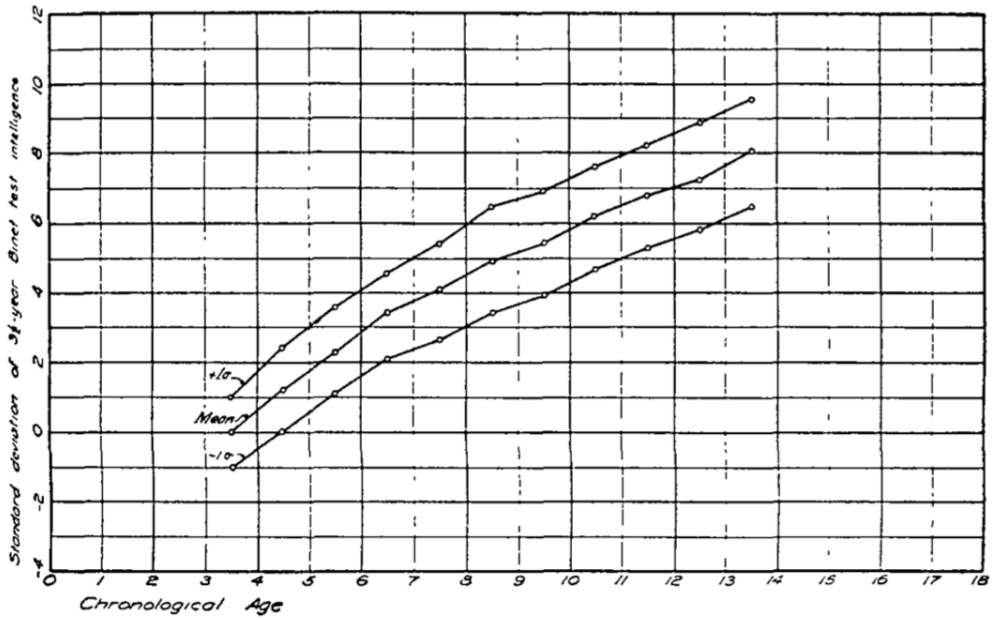


FIG. 4.

- Growth showed no sign of rapid deceleration
- Variance increased over time (notice vertical distance from middle line as grade increases)
- Missing evidence on growth in intelligence into adulthood

A “Map” of Binet Items on an Absolute Scale

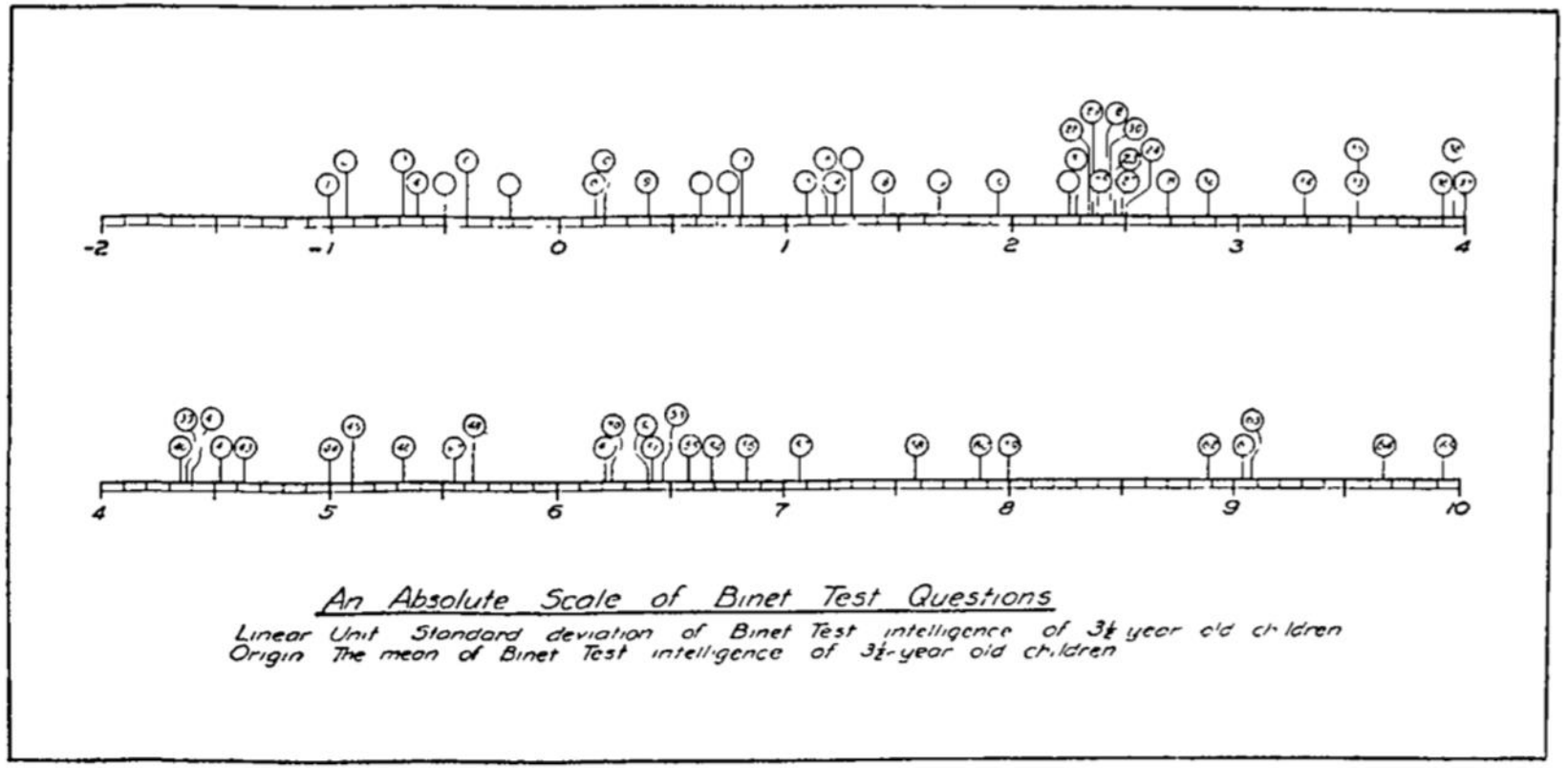
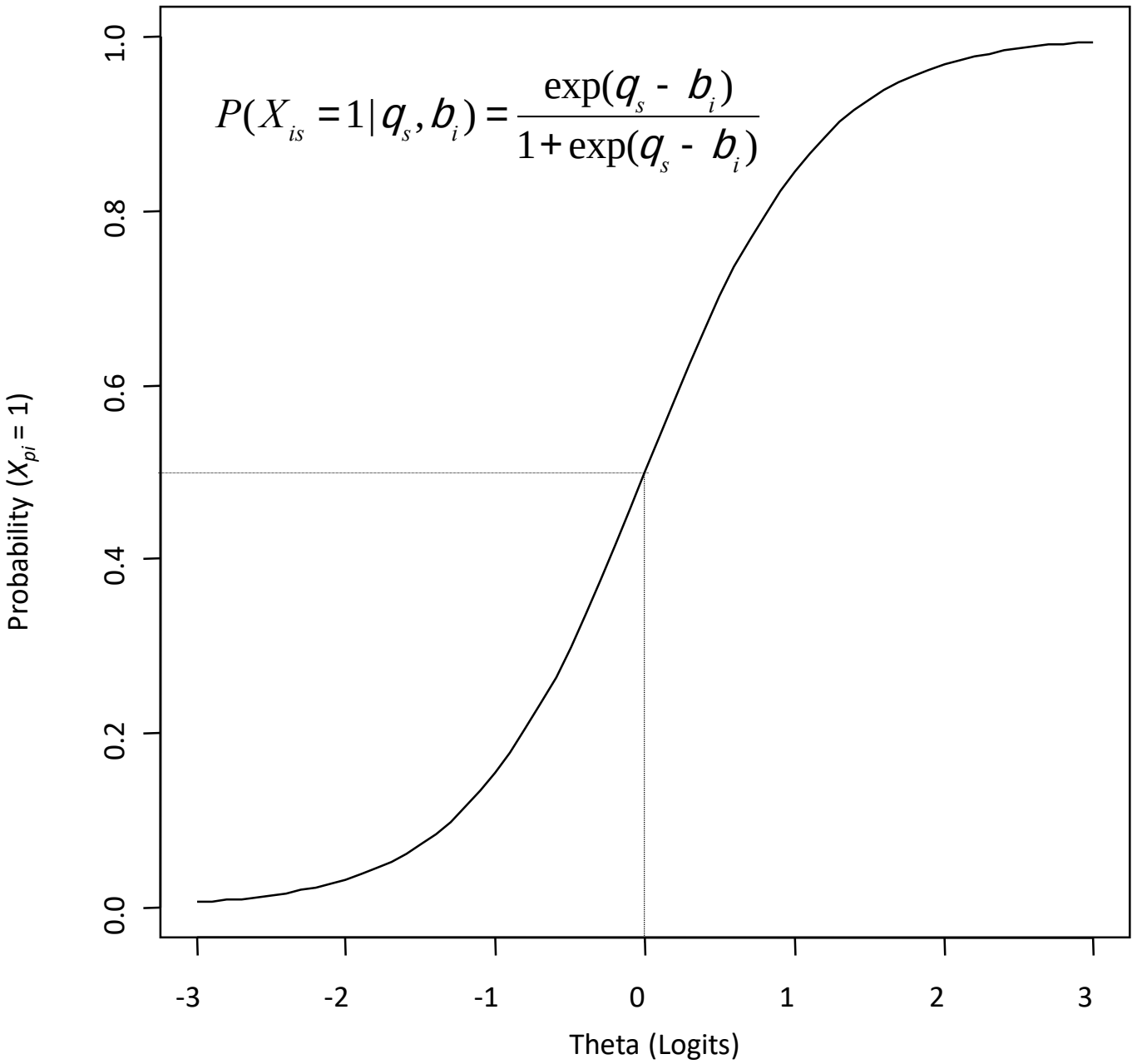


Fig. 6.

Georg Rasch (1901-1980)

Georg Rasch (1901-1980)

- A group of students (struggling academically) in grades 3-7 attend “special reading classes”.
- We wish to evaluate “the benefit these pupils may have drawn from this sort of education.”
- Students are given reading tests “before and after the transfer to reading classes.”
- “Now, of course the tests used changed from one occasion to another, but nonetheless our aim was to evaluate the progress of each pupil.”

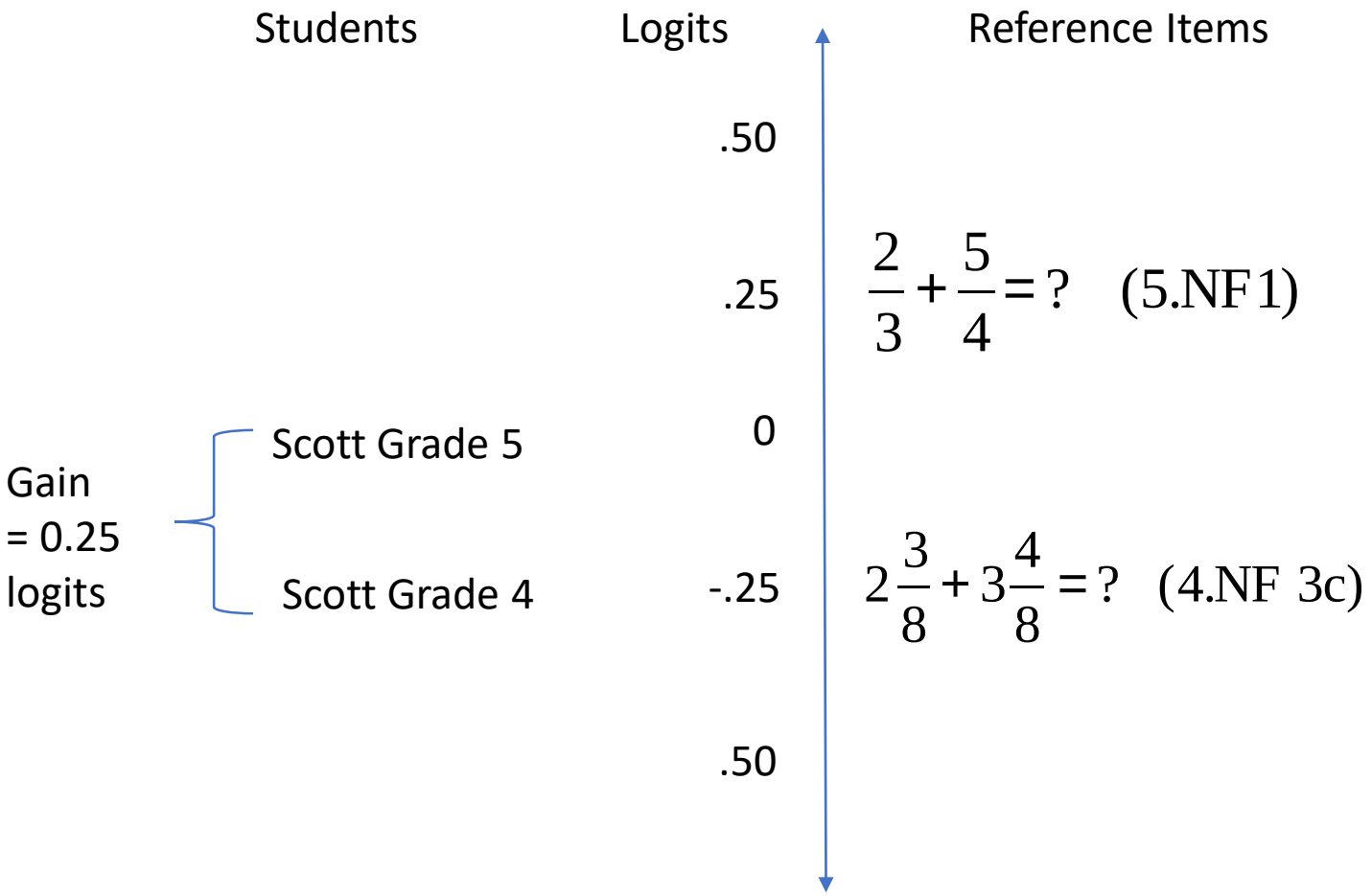


The Rasch Model

The probability of a correct response is a function of two parameters, the difficulty of the item *and* the ability of the student.

Both parameters can be expressed on a common logit scale.

Utility of the Rasch Model: Item-Person Map



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

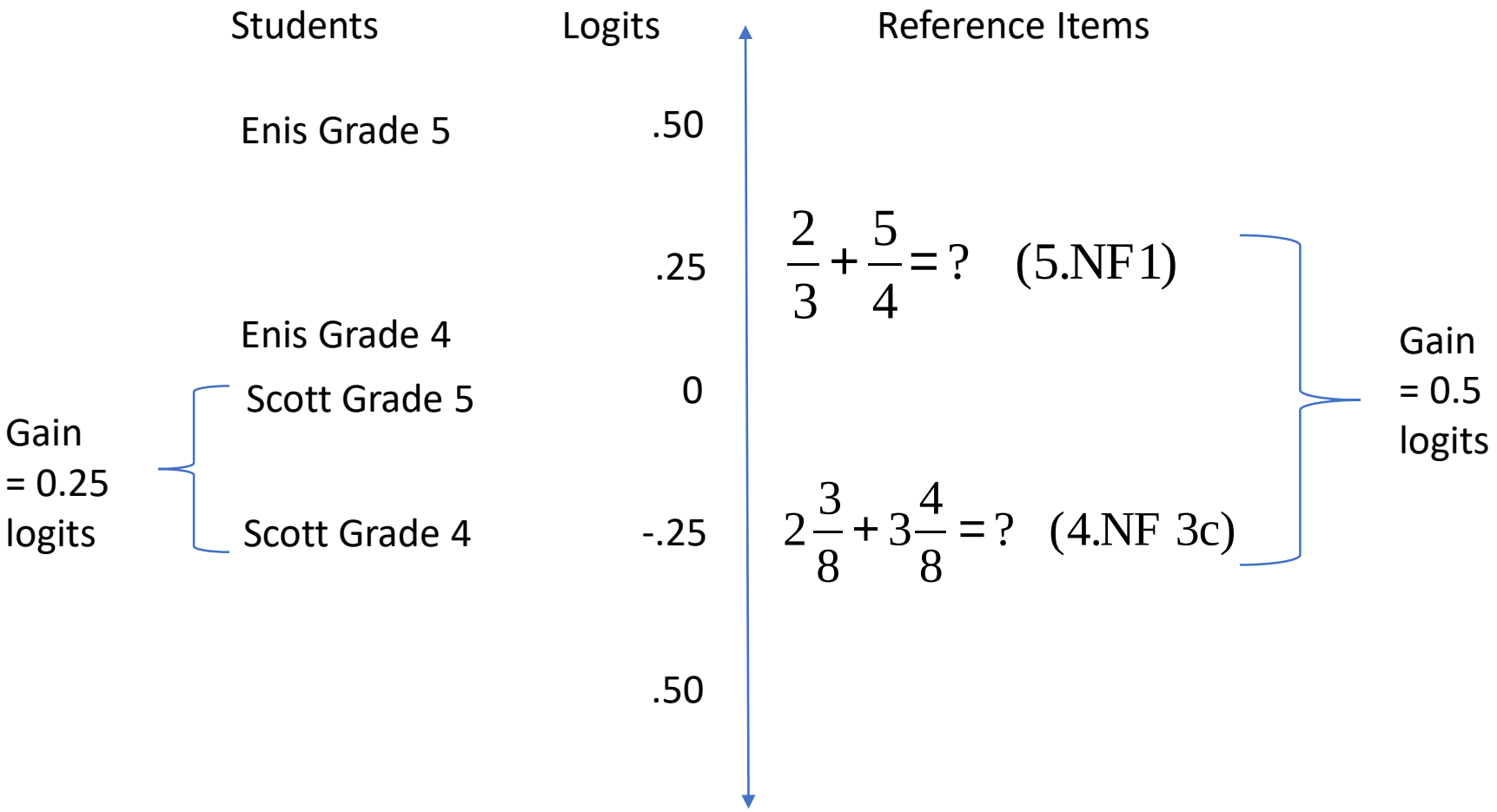
$$2\frac{3}{8}+3\frac{4}{8}=?\quad(4.NF\ 3c)$$

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)

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

Using Items as Frame of Reference



The Value of Invariant Comparisons

If data fits the Rasch Model, then

- Estimates of growth across grades for each student has the same magnitude no matter what item (or item difference) is used as a frame of reference, because when logits are subtracted

$$(\theta_2 - \beta_i) - (\theta_1 - \beta_i) = \theta_2 - \theta_1$$

- Expected growth can be defined by differences in reference item locations, and this does not vary by initial ability of student.

Scales with Substantive Measurement Units

The process of defining a scale can be conceptually and practically separable into at least two parts

a. the definition or determination of the relative interpoint distances on the scale; and
b. The assignment of a system of numerical values to the benchmarks of the scale. (Angoff, 1971, p. 352)

"The long-term value of a test and the scale on which its scores are expressed will depend more on the measurement qualities built into the test…than on any normative properties which might be embodied in the scale and appropriate in the short term." (Angoff, 1971, p. 353)

Angoff, R. (1971). Scales, norms, and equivalent scores. In R. Thorndike (Ed.) *Educational Measurement*, 2nd Edition.

Summary of Historical Lessons

1. A temporal (age) scale is highly seductive…
2. but psychometrically problematic for group comparisons
3. Thurstone's approach defined scale units in terms of variance, but
   - Established criterion of invariance to evaluate the scale
   - Used item locations to interpret the scale
4. The Rasch Model took this the next step
5. Provides opportunity to understand person growth with respect to differences in item content

Toward Content-Referenced Units

Creating a Content-Referenced Unit of Measurement

1. Start with a developmental hypothesis (specify an intended learning progression for an educational/psychological attribute)
2. Write test items with difficulty that is predictable by design
3. Collect data that spans the intended range of the LP
4. Calibrate the data to a model in the Rasch family
5. Choose two anchor locations on the logit scale that define a reference distance that is instructionally meaningful: unit of measurement for growth
6. Express differences in magnitudes of growth as a ratio to this reference distance.

Quick Digression: Mechanics of a Change of Scale

- Consider temperature. In the US we use the Fahrenheit Scale.
  - 0 = temperature of a solution of brine made of equal parts ice, water and salt.
  - 32 = melting point of ice
  - 212 = boiling point of water
  - 1 degree represents 1/180 of the interval between melting of ice and boiling of water.

- Converting to a Celsius scale
- Solve the two eqns below for a and b:

$a + b*32 = 0$

$a + b*212 = 100$

$b = .556, a = -17.78$

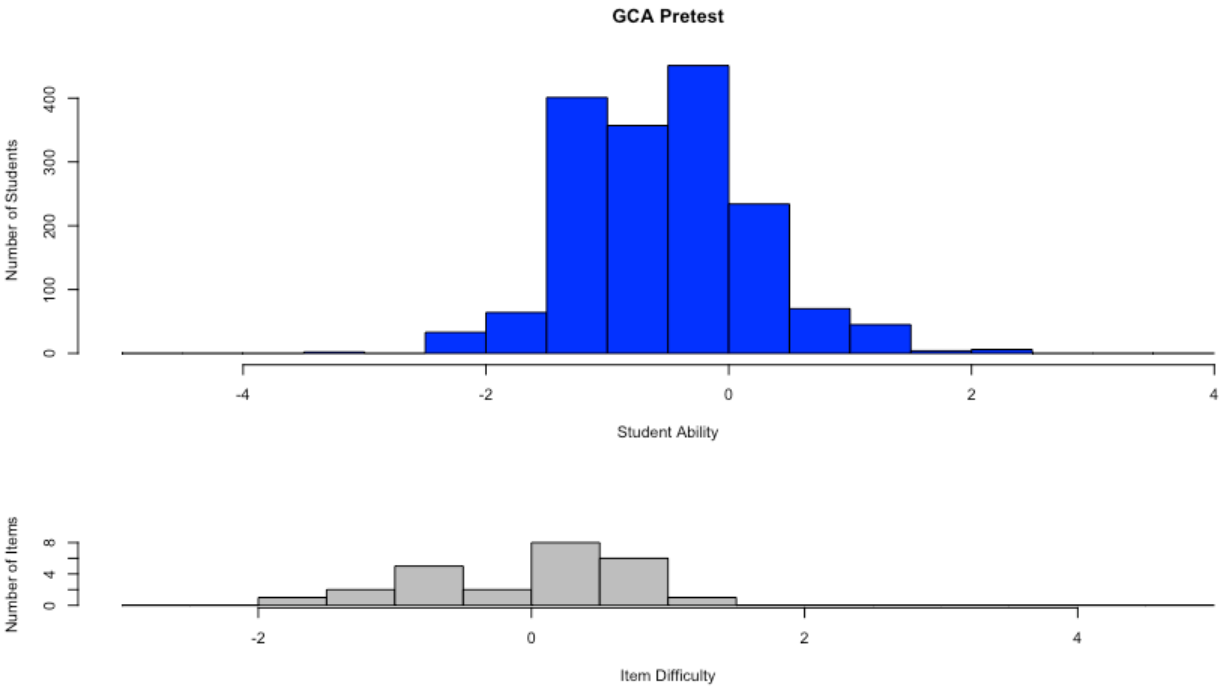
Now we have a new origin and a new reference unit: 1 degree represents 1/100 distance between melting of ice and boiling of water.

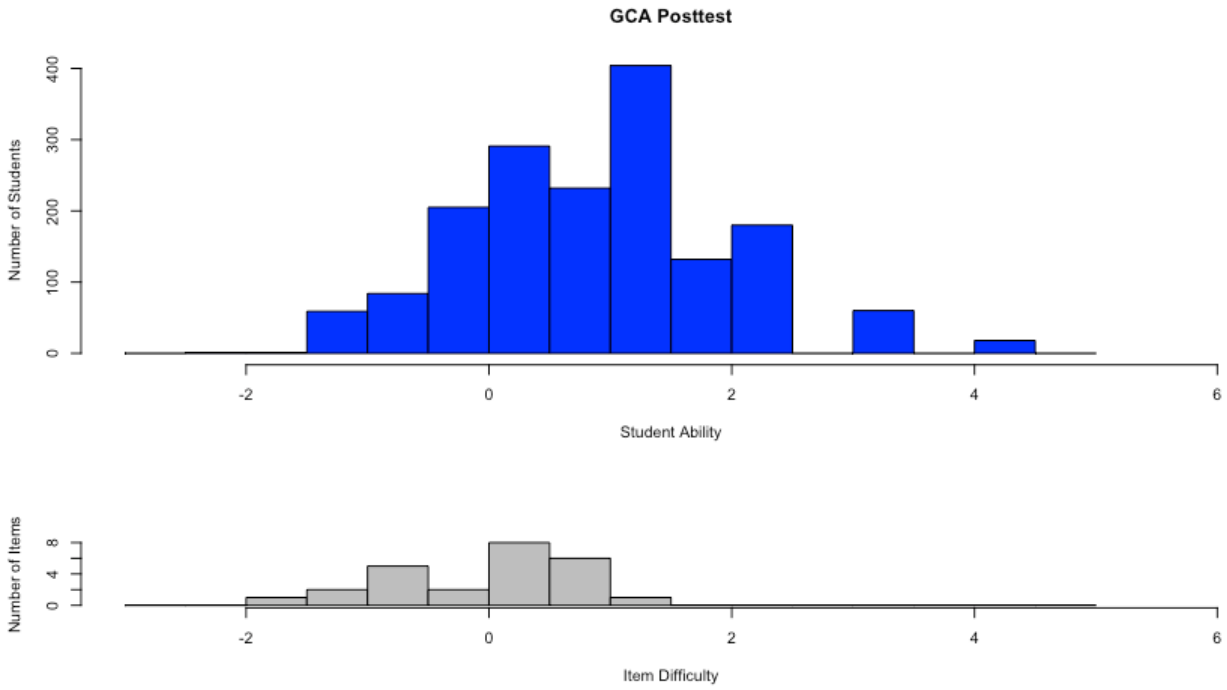
Example 1

Project: Comparing Performance of University Students on a Genetics Concept Inventory

- Data: Students enrolled in 9 different introductory genetics courses at 6 unique American post-secondary institutions. (N = 1,667)
- Four different semesters from Fall 2015 to Spring 2017.
- All students take a 25 item “concept inventory” known as the Genetics Concept Assessment (GCA) at beginning (pre) and end of semester (post).
- All items mapped to 8 specific genetics learning goals and coded for cognitive complexity according to 3 levels Bloom’s Taxonomy (Understand, Apply, Analyze).
- Pre and post test performance calibrated to a common logit scale using the Rasch Model.

Comparing Pre vs Post Administration





The units of this scale are in logits. The mean difference from pre to post was 1.4 logits. But without additional information, this is mostly uninterpretable

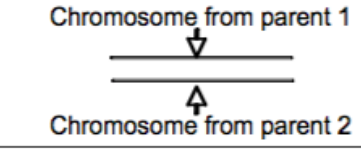
Genetics Concept Inventory (GCA; Smith, Wood, Knight, 2008)

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.

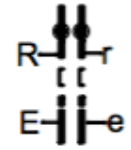


a) #1 only
b) #2 only
c) #3 only
d) #2 and #3 only
e) #1, #2 and #3

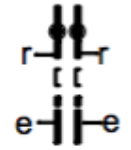
Difficulty = -1.074

Bloom Level 3: Apply

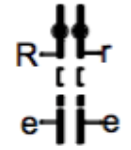
24. Two different genes are located on the same chromosomal pair in rabbits. A particular female rabbit is heterozygous for alleles of both these genes, with the alleles arranged as shown in the diagram to the right. Scientists know that the two genes are on the same chromosome, but do not know their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

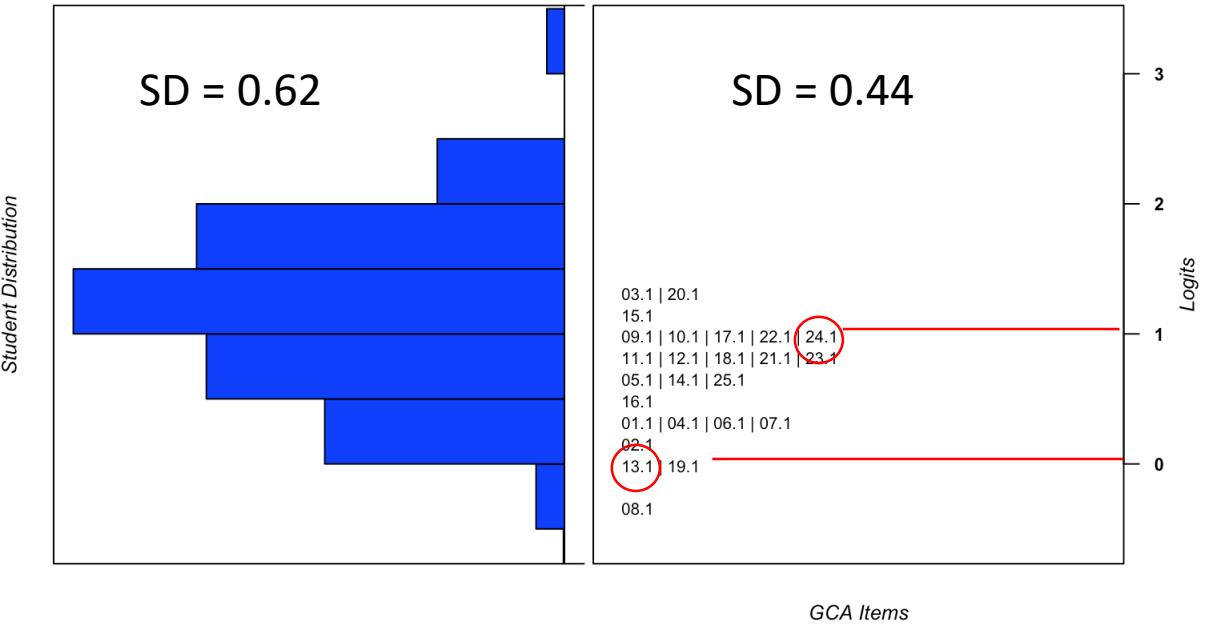


Difficulty = 0.608

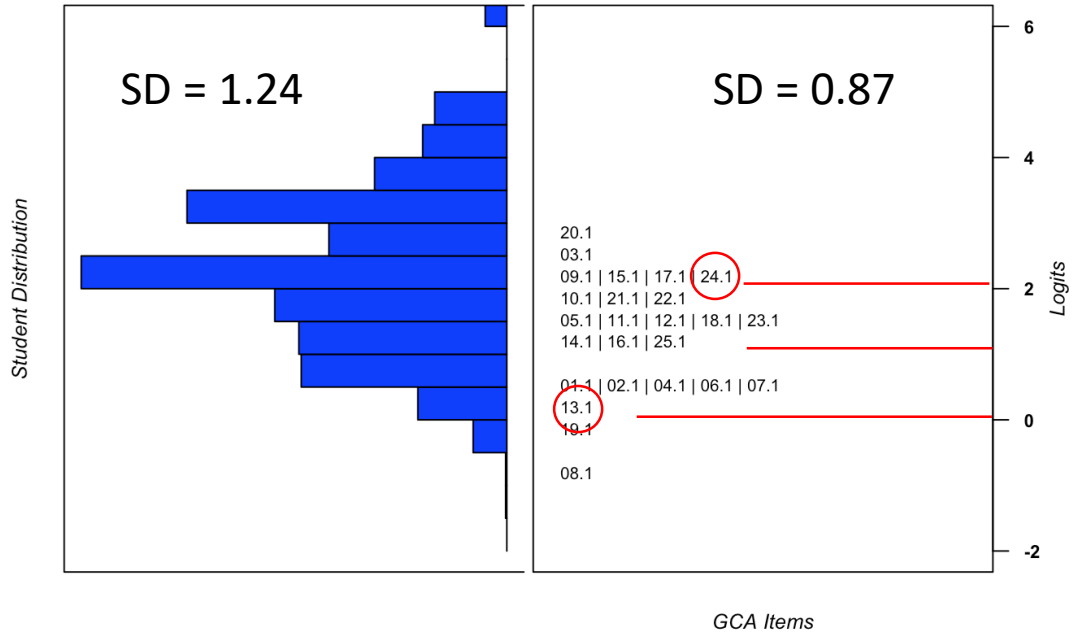
a) Not likely, because the R and e alleles are not on the same chromosome in either parent.
b) Very likely, because the random assortment of chromosomes during cell division to make sperm or eggs allows for the mixing of all alleles.
c) More likely if the two genes are very close together on the chromosome.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

Changing Reference Unit of Scale

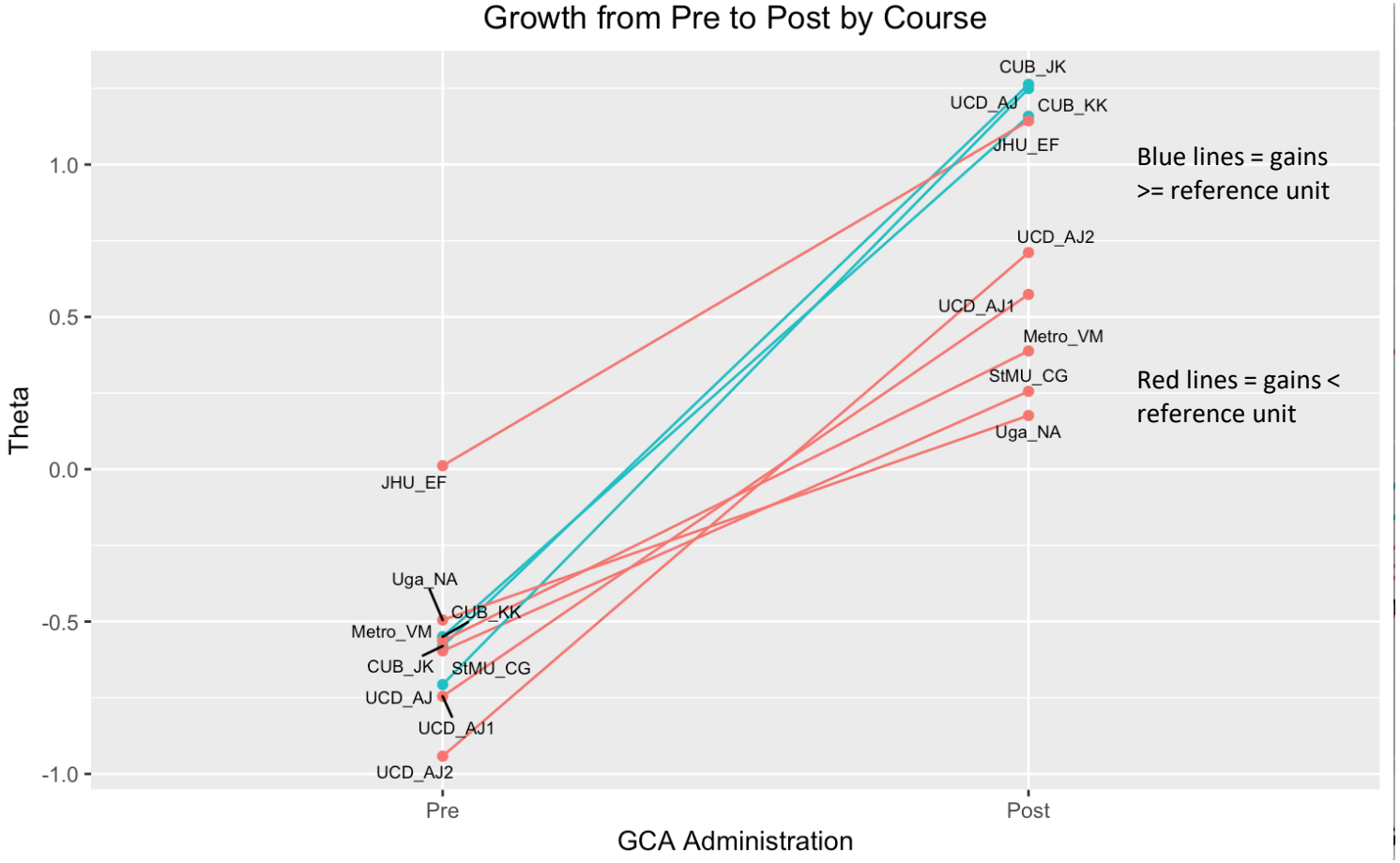
1 unit = difference in location between items 13 and 24.



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



Comparing GCA Gains by Course: Interepting Magnitude by content-based Reference Unit



From this we see that the average student growth at three institutions (CUB_JK, CUB_KK, JHU_EF) was in a substantive—not just statistical—sense more significant than at the others.

Different Magnitude Interpretations

Magnitude Metrics	Pre to Post GCA Gain	Interpretation
Raw Gain	6.8	Average student answered about 7 more GCA items correctly from pre to post
Gain in Logits	1.38	Average student increased GCA by 1.38 logits
Logit Gain/Item-based Reference Unit	.90	Average student increased GCA by 90% of the difference in difficulty between scale anchors
Logit Gain/SD of Post	1.50	Average student increased GCA score by an amount 1.5 times greater than the SD of the post GCA score distribution

Not as Easy as it Looks

- This example did not meet all the requirements I layed out earlier (especially steps 1 and 2)
- The GCA was not designed according to an LP hypothesis
- The instructionally relevant anchor points were defined post hoc
- The invariance of this distance is an open question
- Would we get the same distance if we
  - used different items?
  - used different students?

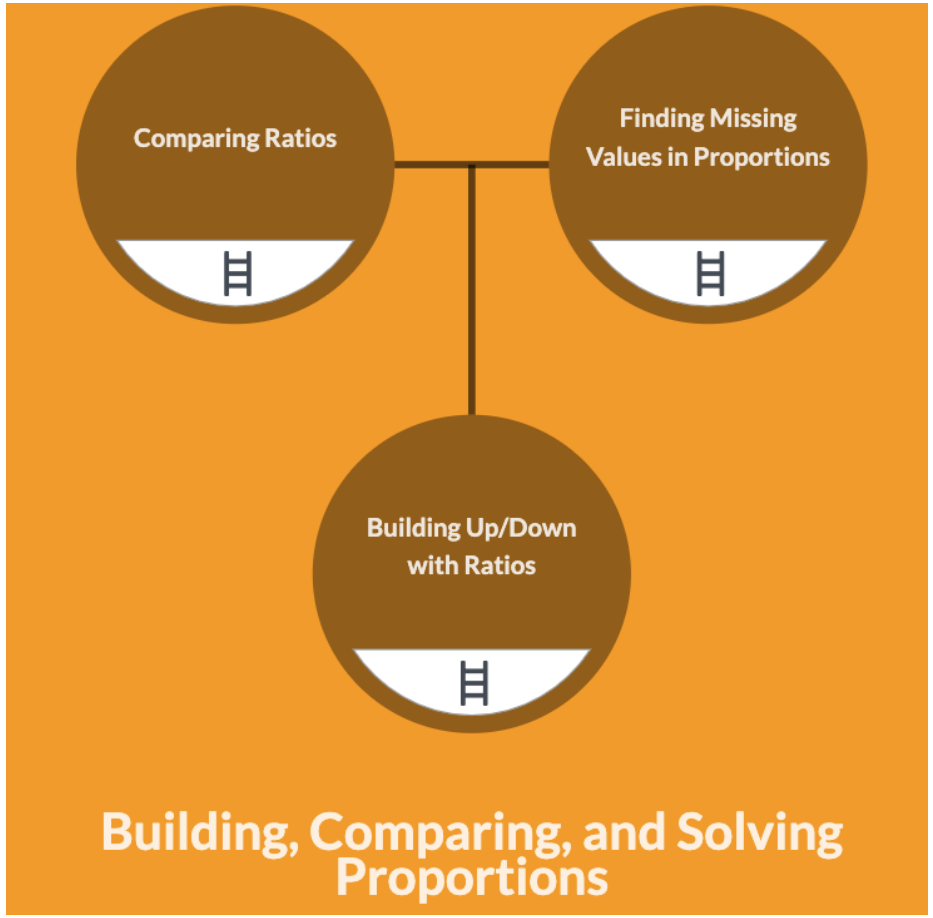
Is distance between these items invariant to Genetics Course (DIF by Course)?

Course	Item 13	Item 24	Difference
StMU_CG	-1.19	0.44	-1.63
UCD_AJ1	-1.39	0.54	-1.94
UCD_AJ2	-0.44	0.60	-1.04
Metro_VM	-0.44	0.60	-1.04
CUB_KK	-1.23	0.92	-2.15
JHU_EF	-1.76	1.37	-3.14
CUB_JK	-1.19	0.81	-2.01
Uga_NA	-0.76	0.16	-0.91
UCD_AJ	-1.99	0.27	-2.26
Combined	-1.07	0.61	-1.68

It doesn't look like it! Probably not surprising, but shows interaction between item difficulty and opportunity of learn in achievement contexts….

Example 2

“Math Mapper” Learning Trajectories



Level	Finding Missing Values in Proportions
7	Given a set of values in a proportional relationship in tables and graphs, identifies the constant of proportionality and relates it to the unit ratio/rate (1/k) and to the equation y=kx
6	Distinguishes proportional from non-proportional relationships
5	Uses 2 × 2 ratio tables that include fractional entries to solve for missing value
4	Finds missing values using multiple methods (rational number multiplication, combinations of whole number multiplication and division, and build up) and equates
3	Describes combinations of multiplication (by ***a***) and division (by ***b***) as multiplication by a rational number, both between quantities and across ratios
2	Finds missing value in a 2 × 2 ratio table, recognizing that equivalence is preserved by combinations of multiplication and division both between the quantities and across ratios
1	Finds missing value in a 2 × 2 ratio table, recognizing that equivalence is preserved by the whole number multiplicative relationships both between the quantities and across ratios

https://www.sudds.co/map

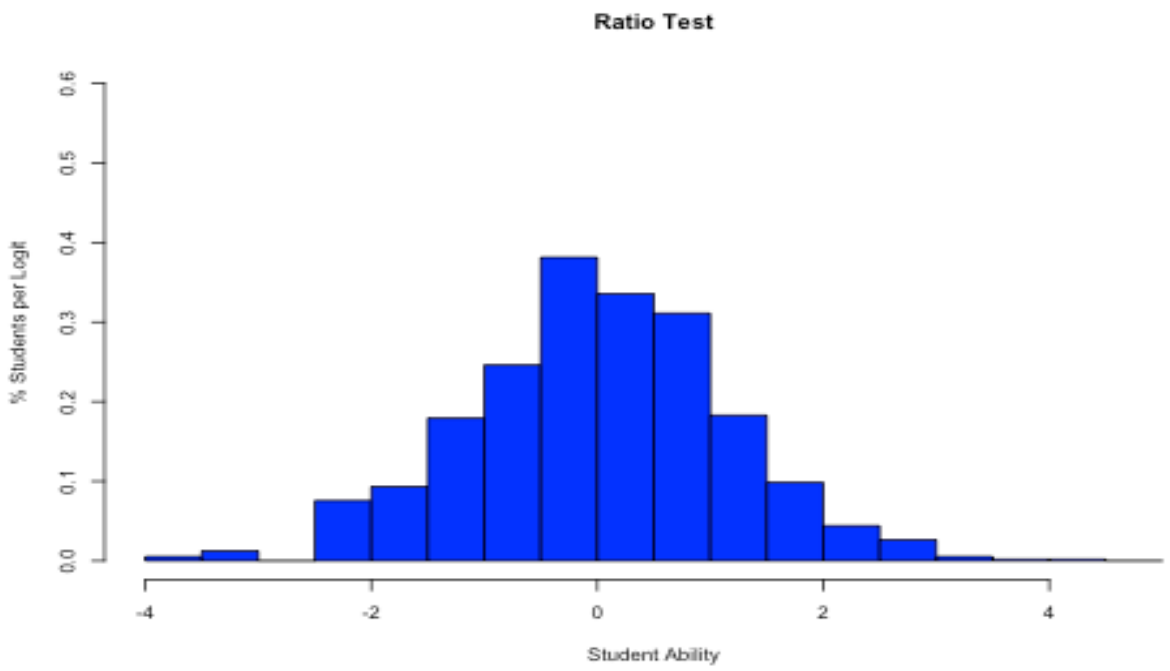
Data

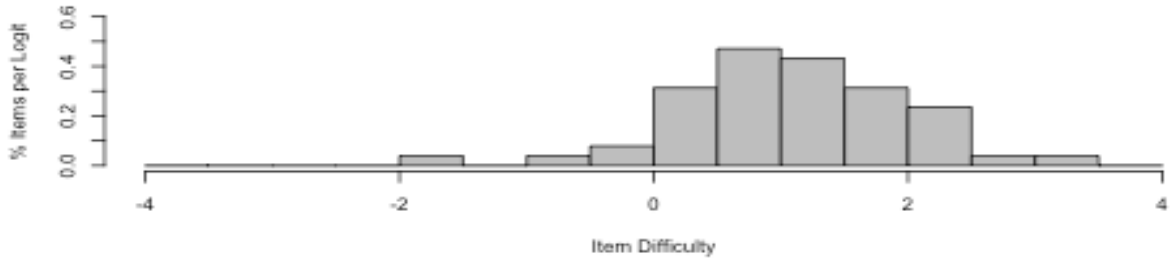
- 5 test forms administered, 3 to grade 7 students, 2 to grade 6 students

Grade	Form	Sample Size	Items
7	1	184	24
7	2	184	21
7	3	184	27
6	4	300	28
6	5	286	28

- Each form contains items that overlap across forms.

Results from Fitting Rasch Model



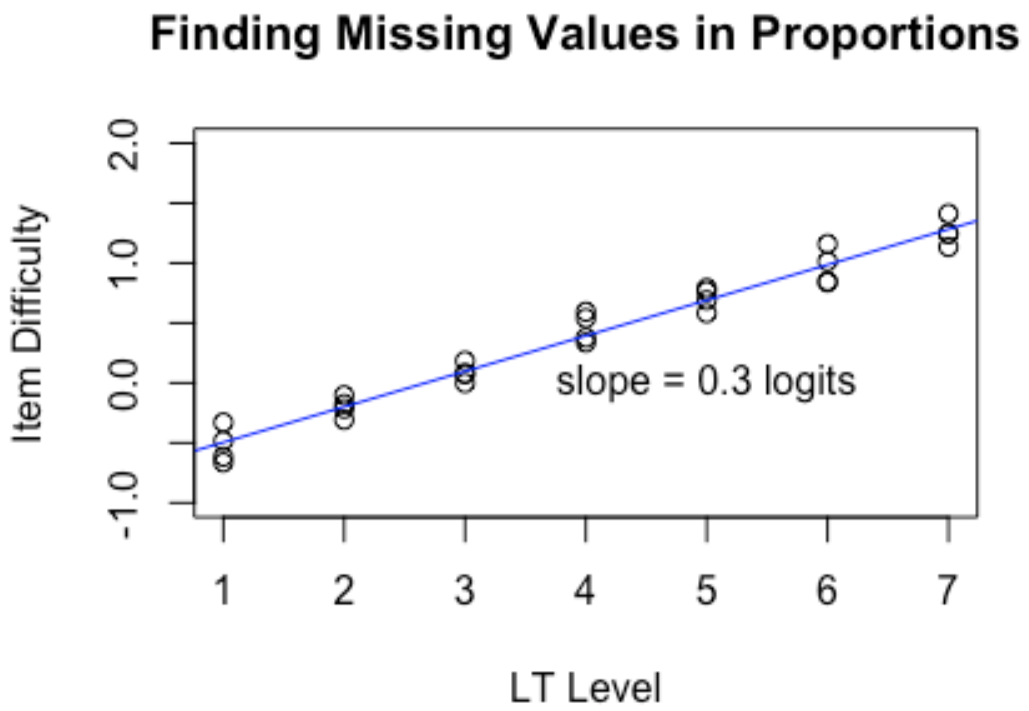


Stat	Student	Item
Mean	0	1.1
SD	1.1	.91
25%ile	-.82	.58
75%ile	.76	1.7
Min	-3.7	-1.6
Max	4.1	3.1

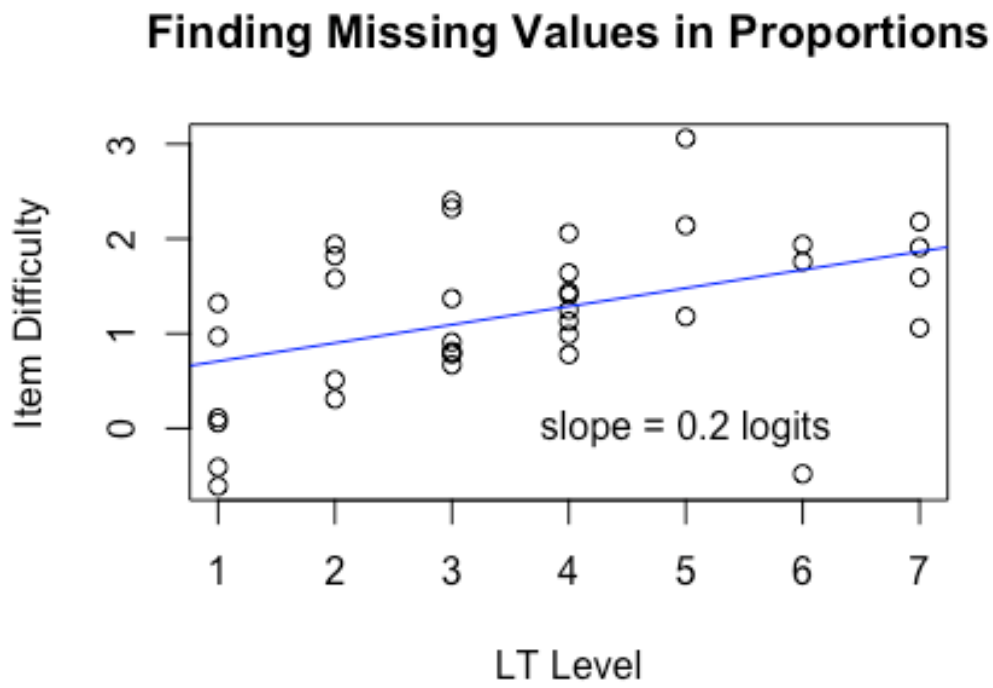
Note: RM estimated under the identification constraint that average of student distribution = 0

Theory-Based Reference Units

A Theoretical Ideal?

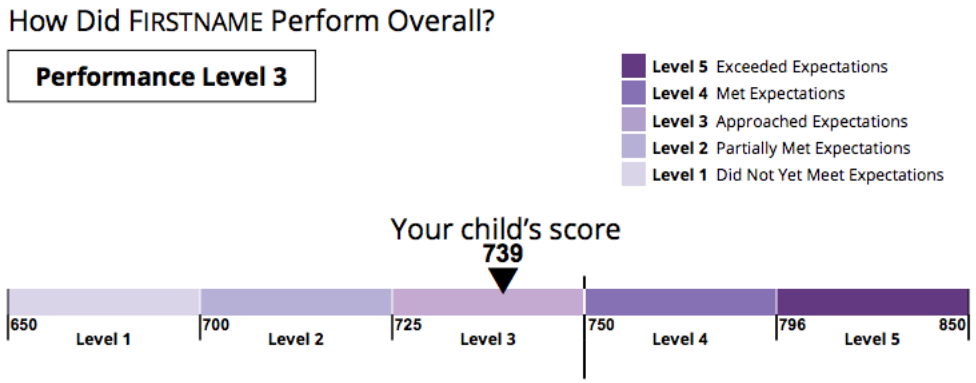


The Empirical Reality



Example 3

“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?

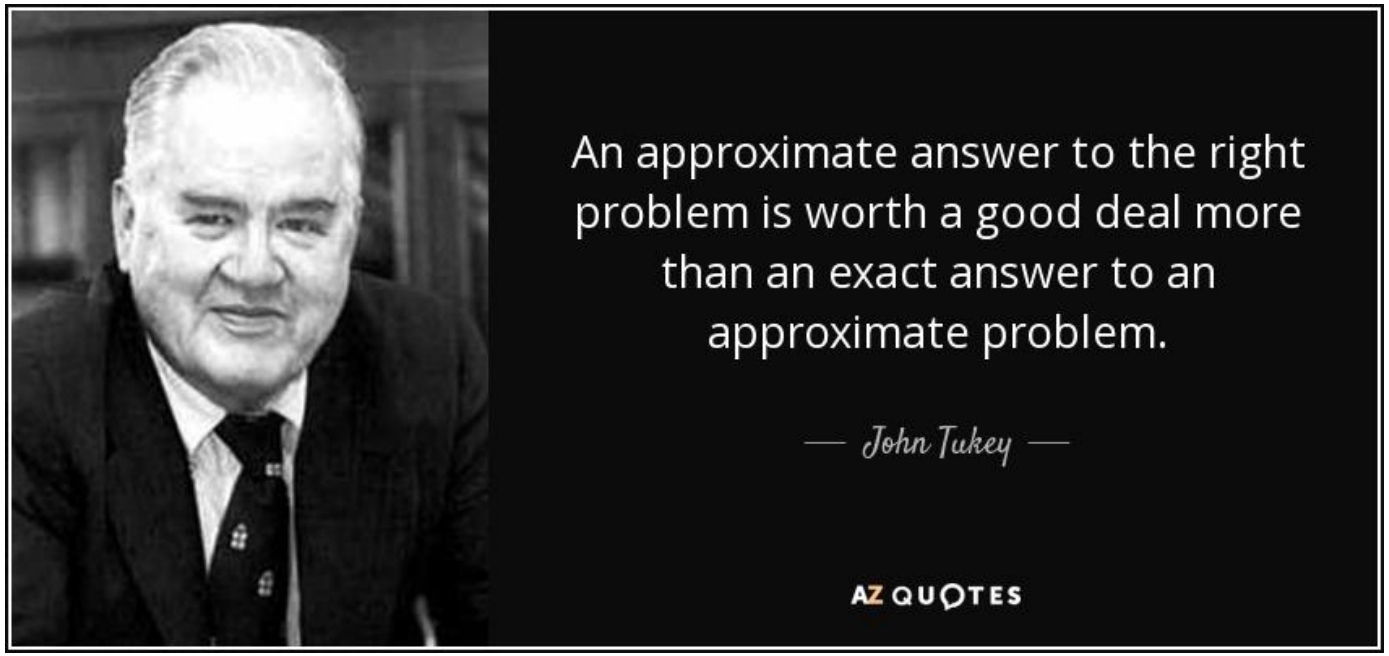
School A: Mean Score of 755
School B: Mean Score of 740

Is this difference significant?

Back to Thesis of this Talk

- Neither SD units nor transformations into temporal units provide a good interpretation of magnitude.
- SD units are too opaque, temporal units suggest misleading precision.
- I suggest an approach that can be used to establish “content-referenced” units of measurement.
- This unit can also be given a spatiotemporal representation that is a prerequisite for a cognitively meaningful interpretation of magnitude.

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

Some References

Briggs, D. C. (2019). Interpreting and visualizing the unit of measurement in the Rasch Model. *Measurement*, 46 (2019) 961–971. https://doi.org/10.1016/j.measurement.2019.07.035

Briggs, D. C. & Peck, F. A. (2015). Using learning progressions to design vertical scales that support coherent inferences about student growth. *Measurement: Interdisciplinary Research & Perspectives*, 13, 75-99

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, 50(2), 204-226.