Looking at Subscores on Summative Assessments…

…Bad Idea or Worst Idea Ever? ☺

First, What We’re NOT Saying…

- We’re not saying that subscore level reports are not desirable or useful
    - We believe that there is legitimate use for subscore level reports that present *actual* subscore level information.
- We’re not saying that tests cannot be constructed to provide valid subscore level information
    - This is just a different purpose for tests and may be at odds with the typical purposes of summative assessments
- We’re not saying that there are no summative assessments that warrant providing subscore level scores
    - There may be, but in general, the information needed to establish the validity of these subscores is either not calculated or not reported

For our portion of the presentation, we’d like to present

- Some theoretical and philosophical reasons to be skeptical about strand scores on summative assessments.
- What some of the current research says about reporting subscores from summative assessments
- Results from looking at some of the above applied to extant summative assessment data

Psychometric Models used for Summative Assessments

- If a summative assessment is scaled using IRT, we need to be skeptical of strand level reports from that assessment
    - IRT models have desirable properties for measuring things like proficiency and growth…many summative assessment use an IRT model for scaling
    - The IRT models in common use, currently, are unidimensional models
        - Isn't unidimensionality at odds with strand scores?
        - The less unidimensional the items in the model, the poorer the model fit
        - There ARE multidimensional IRT models, but they're not in common use…yet
        - Note: 1-PL, 2-PL, and 3-PL IRT models refer to the parameters of the model that are estimated…not the dimensions of the resulting scale

Two important considerations for determining the appropriateness of reporting Subscores:

1) subscore reliability
2) inter-subscore correlations

-- Andrich (2016) presents a way to partition variance in α into components from items and subscores

- This is a classical test theory (CTT) application, not IRT
- Presents ways to look at the relationship of items on a test within and among the subscales
- Presents the correlation between the subscales
  - If the subscales are very highly correlated, are they distinct?
- Recommends at least 30 items per subscale
  - Realistic for our summative assessments?

-- Feinberg & Jurich (2017) present some guidelines for when it is appropriate to present subscores

- Provides a statistic that can be used to determine whether a subscore should be reported
- The Value-added Ratio (VAR), comprised of

  r1, subscore reliability

  r2, disattenuated correlation between the subscore and the remainder score
- VAR ≈ 1.15 + 0.5r1 – 0.67r2
- subscores with VARs >= 1.1 are worth reporting

Andrich, D. (2016). Components of variance of scales with a bifactor subscale structure from two calculations of α. Educational measurement: Issues and practice

Feinberg, R. & Jurich, D. (2017). Guidelines for interpreting and reporting subscores. Educational measurement: Issues and practice.

Having said all that, psychometrics isn’t all black-and-white…

…Perhaps there is a way forward in shades of grey.

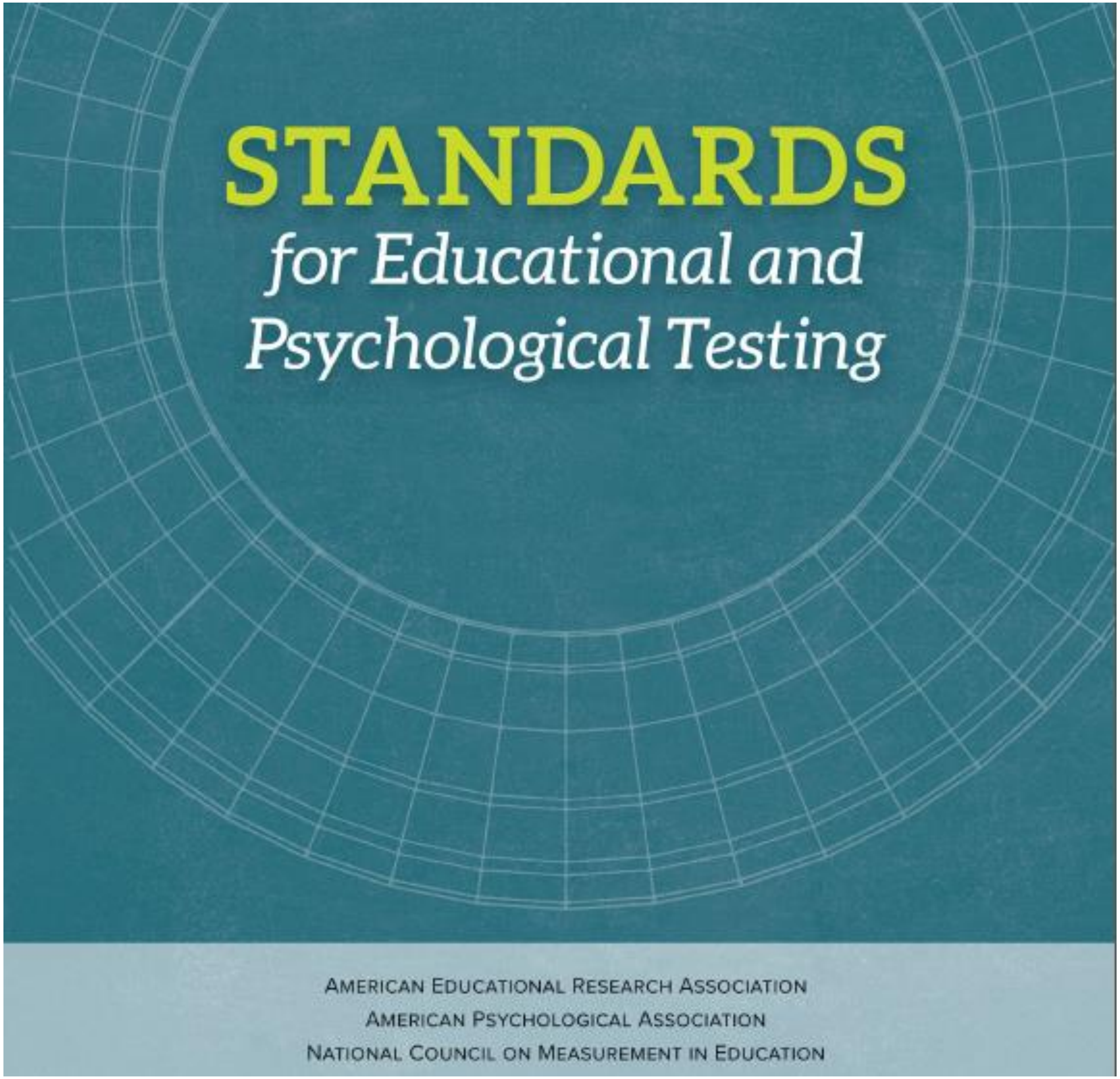
Reporting Subscores

1) What we're *supposed* to do;

2) What we'd *like* to be able to do;

3) What we are *actually* doing.

What we're supposed to do



STANDARDS
for Educational and
Psychological Testing

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION
AMERICAN PSYCHOLOGICAL ASSOCIATION
NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

What we're supposed to do

Cluster 2. Evaluating Reliability/Precision

Standard 2.3: **For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.**

Comment: It is not sufficient to report estimates of reliabilities and standard errors of measurement only for total scores when subscores are also interpreted. The form-to-form and day-to-day consistency of total scores on a test may be acceptably high, yet subscores may have unacceptably low reliability, depending on how they are defined and used. ***Users should be supplied with reliability data for all scores to be interpreted, and these data should be detailed enough to enable the users to judge whether the scores are precise enough for the intended interpretations for use.*** Composites formed from selected subtests within a test battery are frequently proposed for predictive and diagnostic purposes. Users need information about the reliability of such composites

Cluster 3. Specific Forms of Validity Evidence

Standard 1.14: ***When interpretation of subscores* is suggested*, the rationale and relevant evidence in support of such interpretation should be provided*.**

*Comment: When a test provides more than one score, the **distinctiveness and reliability of the separate scores should be demonstrated**, and the interrelationships of those scores should be shown to be consistent with the construct(s) being assessed.*

What we are actually doing

Where we might start:
the M-STEP Technical Manual

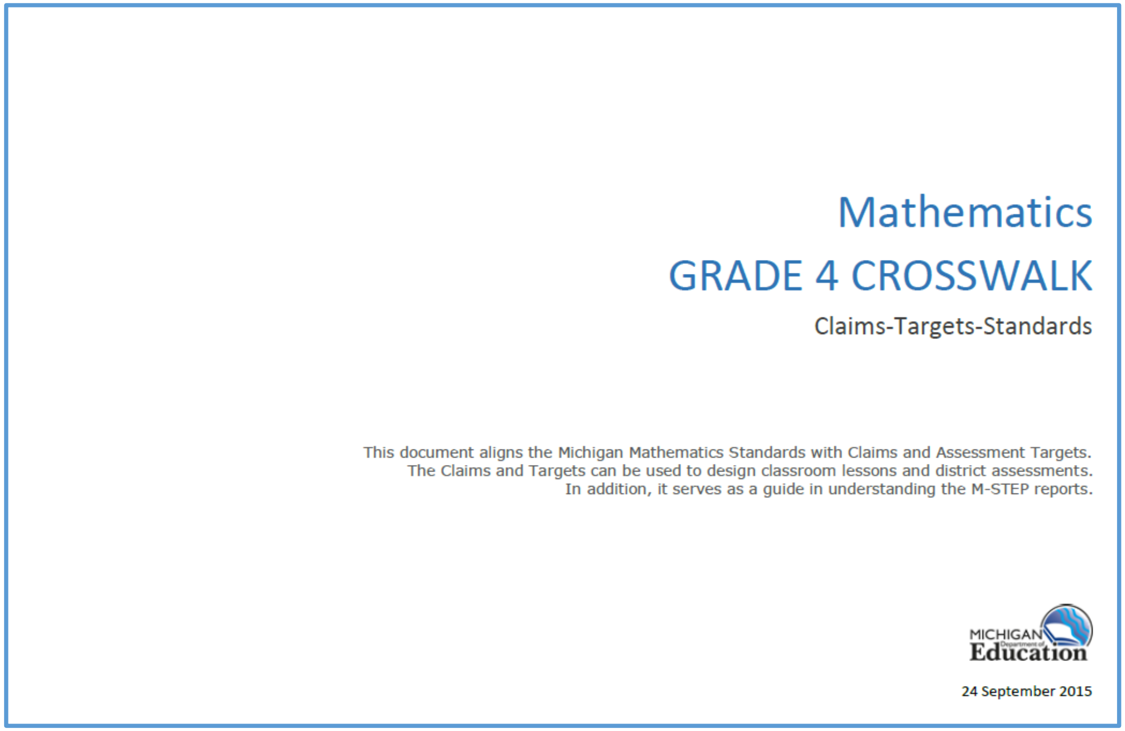
….oops, we don't have one

But, we CAN look at:

1) the Smarter-Balanced Tech Manual,
   and
2) analyses conducted with available data (30,000 kids per grade, about 1/3 of the kids tested in the state)

Before we start, lets look at what we're talking about, when we look at claims

- What is the stuff we're referencing, when we talk about a *Claim Score*?
- Does this underlying stuff provide sufficient specificity, so that *Claim Scores* might be instructionally useful?



Mathematics

GRADE 4 CROSSWALK

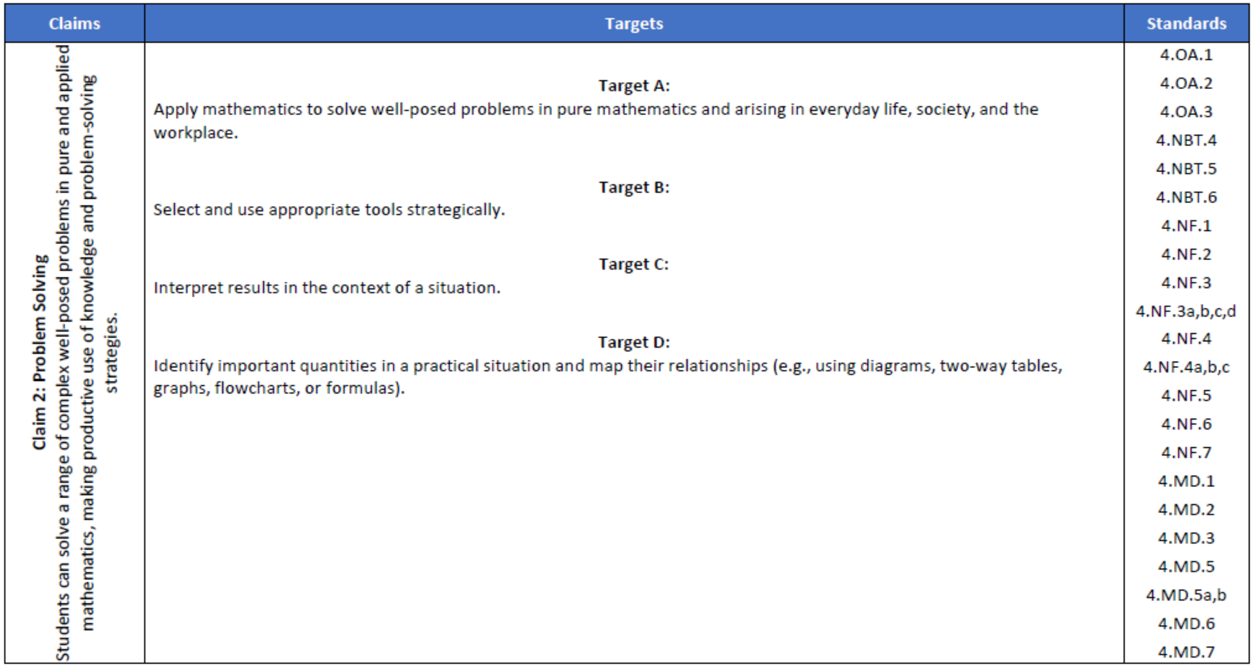
Claims-Targets-Standards

This document aligns the Michigan Mathematics Standards with Claims and Assessment Targets.
The Claims and Targets can be used to design classroom lessons and district assessments.
In addition, it serves as a guide in understanding the M-STEP reports.

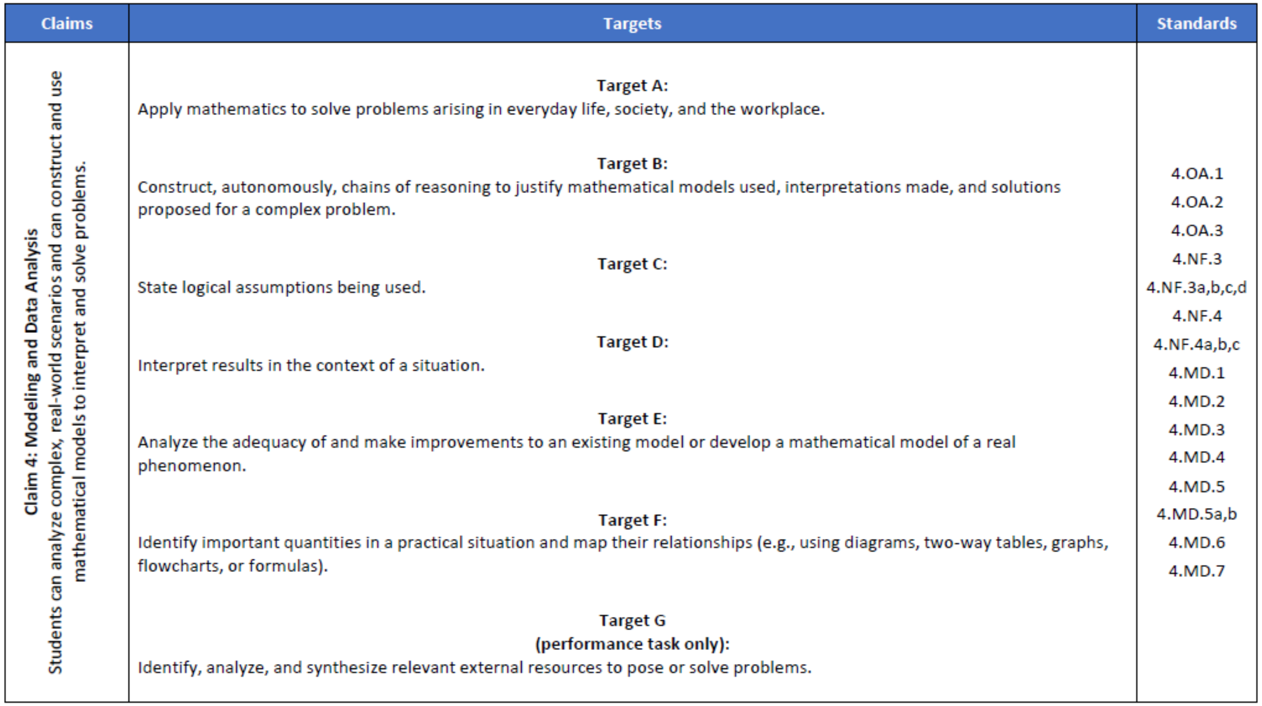
MICHIGAN Department of Education

24 September 2015

Math, Grade 4, Claim 2/4: *Problem Solving and Modeling & Data Analysis*



Claims	Targets	Standards
Claim 2: Problem Solving Students can solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies.	**Target A:** Apply mathematics to solve well-posed problems in pure mathematics and arising in everyday life, society, and the workplace. **Target B:** Select and use appropriate tools strategically. **Target C:** Interpret results in the context of a situation. **Target D:** Identify important quantities in a practical situation and map their relationships (e.g., using diagrams, two-way tables, graphs, flowcharts, or formulas).	4.OA.1 4.OA.2 4.OA.3 4.NBT.4 4.NBT.5 4.NBT.6 4.NF.1 4.NF.2 4.NF.3 4.NF.3a,b,c,d 4.NF.4 4.NF.4a,b,c 4.NF.5 4.NF.6 4.NF.7 4.MD.1 4.MD.2 4.MD.3 4.MD.5 4.MD.5a,b 4.MD.6 4.MD.7



Claims	Targets	Standards
Claim 4: Modeling and Data Analysis Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems.	**Target A:** Apply mathematics to solve problems arising in everyday life, society, and the workplace. **Target B:** Construct, autonomously, chains of reasoning to justify mathematical models used, interpretations made, and solutions proposed for a complex problem. **Target C:** State logical assumptions being used. **Target D:** Interpret results in the context of a situation. **Target E:** Analyze the adequacy of and make improvements to an existing model or develop a mathematical model of a real phenomenon. **Target F:** Identify important quantities in a practical situation and map their relationships (e.g., using diagrams, two-way tables, graphs, flowcharts, or formulas). **Target G (performance task only):** Identify, analyze, and synthesize relevant external resources to pose or solve problems.	4.OA.1 4.OA.2 4.OA.3 4.NF.3 4.NF.3a,b,c,d 4.NF.4 4.NF.4a,b,c 4.MD.1 4.MD.2 4.MD.3 4.MD.4 4.MD.5 4.MD.5a,b 4.MD.6 4.MD.7

Math, Grade 4, Claim 2/4: Problem Solving, Modeling & Data Analysis

(8 to 10 items)

Use the four operations with whole numbers to solve problems.

OA1. Interpret a multiplication equation as a comparison, e.g., interpret35 =5 × 7 as a statement that 35 is 5 times as many as 7 and 7 times as many as 5.Represent verbal statements of multiplicative comparisons as multiplication equations.

OA2. Multiply or divide to solve word problems involving multiplicative comparison, e.g., by using drawings and equations with a symbol for the unknown number to represent the problem, distinguishing multiplicative comparison from additive comparison.[1]

OA3. Solve multistep word problems posed with whole numbers and having whole-number answers using the four operations, including problems in which remainders must be interpreted. Represent these problems using equations with a letter standing for the unknown quantity. Assess the reasonableness of answers using mental computation and estimation strategies including rounding.

Extend understanding of fraction equivalence and ordering

NF1. Explain why a fraction *a*/*b* is equivalent to a fraction(*n* × *a*)/(*n* × *b*) by using visual fraction models, with attention to how the number and size of the parts differ even though the two fractions themselves are the same size. Use this principle to recognize and generate equivalent fractions.

NF2. Compare two fractions with different numerators and different denominators, e.g., by creating common denominators or numerators, or by comparing to a benchmark fraction such as1/2. Recognize that comparisons are valid only when the two fractions refer to the same whole. Record the results of comparisons with symbols >, =, or <, and justify the conclusions, e.g., by using a visual fraction model.

NF3. Understand a fraction *a*/*b* with *a* >1 as a sum of fractions 1/*b*.

a. Understand addition and subtraction of fractions as joining and separating parts referring to the same whole.

b. Decompose a fraction into a sum of fractions with the same denominator in more than one way, recording each decomposition by an equation. Justify decompositions, e.g. , by using a visual fraction model. *Examples: 3/8 = 1/8 + 1/8 + 1/8 ;3/8 = 1/8 + 2/8 ; 2 1/8 = 1 + 1 + 1/8 = 8/8 + 8/8 + 1/8.*

c. Add and subtract mixed numbers with like denominators, e.g., by replacing each mixed number with an equivalent fraction, and/or by using properties of operations and the relationship between addition and subtraction.

d. Solve word problems involving addition and subtraction of fractions referring to the same whole and having like denominators, e.g., by using visual fraction models and equations to represent the problem.

And this:

Extend understanding of fraction equivalence and ordering (continued)

NF4. Apply and extend previous understandings of multiplication to multiply a fraction by a whole number.

a. Understand a fraction *a*/*b* as a multiple of 1/*b*. *For example, use a visual fraction model to represent 5/4 as the product 5 × (1/4), recording the conclusion by the equation 5/4 = 5 × (1/4).*

b. Understand a multiple of *a*/*b* as a multiple of 1/*b*, and use this understanding to multiply a fraction by a whole number. *For example, use a visual fraction model to express 3 × (2/5) as 6 × (1/5), recognizing this product as 6/5. (In general, n × (a/b) = (n × a)/b.)*

c. Solve word problems involving multiplication of a fraction by a whole number, e.g., by using visual fraction models and equations to represent the problem. *For example, if each person at a party will eat 3/8 of a pound of roast beef, and there will be 5 people at the party, how many pounds of roast beef will be needed? Between what two whole numbers does your answer lie?*

Understand decimal notation for fractions, and compare decimal fractions.

NF5. Express a fraction with denominator 10 as an equivalent fraction with denominator 100, and use this technique to add two fractions with respective denominators 10 and 100. *For example, express 3/10 as30/100, and add 3/10 + 4/100 = 34/100.*

NF6. Use decimal notation for fractions with denominators 10 or 100. *For example, rewrite 0.62 as 62/100; describe a length as 0.62 meters; locate 0.62 on a number line diagram.*

Solve problems involving measurement and conversion of measurements from a larger unit to a smaller unit.

MD1. Know relative sizes of measurement units within one system of units including km, m, cm; kg, g; lb, oz.; l, ml; hr, min, sec. Within a single system of measurement, express measurements in a larger unit in terms of a smaller unit. Record measurement equivalents in a two-column table. *For example, know that 1 ft is 12 times as long as 1 in. Express the length of a 4 ft snake as 48 in. Generate a conversion table for feet and inches listing the number pairs (1, 12), (2, 24), (3, 36), ...*

MD2. Use the four operations to solve word problems involving distances, intervals of time, liquid volumes, masses of objects, and money, including problems involving simple fractions or decimals, and problems that require expressing measurements given in a larger unit in terms of a smaller unit. Represent measurement quantities using diagrams such as number line diagrams that feature a measurement scale.

And this:

Solve problems involving measurement and conversion of measurements from a larger unit to a smaller unit. (cont.)

MD3. Apply the area and perimeter formulas for rectangles in real world and mathematical problems. *For example, find the width of a rectangular room given the area of the flooring and the length, by viewing the area formula as a multiplication equation with an unknown factor.*

Represent and interpret data.

MD4. Make a line plot to display a data set of measurements in fractions of a unit (1/2, 1/4, 1/8). Solve problems involving addition and subtraction of fractions by using information presented in line plots. *For example, from a line plot find and interpret the difference in length between the longest and shortest specimens in an insect collection.*

Geometric measurement: understand concepts of angle and measure angles.

MD5. Recognize angles as geometric shapes that are formed wherever two rays share a common endpoint, and understand concepts of angle measurement:

a. An angle is measured with reference to a circle with its center at the common endpoint of the rays, by considering the fraction of the circular arc between the points where the two rays intersect the circle. An angle that turns through 1/360 of a circle is called a "one-degree angle," and can be used to measure angles.

b. An angle that turns through *n* one-degree angles is said to have an angle measure of *n* degrees.

MD6. Measure angles in whole-number degrees using a protractor. Sketch angles of specified measure.

MD7. Recognize angle measure as additive. When an angle is decomposed into non-overlapping parts, the angle measure of the whole is the sum of the angle measures of the parts. Solve addition and subtraction problems to find unknown angles on a diagram in real world and mathematical problems, e.g., by using an equation with a symbol for the unknown angle measure.

Use place value understanding and properties of operations to perform multi-digit arithmetic.

NBT4. Fluently add and subtract multi-digit whole numbers using the standard algorithm.

NBT5. Multiply a whole number of up to four digits by a one-digit whole number, and multiply two two-digit numbers, using strategies based on place value and the properties of operations. Illustrate and explain the calculation by using equations, rectangular arrays, and/or area models.

NBT6. Find whole-number quotients and remainders with up to four-digit dividends and one-digit divisors, using strategies based on place value, the properties of operations, and/or the relationship between multiplication and division. Illustrate and explain the calculation by using equations, rectangular arrays, and/or area models.

SBAC Total Score and Subscore (Claim) Reliability?

Smarter Balanced 2014-15 Technical Report

TABLE 2.12 MATH SUMMATIVE SCALE SCORE MARGINAL RELIABILITY ESTIMATES

Grade	N	Overall	Claim 1	Claims 2/4	Claim 3
3	717,519	0.94	0.88	0.65	0.60
4	702,093	0.94	0.89	0.60	0.70
5	699,713	0.93	0.88	0.56	0.60
6	689,045	0.93	0.87	0.57	0.62
7	681,387	0.91	0.85	0.54	0.49
8	681,197	0.92	0.85	0.54	0.66

[1] Data for the marginal reliability analysis provided by the following Consortium members: Delaware, Hawaii, Idaho, Oregon, South Dakota, US Virgin Islands, Vermont, Washington, West Virginia, California, Montana, Nevada, and North Dakota.

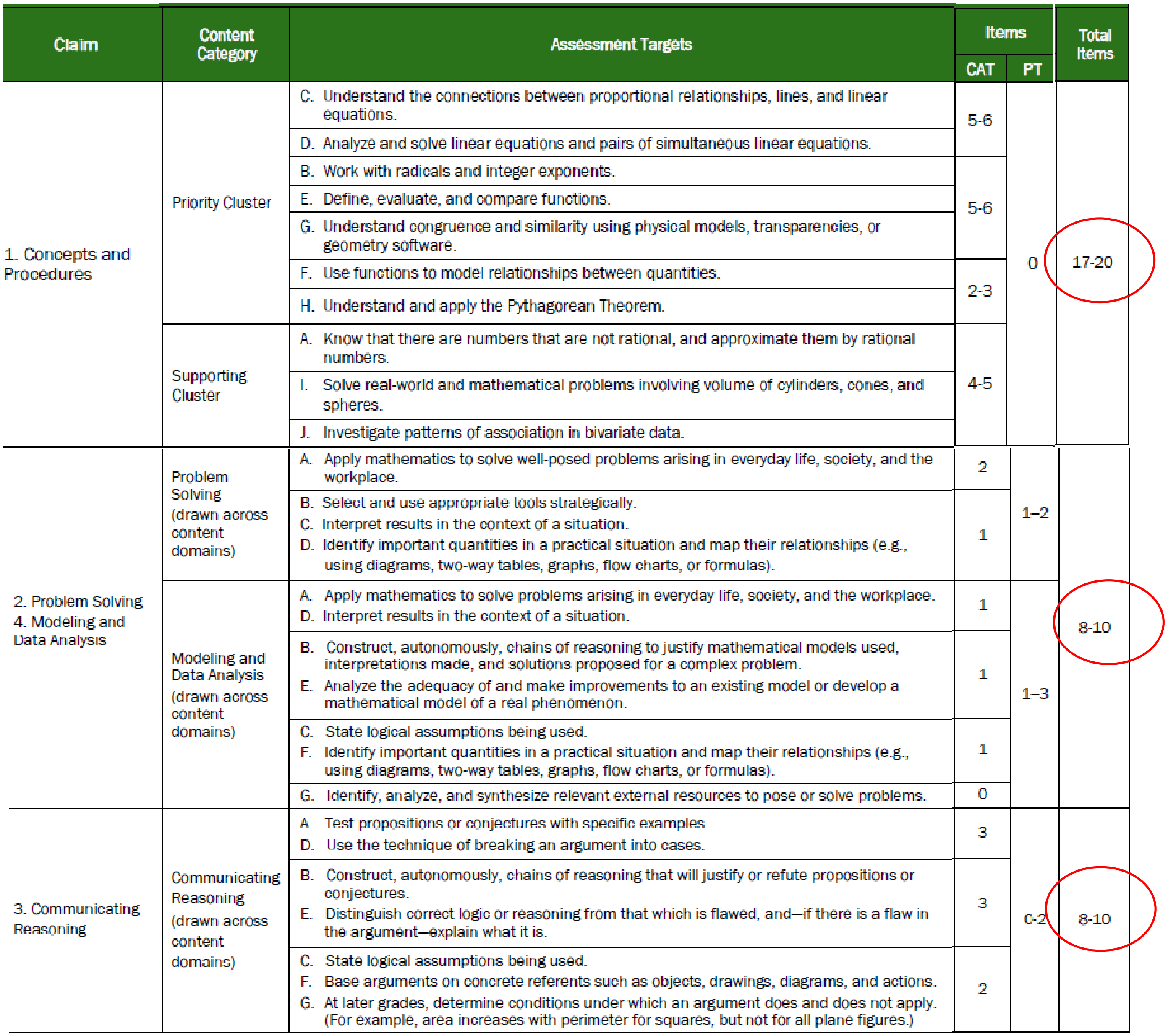
Note: SBAC allows for off-grade items

- *The [adaptive] algorithm proceeds … until a percentage of the test [mathematics, 61%] has been administered, sampling items from all claim areas.*
- *If there is a determination that the student is in either Level 1 or Level 4, the item pool is expanded to include items from no more than two adjacent grades in either direction.*
- *For the remainder of the test, both on-grade and off-grade items can be administered. The item with the best content and measurement characteristics is chosen from the pool.*

Source: Smarter Balanced Summative Assessments, Testing Procedures for Adaptive Item-Selection Algorithm, 2014–15 Smarter Balanced Summative Simulation Report, American Institutes for Research

Number of Items on the Test?

SBAC Math Summative Assessment Blueprint, Grade 8



Claim	Content Category	Assessment Targets	Items CAT	Items PT	Total Items
1. Concepts and Procedures	Priority Cluster	C. Understand the connections between proportional relationships, lines, and linear equations.	5-6	0	17-20
		D. Analyze and solve linear equations and pairs of simultaneous linear equations.			
		B. Work with radicals and integer exponents.	5-6		
		E. Define, evaluate, and compare functions.			
		G. Understand congruence and similarity using physical models, transparencies, or geometry software.			
		F. Use functions to model relationships between quantities.	2-3		
		H. Understand and apply the Pythagorean Theorem.			
	Supporting Cluster	A. Know that there are numbers that are not rational, and approximate them by rational numbers.	4-5		
		I. Solve real-world and mathematical problems involving volume of cylinders, cones, and spheres.			
		J. Investigate patterns of association in bivariate data.			
2. Problem Solving 4. Modeling and Data Analysis	Problem Solving (drawn across content domains)	A. Apply mathematics to solve well-posed problems arising in everyday life, society, and the workplace.	2	1–2	8-10
		B. Select and use appropriate tools strategically. C. Interpret results in the context of a situation. D. Identify important quantities in a practical situation and map their relationships (e.g., using diagrams, two-way tables, graphs, flow charts, or formulas).	1		
	Modeling and Data Analysis (drawn across content domains)	A. Apply mathematics to solve problems arising in everyday life, society, and the workplace. D. Interpret results in the context of a situation.	1	1–3	
		B. Construct, autonomously, chains of reasoning to justify mathematical models used, interpretations made, and solutions proposed for a complex problem. E. Analyze the adequacy of and make improvements to an existing model or develop a mathematical model of a real phenomenon.	1		
		C. State logical assumptions being used. F. Identify important quantities in a practical situation and map their relationships (e.g., using diagrams, two-way tables, graphs, flow charts, or formulas).	1		
		G. Identify, analyze, and synthesize relevant external resources to pose or solve problems.	0		
3. Communicating Reasoning	Communicating Reasoning (drawn across content domains)	A. Test propositions or conjectures with specific examples. D. Use the technique of breaking an argument into cases.	3	0-2	8-10
		B. Construct, autonomously, chains of reasoning that will justify or refute propositions or conjectures. E. Distinguish correct logic or reasoning from that which is flawed, and—if there is a flaw in the argument—explain what it is.	3		
		C. State logical assumptions being used. F. Base arguments on concrete referents such as objects, drawings, diagrams, and actions. G. At later grades, determine conditions under which an argument does and does not apply. (For example, area increases with perimeter for squares, but not for all plane figures.)	2		

Source: Smarter Balanced Assessment Consortium 2014-15 Technical Report

Number of Items on the Test?

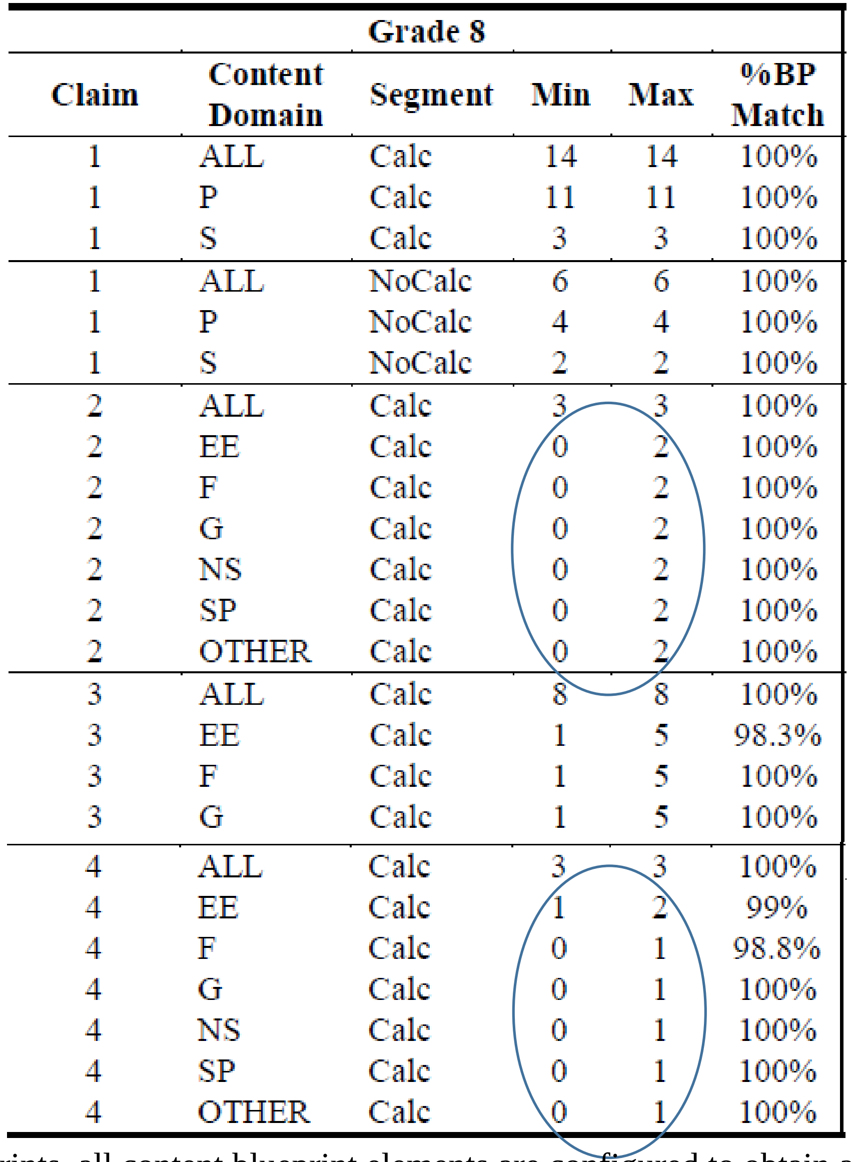
Simulation Study, using rules SBAC Adaptive Item-Selection Algorithm

TABLE 2.6 OVERALL SCORE AND CLAIM SCORE PRECISION/RELIABILITY: MATHEMATICS

	Overall		Claim 1		Claim 2/4		Claim 3	
Grade	Avg # of Items	Marginal Reliability	Avg # of Items	Marginal Reliability	Avg # of Items	Marginal Reliability	Avg # of Items	Marginal Reliability
3	39.7	0.94	20	0.89	9.9	0.74	9.8	0.63
4	39.2	0.93	20	0.88	9.6	0.69	9.6	0.67
5	39.7	0.91	20	0.84	9.8	0.61	9.9	0.63
6	38.8	0.93	19	0.88	9.8	0.67	10.0	0.64
7	39.4	0.90	20	0.83	10.0	0.60	9.4	0.57
8	38.8	0.91	20	0.85	9.1	0.58	9.7	0.66

Number of Items on the Test – which type/target?

Table 7. Percentage of Test Administrations Meeting Blueprint Requirements for Each Claim and Content Domain: Grade 8 Mathematics



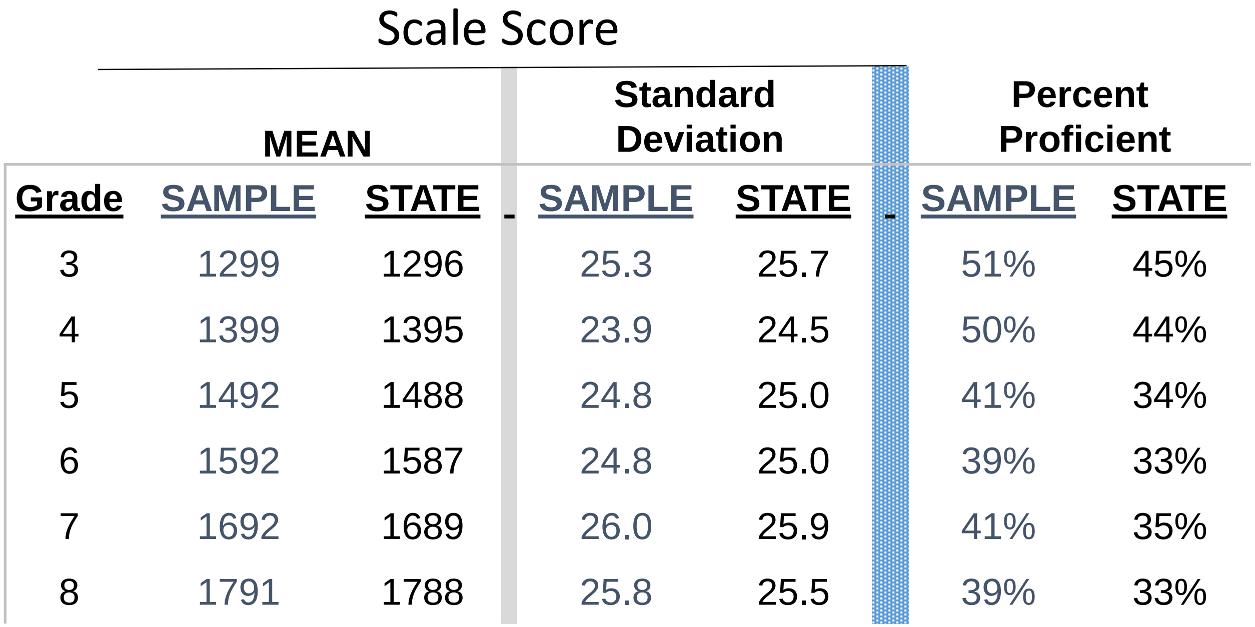
Grade 8					
Claim	**Content Domain**	**Segment**	**Min**	**Max**	**%BP Match**
1	ALL	Calc	14	14	100%
1	P	Calc	11	11	100%
1	S	Calc	3	3	100%
1	ALL	NoCalc	6	6	100%
1	P	NoCalc	4	4	100%
1	S	NoCalc	2	2	100%
2	ALL	Calc	3	3	100%
2	EE	Calc	0	2	100%
2	F	Calc	0	2	100%
2	G	Calc	0	2	100%
2	NS	Calc	0	2	100%
2	SP	Calc	0	2	100%
2	OTHER	Calc	0	2	100%
3	ALL	Calc	8	8	100%
3	EE	Calc	1	5	98.3%
3	F	Calc	1	5	100%
3	G	Calc	1	5	100%
4	ALL	Calc	3	3	100%
4	EE	Calc	1	2	99%
4	F	Calc	0	1	98.8%
4	G	Calc	0	1	100%
4	NS	Calc	0	1	100%
4	SP	Calc	0	1	100%
4	OTHER	Calc	0	1	100%

SBAC Tech Manual: In blueprints, all content blueprint elements are configured to obtain a strictly-enforced range of items administered. The algorithm also seeks to satisfy target level constraints, but these ranges are not strictly enforced.

So, what are marginal reliability indices, for the M-STEP Claim Scores?

Genesee, Lapeer, , Macomb, Oakland, & Ottawa ISDs – around 30,000 kids per grade (out of around 100,000 kids statewide)

Math M-STEP, Spring 2016 – Statewide and My Sample



Scale Score

	MEAN		Standard Deviation		Percent Proficient	
Grade	**SAMPLE**	**STATE**	**SAMPLE**	**STATE**	**SAMPLE**	**STATE**
3	1299	1296	25.3	25.7	51%	45%
4	1399	1395	23.9	24.5	50%	44%
5	1492	1488	24.8	25.0	41%	34%
6	1592	1587	24.8	25.0	39%	33%
7	1692	1689	26.0	25.9	41%	35%
8	1791	1788	25.8	25.5	39%	33%

Computing Marginal Reliability

Pretty simple, with the data available:

Error Variance (average of the squared Scale Score SEs)

Marginal Reliability = $(\sigma_{\theta}^{2} - \sigma_{e}^{2})/ \sigma_{\theta}^{2}$

Total Variance (standard deviation squared)

Total Score and Claim Score Marginal Reliability

M-STEP Math 2016 (approx. 30,000 students per grade)

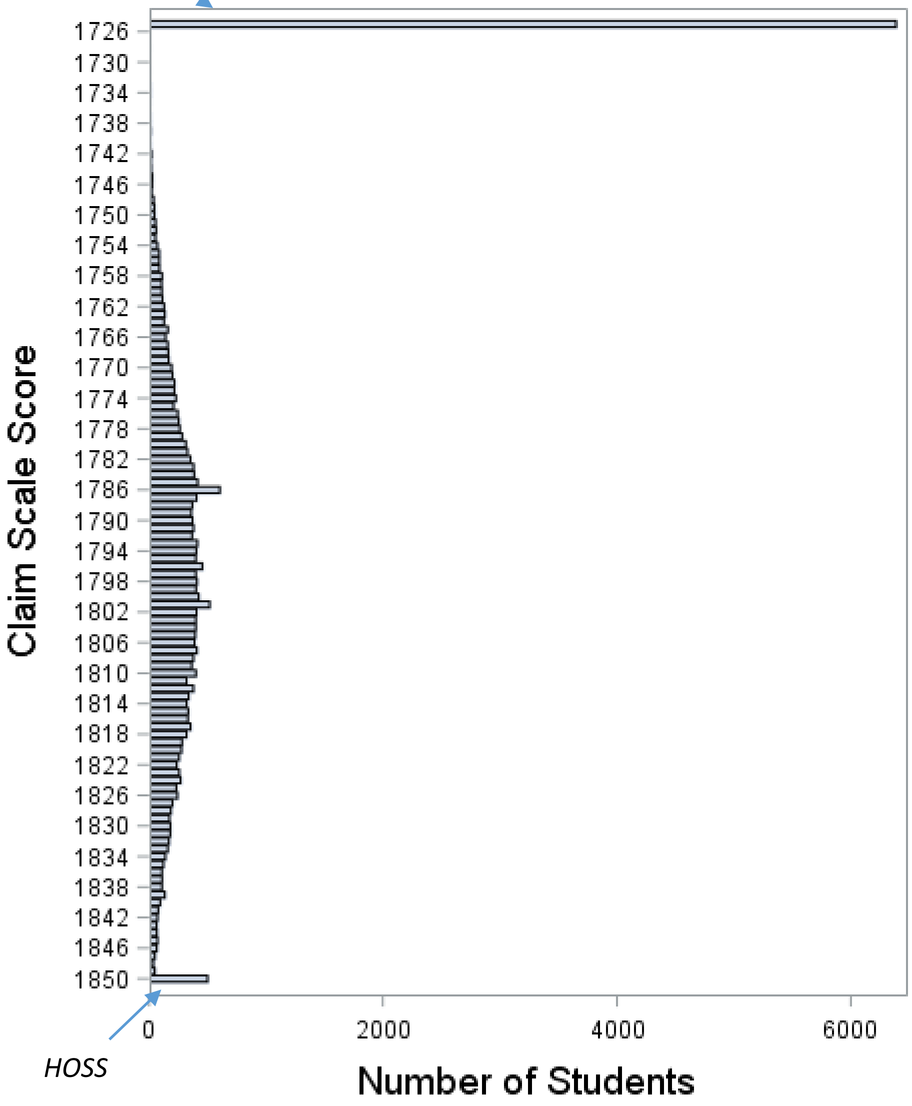
Do any of these seem…*unusual?*

Grade	Total Test	Claim 1 *Concepts & Procedures*	Claim 2/4 *Problem Solving/Modeling & Data Analysis*	Claim 3 *Communicating Reasoning*
3	0.95	0.93	0.65	0.67
4	0.95	0.93	0.70	0.54
5	0.94	0.90	0.06	0.63
6	0.94	0.92	0.59	0.57
7	0.93	0.91	**-0.02**	0.41
8	0.92	0.89	**-0.22**	0.41

Marginal Reliability = (Total Variance – Error Variance) / Total Variance

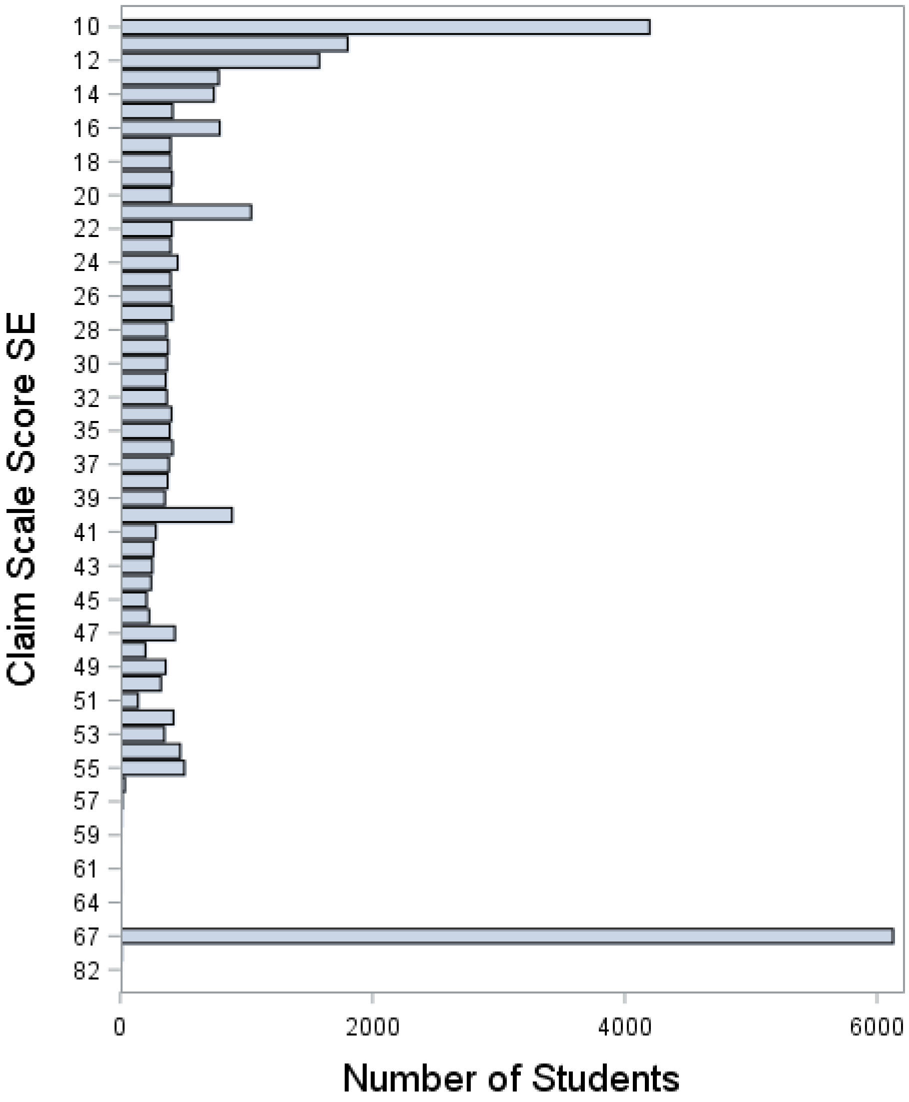
Distribution of Claim Scale Scores and Standard Errors

Grade 8, Claim 2/4



LOSS

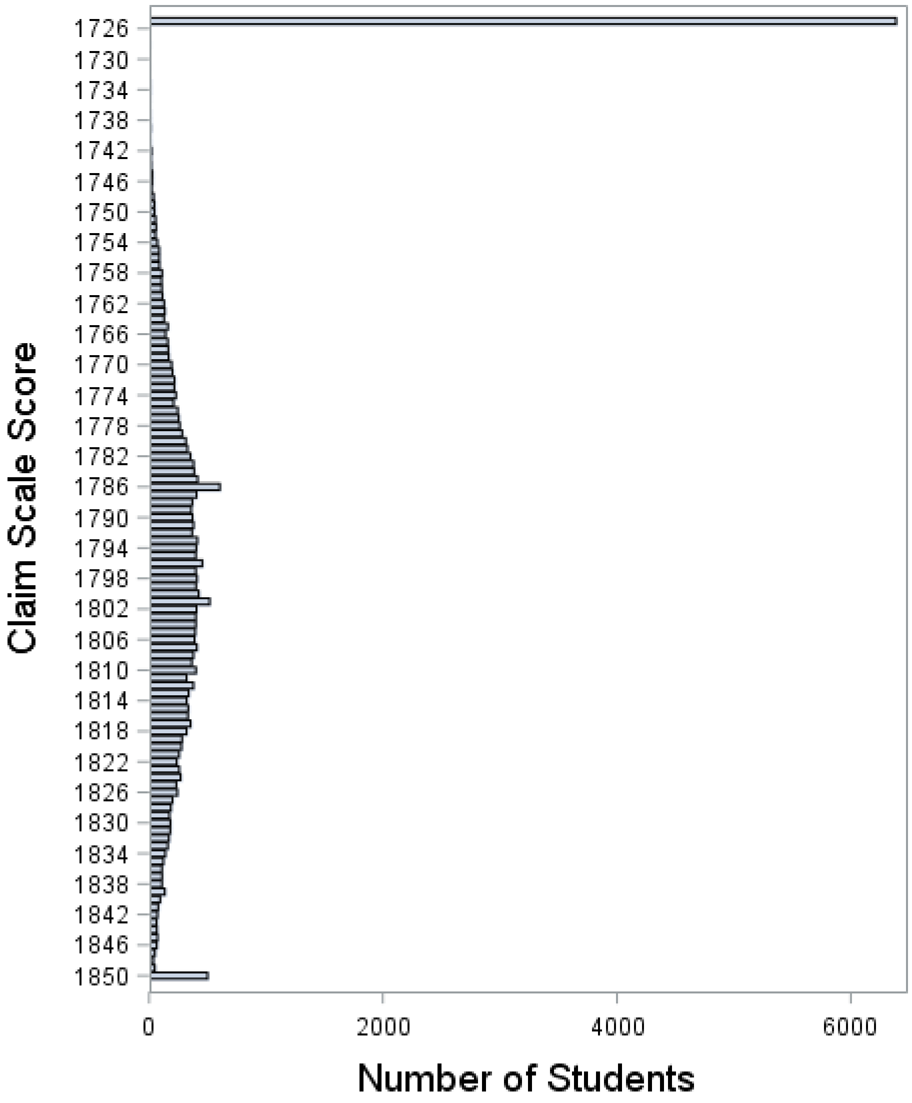
HOSS



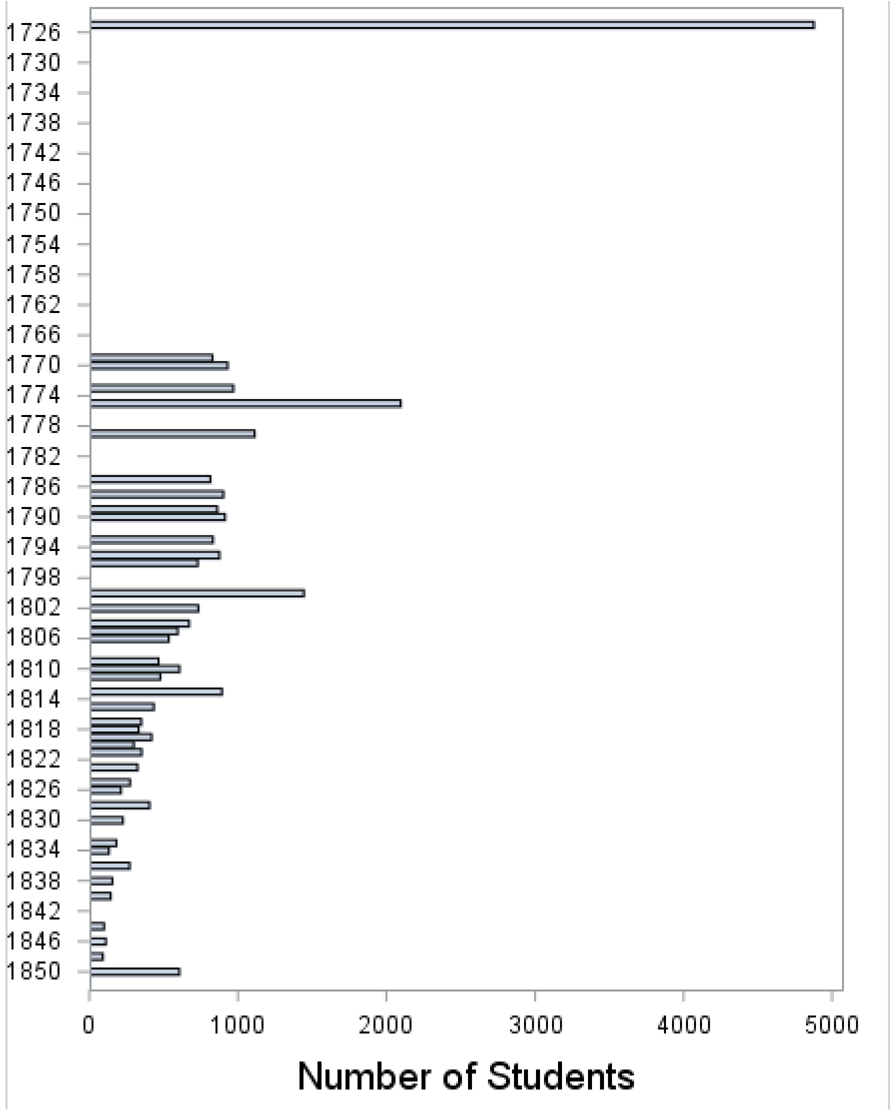
Distribution of Claim Scale Scores and Standard Errors

Grade 8, Claim 2/4

Spring 2016 (CAT)



Spring 2015 (Fixed Form)



Number of students with off-grade items?

SBAC Simulation, based on 1000 Simulees

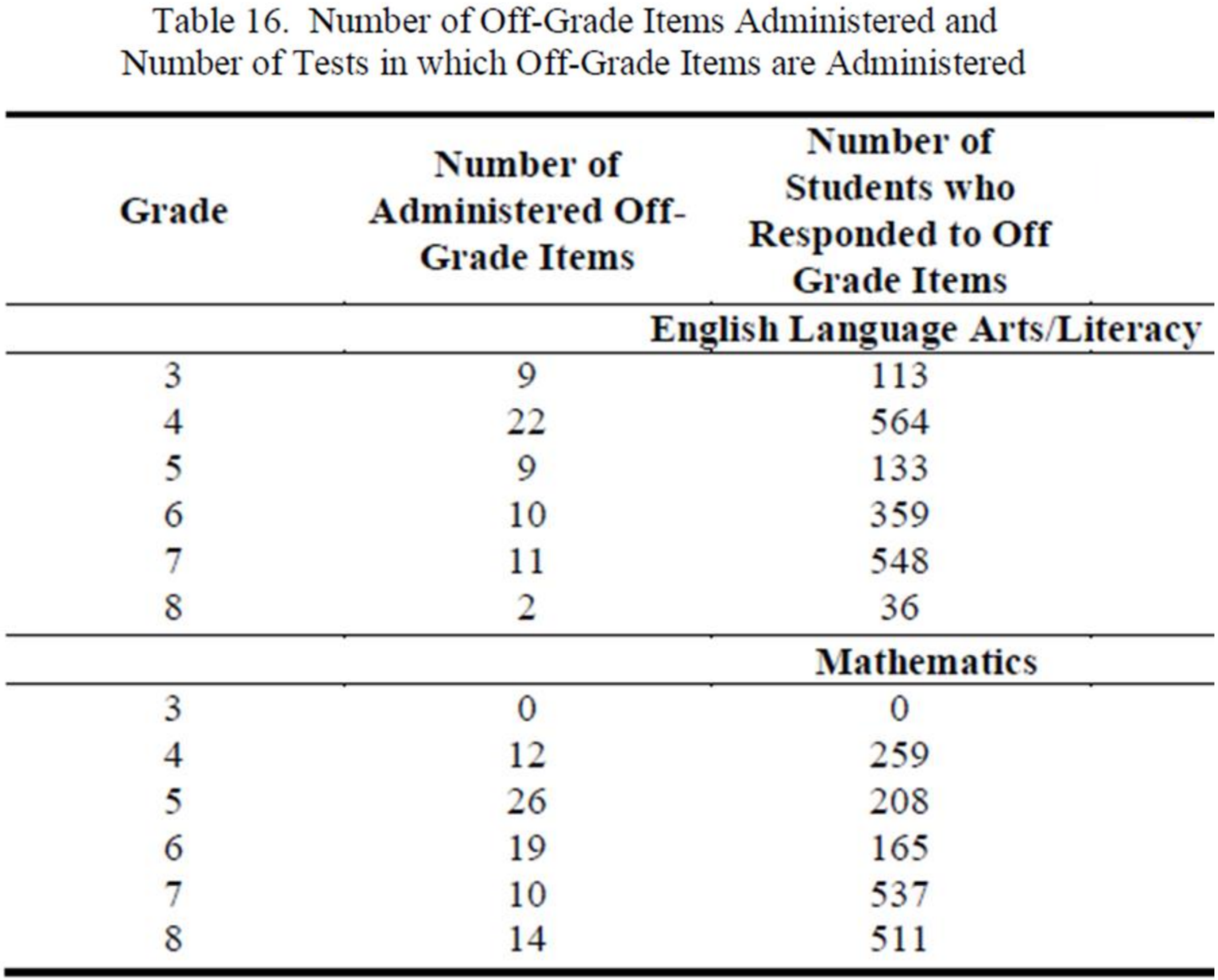


Table 16. Number of Off-Grade Items Administered and Number of Tests in which Off-Grade Items are Administered

Grade	Number of Administered Off-Grade Items	Number of Students who Responded to Off Grade Items
		English Language Arts/Literacy
3	9	113
4	22	564
5	9	133
6	10	359
7	11	548
8	2	36
		Mathematics
3	0	0
4	12	259
5	26	208
6	19	165
7	10	537
8	14	511

Smarter Balanced Summative Assessments - Testing Procedures for Adaptive Item-Selection Algorithm 2014–2015 Test Administrations (AIR)

Number of AVAILABLE Items for the Computer Adaptive Testing?

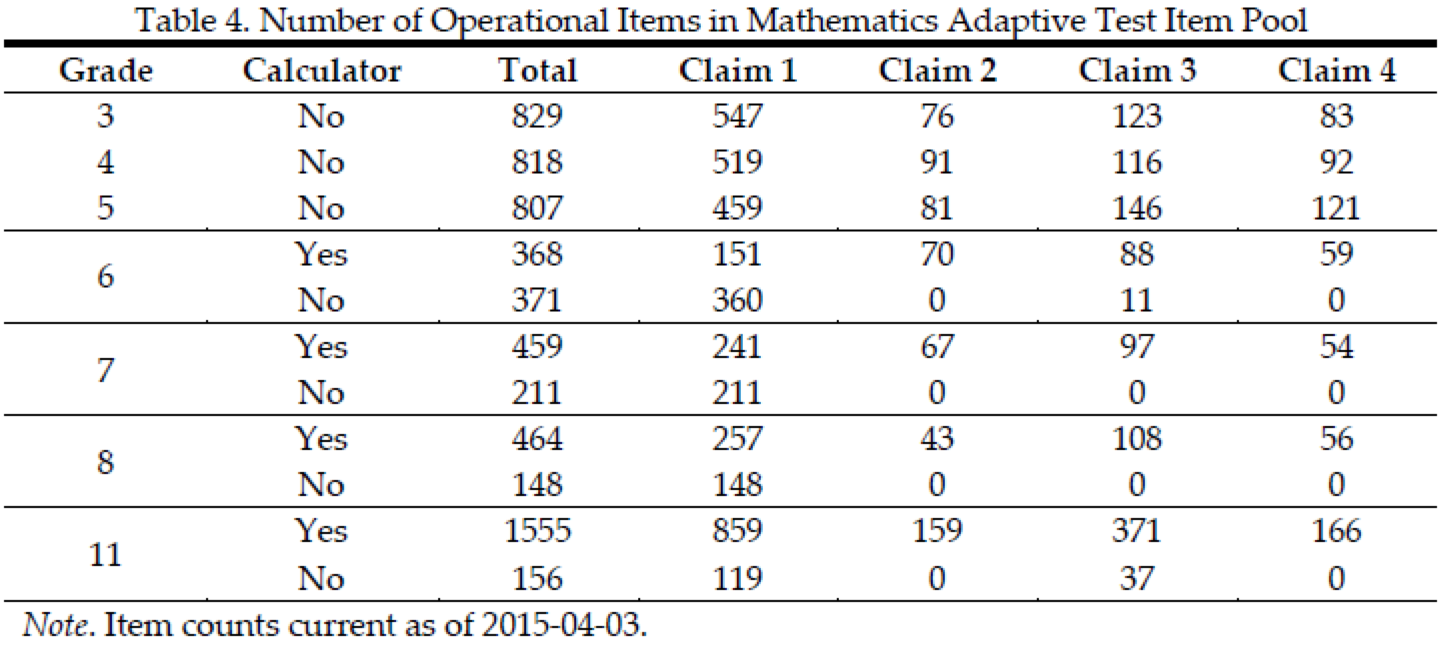


Table 4. Number of Operational Items in Mathematics Adaptive Test Item Pool

Grade	Calculator	Total	Claim 1	Claim 2	Claim 3	Claim 4
3	No	829	547	76	123	83
4	No	818	519	91	116	92
5	No	807	459	81	146	121
6	Yes	368	151	70	88	59
	No	371	360	0	11	0
7	Yes	459	241	67	97	54
	No	211	211	0	0	0
8	Yes	464	257	43	108	56
	No	148	148	0	0	0
11	Yes	1555	859	159	371	166
	No	156	119	0	37	0

Note. Item counts current as of 2015-04-03.

- Number of items available for the M-STEP? This was about the same, as of last year.

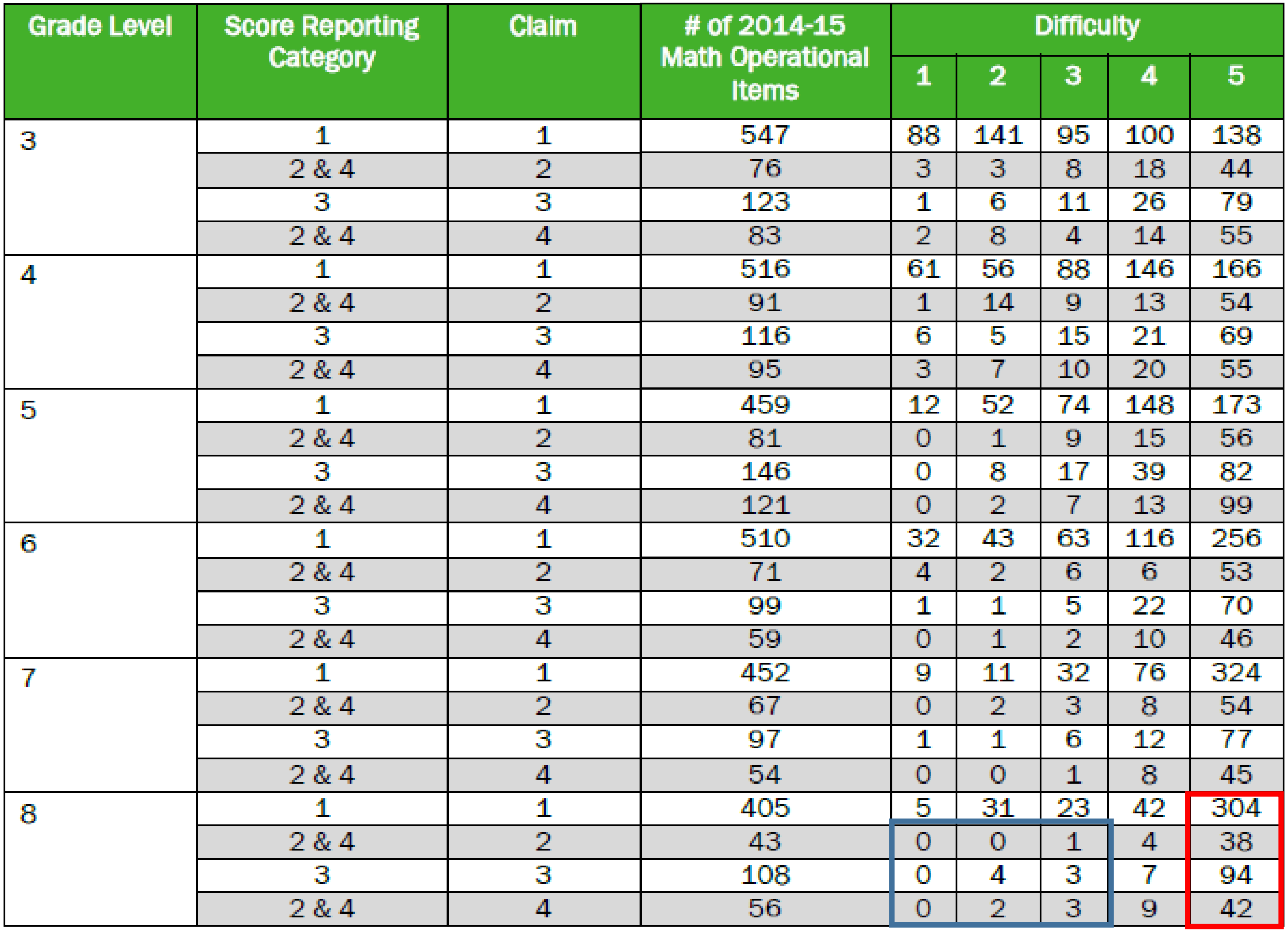
Item Difficulty, Student Ability, and CAT

The content specifications are defined as a combination of item attributes that tests delivered to students should have. There are typically constraints on item content such that they must conform to coverage of a test blueprint. **If there are many content constraints and a limited pool, then it will be difficult to meet the CAT specifications. For a given content target, if the available difficulty/item information targeted at a given level ability is not available, then estimation error cannot be reduced efficiently. A third dimension is that there is usually some need to monitor the exposure of items such that the "best" items are not administered at high rates relative to other ones. Therefore, the quality of the item pools is critical to achieving the benefits that accrue for the CAT over fixed test forms.**

Smarter Balanced used the Reckase "bin" method to evaluate the pool and provide information for new item development. In general, the proportions of items in the pool were written to reflect test blueprints. Although item developers strove to develop items covering the range of examinee achievement levels, the item pool is relatively difficult as compared to the performance that students displayed on the tests.

SBAC 2014-15 Tech Manual, pg 4-33

Math - 2014-15 OPERATIONAL SUMMATIVE POOLS FOR MATHEMATICS



Grade Level	Score Reporting Category	Claim	# of 2014-15 Math Operational Items	Difficulty				
				1	2	3	4	5
3	1	1	547	88	141	95	100	138
	2 & 4	2	76	3	3	8	18	44
	3	3	123	1	6	11	26	79
	2 & 4	4	83	2	8	4	14	55
4	1	1	516	61	56	88	146	166
	2 & 4	2	91	1	14	9	13	54
	3	3	116	6	5	15	21	69
	2 & 4	4	95	3	7	10	20	55
5	1	1	459	12	52	74	148	173
	2 & 4	2	81	0	1	9	15	56
	3	3	146	0	8	17	39	82
	2 & 4	4	121	0	2	7	13	99
6	1	1	510	32	43	63	116	256
	2 & 4	2	71	4	2	6	6	53
	3	3	99	1	1	5	22	70
	2 & 4	4	59	0	1	2	10	46
7	1	1	452	9	11	32	76	324
	2 & 4	2	67	0	2	3	8	54
	3	3	97	1	1	6	12	77
	2 & 4	4	54	0	0	1	8	45
8	1	1	405	5	31	23	42	304
	2 & 4	2	43	0	0	1	4	38
	3	3	108	0	4	3	7	94
	2 & 4	4	56	0	2	3	9	42

478 out of 612 (78%)

"Although there is a wide distribution of item difficulty, pools tend to be difficult in relation to the population and to proficiency cut scores" (Smarter Balanced 2014-15 Technical Report)

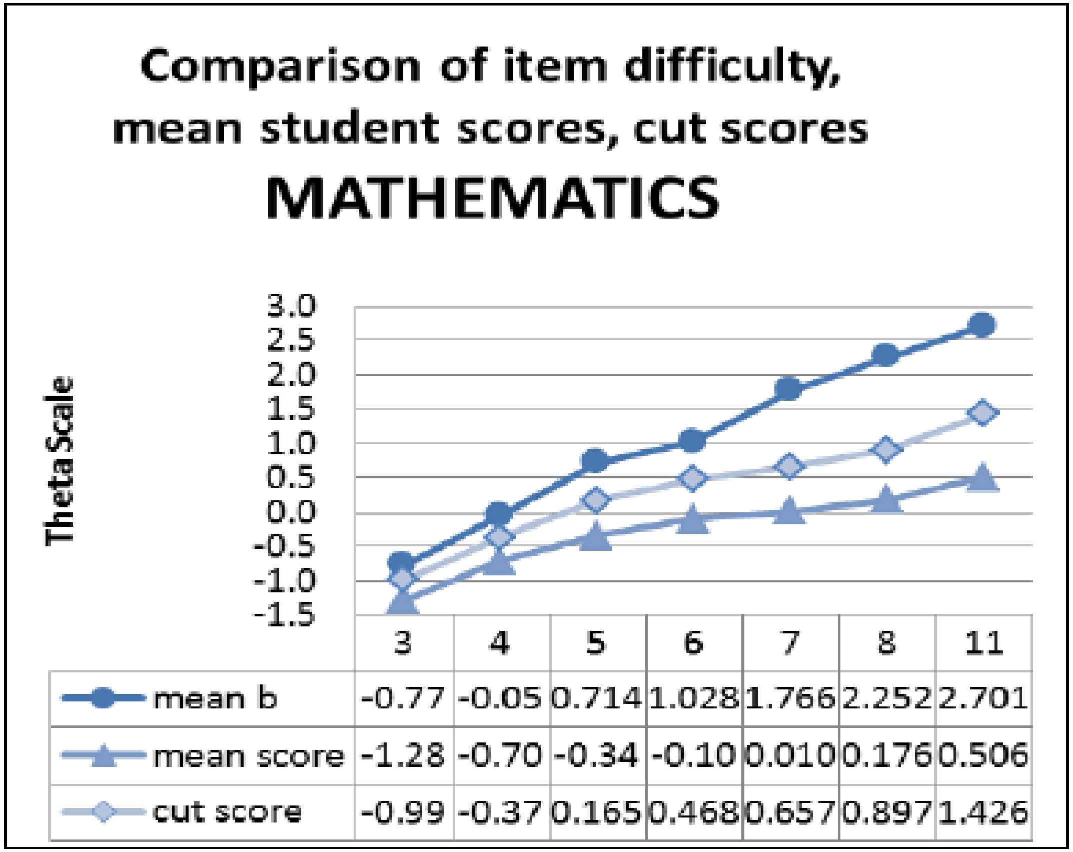
Number of "appropriate" items available, for the bottom 60% of kids taking the M-STEP (this means EVERY non-proficient kid, grades 5-8)

Grade	Claim 3	Claim 2/4
5	25	19
6	7	15
7	8	6
8	7	6

The SB CAT algorithm test does NOT require targets (standards) evenly sampled (or, sampled at all)
nor is information provided on

Item Difficulty, Student Ability, and CAT

In general, the proportions of items in the pool were written to reflect test blueprints. Although item developers strove to develop items covering the range of examinee achievement levels, the item pool is relatively difficult as compared to the performance that students displayed on the tests.



Comparison of item difficulty, mean student scores, cut scores

MATHEMATICS

	3	4	5	6	7	8	11
● mean b	-0.77	-0.05	0.714	1.028	1.766	2.252	2.701
▲ mean score	-1.28	-0.70	-0.34	-0.10	0.010	0.176	0.506
◇ cut score	-0.99	-0.37	0.165	0.468	0.657	0.897	1.426

Interrelationships (correlations) between Claims?

M-STEP Math, Spring 2016 – Between-Claim Correlations

Grade	Claims1 – 2/4	Claims1 -3	Claims 2/4 - 3
3	0.74	0.72	0.70
4	0.79	0.73	0.71
5	0.69	0.75	0.64
6	0.81	0.76	0.69
7	0.72	0.72	0.62
8	0.68	0.73	0.60

OK, what does this mean?

Interrelationships (correlations) between Claims

There are a number of methods for determining if a subscore adds value (to the reporting of the Total Score)…and/or whether reporting subscores can be misleading

Providing Subscale Scores for Diagnostic Information: A Case Study When the Test is Essentially Unidimensional. (2009) *Applied Measurement in Education*, Stone, S et al.

How often do Subscores Have Added Value? Results From Operational and Simulated Data. (2011) *Educational and Psychological Measurement*, Sinharay, S.

Why the Major Field Test in Business Does Not Report Subscores: Reliability and Construct Validity Evidence (2012) *ETS Research Report*, Ling, G.

Guidelines for Interpreting and Reporting Subscores (2016) *Educational Measurement: Issues and Practice*, Feinberg, R. & Jurich, D

Methods for Examining the Psychometric Quality of Subscores: A Review and Application (2015) *Practical Assessment, Research & Evaluation,* Wedman, J & Lyren, P

These methods utilize Subscore Reliability (see previous slide) and between-Subscore Correlations – amounts to a correlation corrected for Measurement Error:

$$\text{Disattenuated Corr} = \text{Corr}_{12} \,/\, \sqrt{\text{REL}_1 * \text{REL}_2}$$

Interrelationships (correlations) between Claims?

M-STEP Math, Spring 2016 – Disattenuated Correlations

Grade	Claims1 – 2/4	Claims1 -3	Claims 2/4 - 3
3	0.96	0.92	1.06
4	0.98	1.02	1.15
5	2.87	1.00	3.19
6	1.10	1.04	1.20
7	.	1.17	.
8	.	1.21	.

Interrelationships (correlations) between Claims?

Based on these findings [where the correlations between observed subscales corrected for attenuation were approximately 1], **subscale scores for tests that are essentially unidimensional provide little if any unique measurement information, and reporting these scores should be reconsidered as they could be misleading and over-interpreted**.

Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. (2010). Applied Measurement in Education, Stone et al.

- NOTE: the SBAC tech manuals discuss extensively the decision to scale the test as unidimensional

  i.e. *A unidimensional scale was conceptualized that combines both CAT and performance tasks. The results from the Pilot Test supported the use of a unidimensional scale, both within a grade and across grades. Since no pervasive evidence of multidimensionality was shown, the decision was to adopt a unidimensional model for scaling and linking.*

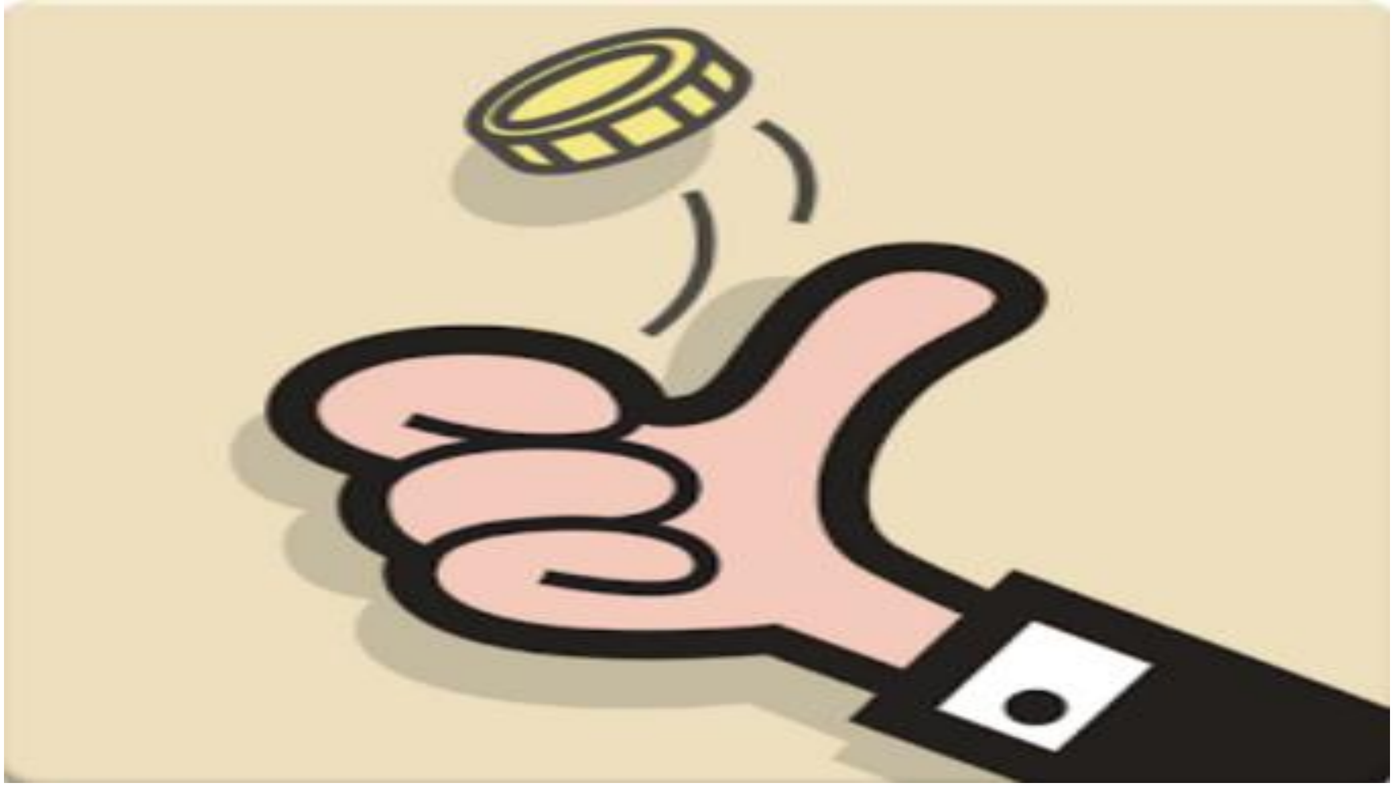
The most important finding is that it **is not easy to have subscores that have added value**. Based on our results, **the subscores have to consist of at least 20 items and have to be sufficiently distinct from each other to have any hope of having added value**. ….**Subscores composed of 10 items were not of any added value even for a realistically extreme (low) disattenuated correlation of 0.70**. The practical implication of this finding is that the test developers have to work hard (to make the subscores long and/or distinct) if they want subscores that have added value.

When Can Subscores Be Expected To Have Added Value? Results From Operational and Simulated Data (2010) Journal of Educational Measurement, 2010, Sinharay, S.

There is no psychometric justification for reporting subscores if the total scores satisfy unidimensionality assumptions, especially the very strict assumptions of unidimensional Item Response Theory. If all items measure the same trait or proficiency, no subset of items provides a measure of anything other than that trait or proficiency **(or random noise).**

Utility Indexes for Decisions about Subscores (2012) Center for Advanced Studies in Measurement and Assessment Research Report, Robert Brennan.

If you want to identify areas of a large-scale, standardized test*, where you should focus attention:



*This includes M-STEP, NWEA, PSAT, SAT, and all others like them

An Argument that Subscores are an Unfairly Maligned Source of Valuable Nuggets(?)

Joseph Martineau
Senior Associate

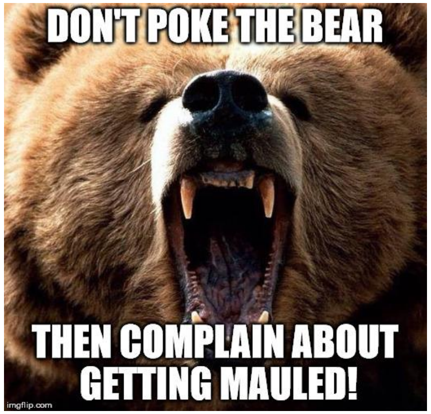
Center for Assessment

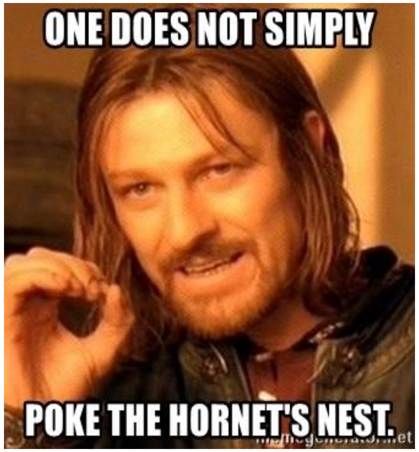
If All the World's a Stage, and You Enjoy Refined Theatrics

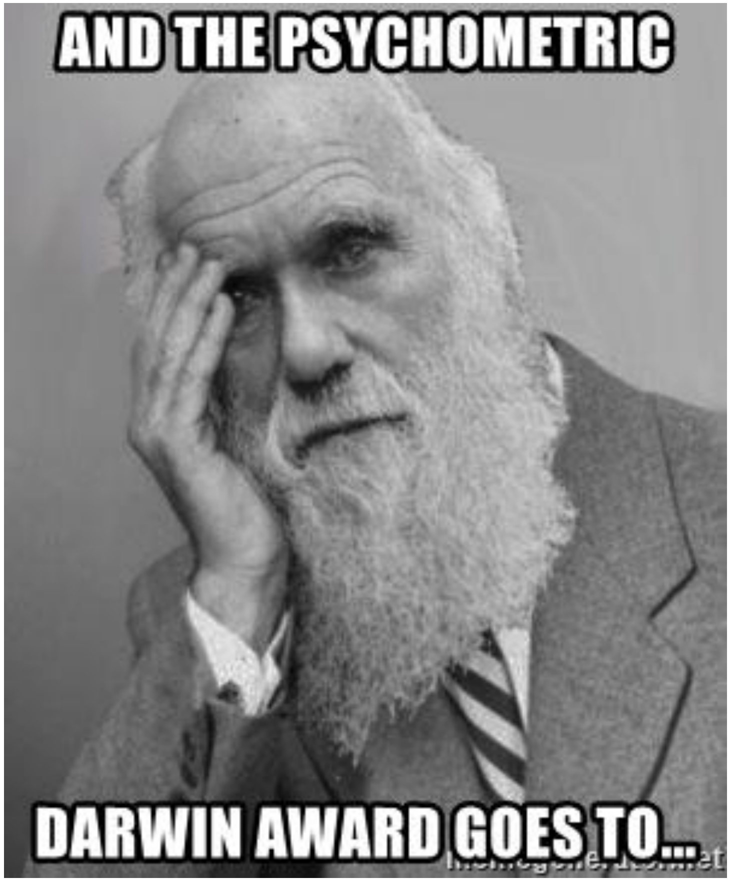
Bring subscores to a TAC meeting to see a well-rehearsed, consistent, passionate, theatrical performance.

- Be prepared for
  - *N* TAC members to give *N* + *M* opinions
  - TAC members to contradict each other
  - TAC members to contradict themselves
  - For the majority of opinions to support only overall scoring

Bringing Subscores to a Technical Advisory Committee Meeting







Common Arguments against Subscores

- Nothing’s ever multidimensional: “g” overwhelms everything

- Multidimensionality is fickle and inconsistent. What you get from one administration is unlikely ever to be repeated

- Multidimensionality is consistent across developmental levels (we can create vertical scales that span K-12 and keep them stable over time)

- Subscores are just less reliable estimates of total scores

Have I Heard All of These Statements in a Single TAC Session?

Everything’s Unidimensional	Nothing’s Unidimensional	Well…
Intelligence (g) explains almost everything. No matter how hard we try, we can’t ever find truly separable dimensions.	The world’s just not that simple.	Things might be multidimensional, but our current measures don’t capture it.
Subscores are useless in practice	Subscores are already useful in practice	Well…
They’re less-reliable estimates of overall scores.	They can identify what we should be looking for (outliers) given similar educational experiences within a subject area.	Maybe we should deliberately construct them rather than privileging item total correlations in test development.
Dimensionality is Consistent Across Developmental Levels (e.g., grades)	Dimensionality Depends on Developmental Level	Well…
We’ve been developing K-12 vertical scales successfully for decades. Subject matter theorists do a good job of defining a domain, but not dimensionality. We just can’t find it.	Standards purposely include multiple dimensions, with representation differing substantially over grades. Psychometricians dismiss subject matter theorists too quickly.	It could be that subject matter experts tend to overestimate distinctness of dimensions and psychometricians tend to underestimate it. Can’t we meet in the middle and get along?
Dimensionality Increases with Development	Dimensionality Decreases with Development	Well…
Literally. As students progress through a subject matter, it fractures into related, but clearly distinct disciplines.	Literally. In early childhood, letter recognition, oral language, concepts of print, and phonemic awareness are actually different skills. As students progress, they just become “reading.”	Whether it increases or decreases probably depends on the situation.

Why Yes, I Have

Everything’s Unidimensional	Nothing’s Unidimensional	Well…
Intelligence (g) explains almost everything. No matter how hard we try, we can’t ever find truly separable dimensions.	The world’s just not that simple.	Things might be multidimensional, but our current measures don’t capture it.
Subscores are useless in practice	Subscores are already useful in practice	Well…
They’re less-reliable estimates of overall scores.	They can identify what we should be looking for (outliers) given similar educational experiences within a subject area.	Maybe we should deliberately construct them rather than privileging item total correlations in test development.
Dimensionality is Consistent Across Developmental Levels (e.g., grades)	Dimensionality Depends on Developmental Level	Well…
We’ve been developing K-12 vertical scales successfully for decades. Subject matter theorists do a good job of defining a domain, but not dimensionality. We just can’t find it.	Standards purposely include multiple dimensions, with representation differing substantially over grades. Psychometricians dismiss subject matter theorists too quickly.	It could be that subject matter experts tend to overestimate distinctness of dimensions and psychometricians tend to underestimate it. Can’t we meet in the middle and get along?
Dimensionality Increases with Development	Dimensionality Decreases with Development	Well…
Literally. As students progress through a subject matter, it fractures into related, but clearly distinct disciplines.	Literally. In early childhood, letter recognition, oral language, concepts of print, and phonemic awareness are actually different skills. As students progress, they just become “reading.”	Whether it increases or decreases probably depends on the situation.

And (distressingly) Sometimes from Myself

Everything's Unidimensional	Nothing's Unidimensional	Well…
Intelligence (g) explains almost everything. No matter how hard we try, we can't find truly separable dimensions.	The world's just not that simple.	Things might be multidimensional, but our current measures don't capture it.
Subscores are useless in practice	Subscores are already useful in practice	Well…
They're less-reliable estimates of overall scores.	They can identify what we should be looking for (outliers) given similar educational experiences within a subject area.	Maybe we should deliberately construct them rather than privileging item total correlations in test development.
Dimensionality is Consistent Across Developmental Levels (e.g., grades)	Dimensionality Depends on Developmental Level	Well…
We've been developing K-12 vertical scales successfully for decades. Subject matter theorists do a good job of defining a domain, but not dimensionality. We just can't find it.	Standards purposely include multiple dimensions, with representation differing substantially over grades. Psychometricians dismiss subject matter theorists too quickly.	It could be that subject matter experts tend to overestimate distinctness of dimensions and psychometricians tend to underestimate it. Can't we meet in the middle and get along?
Dimensionality Increases with Development	Dimensionality Decreases with Development	Well…
Literally. As students progress through a subject matter, it fractures into related, but clearly distinct disciplines.	Literally. In early childhood, letter recognition, oral language, concepts of print, and phonemic awareness are actually different skills. As students progress, they just become "reading."	Whether it increases or decreases probably depends on the situation.

Two Options for Assessing Dimensionality

Option 1

- Start with psychometric/statistical definition of dimensions, then (in the unlikely event that popularly used methods indicate multidimensionality) incorporate SME expertise

- Option 2
  - Start with SME expertise to define dimensions, then perform confirmatory analysis and together with SMEs revise the definition of dimensions to incorporate both substantive and statistical information

- Experience tells me:
  - Option 2 is better
  - My conclusion is that we (psychometricians/statisticians) routinely undervalue subject matter experts’ (SMEs) understanding of what constitutes a separate dimension within a content area

Why Do I Distrust Option 1?

- Option 1: Quantitative Analysis First, SME Engagement Only if Analytical results indicate multidimensionality
  - The most popularly used methods for assessing dimensionality almost always indicates unidimensionality
  - So we never involve the SMEs; we have substituted quantitative methodology for substantive expertise
  - In my decidedly minority opinion, the psychometric/statistical community too easily dismisses substantive experts in the arena of the structure of academic knowledge and skills.

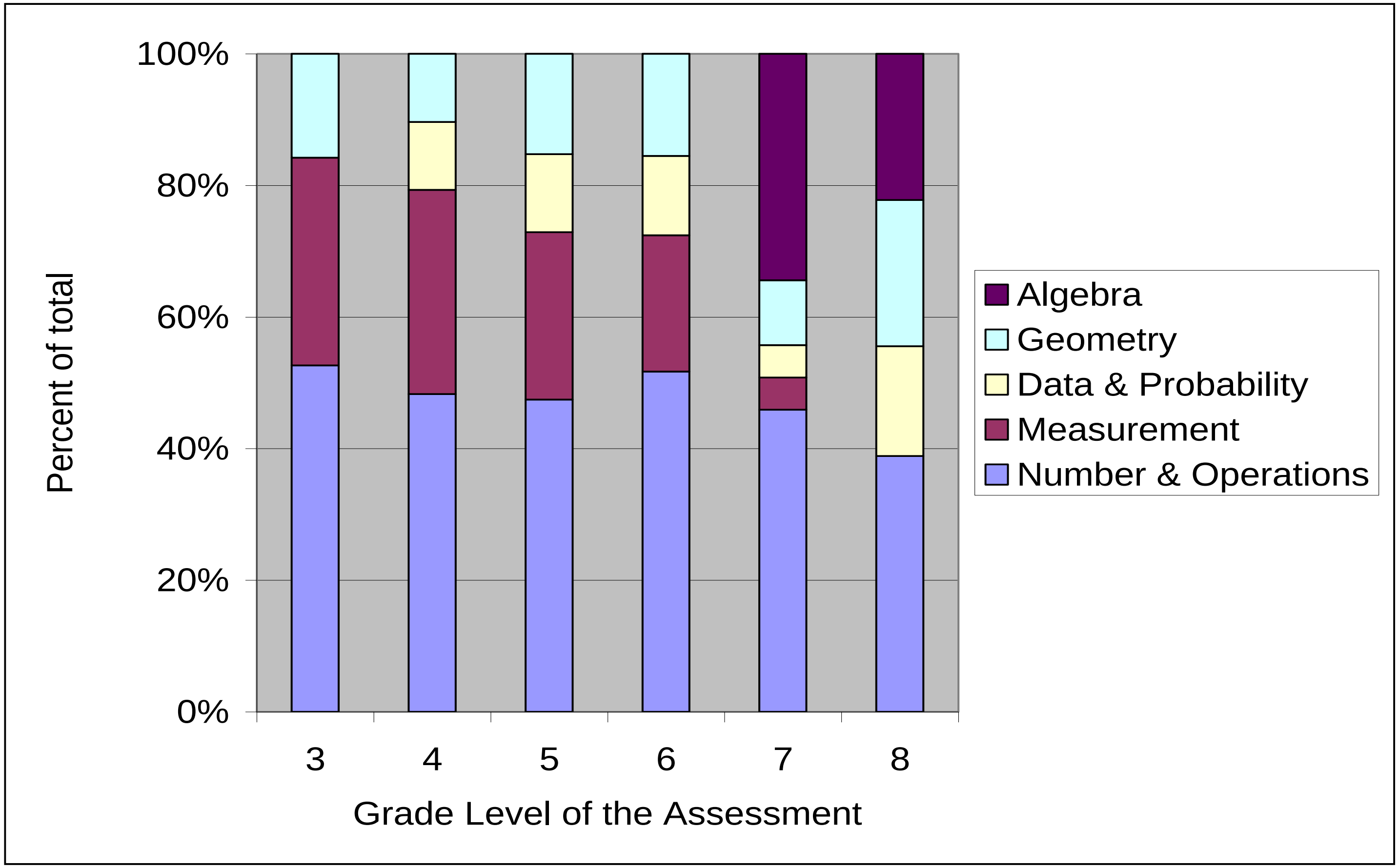
Why Do I Better Trust Option 2?

- Option 2: Start with SME-defined structure, then jointly incorporate analytical results
  - When starting with the SME-defined structure, quantitative analysis is likely to result in some modification
  - The most popularly used methods for assessing dimensionality still almost always indicates unidimensionality
  - However, using more recent methods for assessing dimensionality, we do capture greater degrees of multidimensionality

  - One of the most difficult and controversial parts of dimensionality is using exploratory statistics to define what the subscores should be because the results depend too much on decisions by the analyst
  - Option 2 does not include exploratory approaches

EXAMPLES(?) OF USEFUL SUBSCORES USING OPTION 2

A Blast from the Past (MEAP, circa 2005)



Summarized from Martineau, et al (2007).

Using Purely SME Definitions of Dimensions/Subscores

Grade 3-8 MEAP Math – **Strand Reliabilities**

Strand	Algebra	Data & Probability	Geometry	Measurement	Number & Operations
Algebra	**0.76**				
Data & Probability		**0.52**			
Geometry			**0.59**		
Measurement				**0.62**	
Number & Operations					**0.86**

Using Purely SME Definitions of Dimensions/Subscores

Grade 3-8 MEAP Math – **Raw Correlations**

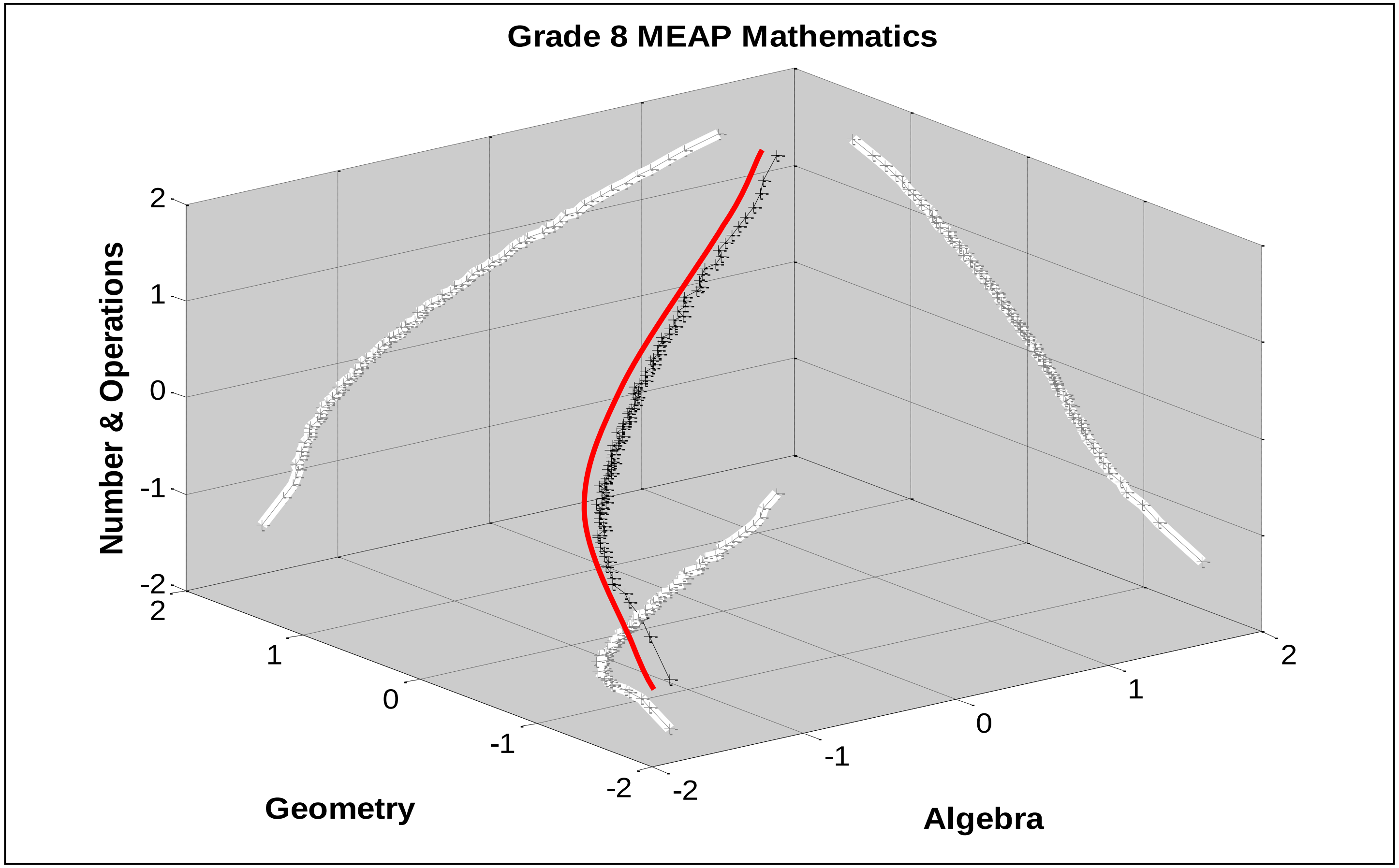
Strand	Algebra	Data & Probability	Geometry	Measurement	Number & Operations
Algebra	0.76				
Data & Probability	**0.55**	0.52			
Geometry	**0.55**	**0.47**	0.59		
Measurement	**0.28**	**0.47**	**0.54**	0.62	
Number & Operations	**0.77**	**0.57**	**0.61**	**0.69**	0.86

Using Purely SME Definitions of Dimensions/Subscores

Grade 3-8 MEAP Math – **Correlations Disattenuated for Unreliability**

Strand	Algebra	Data & Probability	Geometry	Measurement	Number & Operations
Algebra	0.76	**0.87**	**0.82**	**0.41**	**0.95**
Data & Probability	0.55	0.52	**0.85**	**0.83**	**0.85**
Geometry	0.55	0.47	0.59	**0.89**	**0.86**
Measurement	0.28	0.47	0.54	0.62	**0.94**
Number & Operations	0.77	0.57	0.61	0.69	0.86

Using Purely SME Definitions of Dimensions/Subscores

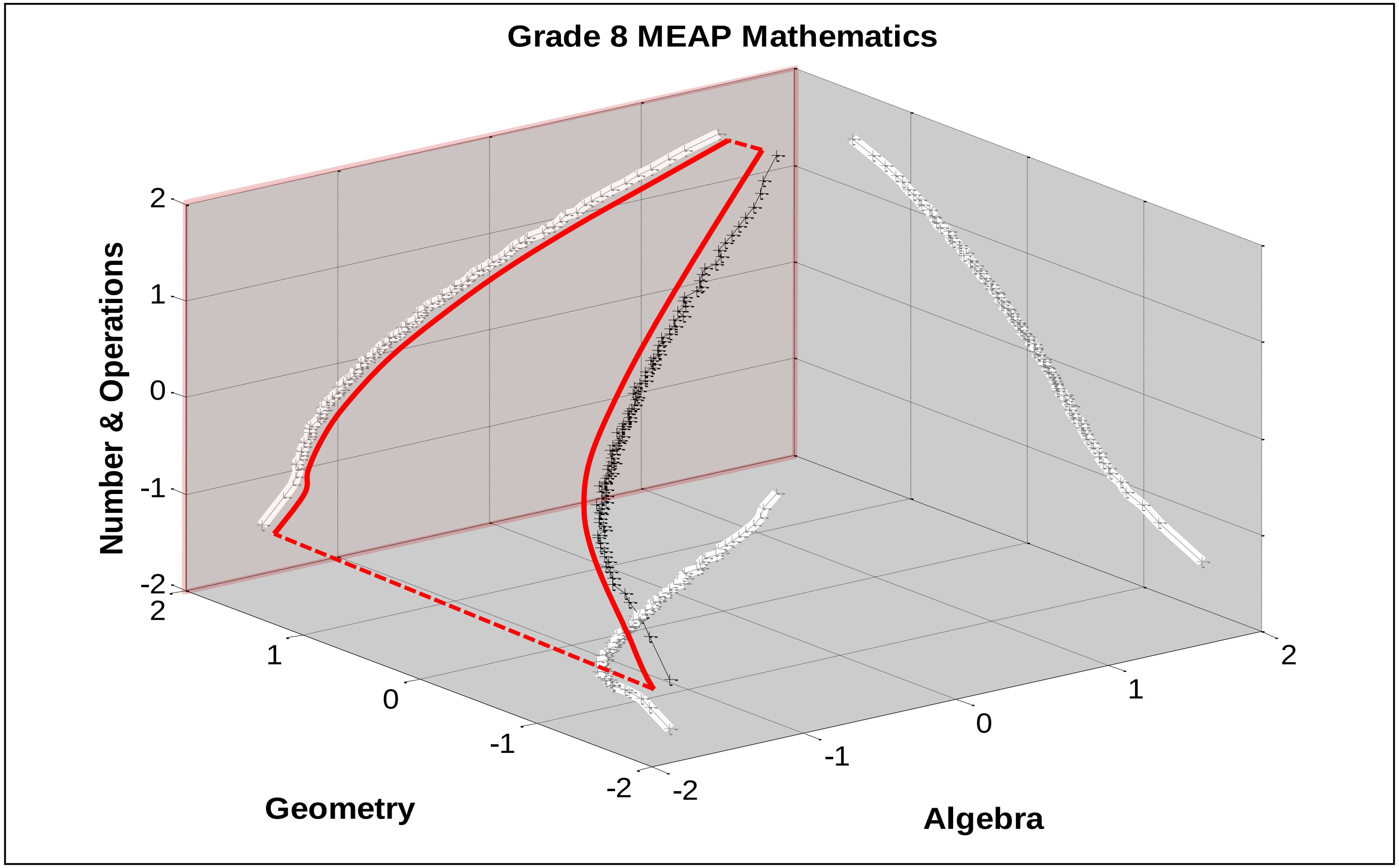


100 equally sized groups of all Michigan 8th graders.

Plotted each groups mean scores on the *Number & Operations*, *Geometry*, and *Algebra* score scales.

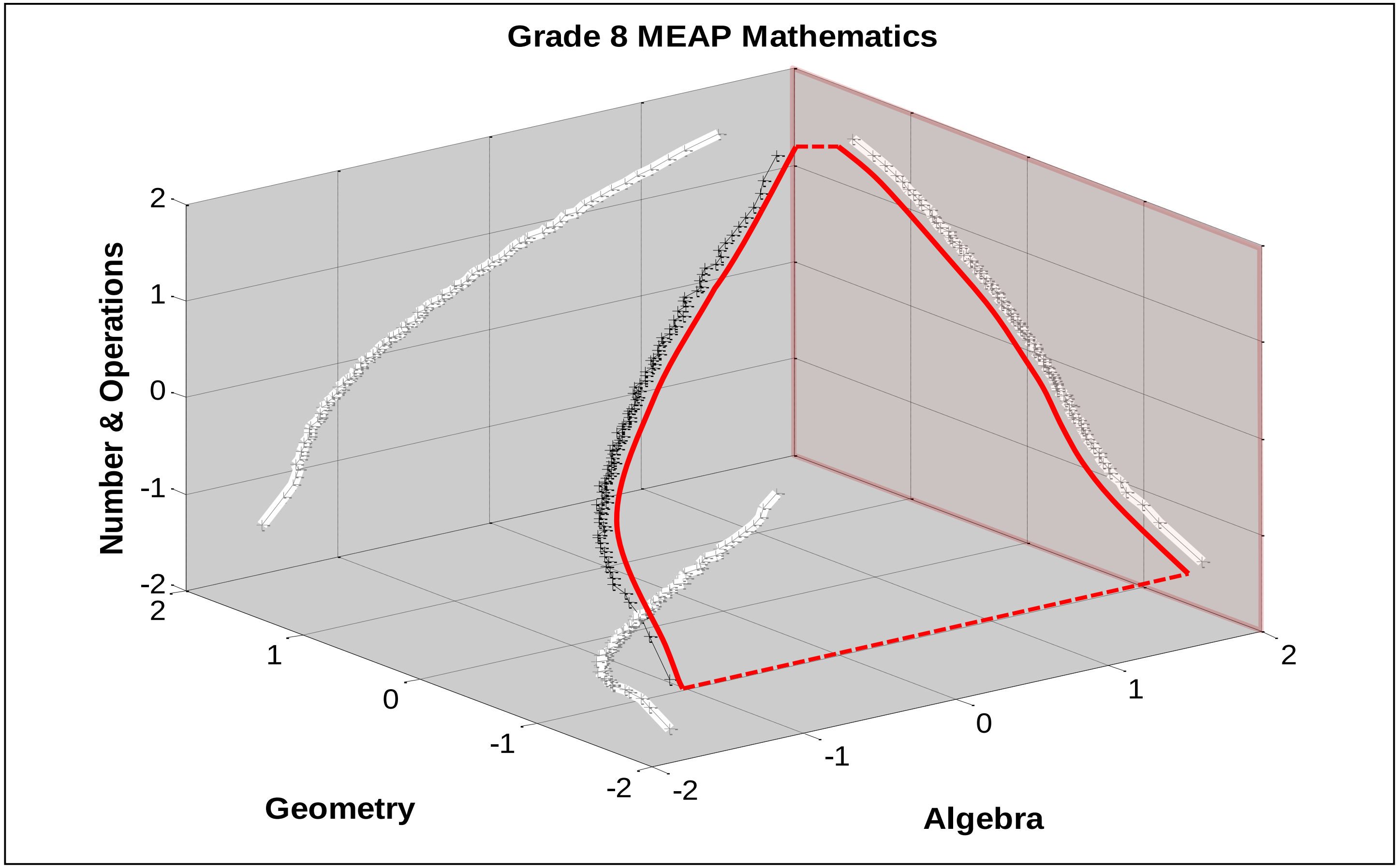
A non-linear typical trajectory through 3 dimensions of mathematics achievement.

Using Purely SME Definitions of Dimensions/Subscores



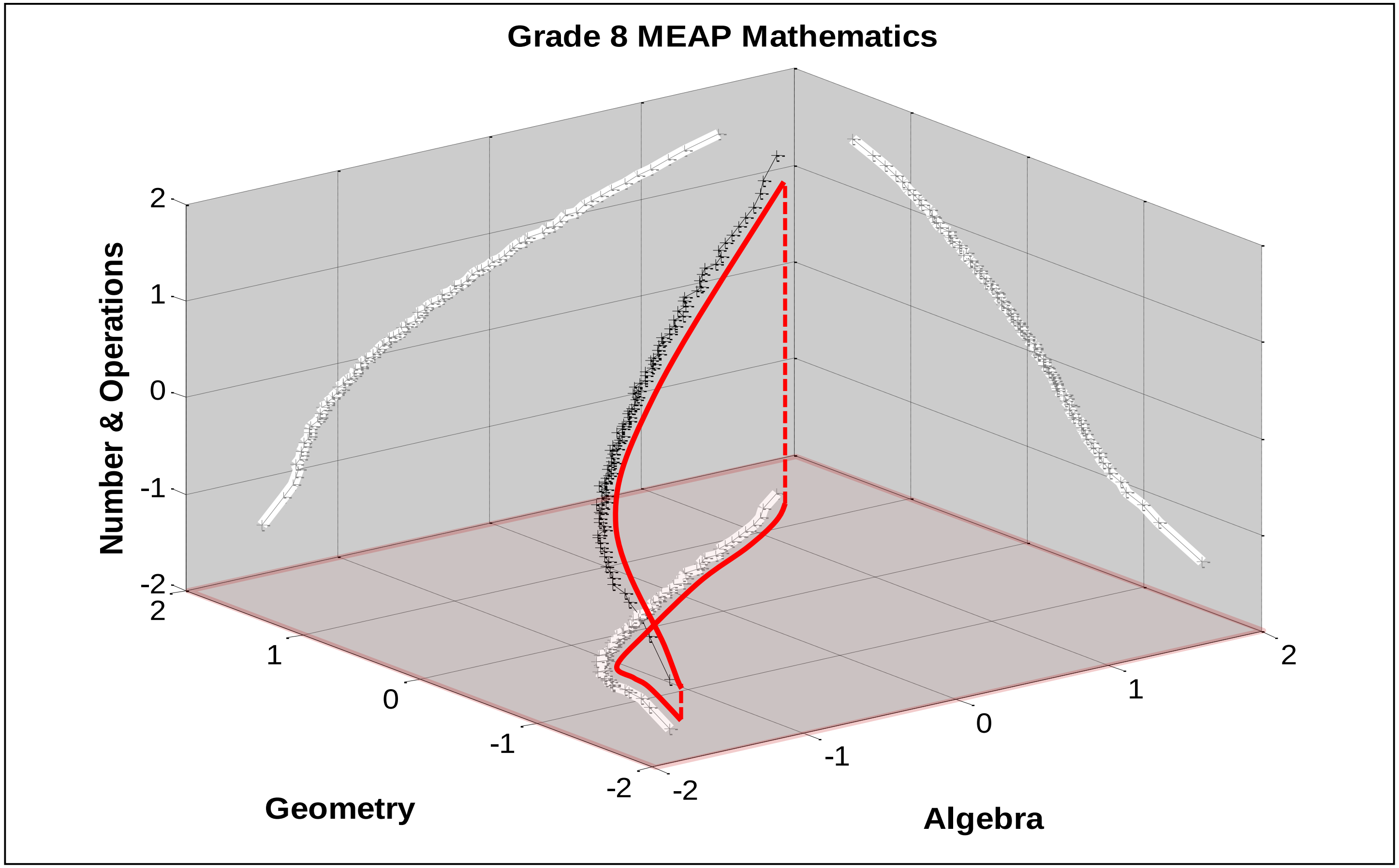
Projection of the 3D curve onto a two-dimensional plane (Algebra vs. Number & Operations)

Using Purely SME Definitions of Dimensions/Subscores



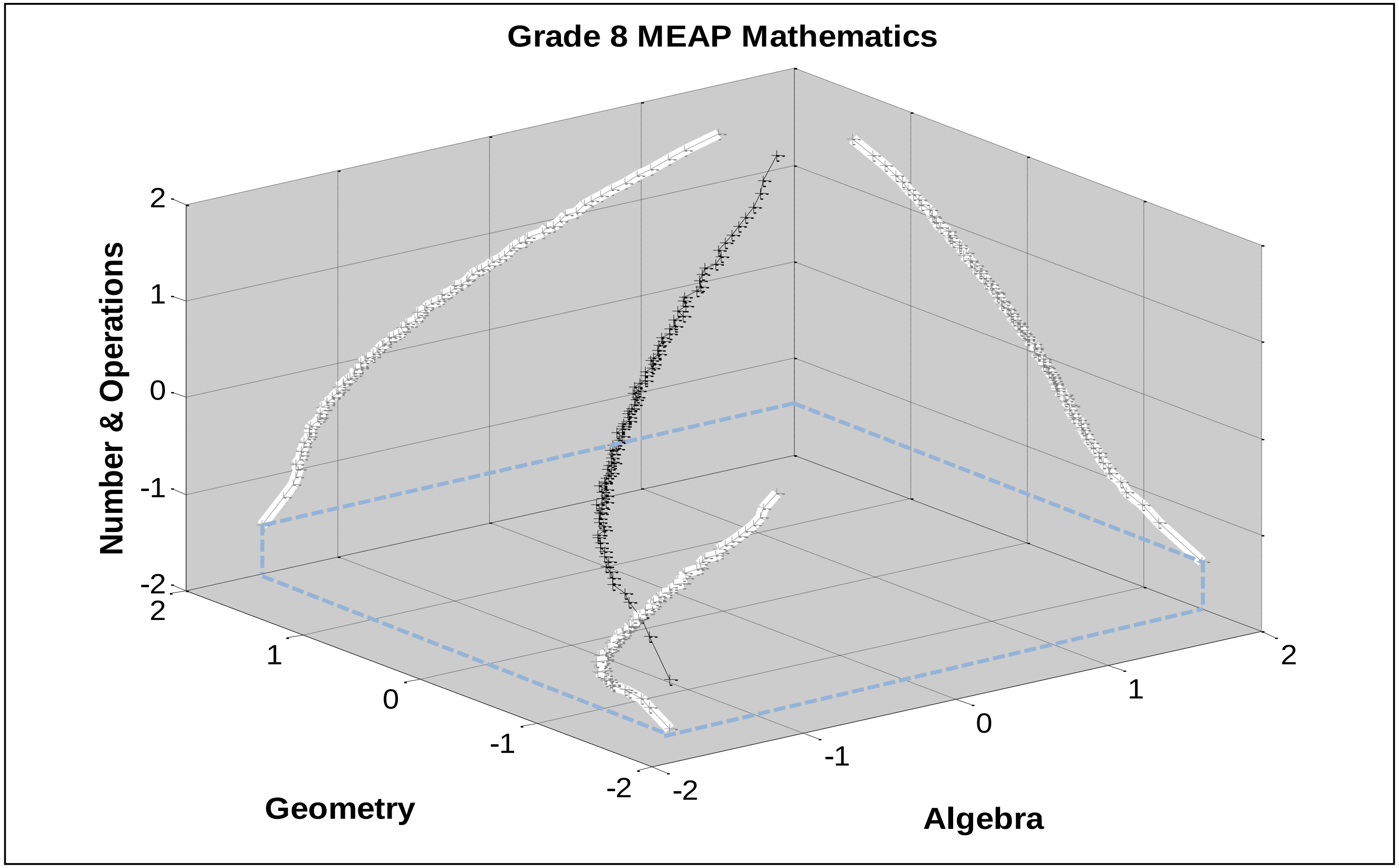
Projection of the 3D curve onto a two-dimensional plane (Geometry vs. Number & Operations)

Using Purely SME Definitions of Dimensions/Subscores



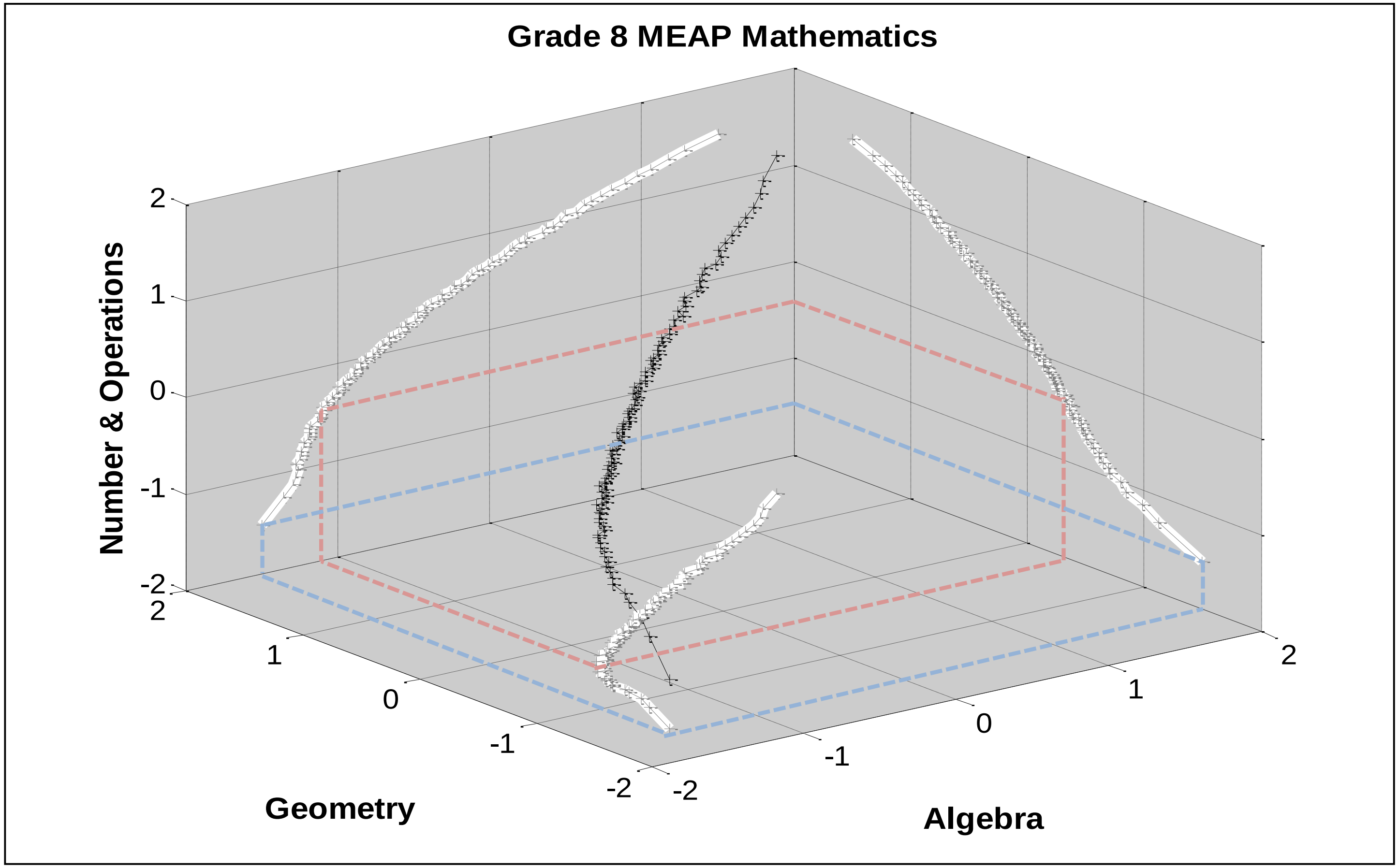
Projection of the 3D curve onto a two-dimensional plane (Algebra vs. Geometry)

Using Purely SME Definitions of Dimensions/Subscores



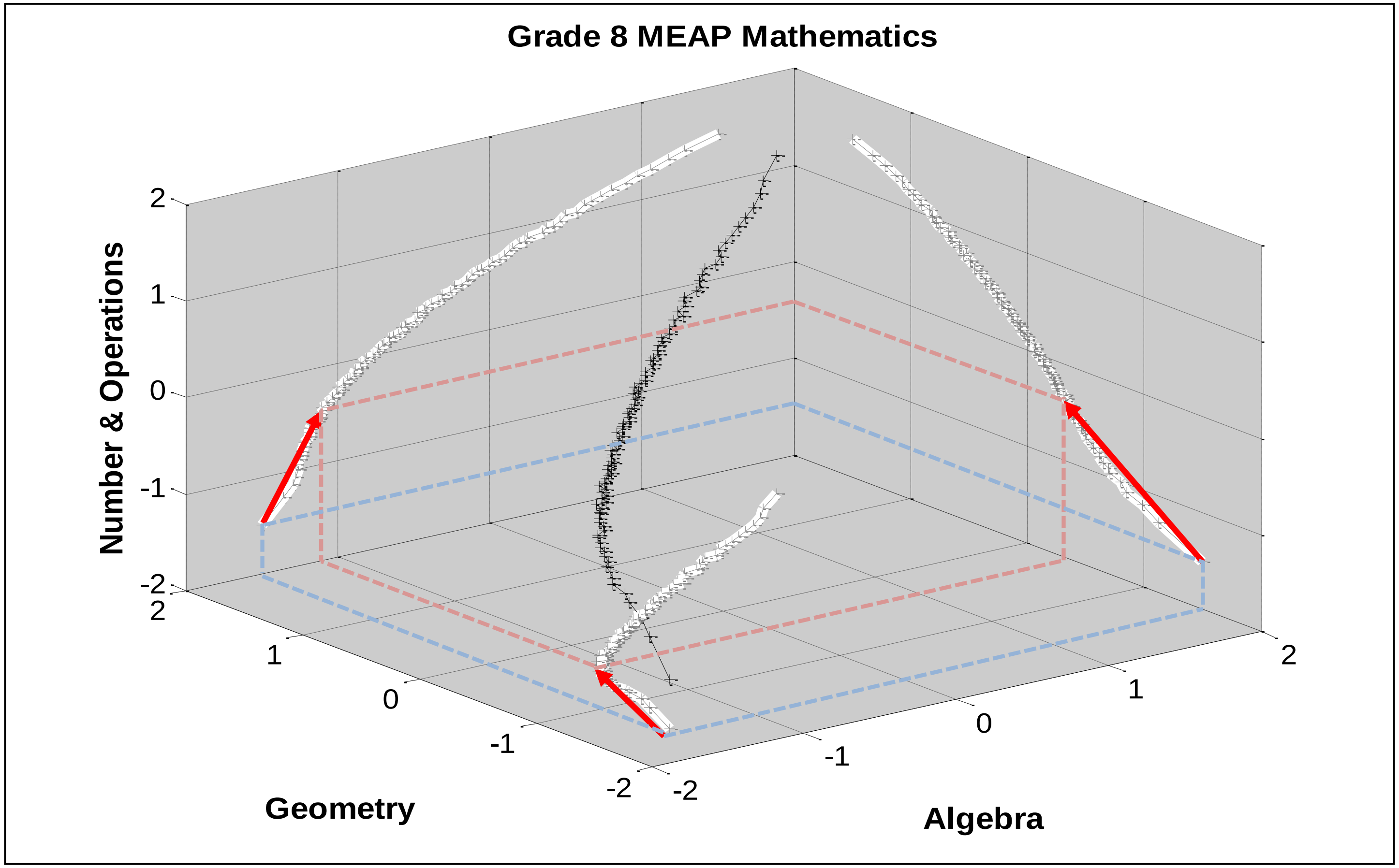
In going from the bottom of the score scales…

Using Purely SME Definitions of Dimensions/Subscores



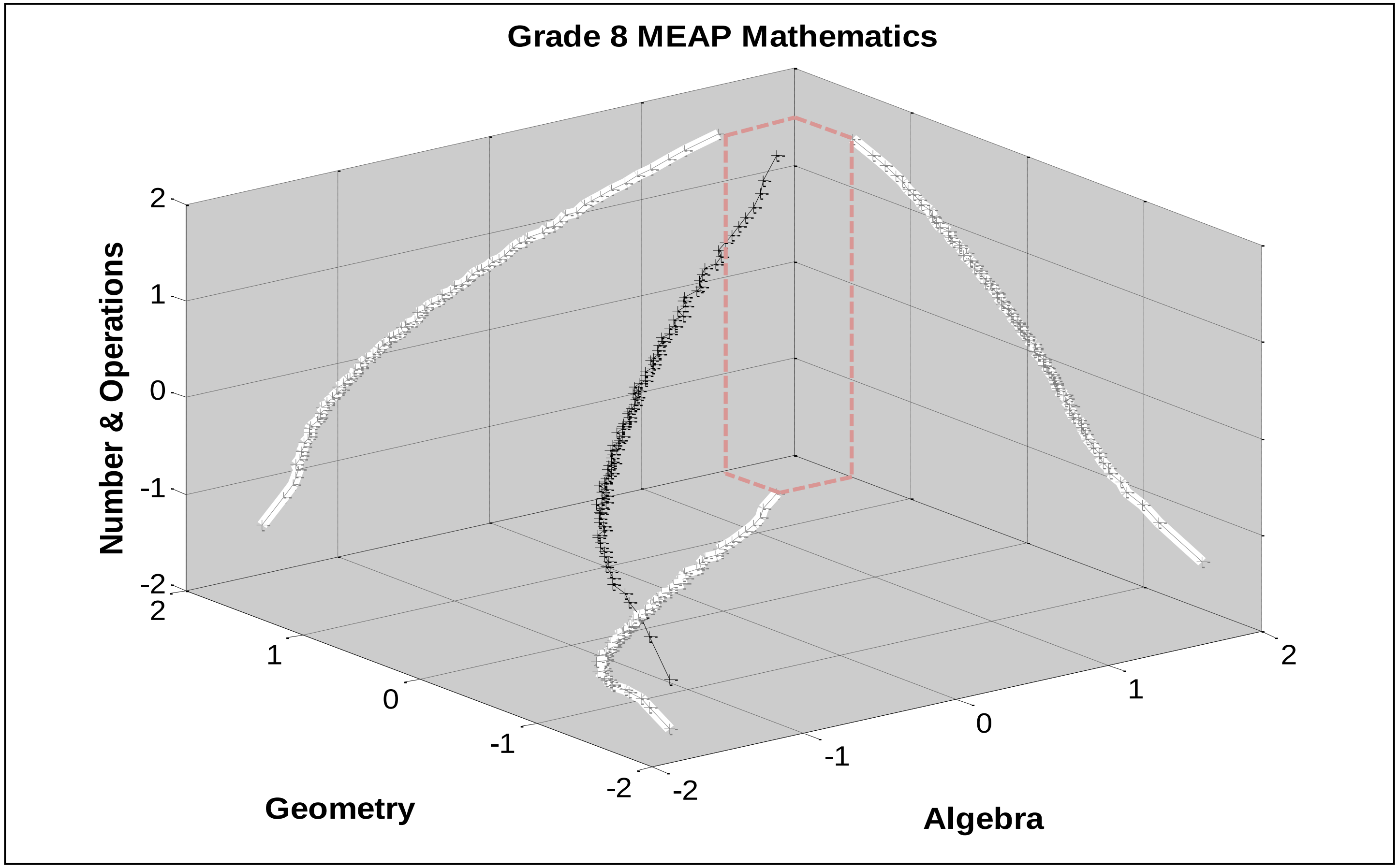
…to part way up the scale

Using Purely SME Definitions of Dimensions/Subscores



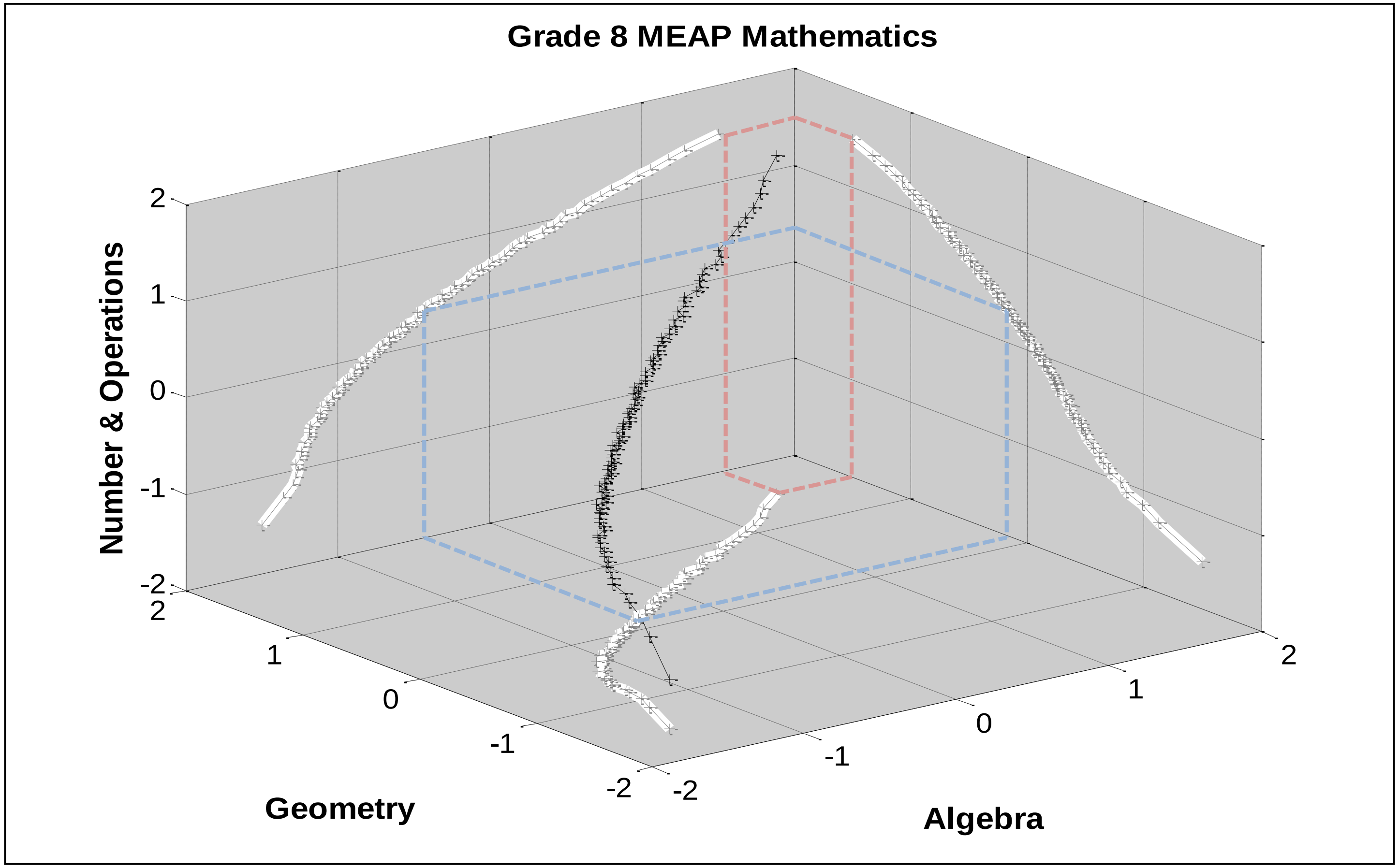
…differences on the low end of the scale in achievement on grade-8 primarily represents changes in student understanding of *Number & Operations* and *Geometry.*

Using Purely SME Definitions of Dimensions/Subscores



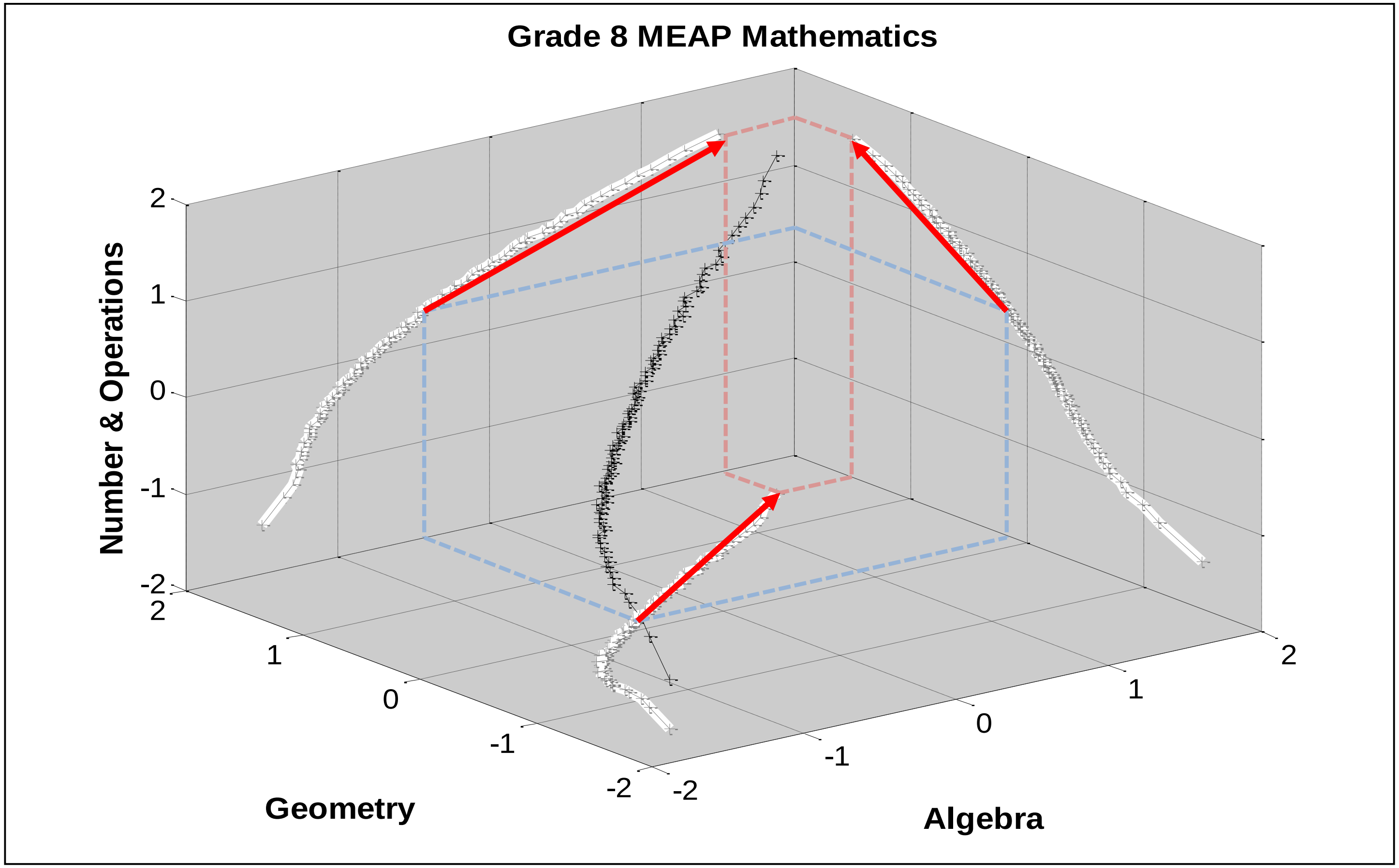
In going from the top of the score scales…

Using Purely SME Definitions of Dimensions/Subscores



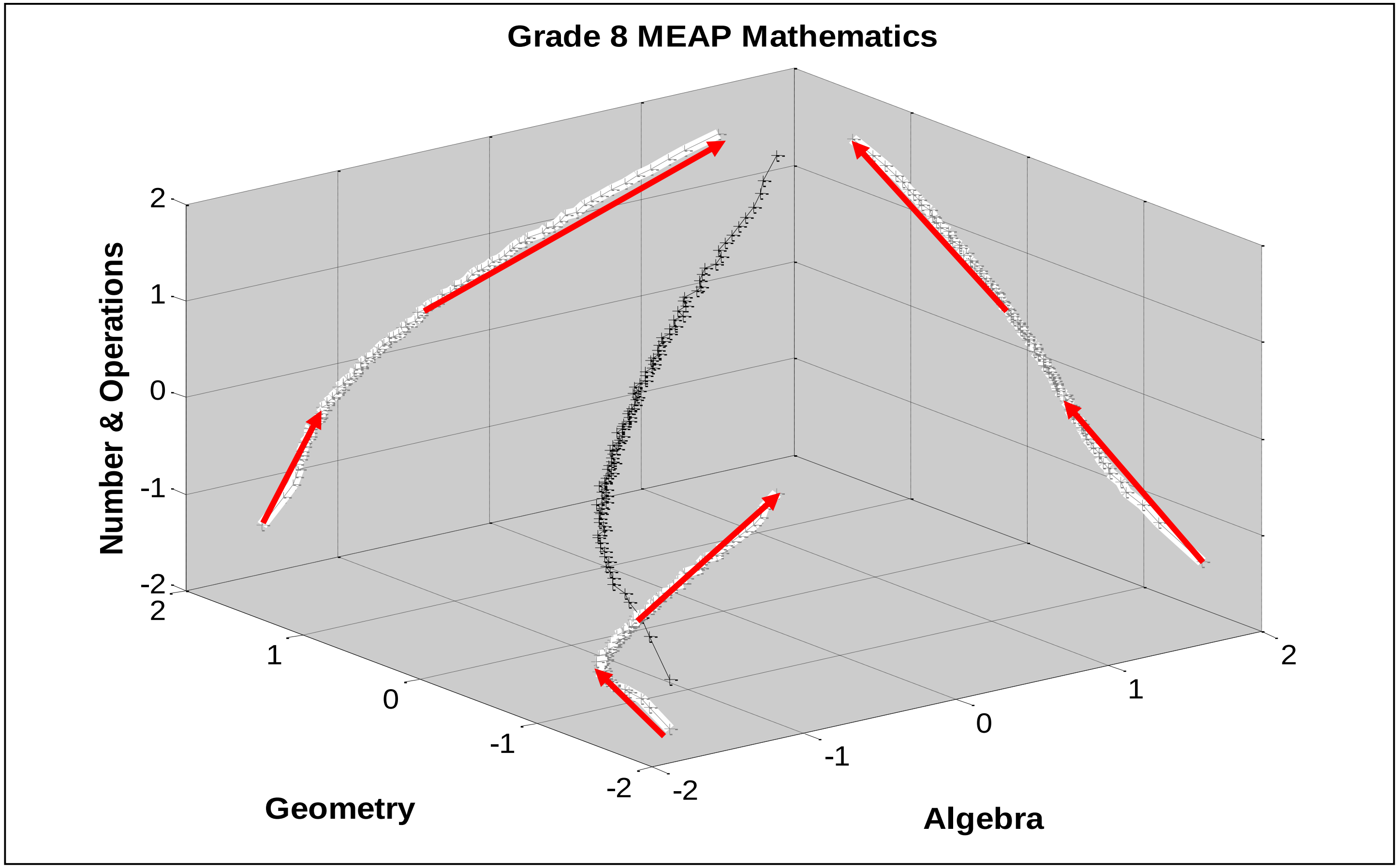
…to about 2/3 of the way down the scale…

Using Purely SME Definitions of Dimensions/Subscores



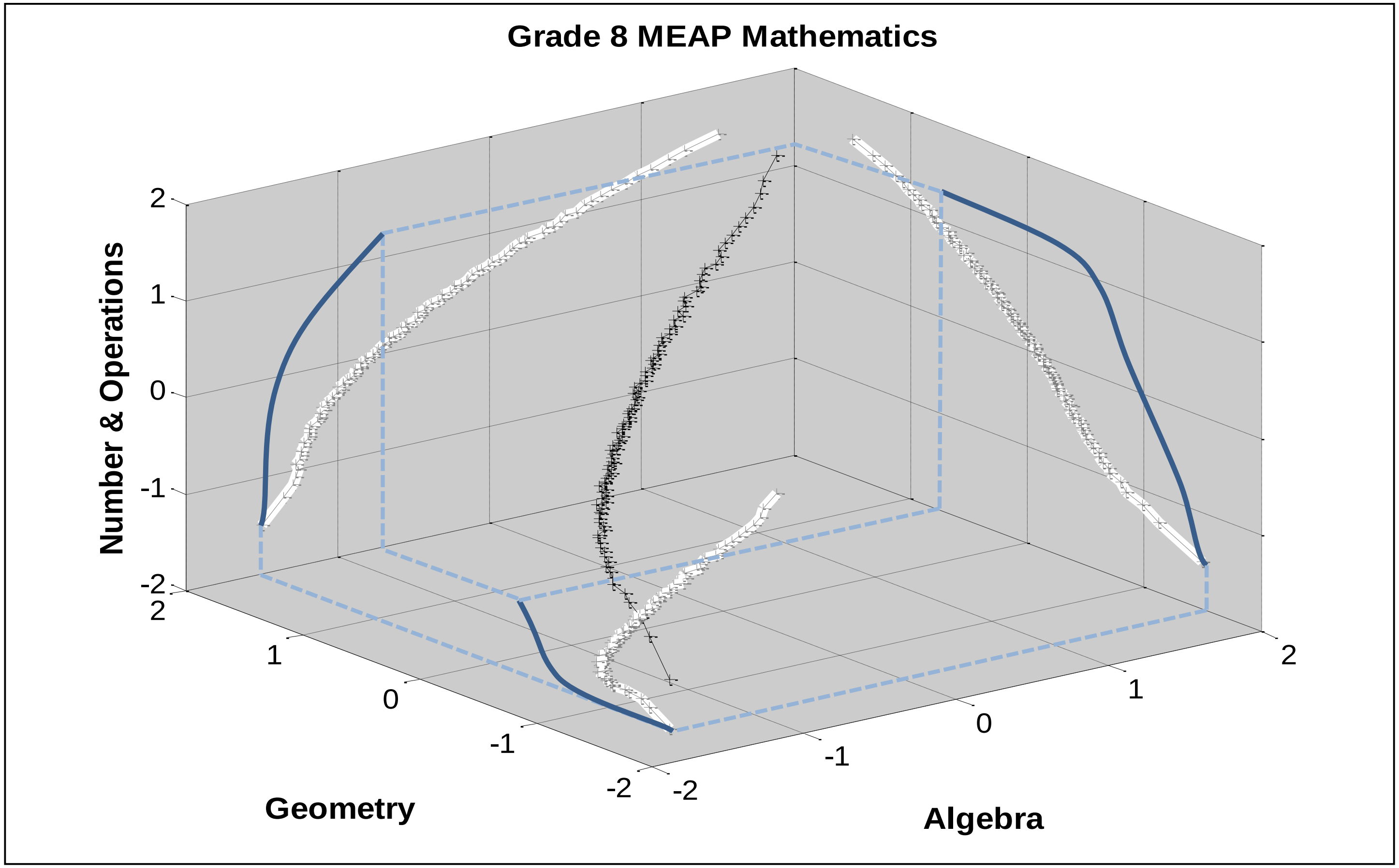
…differences reflect more changes in *Algebra* than in either *Geometry* or *Number & Operations*

Using Purely SME Definitions of Dimensions/Subscores



This is the general statewide picture.

Using Purely SME Definitions of Dimensions/Subscores



But what if a district had subscores which showed a distinctly different picture?

What would this mean if the district wanted to evaluate its new middle school curriculum?

Or target professional learning for math teachers?

A Kindergarten Entry Exam

- Utah State Board of Education
- Kindergarten Entry and Exit Profile (KEEP)
- 20-minute individually administered early literacy & early numeracy assessment
- Given in advance of the beginning of school (or as soon as a student registers)

- I show the results for early numeracy, but the ability to distinguish substantively meaningful results is similar for early literacy.

- Subscores

– Numeral to Quantity	*Mapping from numeral to quantity*
– Sense of Quantity	*Rote counting, quantity comparison, mapping from quantity to numeral*
– Counting and Cardinality	*Counting objects, 1-to-1 correspondence*
– Shape Creation	*Drawing (approximate) circle, square, plus sign, and triangle*
– Numeral Recognition	*Recognizing the numerals 0-10*

A Kindergarten Entry Exam

Subscore	Numeral to Quantity	Sense of Quantity	Counting & Cardinality	Shape Creation	Numeral Recognition	Overall Numeracy
Numeral to Quantity	0.72	**0.71**	**0.56**	**0.40**	**0.56**	0.79
Sense of Quantity	0.49	0.67	**0.92**	**0.62**	**0.86**	1.00
Counting & Cardinality	0.43	0.68	0.82	**0.47**	**0.61**	0.88
Shape Creation	0.33	0.49	0.41	0.92	**0.43**	0.64
Numeral Recognition	0.46	0.68	0.53	0.40	0.93	0.97
Overall Numeracy	0.64	0.88	0.76	0.59	0.89	0.92

A Kindergarten Entry Exam

With differences in profiles for individual students and for a whole classroom, this offers Utah kindergarten teachers that can be useful in planning the first few weeks of instruction.

The problem?

- A typical scree-plot analyses would likely recommend 1-2 dimensions
- Used a procedure I have been working on with colleagues for 10+ years, other procedures would be similar (though not as sensitive)
- If there is time, I will get to that.

Center for
Assessmen

A PRIMER ON DIMENSIONALITY

Or Why Do We Think Dimensions of Achievement are Not Distinct if They are Highly Correlated?

A Graphical Grammar of Dimensionality Analyses in Achievement Testing

Symbol(s)	Meaning
item	An observed score on a test item

Drawn from Conventions in Structural Equation Modeling

A Graphical Grammar of Dimensionality Analyses in Achievement Testing

Symbol(s)	Meaning
item	An observed score on a test item
math	An unobservable, latent dimension of achievement (e.g., degree of math knowledge)

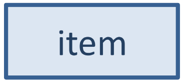
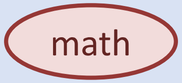
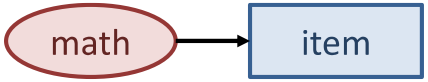
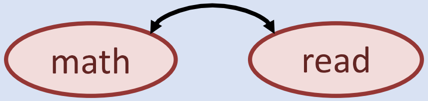
Drawn from Conventions in Structural Equation Modeling

A Graphical Grammar of Dimensionality Analyses in Achievement Testing

Symbol(s)	Meaning
item	An observed score on a test item
math	An unobservable, latent dimension of achievement (e.g., degree of math knowledge)
math → item	A causal relationship (e.g., degree of math knowledge drives answers to a specific test item)

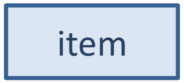
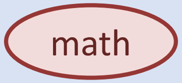
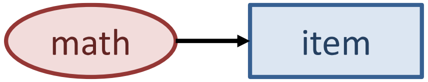
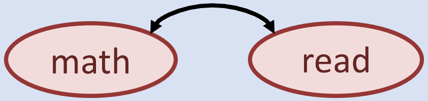
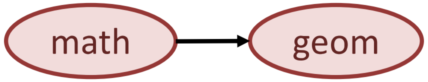
Drawn from Conventions in Structural Equation Modeling

A Graphical Grammar of Dimensionality Analyses in Achievement Testing

Symbol(s)	Meaning
	An observed score on a test item
	An unobservable, latent dimension of achievement (e.g., degree of math knowledge)
	A causal relationship (e.g., degree of math knowledge drives answers to a specific test item)
	A non-causal relationship (e.g., degrees of math knowledge and reading knowledge are correlated)

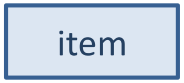
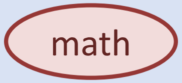
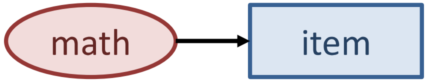
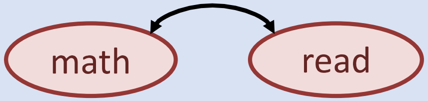
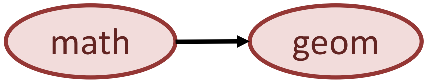
Drawn from Conventions in Structural Equation Modeling

A Graphical Grammar of Dimensionality Analyses in Achievement Testing

Symbol(s)	Meaning
	An observed score on a test item
	An unobservable, latent dimension of achievement (e.g., degree of math knowledge)
	A causal relationship (e.g., degree of math knowledge drives answers to a specific test item)
	A non-causal relationship (e.g., degrees of math knowledge and reading knowledge are correlated)
	Another causal relationship (e.g., degree of general math knowledge causes/contributes to degree of number sense knowledge)

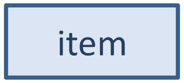
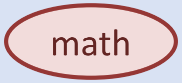
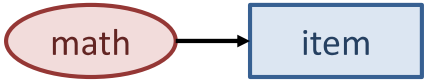
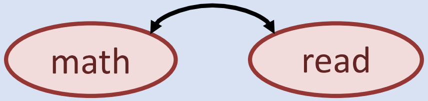
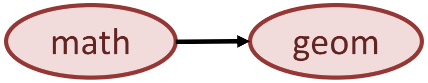
Drawn from Conventions in Structural Equation Modeling

A Graphical Grammar of Dimensionality Analyses in Achievement Testing

Symbol(s)	Meaning
 item	An observed score on a test item
 math	An unobservable, latent dimension of achievement (e.g., degree of math knowledge)
 math → item	A causal relationship (e.g., degree of math knowledge drives answers to a specific test item)
 math ↔ read	A non-causal relationship (e.g., degrees of math knowledge and reading knowledge are correlated)
 math → geom	Another causal relationship (e.g., degree of general math knowledge causes/contributes to degree of number sense knowledge)
r_{xx}	Reliability of the estimate of a dimension of achievement

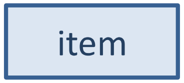
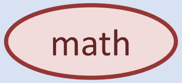
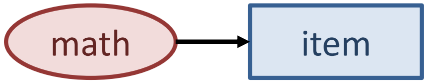
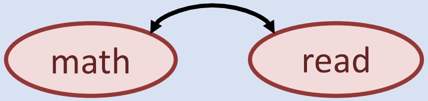
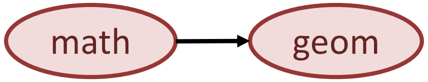
Drawn from Conventions in Structural Equation Modeling

A Graphical Grammar of Dimensionality Analyses in Achievement Testing

Symbol(s)	Meaning
 item	An observed score on a test item
 math	An unobservable, latent dimension of achievement (e.g., degree of math knowledge)
 math item	A causal relationship (e.g., degree of math knowledge drives answers to a specific test item)
 math read	A non-causal relationship (e.g., degrees of math knowledge and reading knowledge are correlated)
 math geom	Another causal relationship (e.g., degree of general math knowledge causes/contributes to degree of number sense knowledge)
r_{xx}	Reliability of the estimate of a dimension of achievement
r_{xy}	Correlation of the estimated scores between two dimensions of achievement

Drawn from Conventions in Structural Equation Modeling

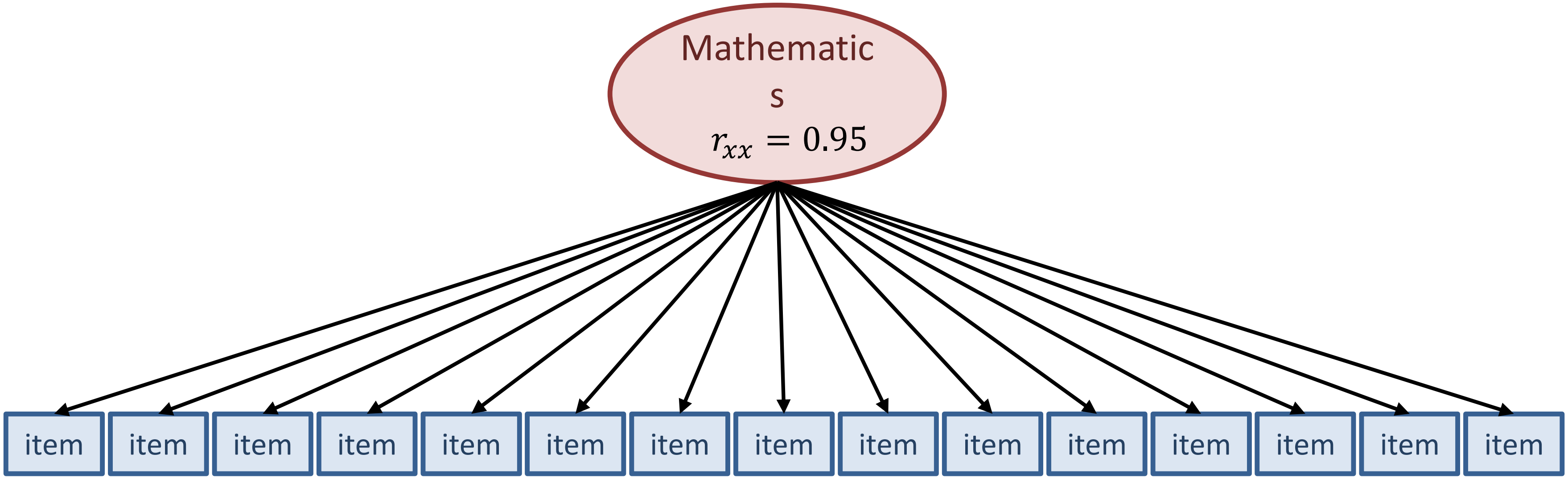
A Graphical Grammar of Dimensionality Analyses in Achievement Testing

Symbol(s)	Meaning
 item	An observed score on a test item
 math	An unobservable, latent dimension of achievement (e.g., degree of math knowledge)
 math item	A causal relationship (e.g., degree of math knowledge drives answers to a specific test item)
 math read	A non-causal relationship (e.g., degrees of math knowledge and reading knowledge are correlated)
 math geom	Another causal relationship (e.g., degree of general math knowledge causes/contributes to degree of number sense knowledge)
r_{xx}	Reliability of the estimate of a dimension of achievement
r_{xy}	Correlation of the estimated scores between two dimensions of achievement
$\hat{\rho}_{xy}$	Estimated correlation of the true scores between two dimensions of achievement

Drawn from Conventions in Structural Equation Modeling

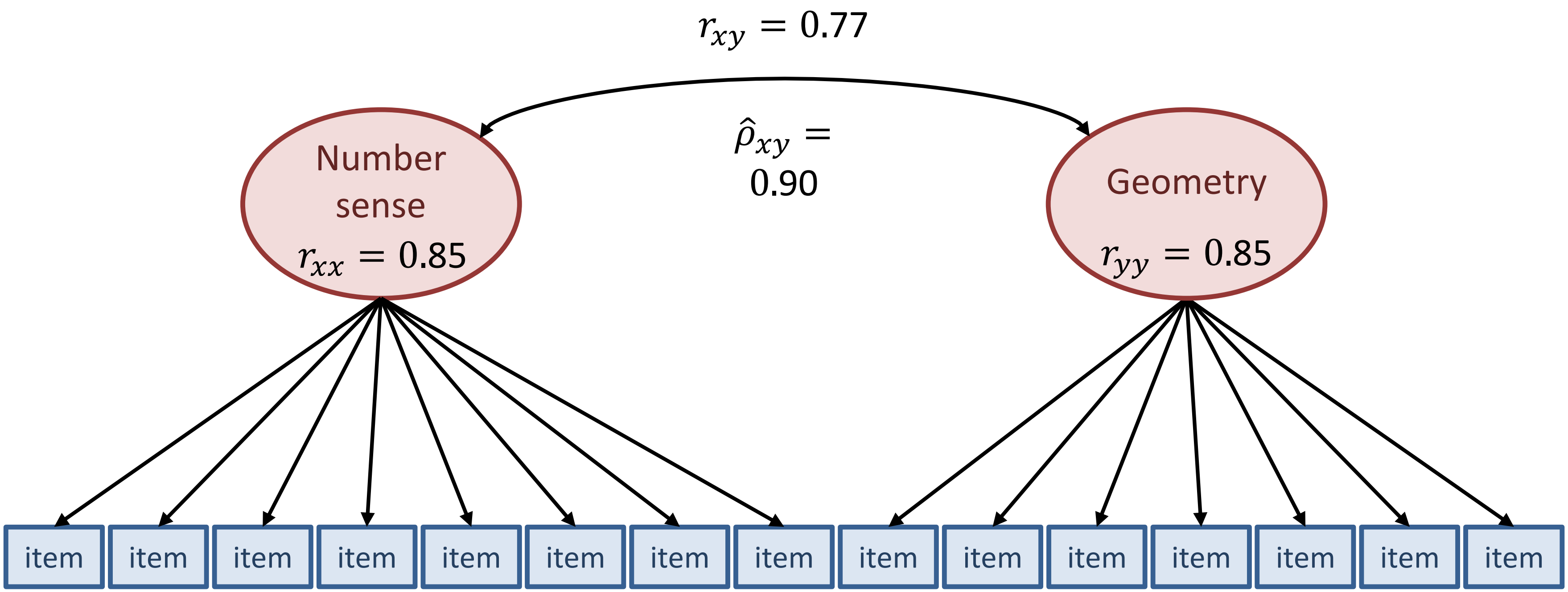
A Fundamental Issue in Dimensionality: Sources of Shared Variance

- A typical measurement model
- All we are trying to do is understand what drives variance in item scores
- We posit a unidimensional model and see if it captures the vast majority of variance, and almost universally, it does



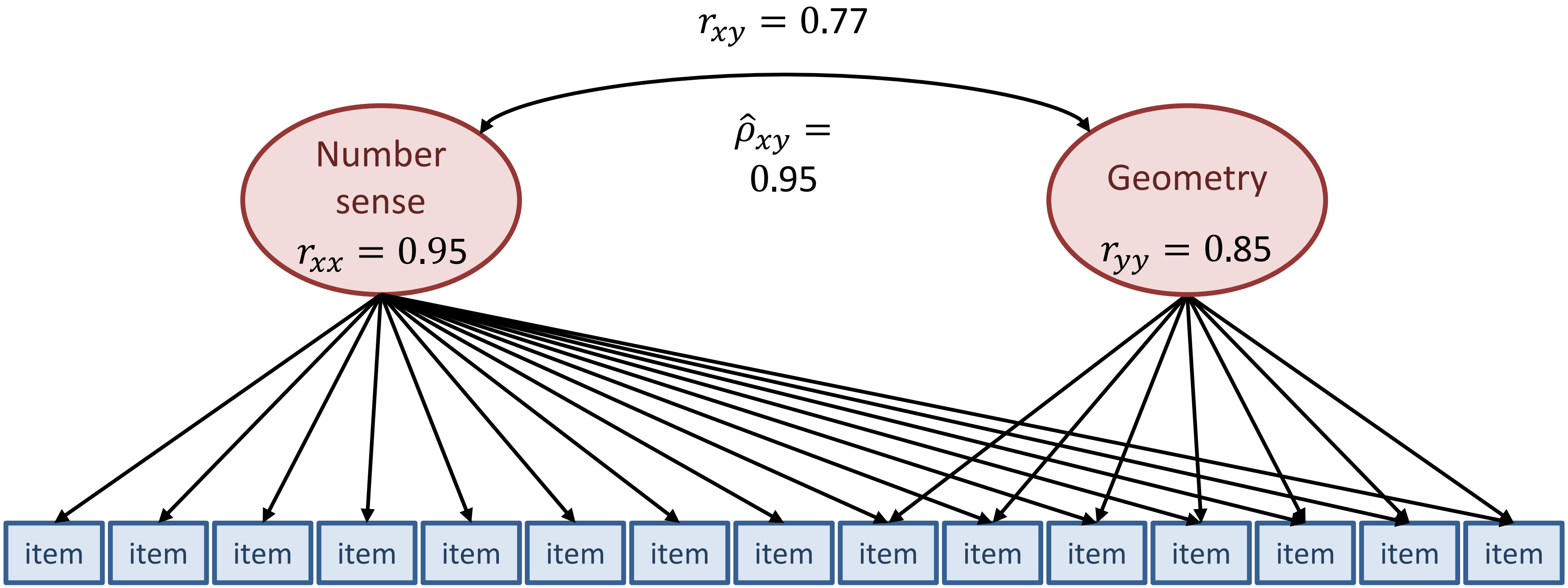
A Fundamental Issue in Dimensionality: Sources of Shared Variance

So we reject this model



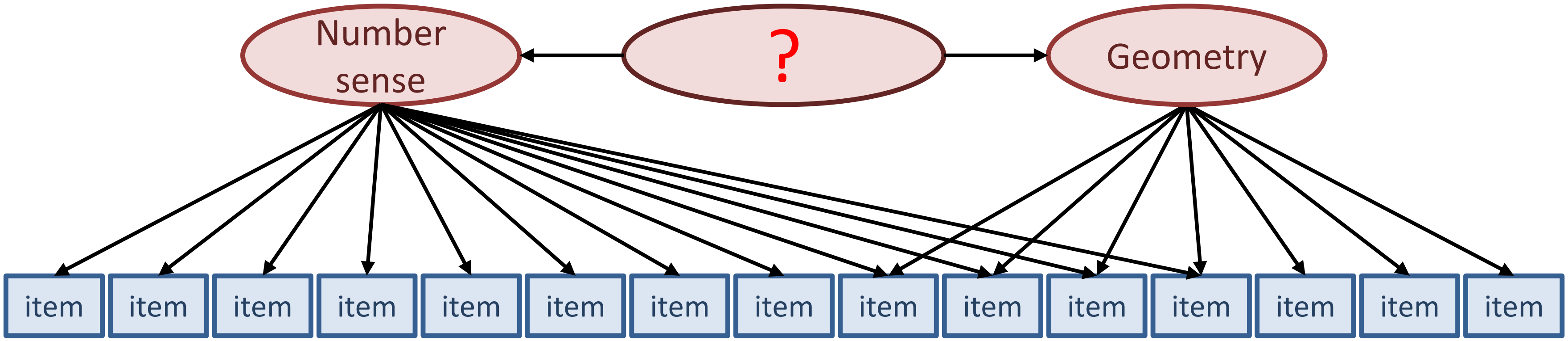
A Fundamental Issue in Dimensionality: Sources of Shared Variance

And we especially reject this model.



A Fundamental Issue in Dimensionality: Sources of Shared Variance

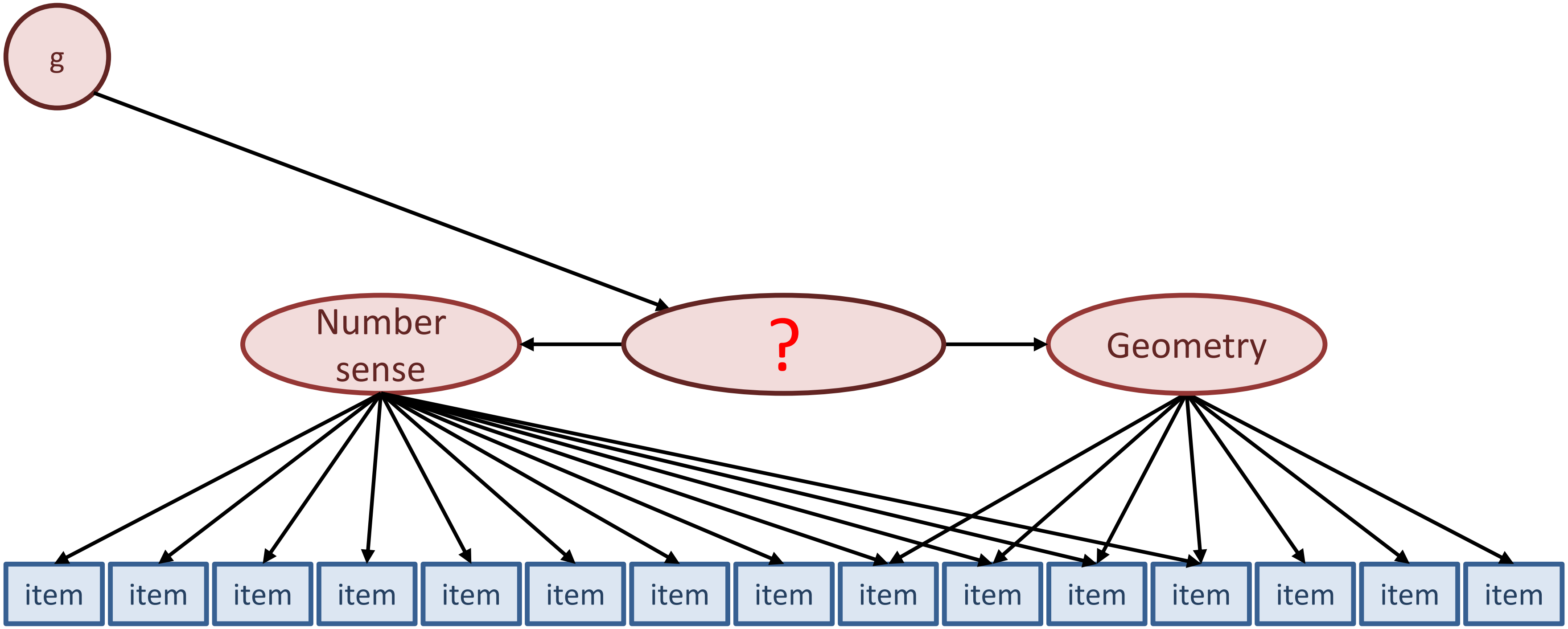
But we ignore this when developing scales (something other than g could be causing high correlations)


Number sense
?
Geometry
item
item
item
item
item
item
item
item
item
item
item
item
item
item
item

Everything about a child's background may create more shared variance.

A Fundamental Issue in Dimensionality: Sources of Shared Variance

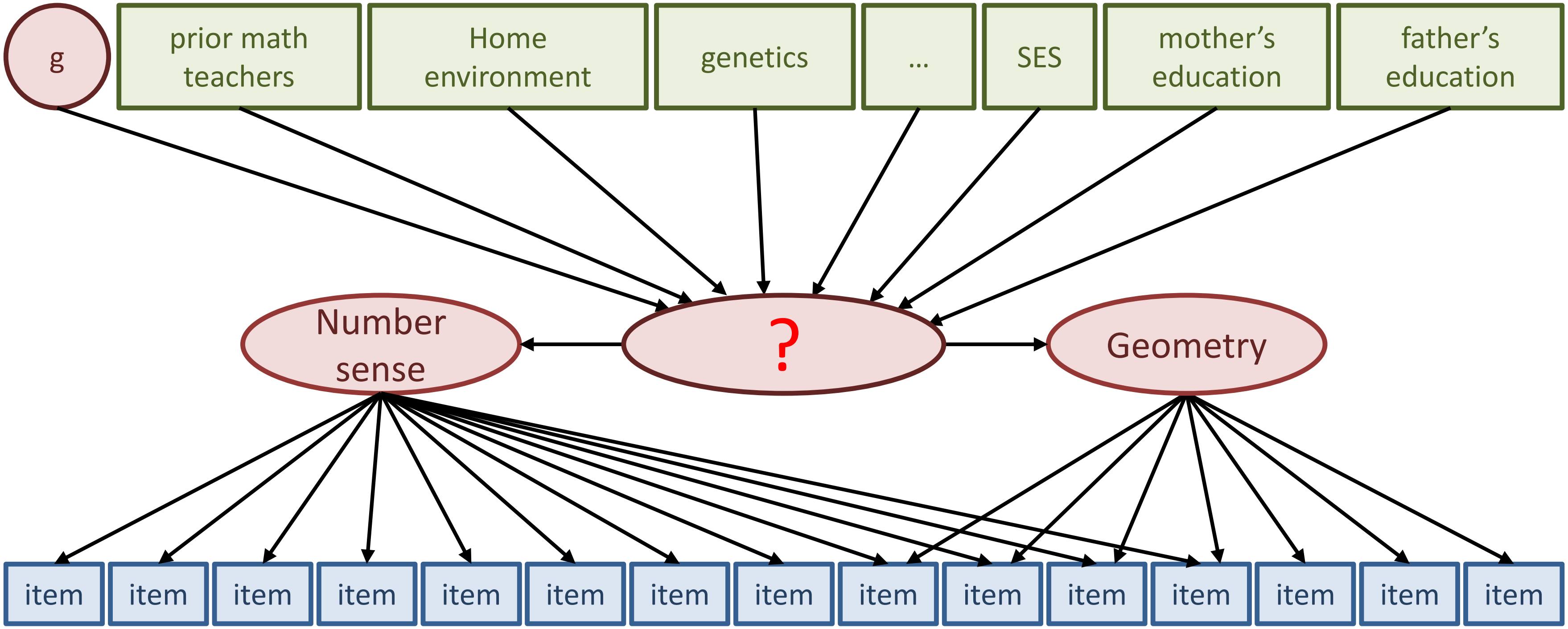
But we ignore this when developing scales (something other than g could be causing high correlations)



So let's start with g

A Fundamental Issue in Dimensionality: Sources of Shared Variance

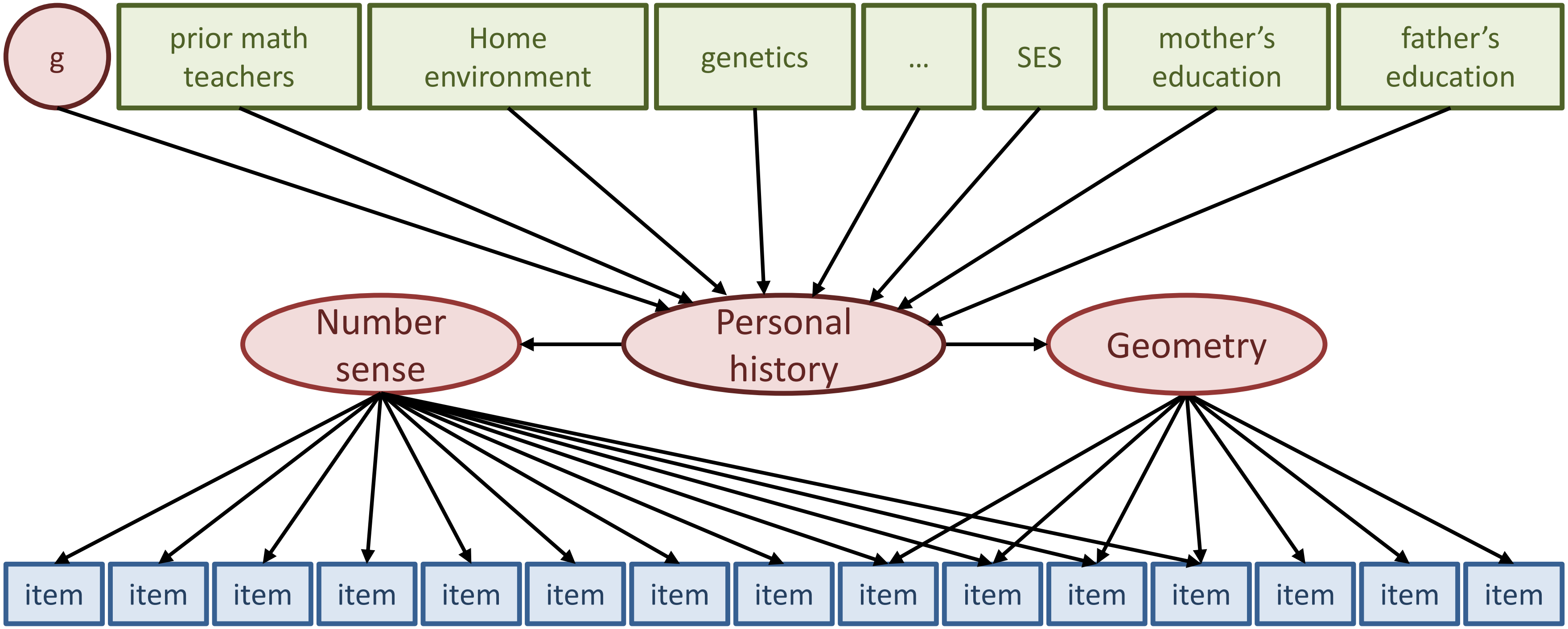
But we ignore this when developing scales (something other than g could be causing high correlations)



And now let's add many other potential influencers (or exogenous variables)

A Fundamental Issue in Dimensionality: Sources of Shared Variance

But we ignore this when developing scales (something other than g could be causing high correlations)



And now let’s add many other potential influencers (or exogenous variables)

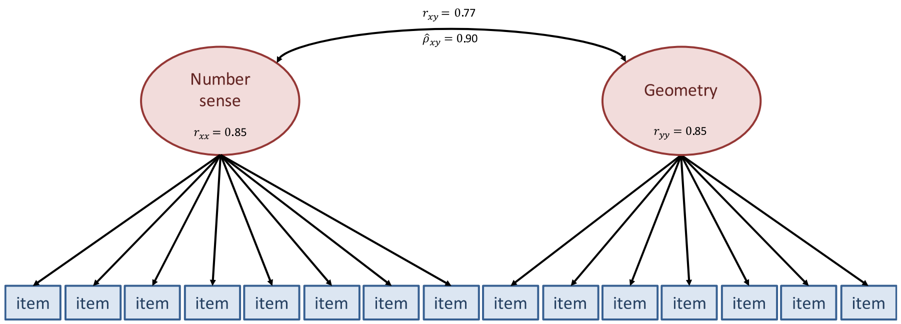
A Fundamental Issue in Dimensionality: Sources of Shared Variance

But that does not mean a student can't excel at *number sense* and perform poorly in *geometry*

she could have had a (somewhat rare) math teacher that was himself good at teaching *number sense* but not *geometry*

A Fundamental Issue in Dimensionality: Sources of Shared Variance

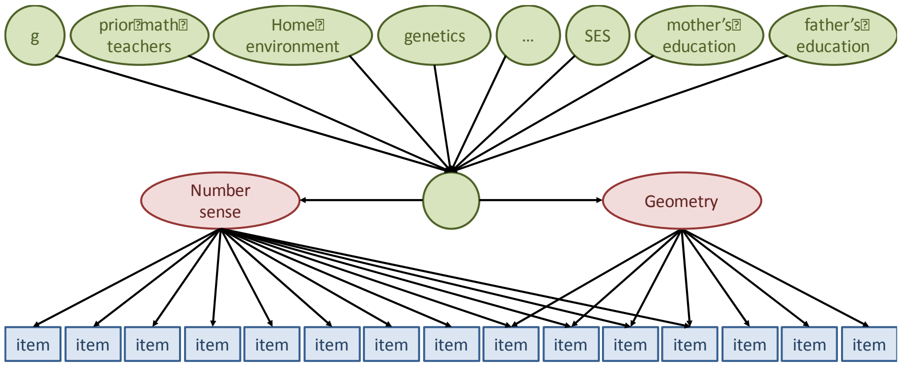
It is understandable to think the disattenuated subscore correlations (0.90 in this example) may be too high to consider them evidence for multiple dimensions when working with this model.



Not so much with this model. We would expect subscores in a subject matter to be very similar for most students. But for any number of reasons, for some students, they may be dissimilar (e.g., missing a large unit covering one of the subscores…)

What we should be looking for is outliers.

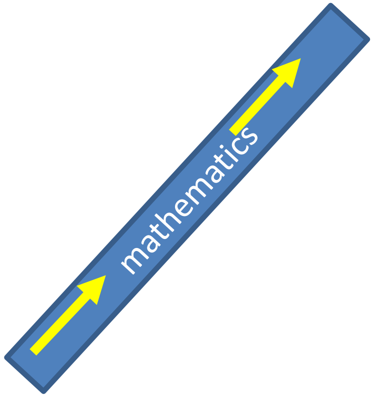
That, or constructing subscores separately (i.e., start privileging item-subscore correlations rather than item-total correlations. We are likely to create much more useful scales that way)



In Either Case, What to Do?

- Looking for a method that is highly sensitive to real differences in dimensions, but does not capitalize on chance because of its sensitivity. A four-paper series:
  - Reckase, M. D., Martineau, J. A., & Kim, J. P. (2000)
  - Zeng, J., & Martineau, J. A. (2009)
  - Zeng, J., Martineau, J. A. (2010).
  - Zeng, J. (2010).
- No published articles (yet), so a grain (or is it boulder?) of salt

Using Angles Between Items to Estimate the Number of Factors

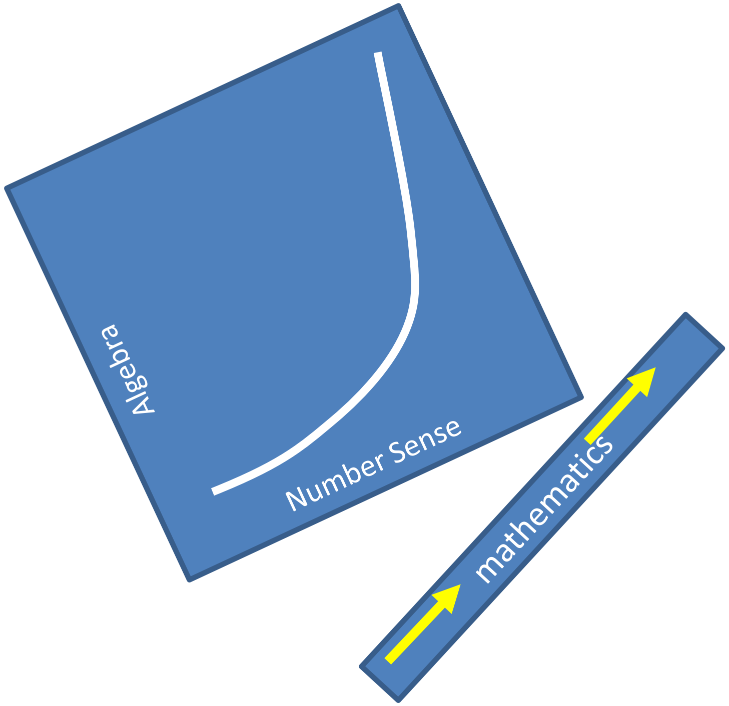


The item-total correlation is represented by the length of the item’s yellow arrow.

The difficulty of the item is represented by the origin of the item’s yellow arrow.

In this case, the two items appear to give a similar amount of information about the student’s achievement even though one is more difficult than the other (the lengths are the same)

Using Angles Between Items to Estimate the Number of Factors

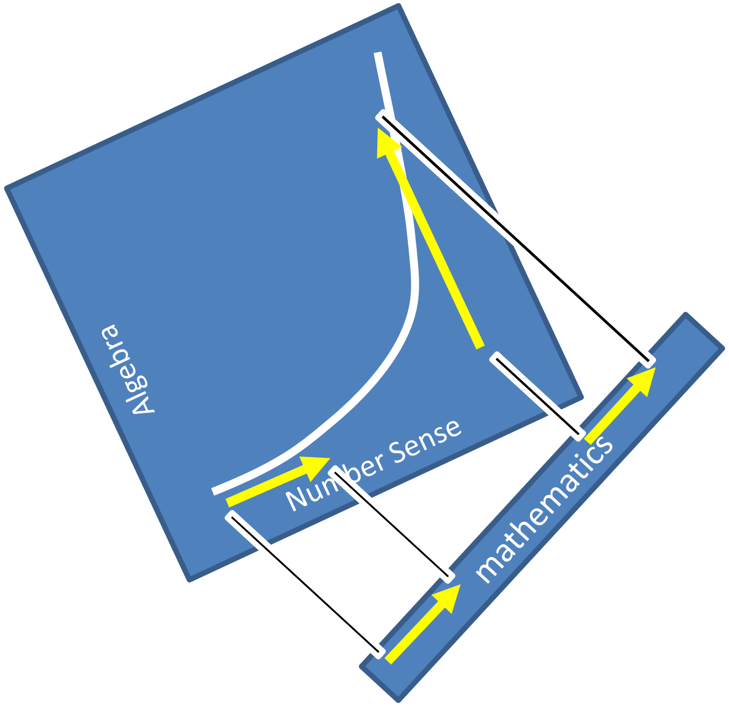


But what if there is a real difference between number sense and geometry achievement? (remember, that was the case with MEAP).

How the unidimensional mathematics score (with the yellow arrows) is oriented with respect to the 2D *number sense/algebra* plane is arbitrary.

Let’s just look at this one for example (remember, we saw a non-linear trajectory like this with MEAP)

Using Angles Between Items to Estimate the Number of Factors

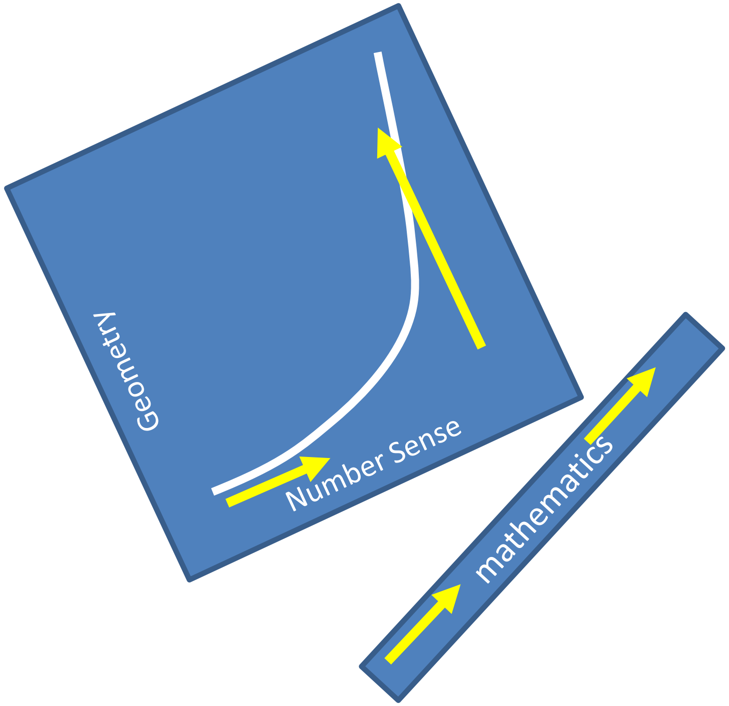


Let’s see how the two-dimensional item-subscore correlations might map onto item-total correlations

In this case, the first item has a similar amount of information about *number sense* as it had about “overall mathematics”

The second item has much greater information about *algebra* as it had about “overall mathematics”

Using Angles Between Items to Estimate the Number of Factors

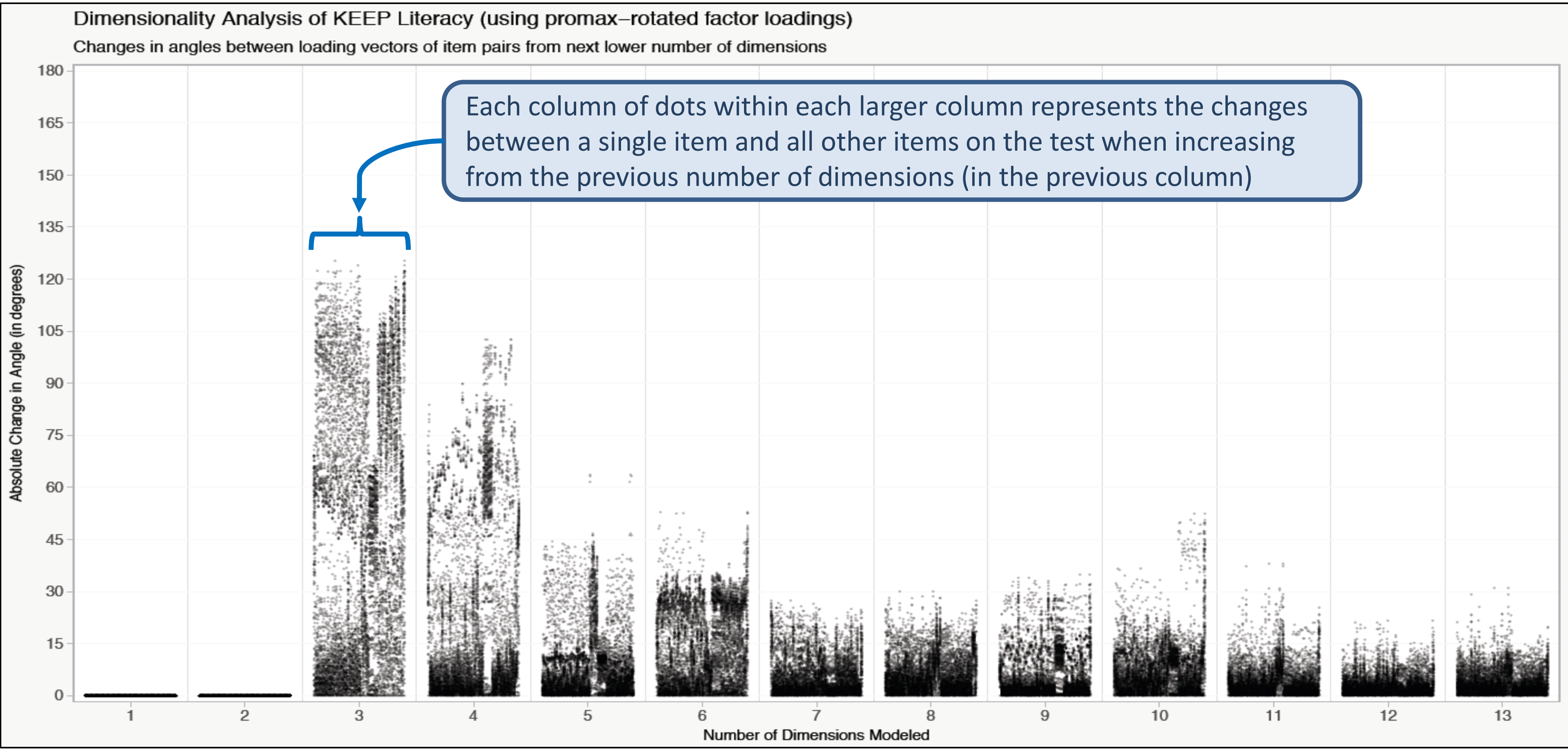


In this extreme case, the angle between the direction of the two arrows changes from 0 to approximately 90 when we create two subscores instead of just an overall score.

In general, angles between items tend to increase when the number of dimensions increases.

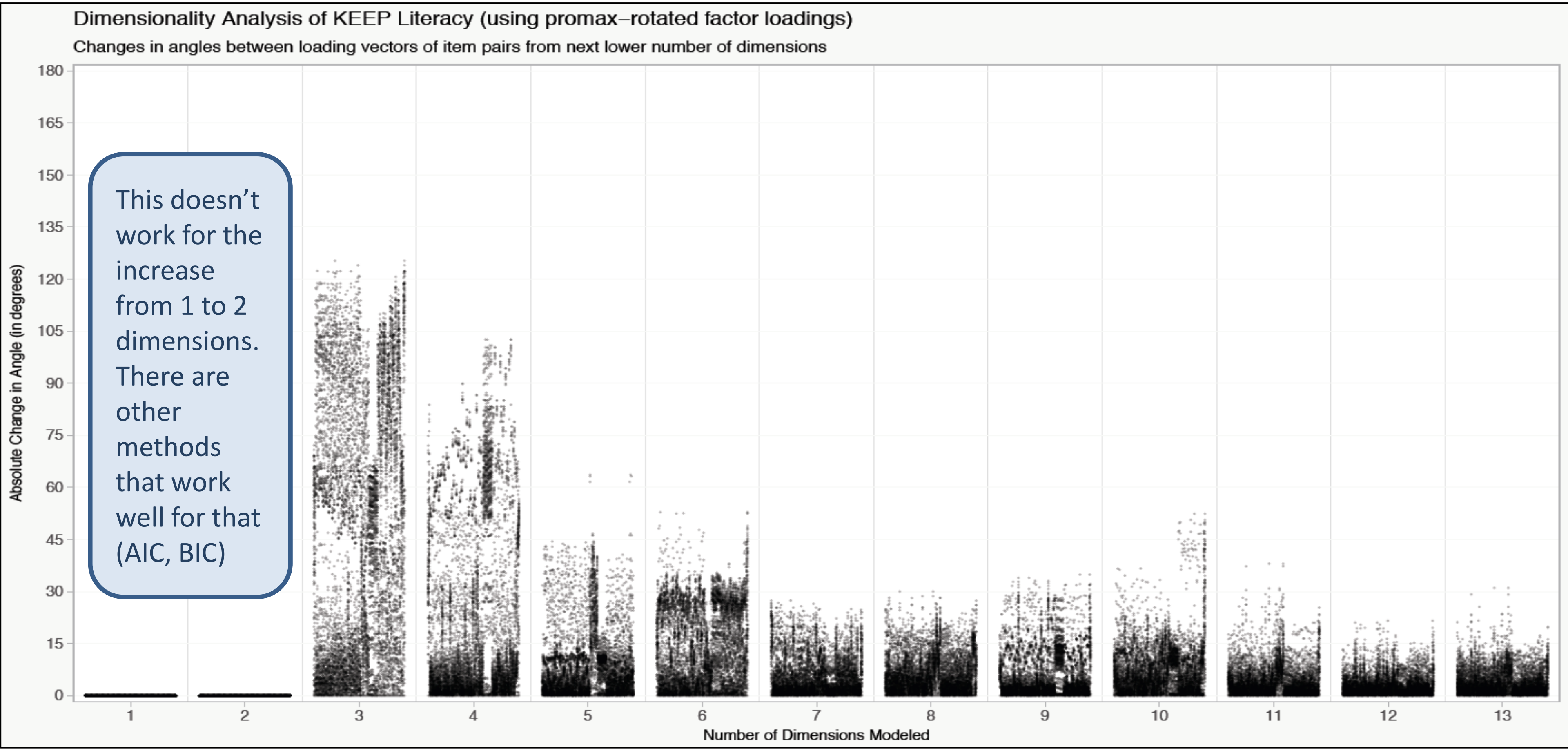
The question is how to tell when increases in angles represent real differences and when they are just adding idiosyncratic noise.

Evaluating Dimensionality – Early Literacy



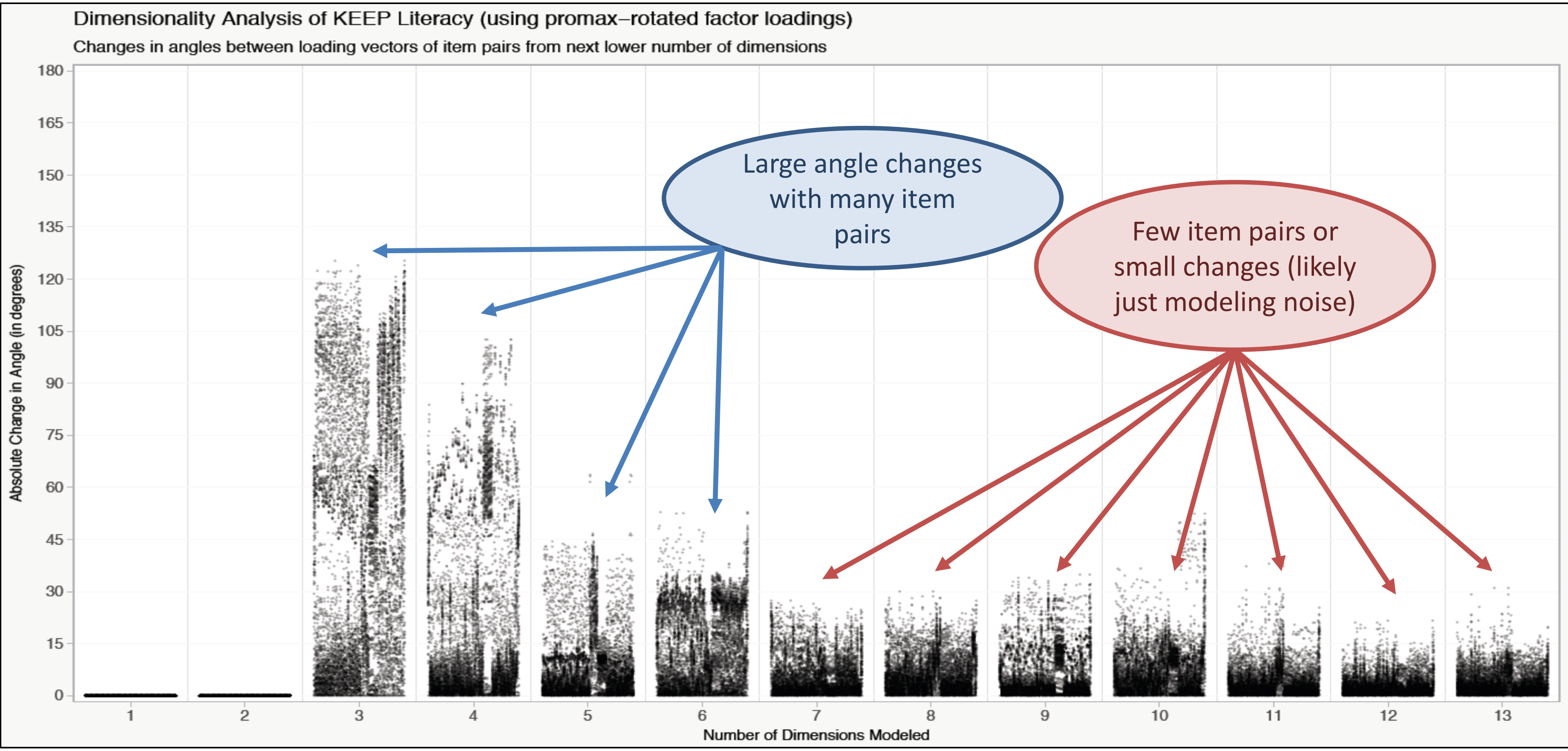
Each column of dots within each larger column represents the changes between a single item and all other items on the test when increasing from the previous number of dimensions (in the previous column)

Evaluating Dimensionality – Early Literacy



This doesn't work for the increase from 1 to 2 dimensions. There are other methods that work well for that (AIC, BIC)

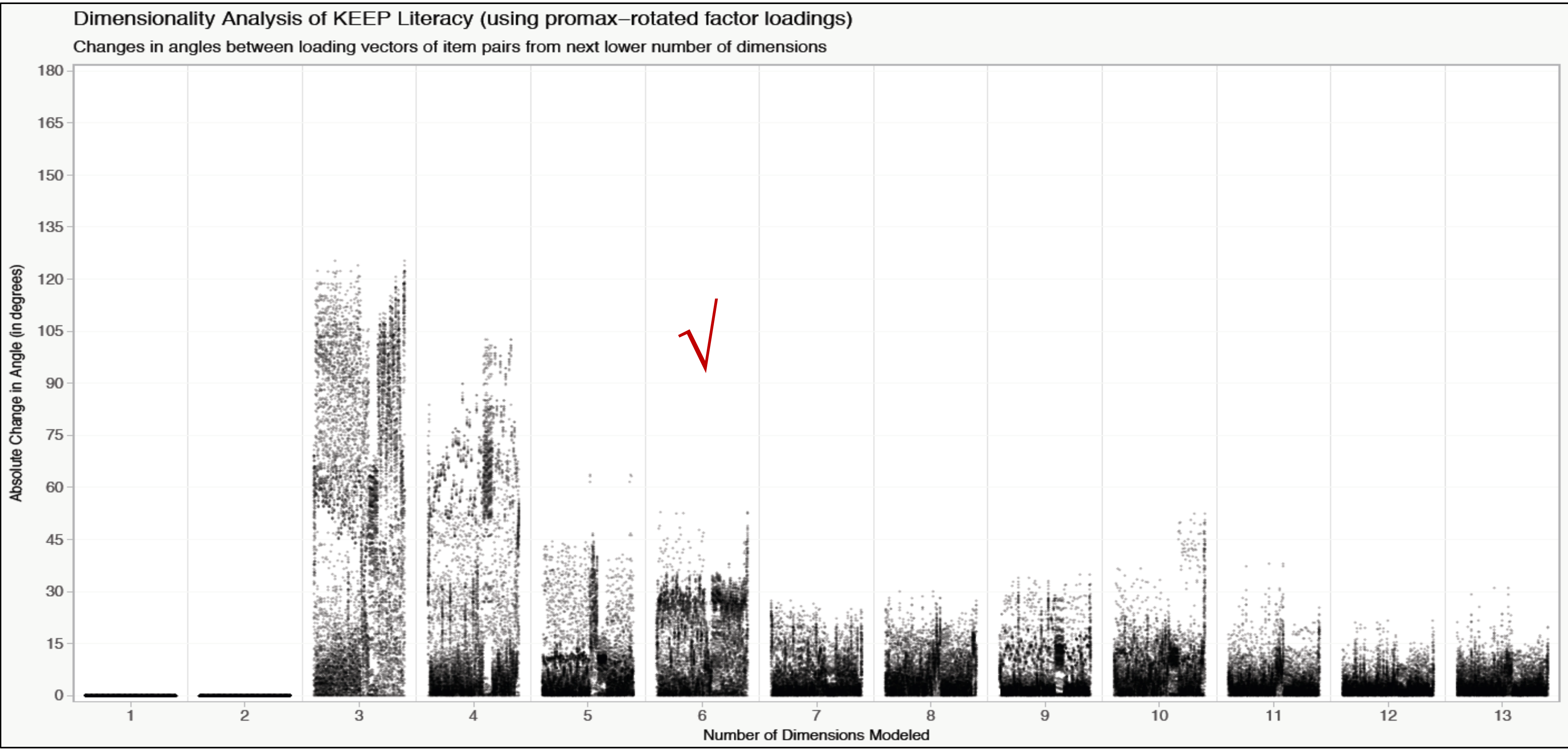
Evaluating Dimensionality – Early Literacy



Large angle changes with many item pairs

Few item pairs or small changes (likely just modeling noise)

Evaluating Dimensionality – Early Literacy

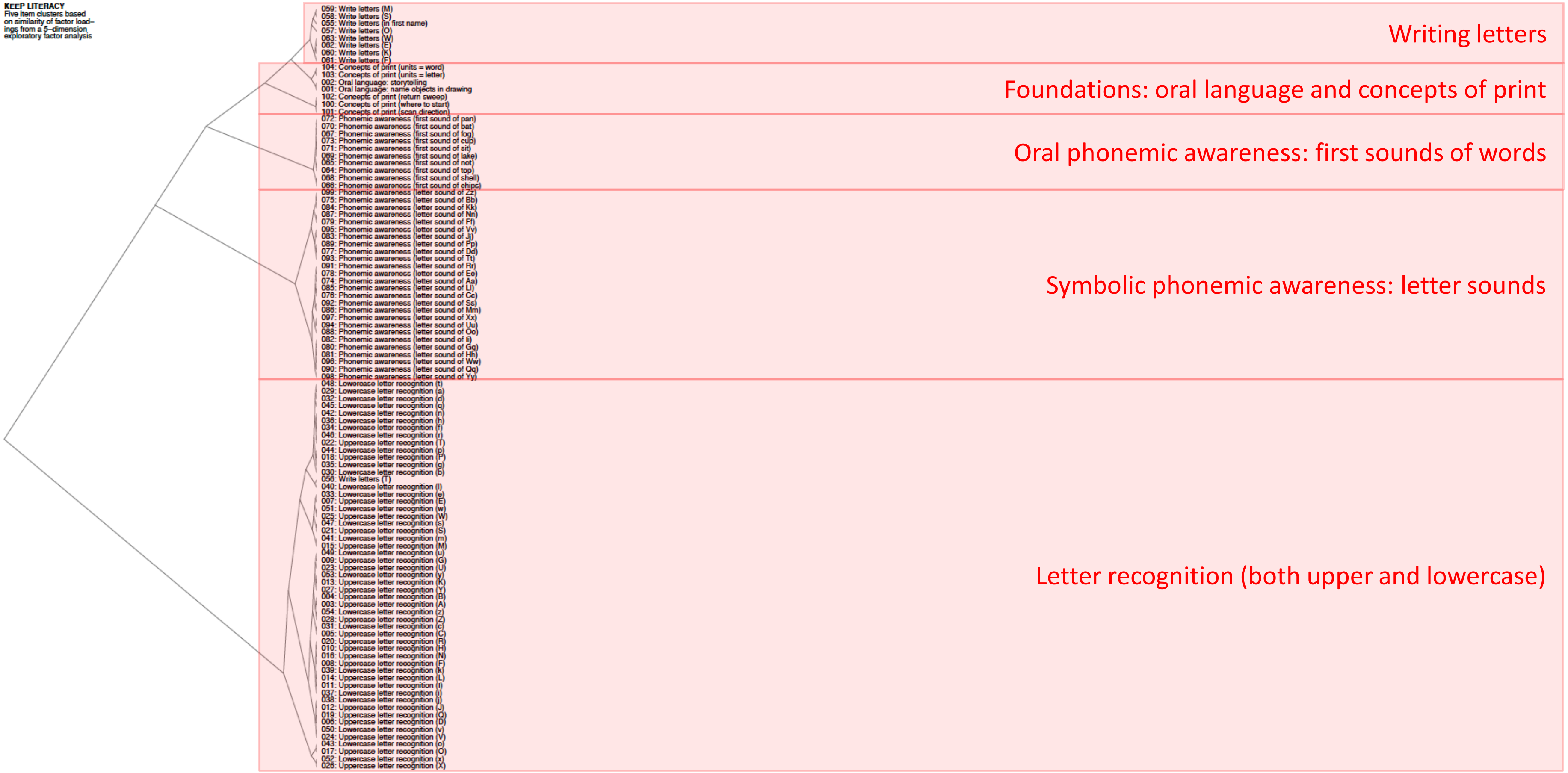


Identifying Which Items are Part of Which Dimension

- Because the KEEP items are all so focused, we can use cluster analysis to identify similar items, without having to worry that they measure a mix of dimensions.

- This approach is unlikely to work with many items that measure multiple dimensions. This is from experience.

Clusters - Literacy


KEEP LITERACY
Five item clusters based on similarity of factor loadings from a 5–dimension exploratory factor analysis
059: Write letters (M)
058: Write letters (S)
055: Write letters (in first name)
057: Write letters (O)
063: Write letters (W)
062: Write letters (E)
060: Write letters (K)
061: Write letters (F)
104: Concepts of print (units = word)
103: Concepts of print (units = letter)
002: Oral language: storytelling
001: Oral language: name objects in drawing
102: Concepts of print (return sweep)
100: Concepts of print (where to start)
101: Concepts of print (scan direction)
072: Phonemic awareness (first sound of pan)
070: Phonemic awareness (first sound of bat)
067: Phonemic awareness (first sound of fog)
073: Phonemic awareness (first sound of cup)
071: Phonemic awareness (first sound of sit)
069: Phonemic awareness (first sound of lake)
065: Phonemic awareness (first sound of not)
064: Phonemic awareness (first sound of top)
068: Phonemic awareness (first sound of shell)
066: Phonemic awareness (first sound of chips)
099: Phonemic awareness (letter sound of Zz)
075: Phonemic awareness (letter sound of Bb)
084: Phonemic awareness (letter sound of Kk)
087: Phonemic awareness (letter sound of Nn)
079: Phonemic awareness (letter sound of Ff)
095: Phonemic awareness (letter sound of Vv)
083: Phonemic awareness (letter sound of Jj)
089: Phonemic awareness (letter sound of Pp)
077: Phonemic awareness (letter sound of Dd)
093: Phonemic awareness (letter sound of Tt)
091: Phonemic awareness (letter sound of Rr)
078: Phonemic awareness (letter sound of Ee)
074: Phonemic awareness (letter sound of Aa)
085: Phonemic awareness (letter sound of Ll)
076: Phonemic awareness (letter sound of Cc)
092: Phonemic awareness (letter sound of Ss)
086: Phonemic awareness (letter sound of Mm)
097: Phonemic awareness (letter sound of Xx)
094: Phonemic awareness (letter sound of Uu)
088: Phonemic awareness (letter sound of Oo)
082: Phonemic awareness (letter sound of Ii)
080: Phonemic awareness (letter sound of Gg)
081: Phonemic awareness (letter sound of Hh)
096: Phonemic awareness (letter sound of Ww)
090: Phonemic awareness (letter sound of Qq)
098: Phonemic awareness (letter sound of Yy)
048: Lowercase letter recognition (t)
029: Lowercase letter recognition (a)
032: Lowercase letter recognition (d)
045: Lowercase letter recognition (q)
042: Lowercase letter recognition (n)
036: Lowercase letter recognition (h)
034: Lowercase letter recognition (f)
046: Lowercase letter recognition (r)
022: Uppercase letter recognition (T)
044: Lowercase letter recognition (p)
018: Uppercase letter recognition (P)
035: Lowercase letter recognition (g)
030: Lowercase letter recognition (b)
056: Write letters (T)
040: Lowercase letter recognition (l)
033: Lowercase letter recognition (e)
007: Uppercase letter recognition (E)
051: Lowercase letter recognition (w)
025: Uppercase letter recognition (W)
047: Lowercase letter recognition (s)
021: Uppercase letter recognition (S)
041: Lowercase letter recognition (m)
015: Uppercase letter recognition (M)
049: Lowercase letter recognition (u)
009: Uppercase letter recognition (G)
023: Uppercase letter recognition (U)
053: Lowercase letter recognition (y)
013: Uppercase letter recognition (K)
027: Uppercase letter recognition (Y)
004: Uppercase letter recognition (B)
003: Uppercase letter recognition (A)
054: Lowercase letter recognition (z)
028: Uppercase letter recognition (Z)
031: Lowercase letter recognition (c)
005: Uppercase letter recognition (C)
020: Uppercase letter recognition (R)
010: Uppercase letter recognition (H)
016: Uppercase letter recognition (N)
008: Uppercase letter recognition (F)
039: Lowercase letter recognition (k)
014: Uppercase letter recognition (L)
011: Uppercase letter recognition (I)
037: Lowercase letter recognition (i)
038: Lowercase letter recognition (j)
012: Uppercase letter recognition (J)
019: Uppercase letter recognition (Q)
006: Uppercase letter recognition (D)
050: Lowercase letter recognition (v)
024: Uppercase letter recognition (V)
043: Lowercase letter recognition (o)
017: Uppercase letter recognition (O)
052: Lowercase letter recognition (x)
026: Uppercase letter recognition (X)

Writing letters

Foundations: oral language and concepts of print

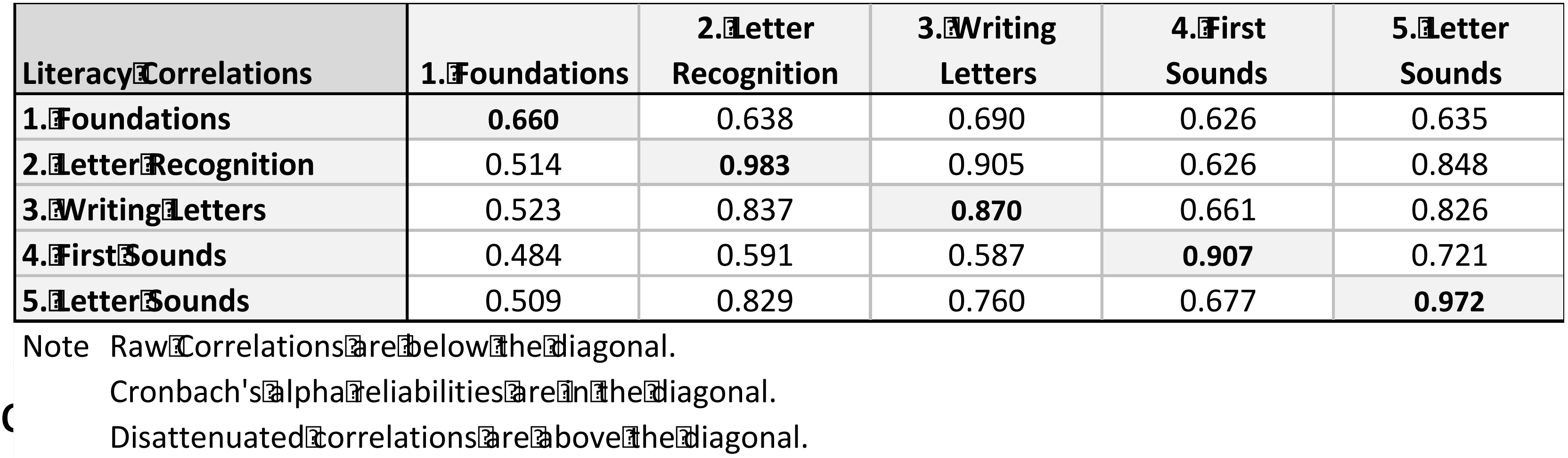
Oral phonemic awareness: first sounds of words

Symbolic phonemic awareness: letter sounds

Letter recognition (both upper and lowercase)

A Remarkably Well-behaved Dataset

- Allows for highly reliable subscore reporting



Literacy Correlations	1. Foundations	2. Letter Recognition	3. Writing Letters	4. First Sounds	5. Letter Sounds
1. Foundations	**0.660**	0.638	0.690	0.626	0.635
2. Letter Recognition	0.514	**0.983**	0.905	0.626	0.848
3. Writing Letters	0.523	0.837	**0.870**	0.661	0.826
4. First Sounds	0.484	0.591	0.587	**0.907**	0.721
5. Letter Sounds	0.509	0.829	0.760	0.677	**0.972**

Note Raw Correlations are below the diagonal.
Cronbach's alpha reliabilities are in the diagonal.
Disattenuated correlations are above the diagonal.

- Pr

- Can this be extended to a complex case?

REFERENCES

Martineau, J. A., Subedi, D., Ward, K., Li, T., Diao, Q., Drake, S., Kao, S.-C., Li, X., Lu, Y., Pang, F.-H., Song, T., Zheng, Y. (2007). Non-Linear Scale Trajectories through Multidimensional Content Spaces: A Critical Examination of the Common Psychometric Claims of Unidimensionality, Linearity, and Interval-Level Measurement. In Lissitz R. W. (Ed.). *Assessing and Modeling Cognitive Development in School: Intellectual Growth and Standard Setting*. JAM Press, Maple Grove, MN.

Reckase, M. D., Martineau, J. A., & Kim, J. P. (2000, June). *A Vector Approach to Determining the Dimensionality of a Data Set*. Presentation at the Psychometric Society, Seattle, WA.

Zeng, J., & Martineau, J. A. (2009, April). *Objective Extension and Evaluation of a Vector-Based Approach to Dimensionality Assessment*. Poster at NCME, San Diego, CA.

Zeng, J., Martineau, J. A. (2010, May). *A Method for Dimensionality Identification with Correlated Underlying Dimensions*. Presentation at NCME, Denver, CO.

Zeng, J. (2010). *Development of a Hybrid Method for Dimensionality Identification Incorporating an Angle-Based Approach*. Unpublished Dissertation. Ann Arbor: University of Michigan.