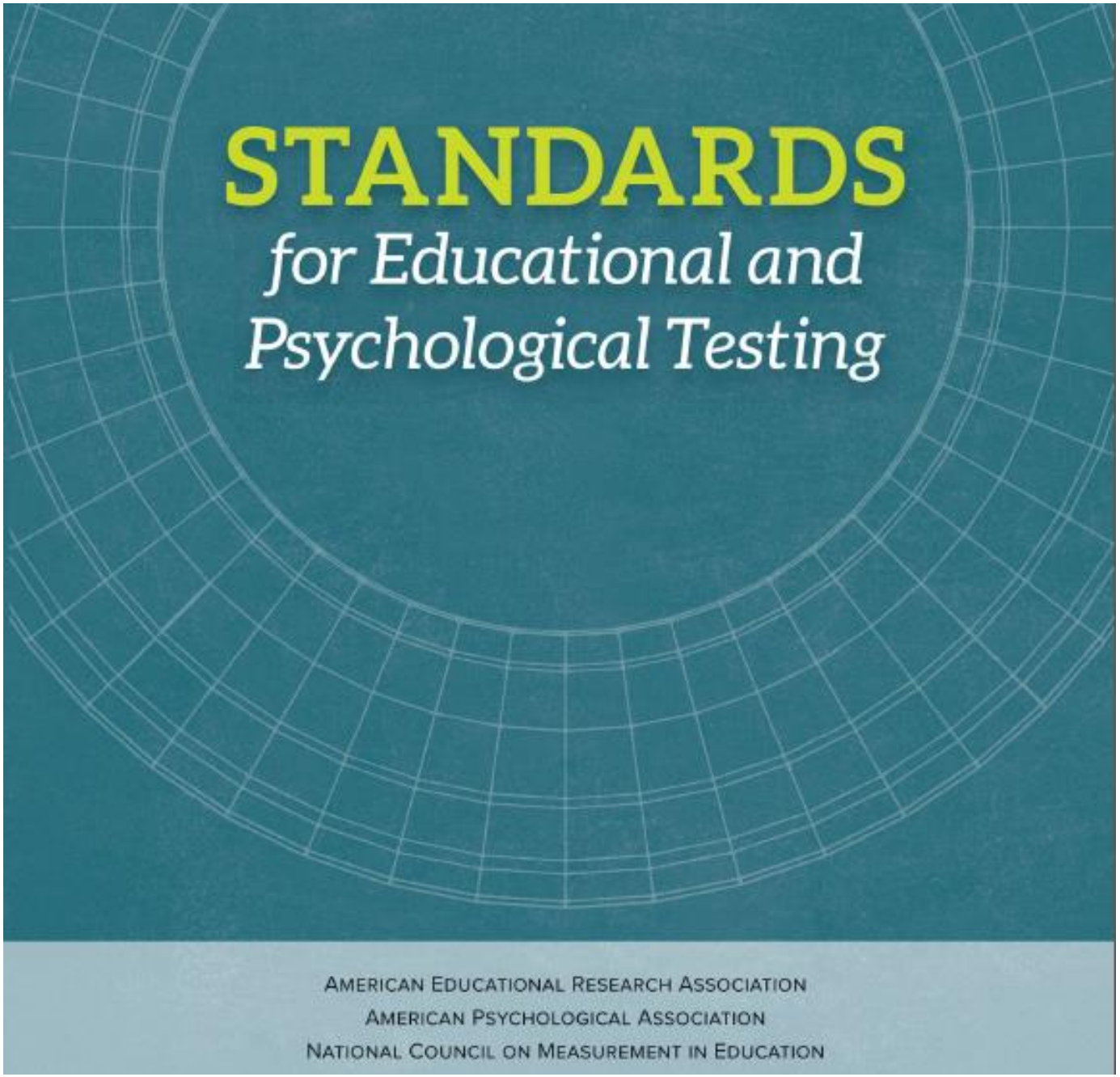
A look at subscores on large-scale assessments - *Can Claim Score results really do what they claim they do*?

MERA Conference – Spring 2017

Dave Treder
Genesee ISD


STANDARDS
for Educational and
Psychological Testing
American Educational Research Association
American Psychological Association
National Council on Measurement in Education

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

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.*

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. E. Define, evaluate, and compare functions. G. Understand congruence and similarity using physical models, transparencies, or geometry software.	5-6		
		F. Use functions to model relationships between quantities. H. Understand and apply the Pythagorean Theorem.	2-3		
	Supporting Cluster	A. Know that there are numbers that are not rational, and approximate them by rational numbers. I. Solve real-world and mathematical problems involving volume of cylinders, cones, and spheres. J. Investigate patterns of association in bivariate data.	4-5		
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		

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)

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

Total Variance (standard deviation squared)

Total Score and Claim Score Marginal Reliability

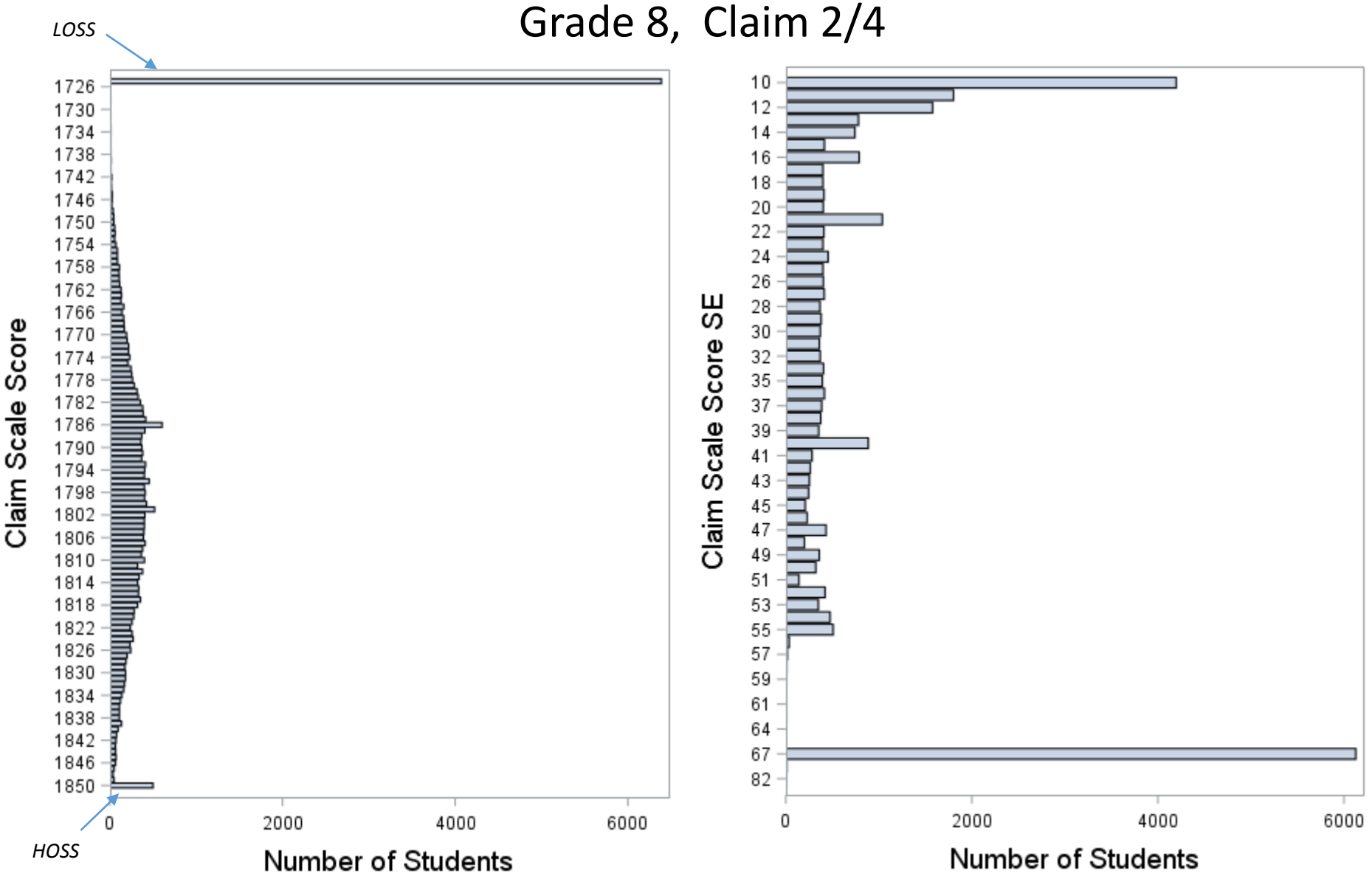
M-STEP Math 2016 (approx. 30,000 students per grade)

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

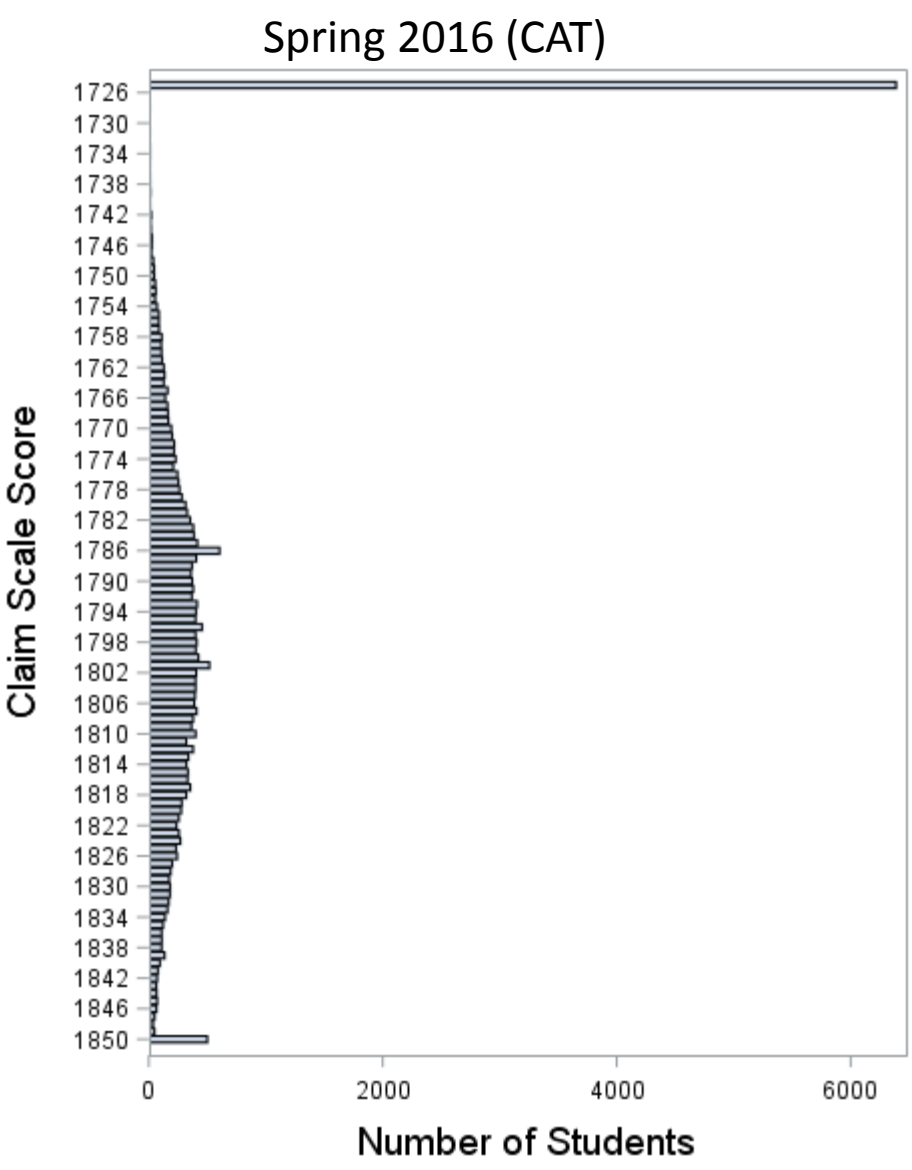
Marginal Reliability = (Total Variance – Error Variance) / Total Variance

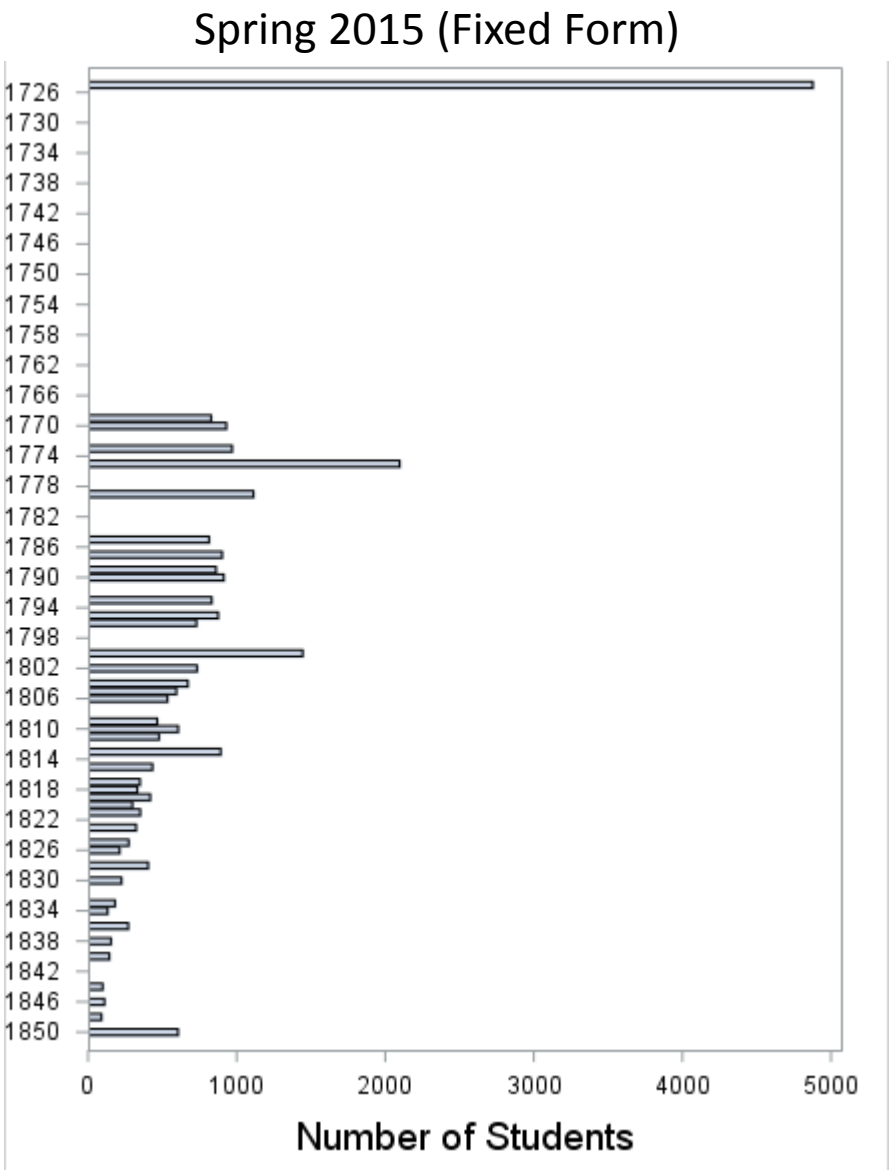
Distribution of Claim Scale Scores and Standard Errors



Distribution of Claim Scale Scores and Standard Errors

Grade 8, Claim 2/4





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	Number of Proficient Students with Above Grade Items	Number of not-Proficient Students with Below Grade Items
English Language Arts/Literacy				
3	9	113	113	0
4	22	564	183	381
5	9	133	129	4
6	10	359	95	264
7	11	548	51	497
8	2	36	36	0
Mathematics				
3	0	0		0
4	12	259		259
5	26	208		208
6	19	165		165
7	10	537		537
8	14	511		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.

- Similar to the number of items available for the M-STEP

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

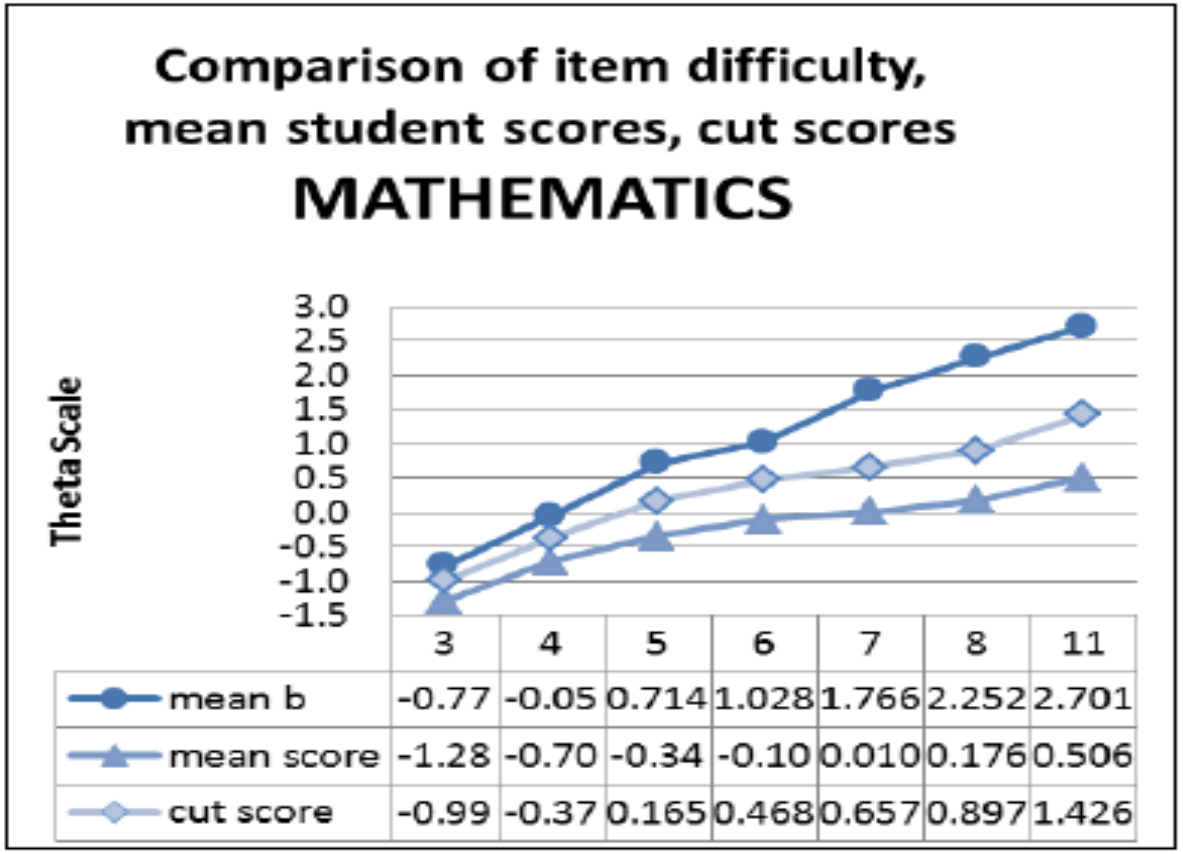
Math - 2014-15 OPERATIONAL SUMMATIVE POOLS FOR MATHEMATICS

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)

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 Tech Manual)



	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

Table 2.8 shows the distribution of items across simulated test events. Exposure rates represent the number of test events in which items appeared…. Most items show a desired moderate exposure. (SBAC Tech Manual)

TABLE 2.8 PERCENT OF ITEMS BY EXPOSURE RATE

Grade	Total Items	Exposure Rate					
		Unused	0%-20%	21%-40%	41%-60%	61%-80%	81%-100%
Mathematics							
3	829	0.48	99.16	0.36	0	0	0
4	818	0.12	99.14	0.73	0	0	0
5	807	0.12	99.38	0.50	0	0	0
6	739	0.14	99.05	0.81	0	0	0
7	670	0.15	98.66	1.19	0	0	0
8	612	0.00	98.04	1.80	0.16	0	0

Claim	Difficulty				
	1	2	3	4	5
1	5	31	23	42	304
2	0	0	1	4	38
3	0	4	3	7	94
4	0	2	3	9	42

11 Items (1.8% of 612), given to (up to) 40% of the test takers…
which students, and which strands all had the same items?

Standards for Educational
& Psychological Testing

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.

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?

There are a number of methods for determining if Subscores add 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.

A simple equation to predict a subscore's value. (2014) *Educational Measurement: Issues and Practice*, Feinberg, R. & Wainer, H.

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

These methods utilize Subscore Reliability (see slide 10) 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.

A brief note on Adjusted/Augmented Subscores

- Adjusted/Augmented Subscores: methods that increase the precision of subscores by borrowing information from the total test score or other subscores.

A Review and Empirical Comparison of Approaches for Improving the Reliability of Objective Level Scores *(2010)* Educational and Psychological Measurement, Skorupski & Carvajal

The comparison of subscore augmentation approaches found that generally **all methods were very successful in dramatically increasing the reliability of subscore estimates**. However, **this increase was accompanied by near-perfect correlations among the subscore estimates**. This finding called into question the validity of the resultant subscores, and therefore the usefulness of the subscore augmentation process.

-- and, the response by the authors who had been "maligned"
(research journal spats are always fun)

Do Adjusted Subscores Lack Validity? Don't Blame the Messenger (2011) Educational and Psychological Measurement, 2011, Sinharay , Haberman, & Wainer

The tests considered in [in the critique] were unidimensional and were incapable of producing diagnostic scores of any kind. So it is no wonder that the adjusted subscores computed from these data are not valid. However, **responsibility for the lack of validity lies not with the adjusted subscores but rather with the tests and those who try to report any diagnostic subscores from the tests in the first place. The adjusted subscores are just the messengers of the bad news that the data are not appropriate for diagnostic score reporting**.

Summary / Conclusions

- Test Blueprints, while essential, DO NOT mean that you can reliably/validly report sub-stuff level scores
  (in fact, in most cases, you can't)

- There are different levels of CATness
  - depends on:
    > The purpose of the Test (proficiency, off-grade items?)
    > The amount of stuff you're required to sample;
    > How many items you have for each of the stuff;
    and
    > How well the items are distributed across different ability levels

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

*This includes M-STEP, NWEA, PSAT, SAT, etc.

Or, paying homage to our retired colleague, Dr Ernie Bauer -- five closing slides he used in a presentation on Claim Scores

Parting words…

- Total math scores tell the story
- Math claim scores add nothing useful
- Math claim scores may add misleading noise
- Please ignore math claim scores
- What should we focus on?
  - Solid math instruction

So want to do something useful with the M-Step ELA and Math ***Claims*** information…