Stuff of Interest to *Me*

Dave Treder

MERA Fall 2019 Conference

https://tinyurl.com/MERA-TREDER

Couple Things that stuck in my craw

- The mis-match of M-STEP item difficulty & student ability (and the resulting impact on measuring – or, inaccurately measuring – growth)
- The appropriateness of using Adequate Growth Percentiles
- The New A-F Report Card

MICHIGAN Department of Education

Technical Report

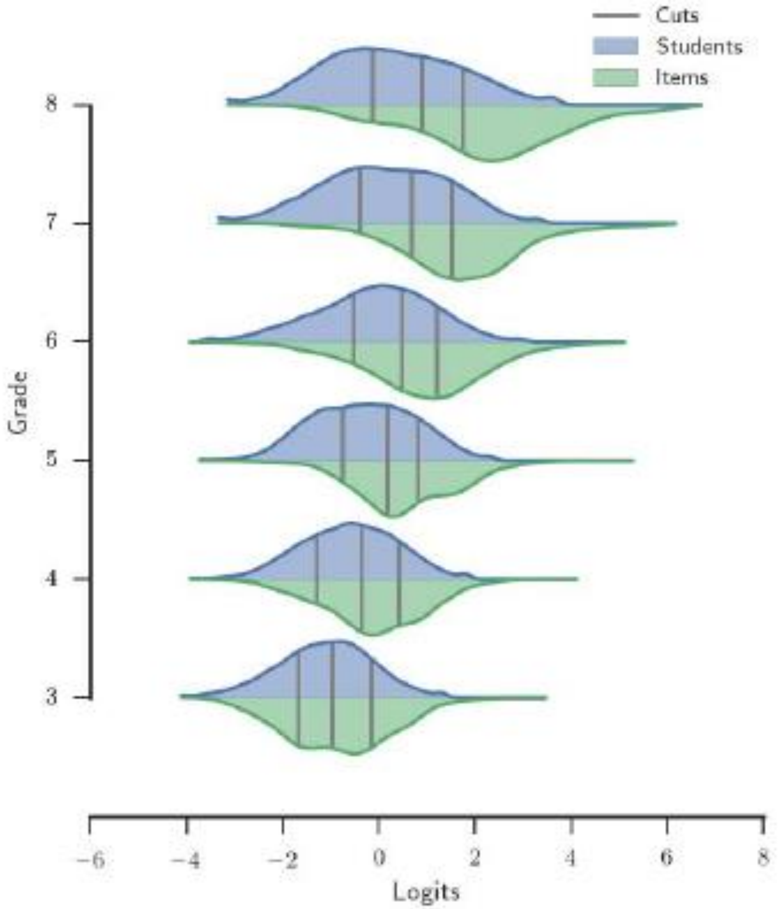
Spring 2017

Michigan Student Test of Educational Progress

(M-STEP)

M-STEP -- Item Difficulty & Student Ability

Figure 8.2. Mathematics Item Pool Difficulty in Comparison to the Student Ability Distribution*



Item
Difficulty

2.22

1.74

0.99

0.60

-0.09

-0.83

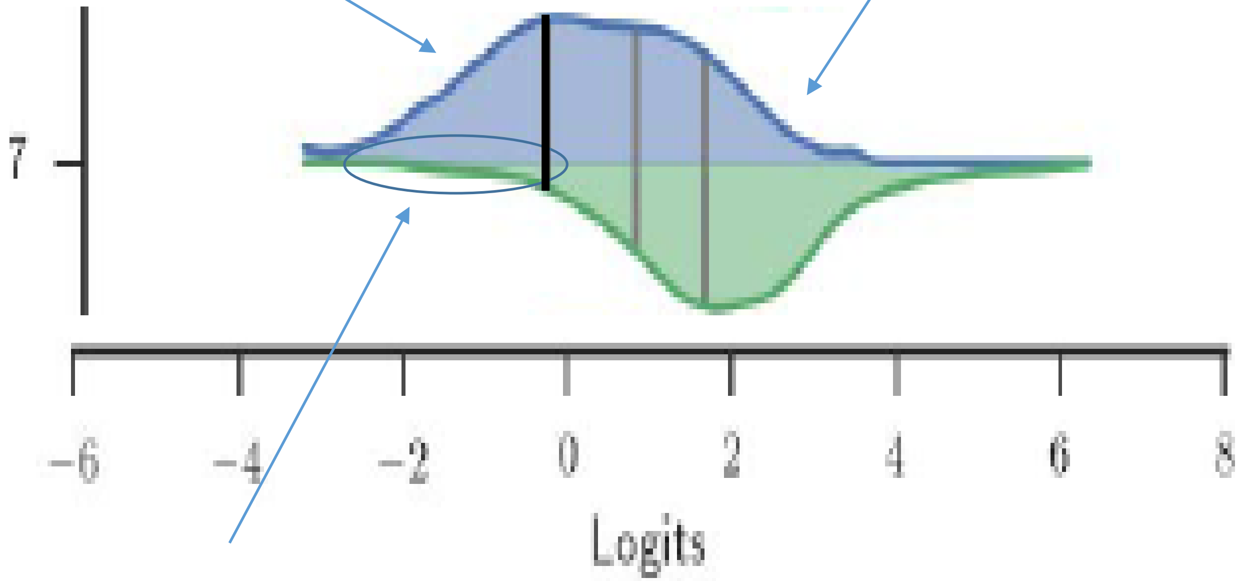
Figure 8-2 show the comparison of item difficulty, student scores, and cut scores for mathematics, by grade For most grades, the item pool has good alignment with the student ability distribution. *However, in grades 6 to 8 for mathematics, the item pool appears to be more difficult when compared to the corresponding student ability distribution*.

* Their method of scaling puts *item difficulty* and *person ability* on the same scale: average item difficulty = 0, and average student ability = 0 **–** because this is a vertical scale, all items & students (all grades) are put on the same scale (together)

M-STEP Grade 7, tem difficulty & student ability

Not Proficient: 36%

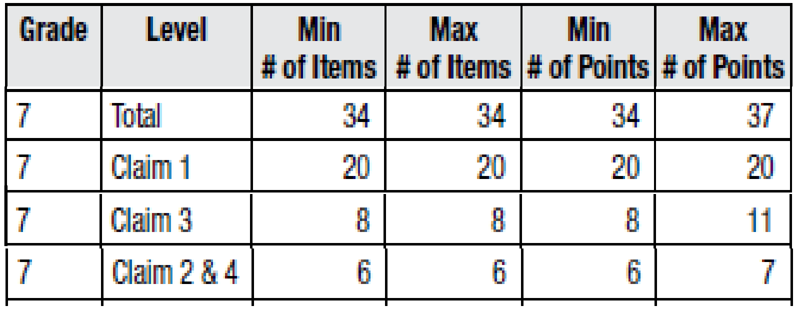
Proficient: 19%
Advanced: 16%



Items whose difficulty "matches"
the ability level of Not Proficient Students

M-STEP Technical Manual

Table 6-6. Summary of Items and Points for Mathematics



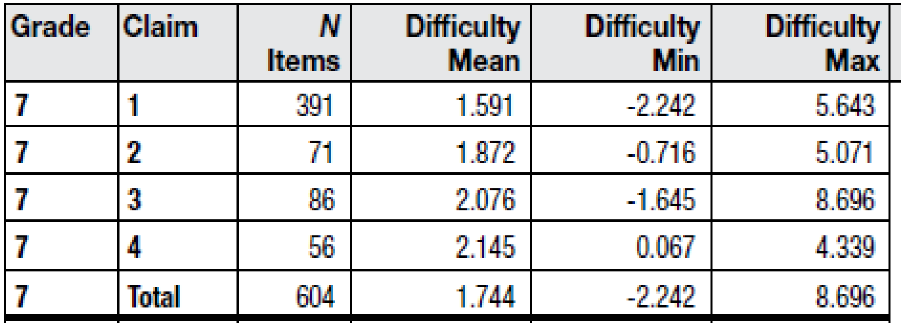
Grade	Level	Min # of Items	Max # of Items	Min # of Points	Max # of Points
7	Total	34	34	34	37
7	Claim 1	20	20	20	20
7	Claim 3	8	8	8	11
7	Claim 2 & 4	6	6	6	7

The requirement of *content coverage* can dramatically affect the CAT-ness of a CAT

In the mathematics assessment,

- Claim 1 (Concepts and Procedures) **consists of 20 items** (MC or TE) in the CAT.
- Claim 2 (Problem Solving) **consists of 4 or 5 items**, assessed in both the CAT and the PT.
- Claim 3 (Communicating Reasoning) **consists of 9 or 10 items**, assessed primarily in the CAT with a couple of Claim 3 items in the PT.
- Claim 4 (Modeling and Data Analysis) section **consists of 5 items** across the CAT and the PT.

Table 8-5. Distribution of Item Level Difficulty

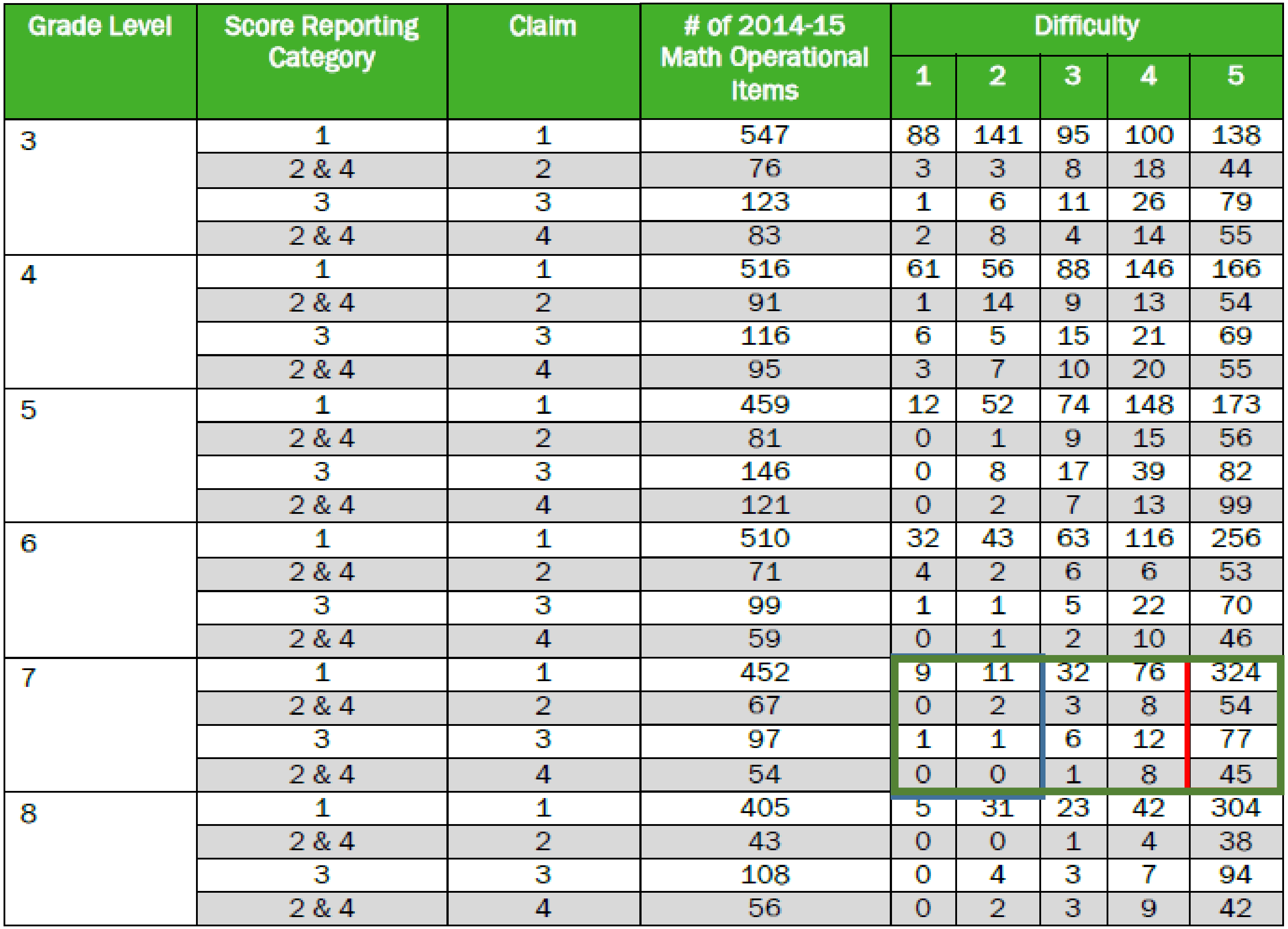


Grade	Claim	*N* Items	Difficulty Mean	Difficulty Min	Difficulty Max
7	**1**	391	1.591	-2.242	5.643
7	**2**	71	1.872	-0.716	5.071
7	**3**	86	2.076	-1.645	8.696
7	**4**	56	2.145	0.067	4.339
7	**Total**	604	1.744	-2.242	8.696

Because Items and Students are on the same scale across ALL grades, comparing item difficulty and student ability, **within grades**, is…a bit tricky (overall, *Theta* values are akin to z-scores – but, in the 265 page Tech Manual, I was unable to find within-grade distributions of student ability *Theta* values, or *Theta* values at the different Performance Levels – other than the item/student graph included previously.

Smarter/Balanced Technical Manual

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

24 out of 656 (3.6%)

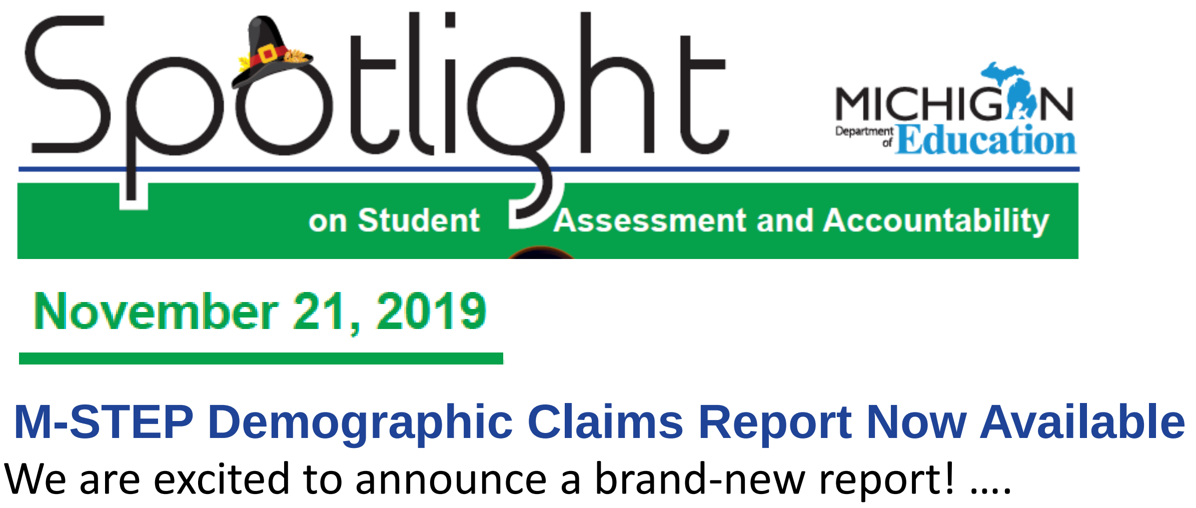
Gr. 7, 656 Items

500 out of 656 (76%)

"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)

This Item-Ability Mismatch has little impact on Proficiency, but can have a significant impact on:

1) Reporting Claim Scores
2) Measuring "Growth"



Spotlight

MICHIGAN Department of Education

on Student Assessment and Accountability

November 21, 2019

M-STEP Demographic Claims Report Now Available

We are excited to announce a brand-new report! ….

Well, glad someone is…

I don't wish to re-litigate the appropriateness of Claim Score Reporting

[if interested, look up my presentations from the Spring 2018 MERA Conference & 2019 MSTC]

MERA: http://merainc.org/spring-2017-presentations/

MSTC: http://gomasa.org/2018-michigan-school-testing-conference-resources/ (session E-1)

BUT, some highlights:

[along with the previously highlighted issues, i.e., the dearth of appropriate items for low-ability students]

For grade 7 Math:

- *Reliability* Coefficients:

  Claim 3: **.15**

  Claim 2& 4: **.48**

  -- not sure how it was justified, to report a score with a reliability index of .15…)

- Between Claims Correlations[1]:

  Claims 1--3: **1.87;** Claims 1—2&4: **.99;** Claims 2&4--3: **1.83**[2]

  -- pretty strong indication that reporting claim scores add no value to the reporting of the total score.[3]

As an interesting aside: the Technical Manual uses these correlations *not to caution against reporting Claim Scores* , but as *support for treating the test as unidimensional.*

[1] Corrected for *attenuation* (measurement error).

[2] For those of you that remember your Stats class, correlations range from -1.0 to 1.0…which make a correlation of 1.87 REALLY LARGE.

[3] And, could actually do harm, leading schools to mistakenly assume the instruction in one claim is weaker than instruction in another claim.

Item-Ability Mismatch

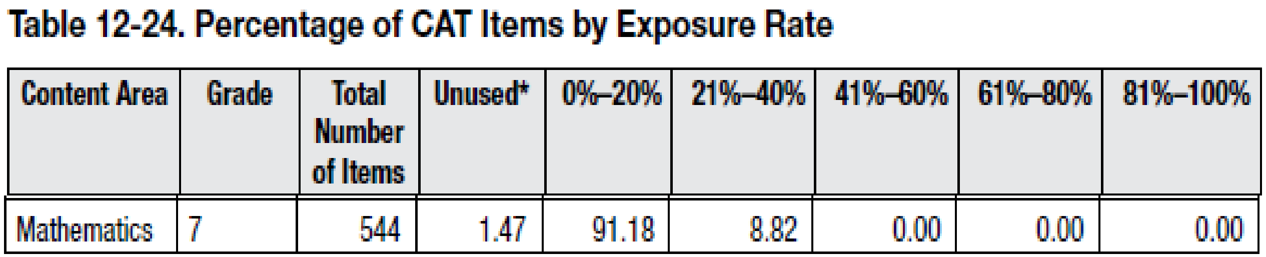


Table 12-24. Percentage of CAT Items by Exposure Rate

Content Area	Grade	Total Number of Items	Unused*	0%–20%	21%–40%	41%–60%	61%–80%	81%–100%
Mathematics	7	544	1.47	91.18	8.82	0.00	0.00	0.00

The Table "Highlights" the small % and number of items that have a large exposure rate

- Well, don't think this is saying much

  (or, don' think it's asking the right question)

  – the question that *should* need answered:

  > What percent of items are repeated to low achieving students?

  (or exposure rates *across achievement bands*)

  > How are standards represented within the Claims?

- 8.8% of the items are administered to (up to) 40% of the test takers

  – a given the small number of items that are "appropriate" for a good portion of the 7th graders, it seems reasonable to assume that the SAME items are administered to ALL low performing students

  -- Probably measuring a very restricted number of Standards

  -- and then, growth would be measured on that restricted number of standards (in "year 1") to a similarly restricted number of standards in "year 2."

Our New A-F Report Card

We're not really going to use AGPs and say we're measuring Growth, are we?...

Well, yes, we are

As with Claim Scores, I'm kinda over playing the "let me convince you" card

-- BUT (like I can help myself), will share a couple pieces from a previous presentation on AGPs

(can be found at: http://merainc.org/spring-2018/)

AGPs

Briefly,

-- AGPs, as a growth metric, represent the percent of students that meet their *Growth Target*

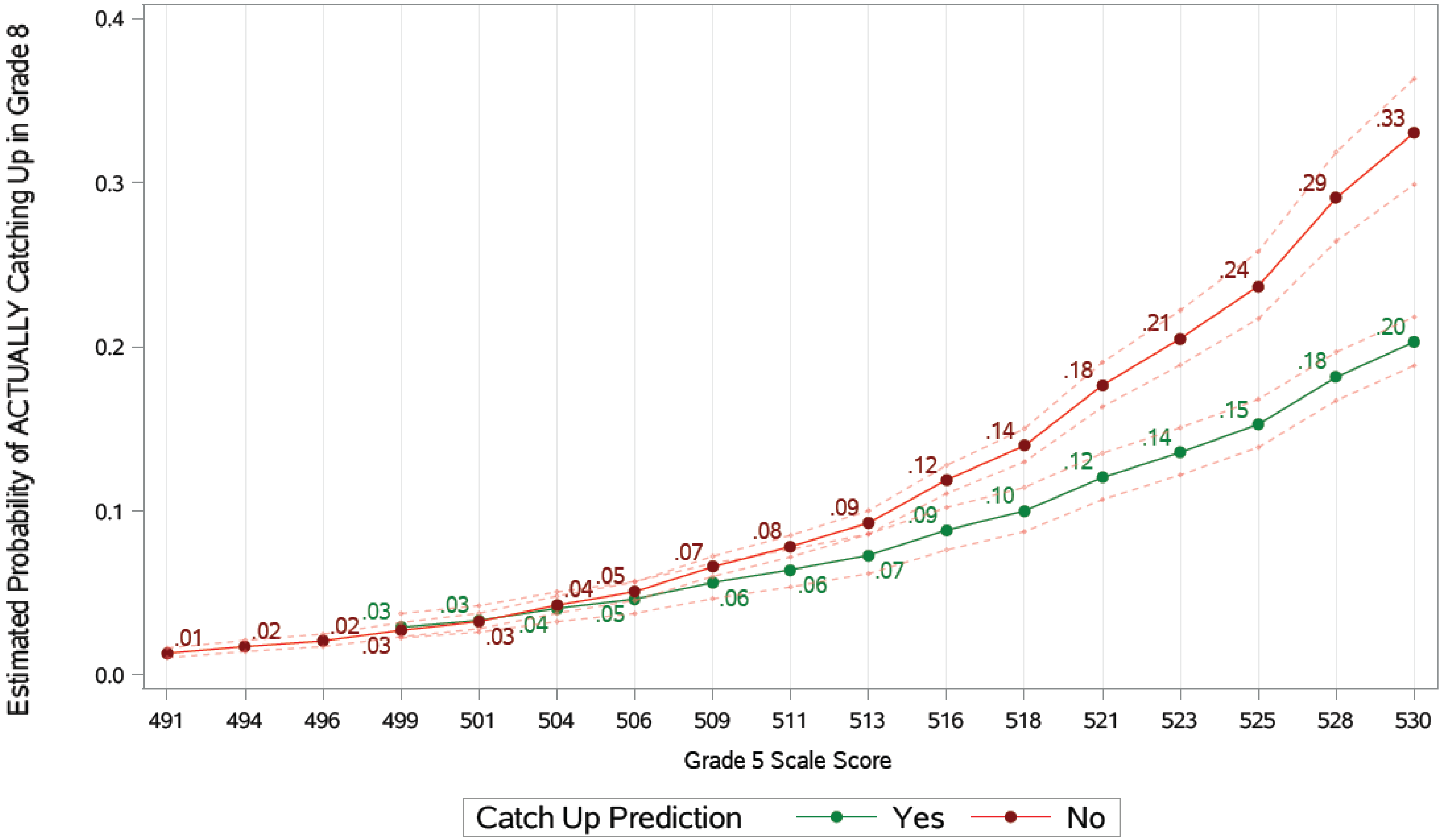
(or, the percent of students that are *On Track to Proficiency*)

-- In the SGP datafile, students are reported in two groups:

- Previous Year ***Not Proficient*** students:
  -- IF they meet their growth target, they are classified as "Catching Up"; if they don't, then "Not Catching Up"
- Previous Year ***Proficient*** Students:
  -- IF they meet their growth target, they are classified as "Keeping Up"; if they don't, then "Not Keeping Up"

Which non-proficient kids had a greater probability of reaching proficiency (actually catching up),
those who were projected to "catch up" or those who weren't?

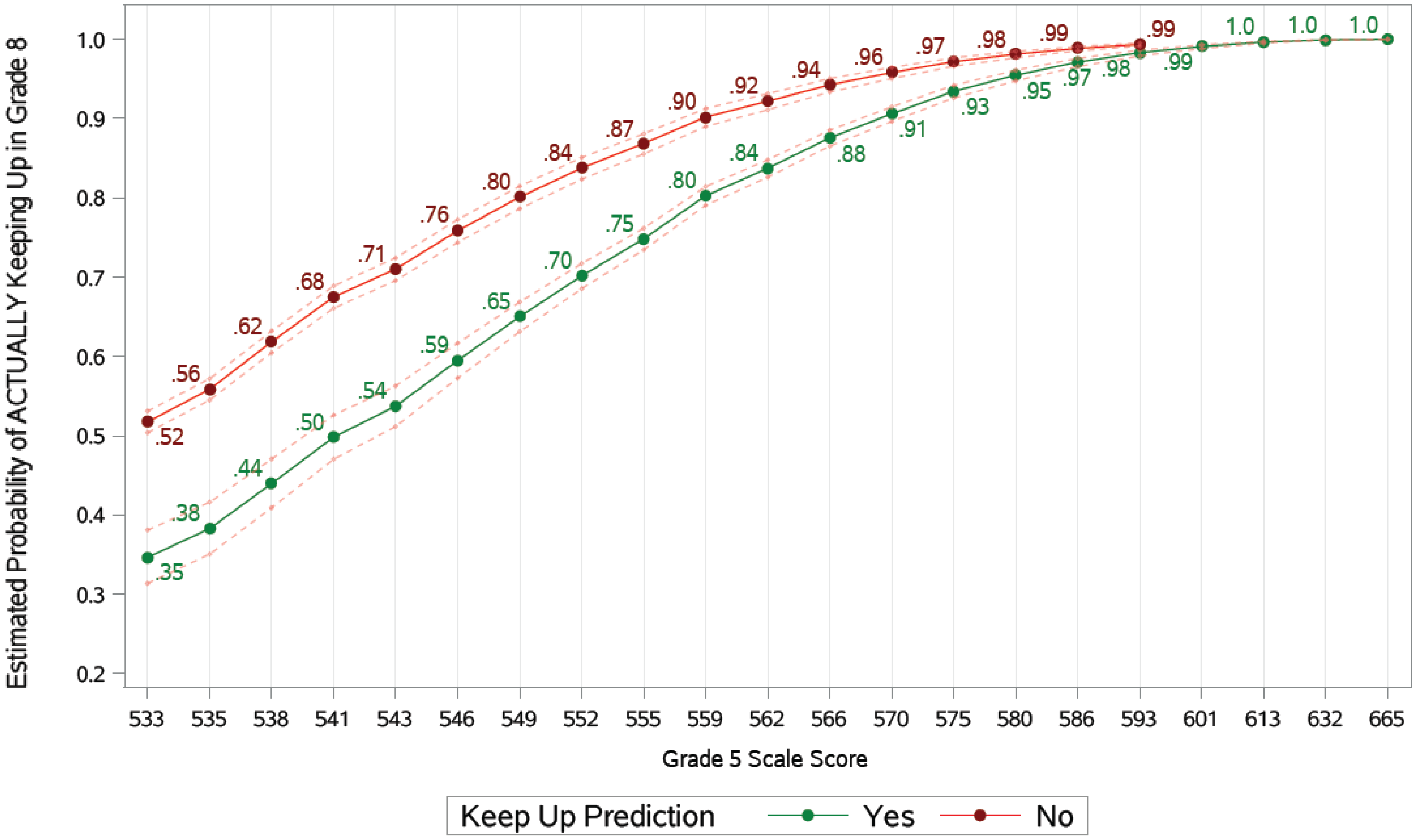
Probability of Proficiency, based on Scale Score and "Catch Up Status" (YES or NO)



Count																	
Catch-Up: YES	.	.	.	1	6	9	16	36	71	71	113	154	184	232	265	319	321
Catch-Up: NO	523	550	559	626	636	666	698	662	616	605	590	512	448	365	314	225	144

Which proficient kids had a greater probability of staying proficient, those who were "keeping up" or those who weren't?

Probability of , based on Scale Score and "Keep Up Status" (YES or NO)



Count

	533	535	538	541	543	546	549	552	555	559	562	566	570	575	580	586	593	601	613	632	665
Keep-Up: YES	41	38	113	190	236	330	334	382	433	496	479	509	537	510	559	510	490	417	327	186	102
Keep-Up: NO	511	509	516	470	446	414	327	257	230	174	146	140	94	58	29	15	1	.	.	.	.

"Catch Up: No" vs "Catch Up: Yes" and Grade 8 Proficiency

- Kids are 2.3 times more likely to reach proficiency in Grade 8, if they were classified AS NOT catching up, compared the to kids who were classified AS catching up.

  -- Kids expected NOT to catch up have, in fact, a 126% greater chance of catching up than the kids expected TO catch up.

"Keep Up: No" vs "Keep Up: Yes" and Grade 8 Proficiency

- *Kids are 1.1 times more likely to stay proficiency in Grade 8, if they were classified as NOT keeping up, compared the to kids who were classified AS keeping up.*

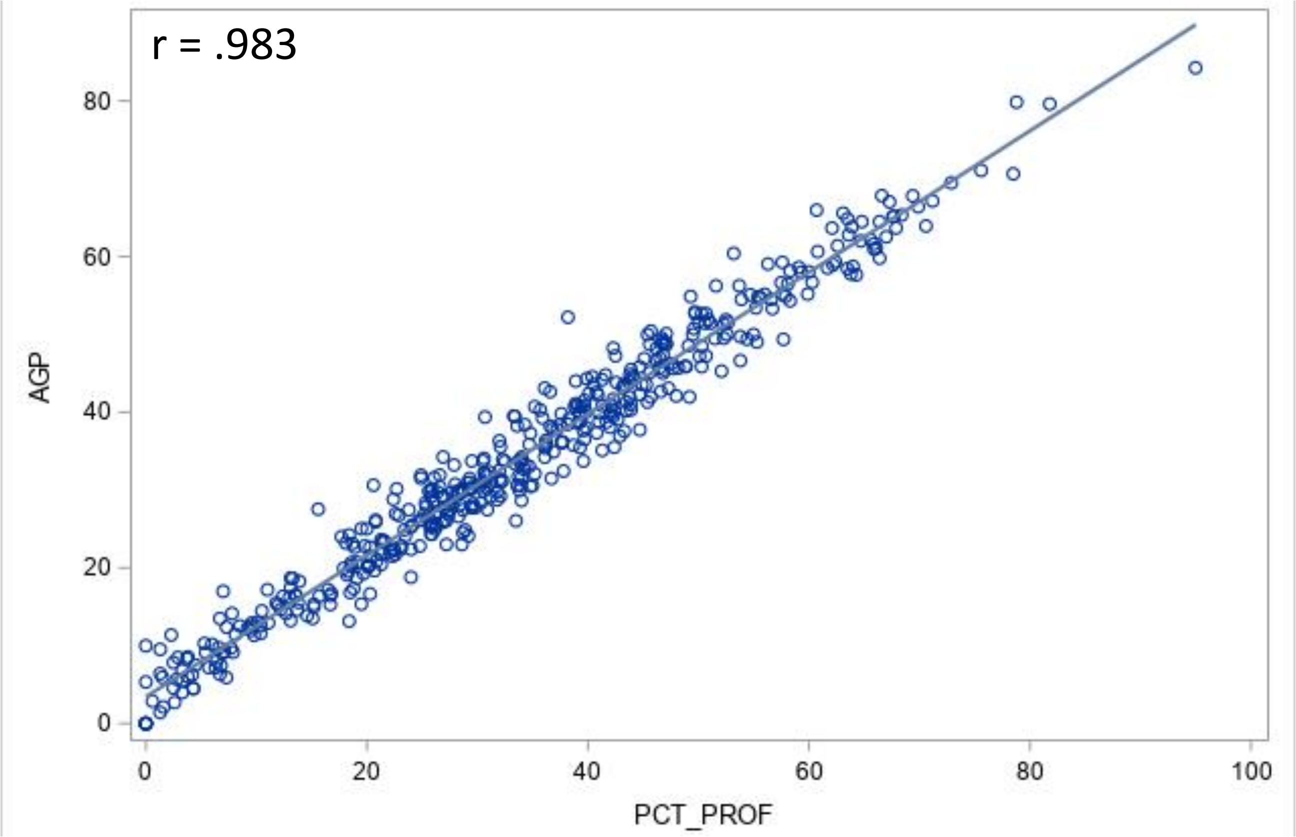
  -- It's nearly a crap-shoot-- with kids projected NOT to be Keeping Up slightly likely to be proficient than kids projected TO BE proficient.

So, What is the Growth Index (AGPs) Measuring?

The Michigan Index System labels this a "Growth Index"

Percent Proficient by ***Percent Growth Target Met***

Statewide, Middle Schools (n=471)



- 97% of the Variance between-School "Growth" can be explained by Percent Proficient

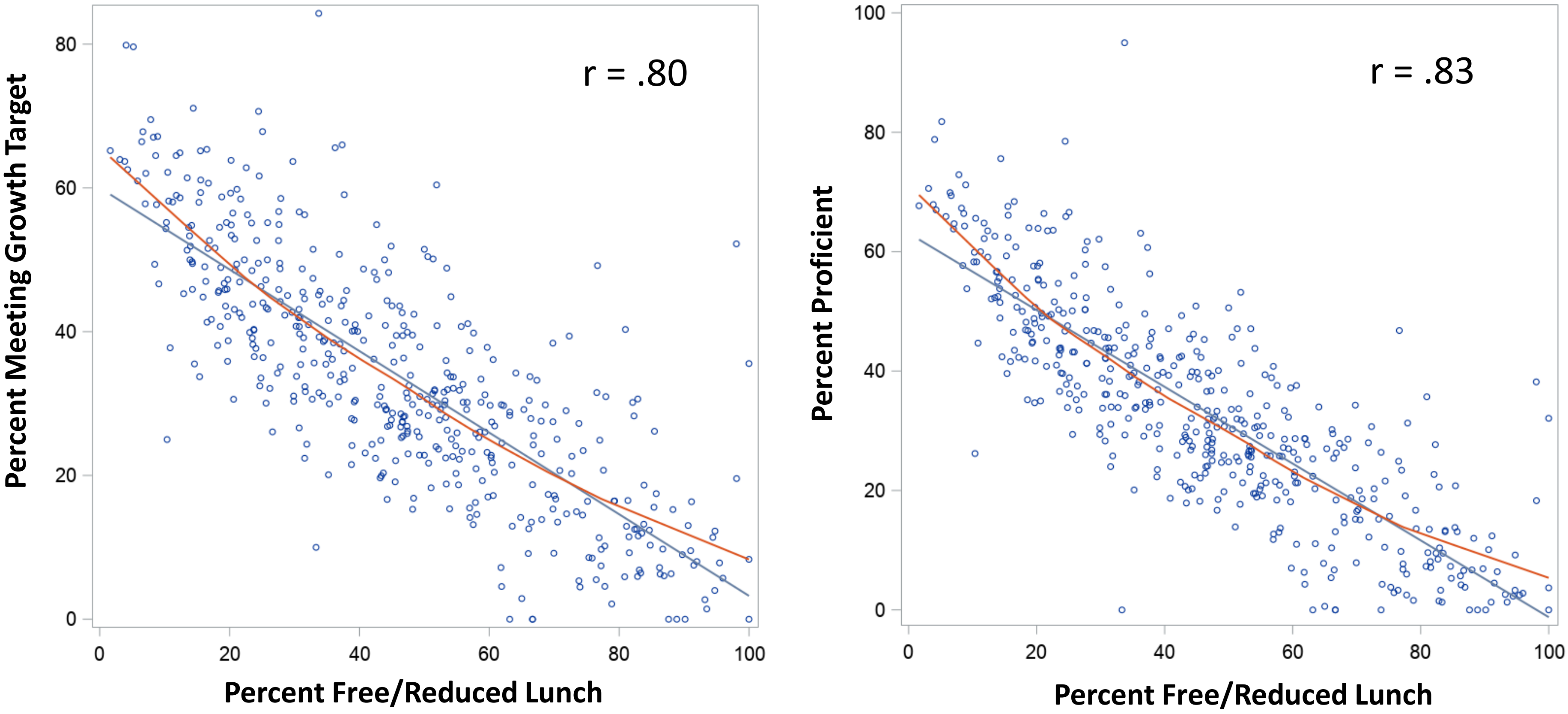
97% of the variance in between-school "Growth" can be explained by percent proficient

- To put this in perspective, a summary of 28 alternate-form reliability estimates of the SAT showed reliability values of .88 to .91
  -- so, only 80% to 83% of the between-test variance, on **parallel forms of the SAT**, is reflected in the other – on **two tests purported to measure the exact same thing.**

Almost axiomatic (given that 97% of stuff measured by "growth" can be explained by proficiency): the **relationship between growth and poverty** is pretty much indistinguishable – i.e., the same, not different, statistically equivalent, etc. – from the **relationship between proficiency and poverty**.

The relationship between growth and poverty and the relationship between proficiency and poverty

Subject = Math Type = Middle School



So, could we please stop the pretense that this Index is measuring "growth"?

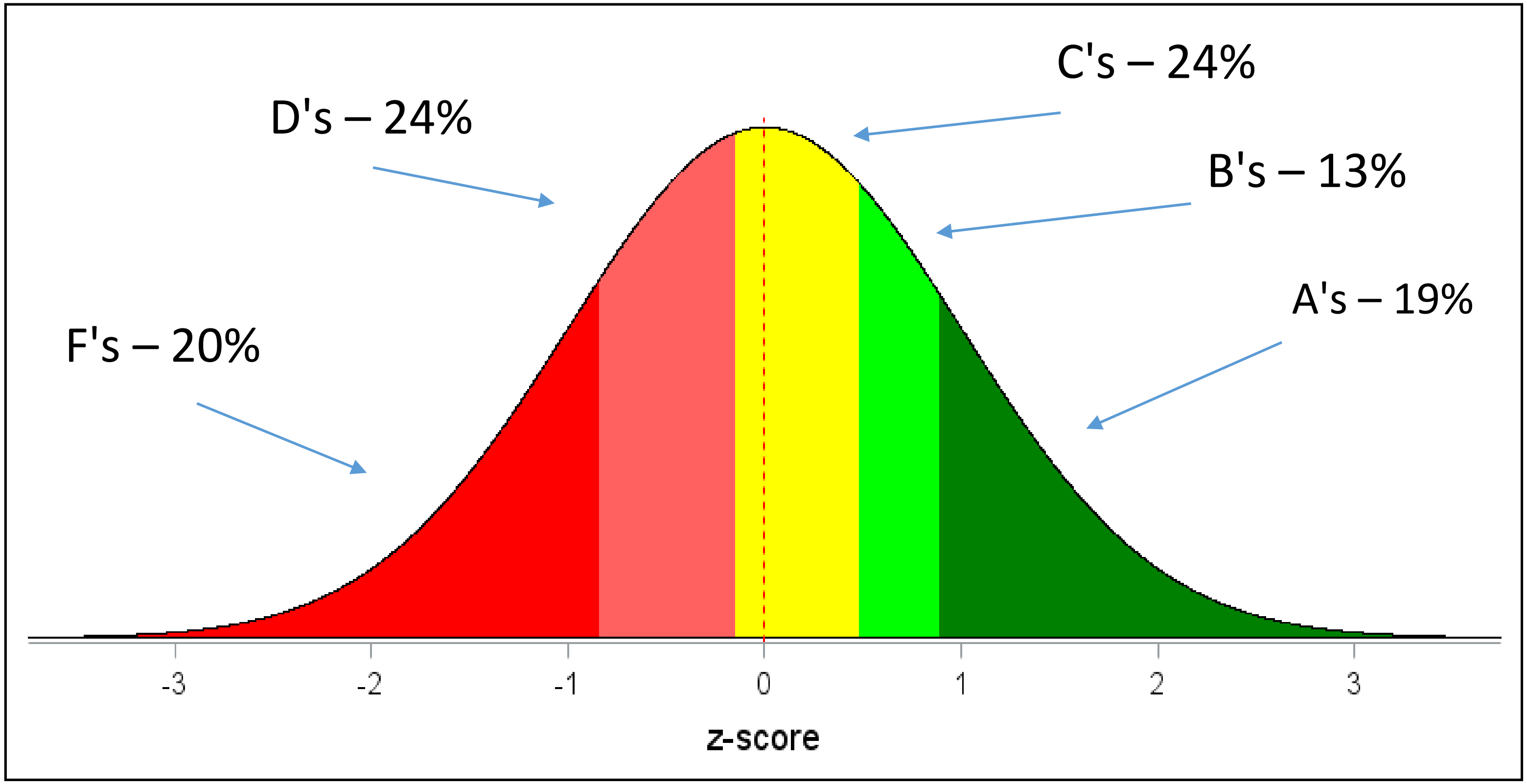
This information was available to MDE, but guessing not shared with the Task Force charged with creating the A-F scale

(and here's a surprise – I wasn't invited to join this get-together…)

Our New A-F Report Card

Comparison of Performance Amongst Peers

Would like to know, with a 100% *normative* metric, why the hell they would decide to have more D's (and F's) than B's (and A's)??



20% F's and 19% A's this year, and next year…and the next year…and into perpetuity…until, of course, things change (and, maybe, this silliness goes away?)

Rhetorical Question:

When was the last time you heard someone advocate for the use of "grading on a bell curve"?

Or, where in educational literature do you find, anywhere, that grading in a bell curve is an appropriate practice?

In closing,

Here's an idea to save taxpayer money

and

own up to what we are actually doing:

Percent Free or Reduced Lunch	Greater than 90%	70%-90%	30%-70%	10%-30%	Less than 10%
Letter Grade	F	D	C	B	A

And, It's Symmetrical!

Did we have fun, or what?