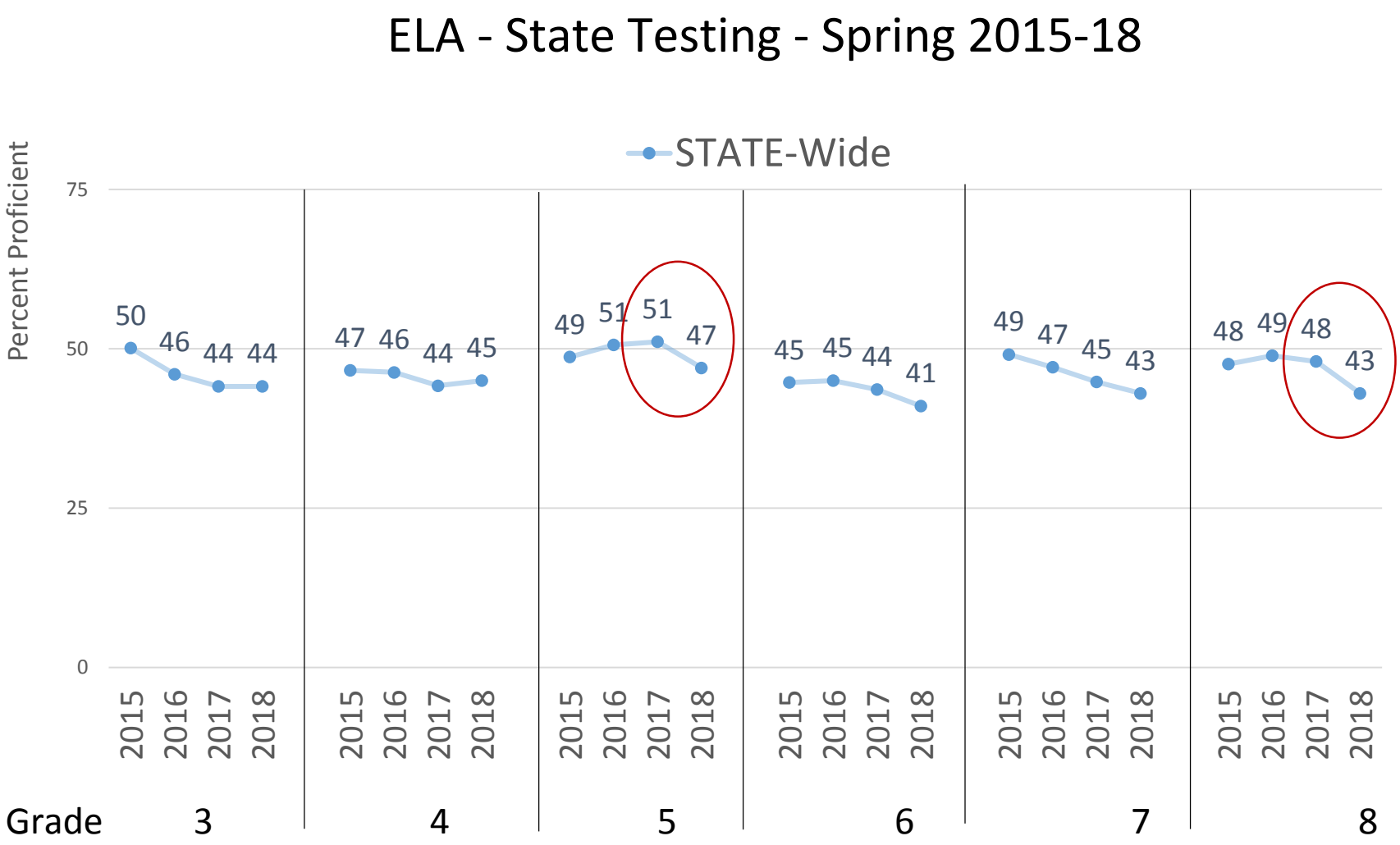
MERA Fall 2018 Conference

Dave Treder

What's up with this?



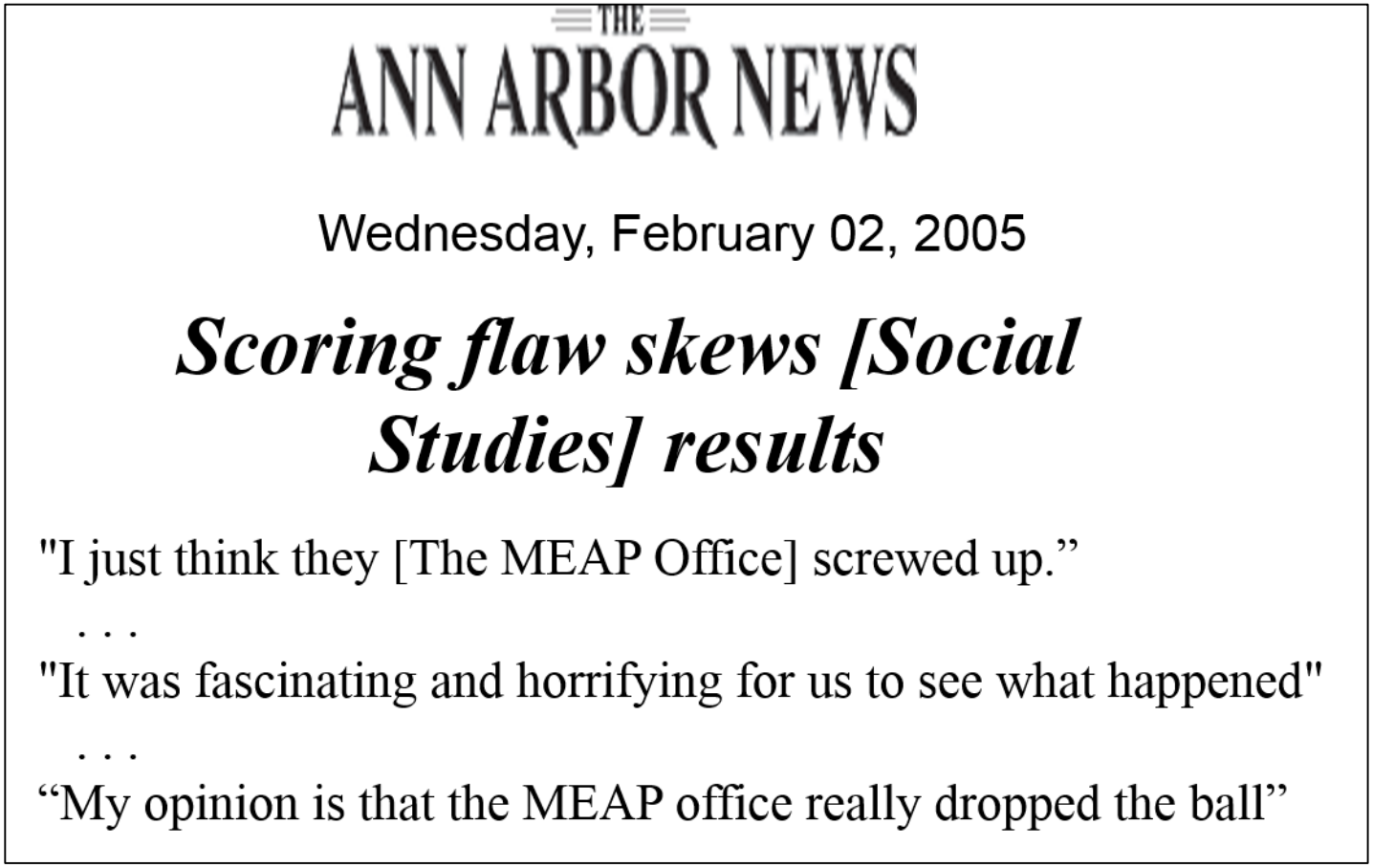
When you see a 5% change, from one year to the next, in 100,000+ kids, what do you think is most likely?

1. The "sample" of kids in year 2 are ~~smarter~~ *higher achieving* (or ~~dumber~~ *lower achieving*) than the kids in year 1?
2. The instruction/teachers in year 2 got better (or worse)?
3. The difference is simply do to the vagaries of the different types of *measurement error?*

OR,

4. *Something changed, in the test or the testing environment, from year 1 to year 2?*

From my *Michigan State Testing Conference 2005 Presentation*



THE ANN ARBOR NEWS

Wednesday, February 02, 2005

Scoring flaw skews [Social Studies] results

"I just think they [The MEAP Office] screwed up."

. . .

"It was fascinating and horrifying for us to see what happened"

. . .

"My opinion is that the MEAP office really dropped the ball"

[Grade 5 *Percent Proficient* dropped, from 31% to 25%]

…a bunch of slides explicating the equating of the 2004 & 2005 tests

From my *Closing Remarks*

- It's important to remember that Equating provides only an ***estimate***.
- As Educational Measurement becomes more complex, these types of procedures become less transparent.
- MEAP Office: "*A thorough review of the Equating was conducted by both the MEAP TAC and outside expert*"
  [NO documentation of these analyses is available to the public]
- It'd be nice if the State made public any possible "issues" and any analyses they conduct.

Opening slide of my 2016 *MERA Fall Conference* Presentation

Spotlight

MICHIGAN Department of Education

on Student Assessment and Accountability

December 17, 2015

Happy Holidays!

Comparability Between Online and Paper/Pencil Modes

Please note: We are aware that there is the potential for differences between scores of schools that administered M-STEP online, and schools that did a paper/pencil administration. This is not unexpected, given both the transitional nature of this first year of assessment and two administration modes. We are going to be conducting a statewide comparability study this winter and will present our findings in a future Spotlight.

Am I the only one who missed the presentation of findings?

…a number of slides explaining the problems with the equating of the On-Line & Paper-&-Pencil tests

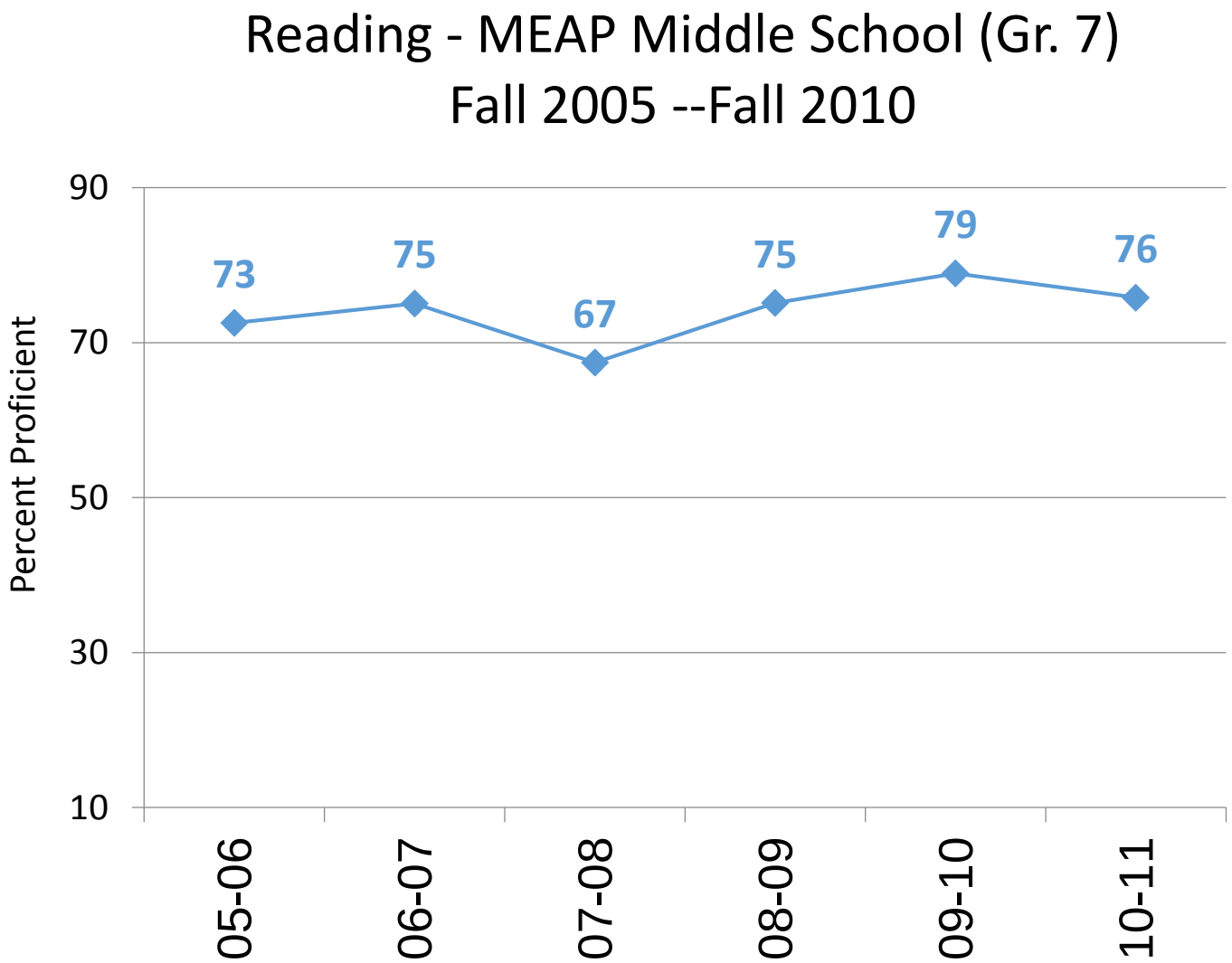
From my *Closing Remarks*

A number of conclusions can be reached (I'm sure more than I've listed here, and these are not meant to be either/or)

1) There was something different about the results, between the two modes of administration, which made the equating….problematic?

2) There was something pretty different in the CONSTRUCT they were actually assessing (which could well be highlighted in the previous point)

Other (relatively) extreme year-to-year Percent Proficient change?



No Presentation on this…was off tilting at other windmills

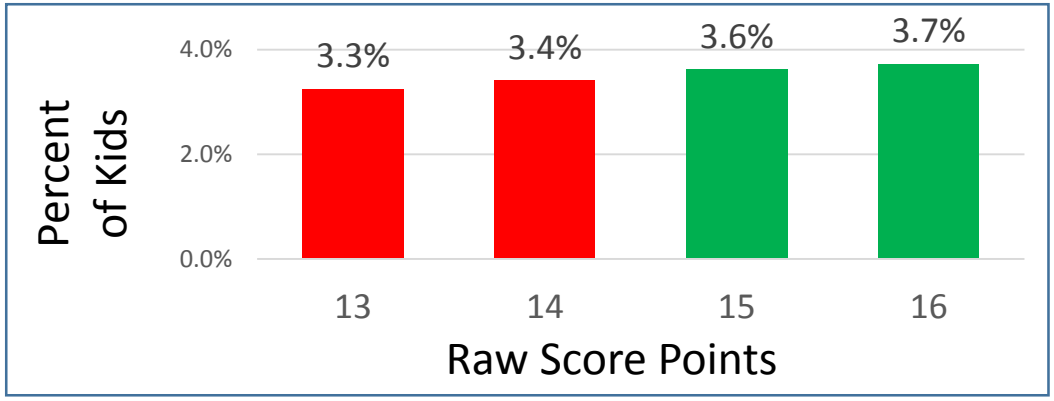
I'm thinking this is probably a more appropriate graphic

Or, Maybe, Just This…..

Anyways...

In looking at this 7 point drop in Percent Proficient:

1. The Fall 2005-09 ELA tests had only 29 items, with 3-4% of kids at each score point (near the cut score)
   - equating is done at the decimal level, while cut scores need to be integers

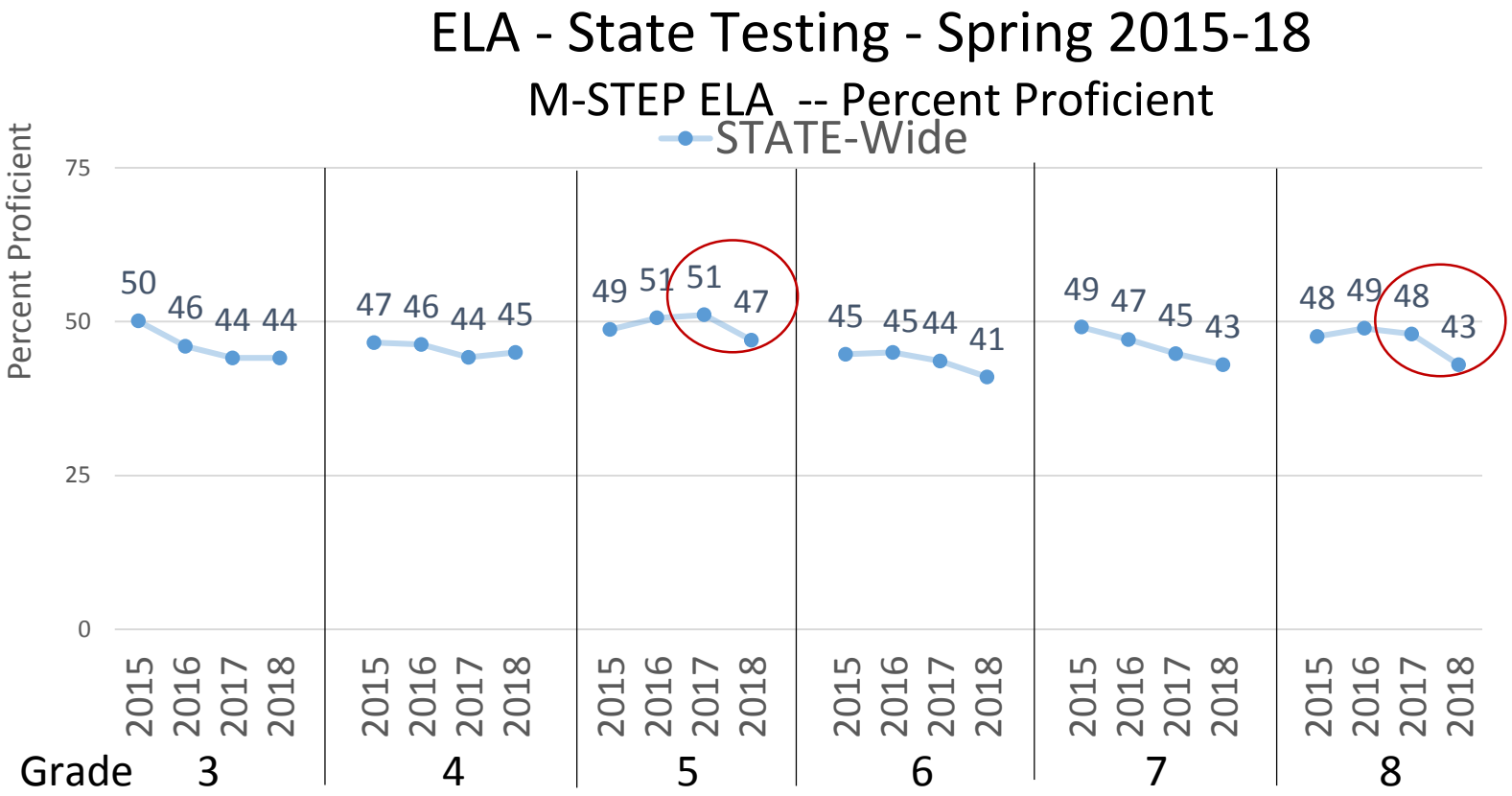


2. Equating error could well add another +/- 1 raw score point difference*
   - Common-item equating is exacerbated by "testlets" (multiple items with a single prompt)

Given (1) & (2), I wouldn't consider a 7 point change a huge deal

*For group-level scores, (1) measurement error in score summaries and (2) examinee sampling equating error both shrink as sample size increases. However, **error due to common-item sampling does not depend on the size of the examinee sample—it is affected by the number of common items used—so it could constitute the dominant source of error for summary scores**. The random selection of common items should be acknowledged in the analysis of a test and the arising error variance calculated for proper reporting of score accuracy

And back to the original point



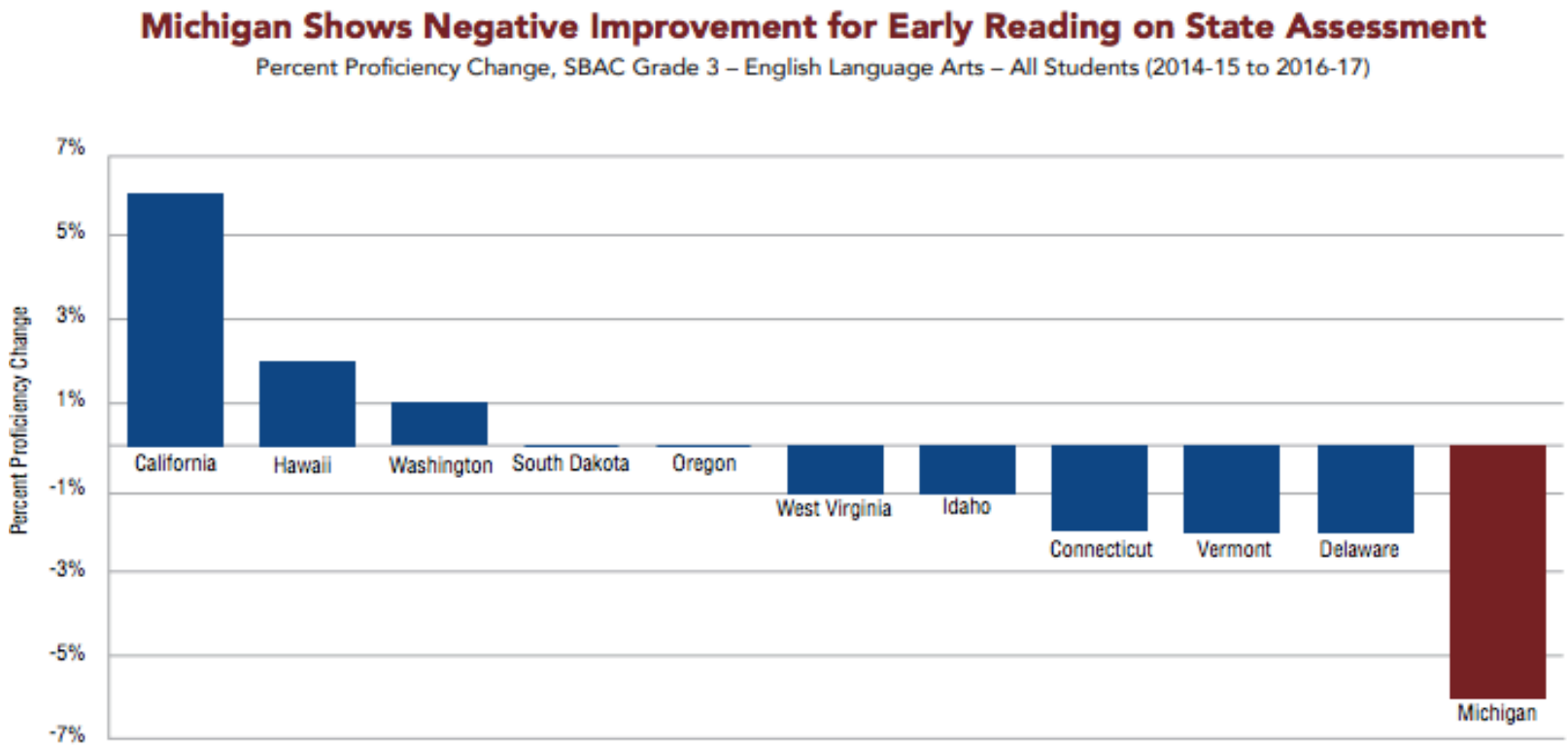
What changed, from Spring 2017 to Spring 2018, in grades 5 & 8?

-- 2017: A stand-alone *Performance Task*

-- 2018: An embedded Essay question (*Text-Dependent Analysis*)

Misleading Analysis Award:

Ed Trust: 2018 State of Michigan Education Report

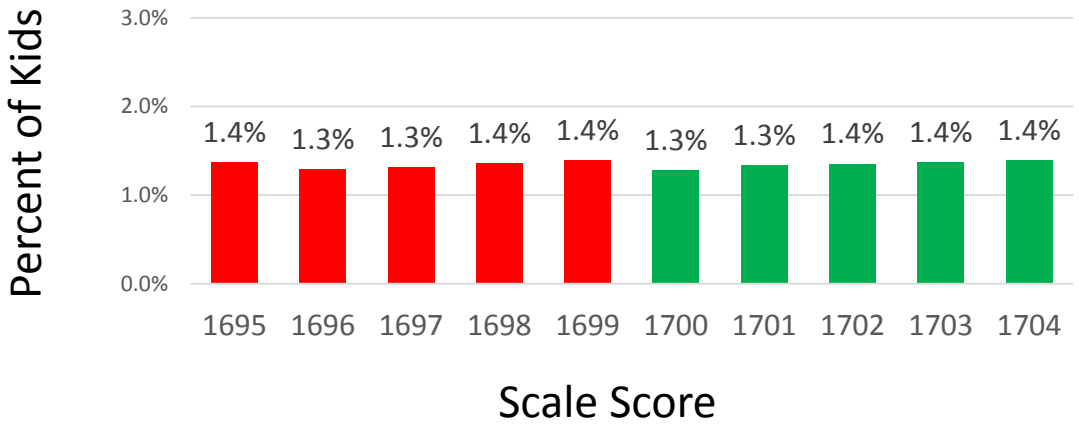


Source: Individual state score releases for 2014-2015 and 2016-2017

Note: Only states with two years of complete data results are included. Montana, Nevada, and North Dakota were excluded due to testing discrepancies in 2015 results. New Hampshire was excluded due to delays in reporting 2017 assessment results. Michigan's current statewide assessment system, the M-STEP, was designed by the Smarter Balanced Assessment Consortium (SBAC) as required by Public Act 94 in June 2014. Michigan is one of the 13 governing members that uses the SBAC assessment. Caution should be used when interpreting individual SBAC results across states, as each state has their own policies and procedures for assessment administration. Additionally, SBAC is also a relatively new assessment system, meaning longitudinal data will be important to continue analyzing longer-term trends.

5 Point Drop, 2017 to 2018

1. The "bunching" of kids at score points is much less a problem



2. "Equating" is all done *a priori*, so *equating error* becomes, in general, much less of an issue
   (the accuracy of the initial setting of item parameters and item parameter drift are a topic for another presentation…by someone smarter than me)

One way to look at the impact of removing the Performance Task (PT) and adding a Text-Dependent Analysis (TDA) essay question:

Preliminary Performance (PL) Level – based on Multiple Choice items only (2017 & 2018)
Final Performance Level (PL) – with the *Performance Task (2017) and the TDA (2018)*

Specifically,

1) *How does the addition of a stand-alone PT change the Preliminary PL in 2017?*

2) *How does the addition of an embedded TDA effect the Preliminary PL in 2018?*

3) *What's the relationship between these?*

point of clarification: in 2018, the item parameters of the multiple choice items are "fixed" to previously determined values; the TDA is "scaled" during the administration – this means that the *average* Scale Score should remain approximately the same (between the Preliminary & Final scale scores), but individual scale scores can fluctuate.

DATA

2017 & 2018 grades 5 & 8 ELA M-STEP Results

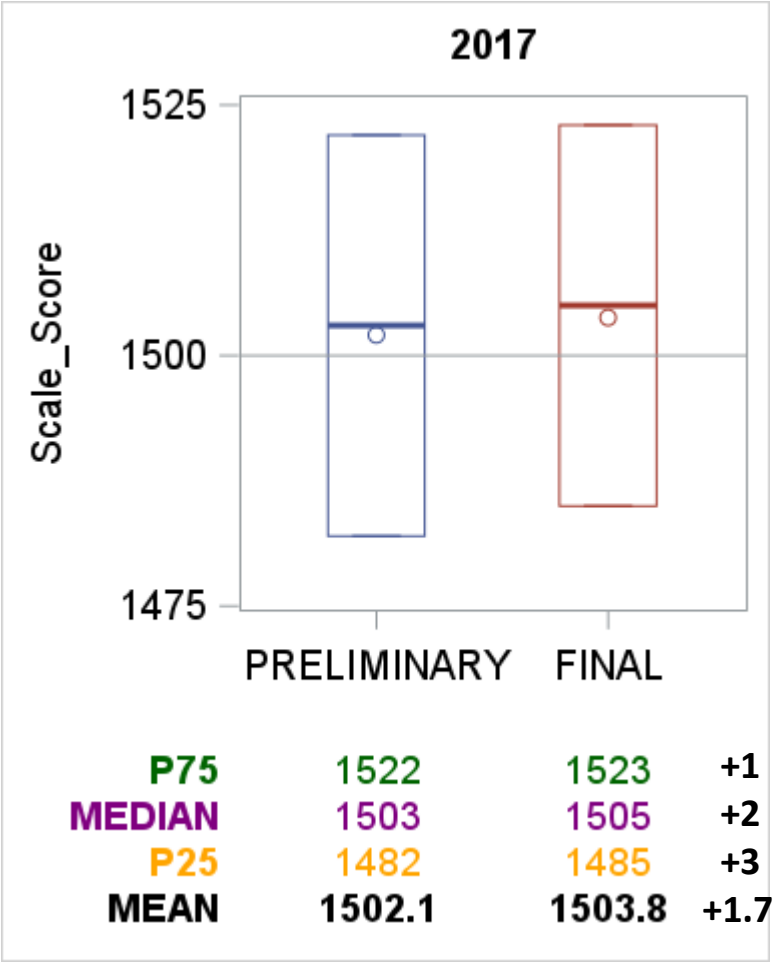
-- From Colleagues in Genesee, Lapeer, Macomb, Oakland, Ottawa, & Shiawassee ISDs

≈ 33,000 kids per grade

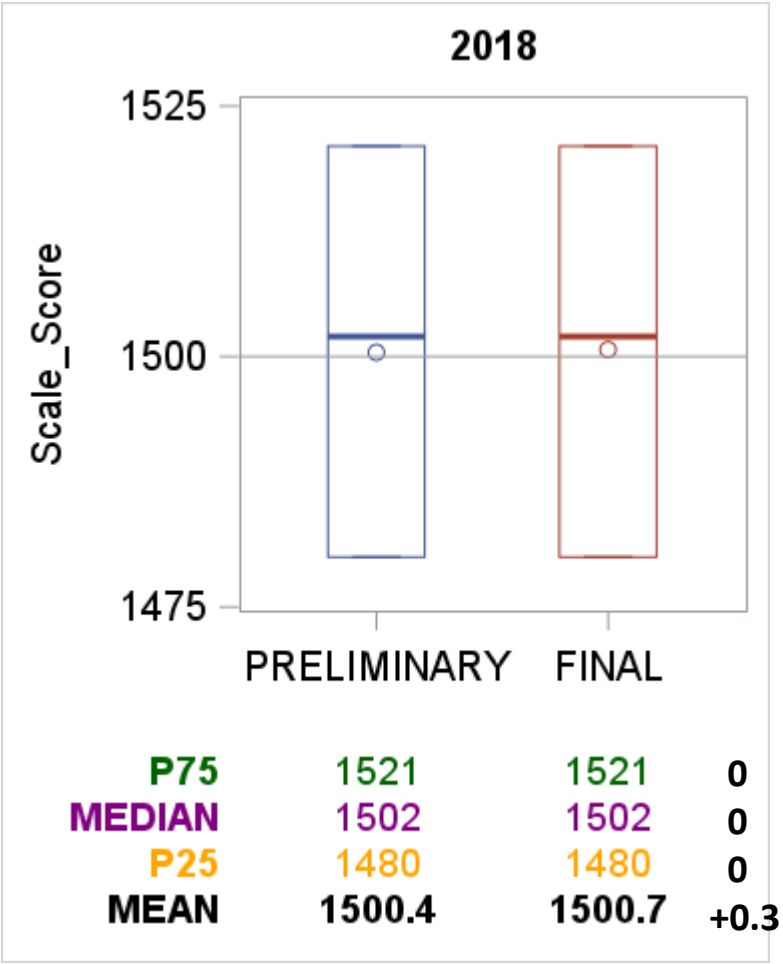
-- From BAA – by-Student Preliminary Scale Scores (Scale Scores that were computed with Mult Choice Items only)

Scale Score Difference, Preliminary to Final

Grade 5



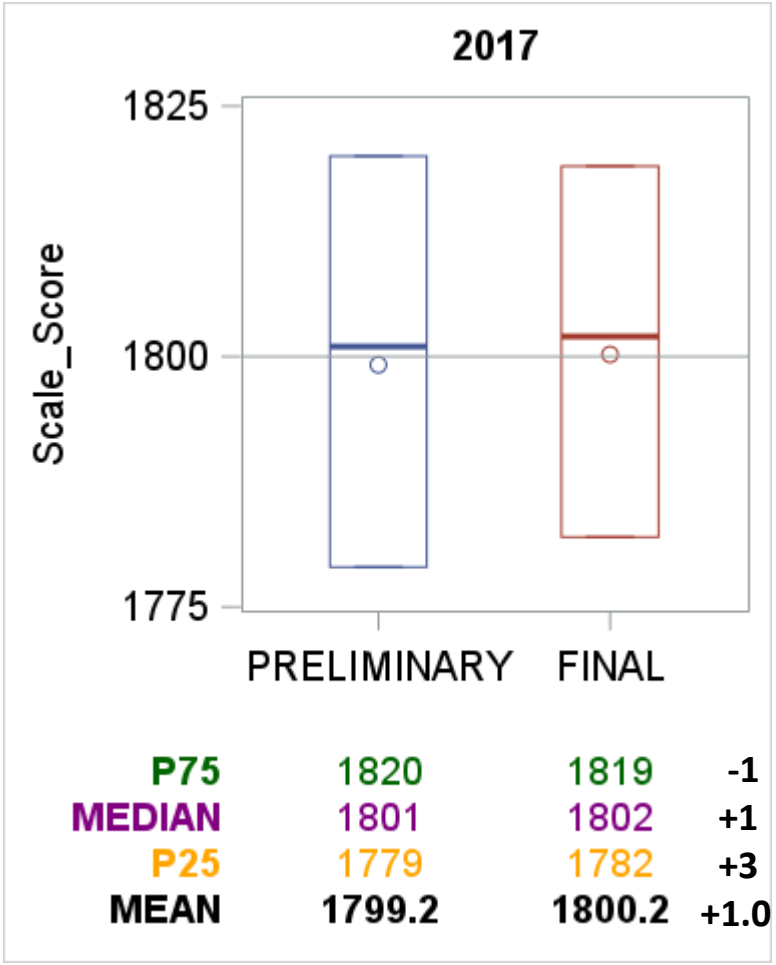
P75	1522	1523	+1
MEDIAN	1503	1505	+2
P25	1482	1485	+3
MEAN	1502.1	1503.8	+1.7



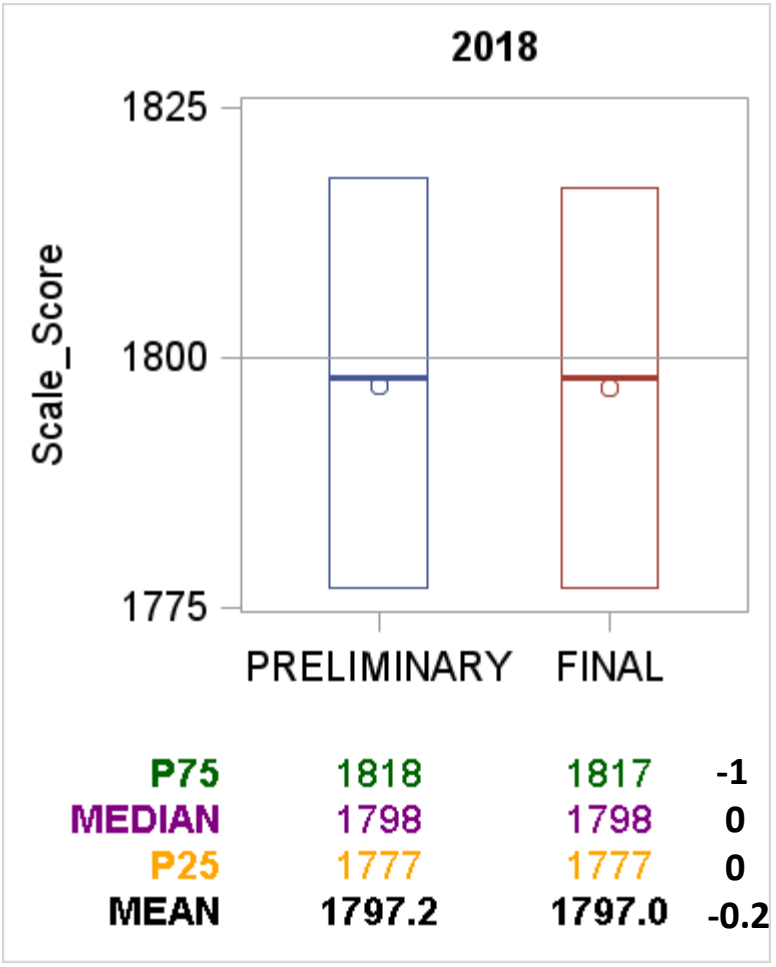
P75	1521	1521	0
MEDIAN	1502	1502	0
P25	1480	1480	0
MEAN	1500.4	1500.7	+0.3

Scale Score Difference, Preliminary to Final

Grade 8

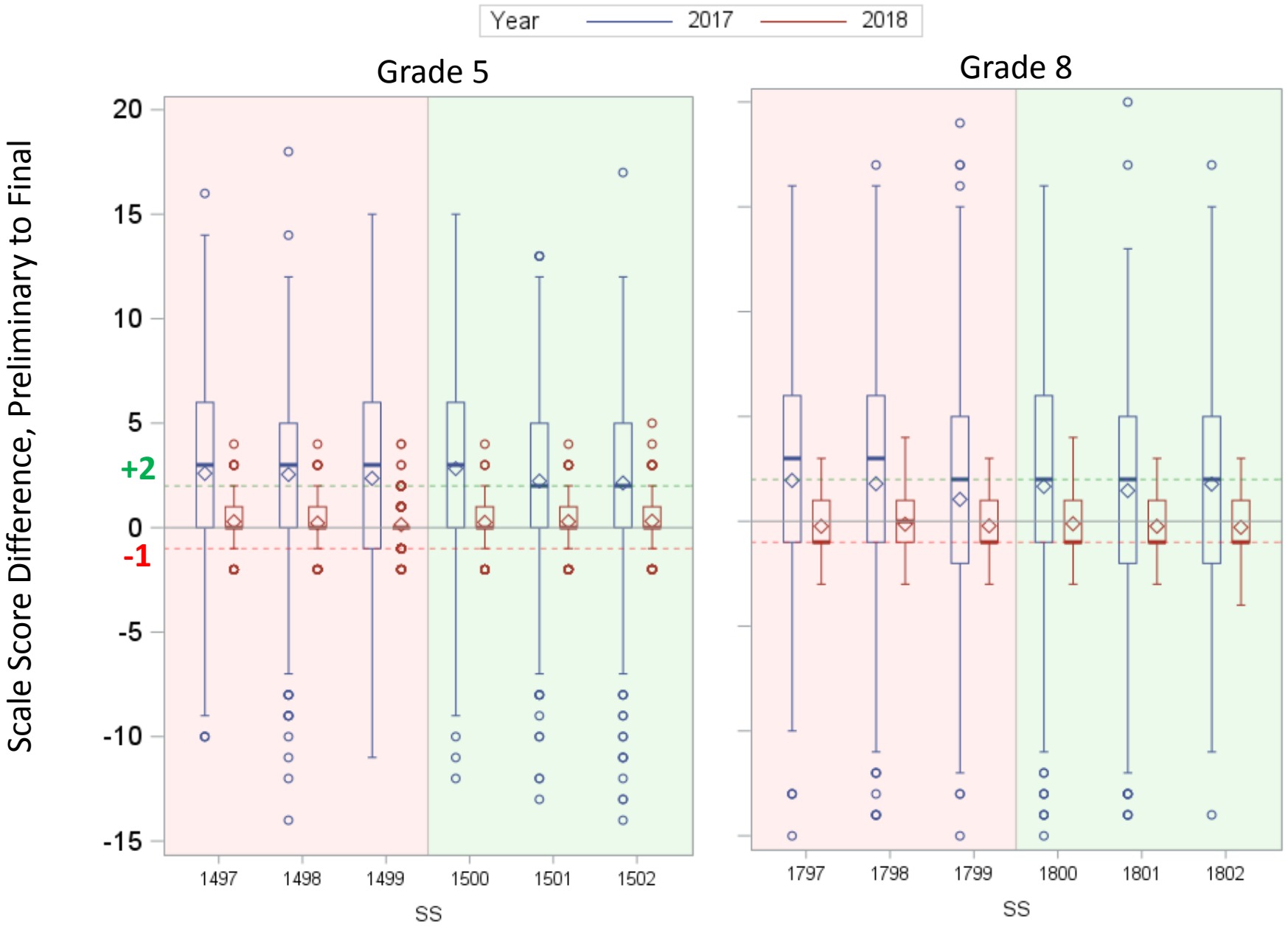


P75	1820	1819	-1
MEDIAN	1801	1802	+1
P25	1779	1782	+3
MEAN	1799.2	1800.2	+1.0



P75	1818	1817	-1
MEDIAN	1798	1798	0
P25	1777	1777	0
MEAN	1797.2	1797.0	-0.2

Change, from Preliminary to Final Scale Score, Close to the Cut-Score?



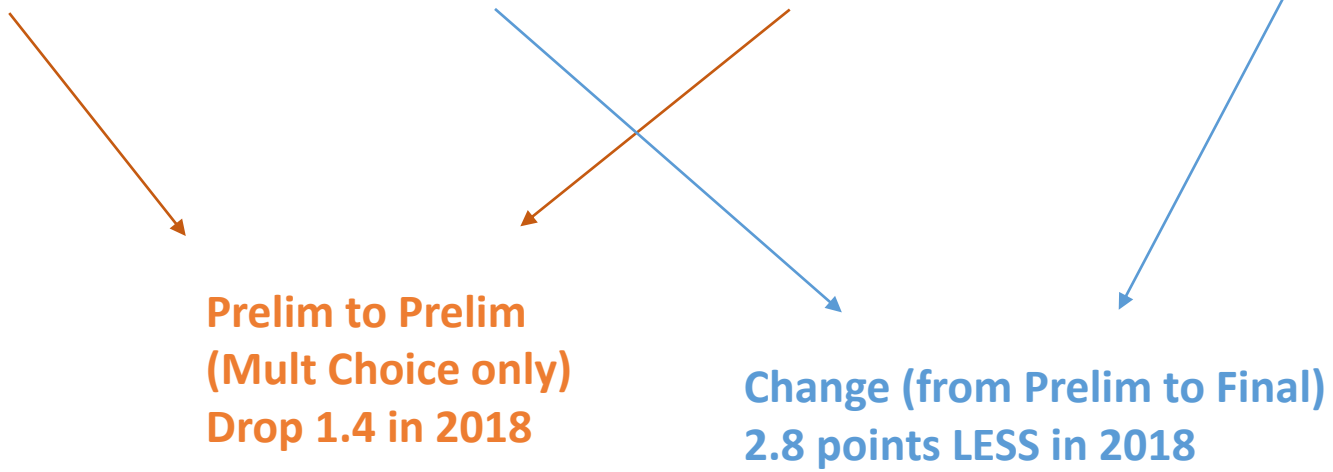
Looking at Percent Proficient, Based on:

1) Preliminary Results (Multiple Choice items only)

2)Final Results

GRADE 5

Spring 2017			Spring 2018		
Preliminary	FINAL	Change	Preliminary	FINAL	Change
54.51%	57.61%	3.10%	53.08%	53.44%	+0.34%



Summary:

1) Percent Proficient dropped 1.4%, on the Mult Choice (only) items, from 2017 to 2018

2) Percent Proficient change, from Prelim to Final, was 2.8% less in 2018 than 2017

Looking at Percent Proficient, Based on:

1) Preliminary Results (Multiple Choice items only)

2)Final Results

GRADE 8

Spring 2017			Spring 2018		
Preliminary	FINAL	Change	Preliminary	FINAL	Change
51.26%	53.09%	+1.83%	48.64%	48.51%	-0.13%

Prelim to Prelim (Mult Choice only) Drop = 2.6

Change (from Prelim to Final) 1.96 points LESS in 2018

Summary:

1) Percent Proficient dropped 2.6%, on the Mult Choice (only) items, from 2017 to 2018

2) Percent Proficient change, from Prelim to Final, was 2% less in 2018 than 2017

In Closing -

What happens when you replace

1) A ***stand-alone*** Performance Task
(on which, based on the data, students did better than the pilot data would indicate)

With

2) An ***embedded*** Essay question
(an item on which, based on the data, students found difficult)

Does it seem reasonable to "equate" the scores on these two tests?

Text Dependent Analysis Scores

Grade	_4	_3	_2	_1	_0		B	I	M	T
5	1%	5%	21%	52%	21%		0%	11%	9%	0%
8	1%	6%	26%	53%	13%		0%	5%	7%	0%

B = Blank I = Insufficient M = Off Purpose T = Off Topic
_0 = sum of B, I, M,