The 2017 NAEP Results: Why A Full, Public Data Release Matters[1]

EMBARGOED prior to the public release of 2017 NAEP Results[2]

The Johns Hopkins Institute for Education Policy
Dr. Alanna Bjorklund-Young, Senior Research and Policy Analyst
Dr. David Steiner, Executive Director

- For the first time in its history, the 2017 National Assessment of Educational Progress (NAEP) was administered online instead of by paper-and-pencil (a "modal change").
- The National Center for Education Statistics (NCES), which conducts the NAEP assessments, also administered a small sample of the 2017 NAEP *via* paper-and-pencil. There were important differences from the results of the two testing modes.
- The modal differences confirm earlier research indicating that lack of prior experience with online assessments correlates with lower scores. As our analysis found, prior to 2017, students in different states had had different exposure to online assessments.
- NCES faced a challenging situation: stakeholders rely on the NAEP to establish the national trend line *and also* state-level results. While NCES equated the two results to maintain the national trend line, this created difficulties for reporting state-level results with full accuracy.
- Our preliminary review of NCES's process indicates that the state-level results, made public today, should be treated with caution.
- So that all stakeholders have access to the full data sets from which to draw educational conclusions, we urge NCES to make public all the data related to the 2017 NAEP assessment.

[1] *The opinions expressed in this memo represent solely the judgments of The Johns Hopkins Institute for Education Policy and do not necessarily reflect the views of The Johns Hopkins University, The Johns Hopkins School of Education, or the Maryland State Board of Education.*

[2] This analysis is embargoed until the NCES releases the 2017 NAEP results.

Introduction

Today, the National Center for Education Statistics (NCES) released the 2017 NAEP results for 4th- and 8th-grade Reading and Math assessments. NAEP is often regarded as the gold standard of America's K-12 academic assessments, and progress or decline in the nation's results over time is a critical measure of our efforts to give the greatest possible educational opportunities to all of our students. For the 2017 NAEP, however, the customary state-level comparisons to prior tests and across states are not straightforward. This is because the NCES changed the exam mode for both Reading and Math from a paper-and-pencil test in 2015 to an online version in 2017. Indeed, it understandably took NCES an additional six months to score the 2017 NAEP, due to this important and timely transition.

During this period, NCES worked hard to minimize the risk that the change in testing mode impacted its capacity to report the national trend line accurately. It reports today that this trend line is essentially flat from 2015-2017 in fourth- and eighth-grade Reading and Math. This is not good news. We further note the deeply disturbing increase in the performance gap between our more privileged, more strongly-performing students, and their less-privileged peers.

We hope that the release of the following technical memo will become at least partially redundant, because NCES will choose to release all of its data - especially the results comparing the performance, by state, of students' results in 2015 with the performance, by state, of the small sample[3] of students in 2017 who took the NAEP in a pencil-and-paper form – just as everyone did in 2015. This is the critically important, apples-to-apples comparison. Our current understanding is that NCES has chosen to share the state-level, pencil-and-paper 2015-2017 results in the following form: each state will receive only its own results, with a note clarifying that they are "not official statistics." Inevitably, for reasons we will explain, this will have the effect of incentivizing some states to release these results, and others, not.

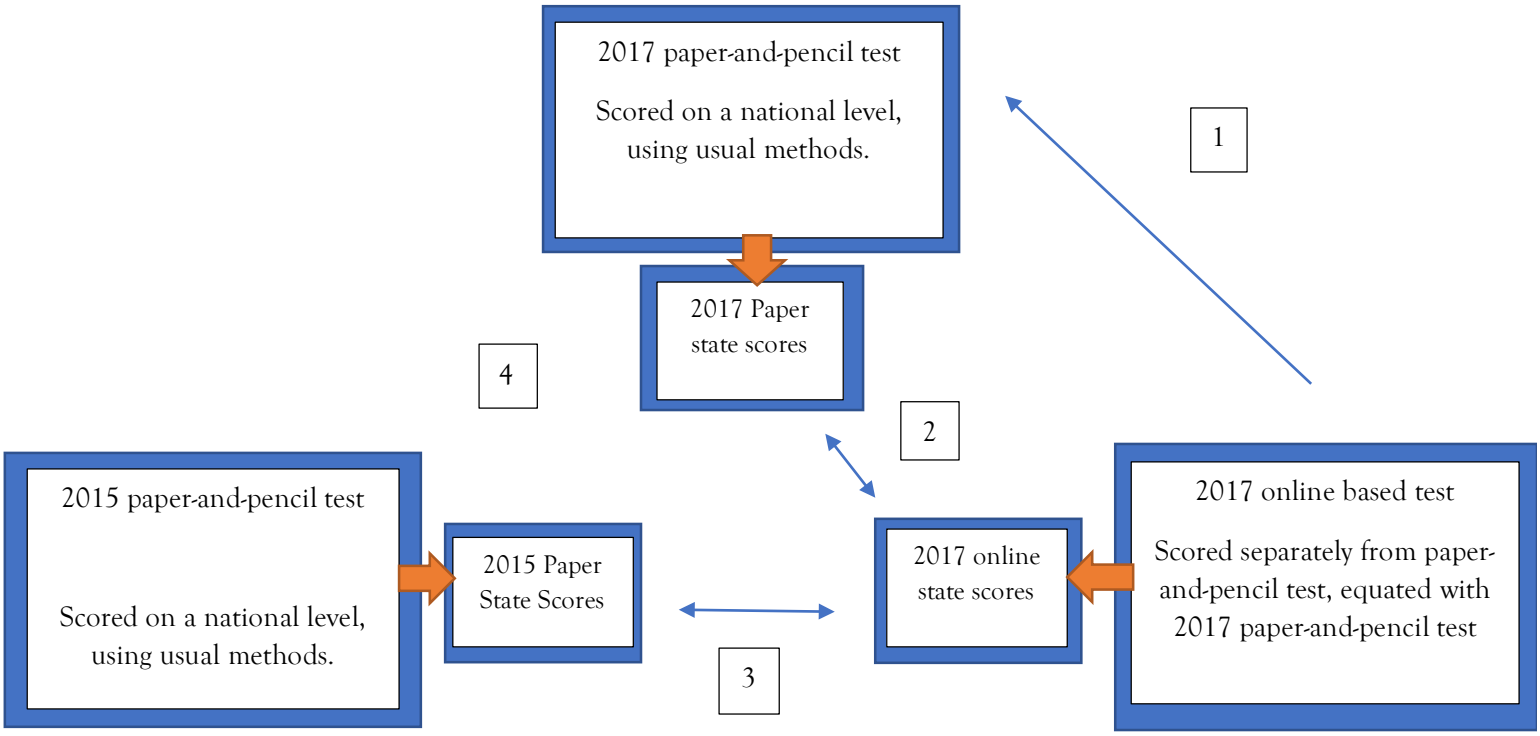
We based our analysis on the available facts as we know them. Naturally, the full public release of all data would enable the policymakers, researchers, and indeed all stakeholders, to undertake the most complete analysis possible.

We fully understand that the management of the modal transition from paper-and-pencil to online testing in 2017 presented NCES with unique challenges, and we are not criticizing the fact that it chose to treat the data in such a way as to preserve its capacity to report on national trends. Our concern, evidenced below, is that if NCES does not release the full, paper-and-pencil, state-by-state results in 2017 (along with the subgroup comparisons of the same data), the 2017 reported results will not fully capture the state-level or sub-group academic outcomes, but rather conflate them with other factors (such as a state's history in the use of online testing).

The second part of our memo examines the relationship between students' experience with online testing and reported changes in NAEP scores from 2015 to 2017. We did all the checking we could to distinguish those states in which few (30% or less) of 4th-grade and 8th-grade 2017 NAEP test-takers had previously experienced an online assessment, and those states in which 70% or more had done so.

[3] We assume that this is a sample of randomly selected students. We have not seen details of this, however.

Part 1: How NCES established and reported 2017 NAEP scores



PROCESS

1. The National Center for Education Statistics (NCES) scored its 2017 paper-and-pencil sample using the usual methods on a national level (i.e., equating across administrations). NCES scored the 2017 online sample of the tests separately, and then used equating to make sure that the distribution of 2017 online scores matched the distribution of 2017 paper-and-pencil scores, on a national level.[4]

2. NCES then compared the state-level, average scores on the paper-and-pencil samples to the state-level, average scores from the newly equated online sample. **It is our understanding that NCES found that twenty-one of these differences, affecting nineteen states, were statistically significant**. NCES points out that fewer than half of these statistically significant differences would actually reverse the direction of the states' results (from negative to positive or positive to negative).

[4] For each grade and subject.

But large movement on either side of the positive/negative axis is still important. For example, to state leaders, being four points versus only one point up, or three points down versus flat, is a very serious difference.

NCES also compared online and paper-and-pencil scores from different subgroups at the national level, and it is our understanding that only one subgroup in one assessment showed a statistically significant difference.

Even beyond the 21 statistically significant differences, differences in results between the online and paper-and- pencil results can be important educationally and politically.

In our hypothetical example: State X had a 3-point difference between its average, paper-based score and its average (adjusted), online-based score. While this might not be a statistically significant difference, this is certainly an educationally and politically meaningful difference.

3. *Because of the modest number of statistically significant differences between 2017 state-level, paper-and-pencil scores and 2017 state-level, adjusted online scores, and more importantly to preserve its capacity to report the national trend data, NCES has only reported the online scores.* These 2017 online scores are then compared to 2015 paper-based scores. NCES has also reported the resulting gains or losses and ranked states accordingly, with indications of statistical significance.

   In our hypothetical example: *In 4th-grade Reading, NAEP's 2017 results will show that State X dropped 4.6 points from its 2015 paper-and-pencil scores to the reported 2017 online scores. This is marked as a significant drop.*

4. *However, NCES is not reporting the apples-to-apples comparison of 2015 paper-based scores to 2017 paper-based scores, except to provide each state with its own results and a note indicating that these are not "official statistics."* We respectfully urge NCES to make all information fully public.

   In our hypothetical example: In *4th-grade Reading, State X dropped 1.6 points from the 2015 paper scores to the 2017 paper scores. This is not a statistically significant drop.*

   ***Key Takeaway**: The difference between a 1.6-point drop and a 4.6-point drop is very important to a state.* The first is a mildly disappointing result, which a state would review by examining education policy in the context of known demographic and economic developments since 2015. A drop of 4.6 points, however, suggests the need for an in-depth look at the strategies a state has employed and risks undermining confidence in all that a state has undertaken.

5. Our hypothetical State X is, in fact, a real state: Louisiana.

   Our story recounts what has actually occurred. Louisiana went from what would have been a non-statistically-significant drop of 1.6 points if its apples-to-apples comparison had been done, to a statistically significant drop of 4.6 points based on the apples-to-oranges comparison. Our understanding of the data we have seen, subject to clarification and correction, is that a number of states have experienced a similar decline on one or more NAEP assessments - from a non-statistically significant drop to a statistically significant one. We also understand that two states would have shown a statistically significant decline in the apples-to-apples comparison but, instead, will be reported as having no statistically significant change at all. This suggests that many states are substantively affected by the differences between their online scores and their paper-based scores. Certainly, some states will show a smaller difference from the testing mode change than did Louisiana. But even a one- or two-point change in either direction matters educationally and politically, even if not statistically, in part because NCES has historically emphasized these rankings and numbers.

6. NAEP will report 2017 results by ranking states' movement from 2015 to 2017. The differences between one state and the next (above or below it) are almost never statistically significant, yet they will be regarded as significant by the broader audience, and state-level gains or losses of three or more points will be taken as evidence of policy successes or failures. Given this reality, we urge state-level policymakers to treat any publicly-reported state results – positive or negative – with real caution.

Part 2: The 2017 Mode Impact

The Johns Hopkins Institute for Education Policy conducted a preliminary analysis of the relationship between the changes in NAEP scores[5] (from 2015 to 2017)[6] and states' prior experience with online test-taking. Compared with states whose students had had prior experience with online testing, states whose students had *not* had prior experience[7] witnessed a larger reduction in average NAEP scores from 2015 to 2017 in 4th-grade Reading and Math and in 8th-grade Reading. These differences are statistically significant.

[5] Note that at the time of writing, NAEP has not yet released the final form of its 2017 data. Therefore, this analysis was conducted with preliminary data that show changes in scores rounded to the nearest quarter of a point. When NCES officially releases the 2017 NAEP scores, we will verify our analysis and provide any necessary adjustments.

[6] Note that all "change in NAEP scores" or "gains on the NAEP" refers to the point change in a state's score from the 2015 NAEP to the 2017 NAEP in the same subject and grade level. Further note that NAEP reports out scores for each state, Washington D.C., and the Department of Defense. For this analysis, we include each state and Washington, D.C.

[7] Defined as 30% or less of students took online exams in 2016. A list is provided in Table 1. We made every effort to confirm the accuracy of this data, including contacting every one of the 50 state education agencies and Washington, D.C., and conducting multiple web searches. We took a conservative view of state allocation. For example, we understand that Alaska administered some tests online in 2015, prior to a highly truncated online test administration in 2016. We thus counted the state as having had experience with online testing, even though including it as a paper-and-pencil state would have strengthened our findings.

We find that the relationship between previous experience with online testing and changes in NAEP scores is stronger in the 4th grade than in 8th grade, and stronger in Reading than in Math. These results suggest that prior, in-school online testing experience—i.e. taking an online state test and the in-school practice leading up to the online test—is particularly important for younger students. This might be because many younger children do not otherwise gain the kinds of skills necessary for online testing. These results also suggest that in-school, online testing experience is especially important in Reading, possibly because of the typing and editing skills required by the Reading exams.

Our analysis shows that in 4th grade, only 9% of states[8] that had used paper-and-pencil state testing in 2016 experienced any gains in Reading and 9% of such states showed any gains in Math, between 2015 NAEP and 2017 NAEP. This is striking when compared to the 48% of states in Reading and 35% states in Math that *had* used online state testing in 2016 and also registered gains on the 2017 NAEP. In fact, the average differences in scores on NAEP from 2015 to 2017 were approximately 2 points higher in Reading and 1.6 points higher in Math for states that had used online testing in 2016 compared with states that had used paper-and-pencil state tests. We find that the relationship between prior state-testing mode and NAEP gains is robust: the association remains statistically significant and stable[9] across different models that control for various state characteristics.[10]

A visual representation of the relationship between states' 2016 testing mode and 2017 NAEP gains is shown in Graph 1 below. *Note that in 4th-grade Reading, whether or not a state had used online testing prior to the 2017 NAEP predicts roughly 15% of the variation in NAEP score changes. Students' prior online state-testing experience explains approximately 11% of the variation in 4th-grade NAEP Math score changes.* By contrast, a state's poverty level is not significantly related to changes in 4th-grade NAEP scores and predicts less than one percent of the variation in 4th-grade NAEP score-changes, as shown in Graph 2. We similarly find that none of the other state characteristics we included in our models produced similar magnitudes and statistical significances, or explained as much variation in 4th-grade NAEP score-changes, as did prior experience with online testing.

The relationship between 2016 state-test mode and 2017 NAEP score gains is neither as strong nor as consistent in 8th grade.[11] Students with prior experience with online state testing do, on average, have approximately 1.1 more points on the 8th-grade Math NAEP test, and this relationship is stable[12] and generally statistically significant when other state characteristics are controlled for.[13] However, there is no significant relationship between state test mode and 8th-grade Math NAEP gains. Graph 1 shows these

[8] This translates into only one state in Reading (Tennessee) and only one such state in Math (Wyoming).

[9] The estimated association between states with experience in online testing in 2016 and changes in the 4th-grade Reading 2017 NAEP test, range from 1.9 to 2.1 across multiple models, depending on the other state controls we include, and are all statistically significant at a 1% level. The estimated association between online testing in 2016 and the change in the 4th grade Math NAEP test range from 1.5 to 2.2 across multiple models and are all statistically significant at least a 5% level.

[10] Including the state's 2016 poverty rate, the percentage of black students in the state, the percentage of Hispanic students in the state, pupil-teacher ratios, the number of students in the state, and the number of charter schools in the state.

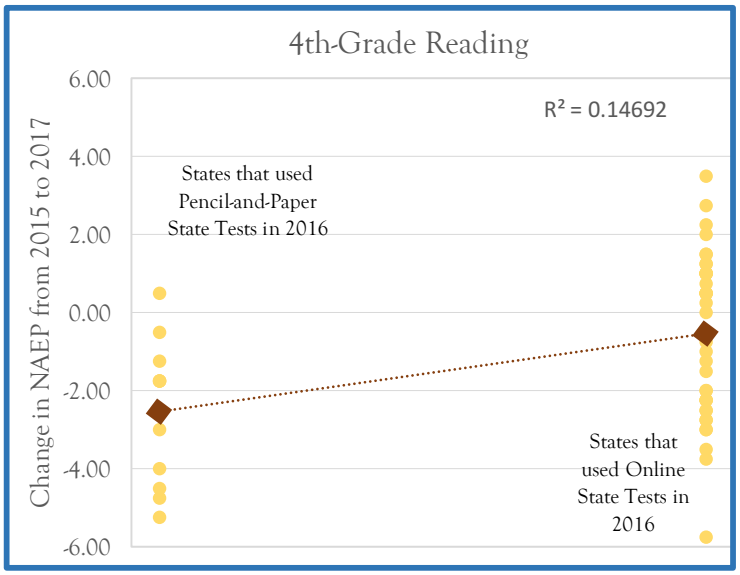
[11] As it is in 4th grade.

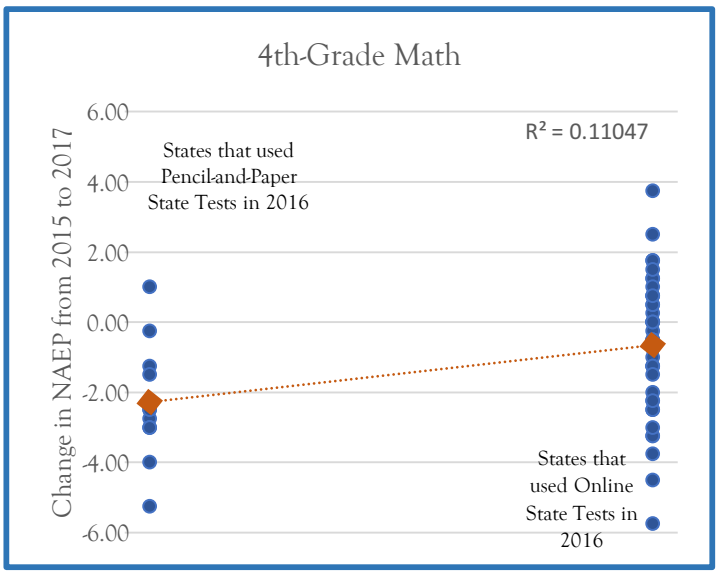
[12] The estimated association ranges from 1.1 to 1.7.

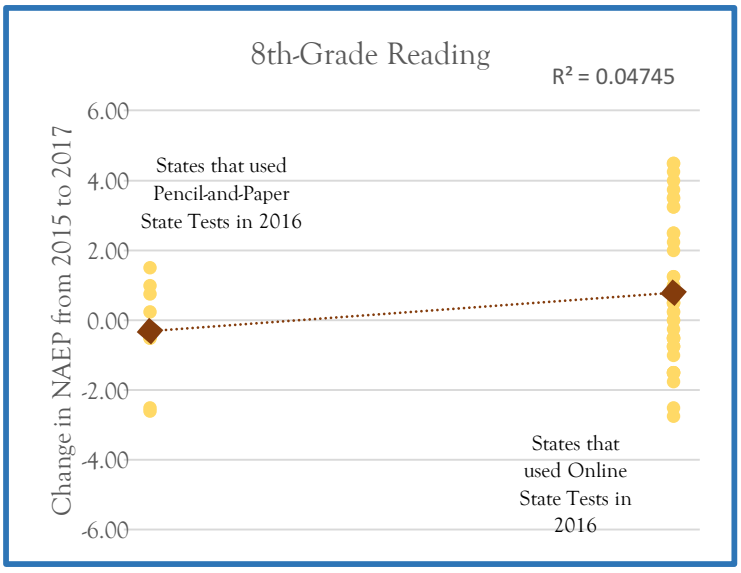
[13] Including the state's 2016 poverty rate, the percentage of black students in the state, the percentage of Hispanic students in the state, pupil-teacher ratios, the number of students in the state, and the number of charter schools in the state.

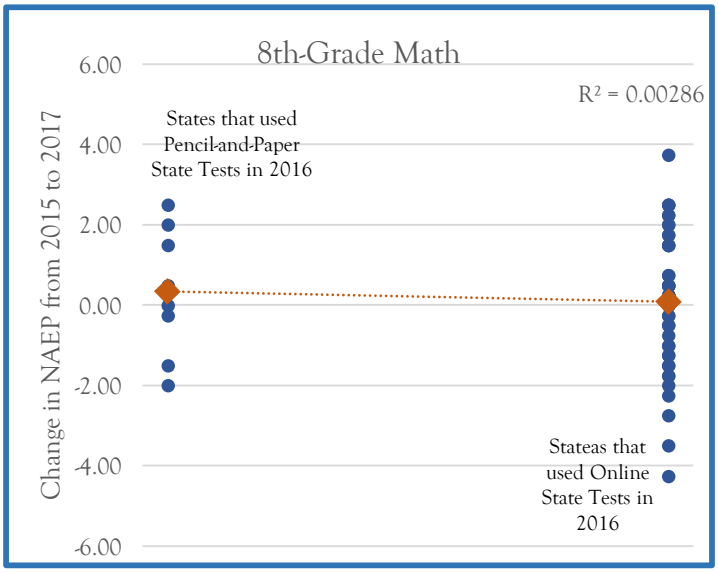
weaker relationships. Approximately 5% of the variation in the 8th-grade Reading NAEP test changes are explained by state test mode in 2016. Test mode explains no variation in the 8th-grade Math NAEP changes.

Figure 1 Graphs: Changes in NAEP Scores from 2015-2017: Comparing States with Pencil-and-Paper State Tests to States with Online State Tests Pre-2017 NAEP[14]



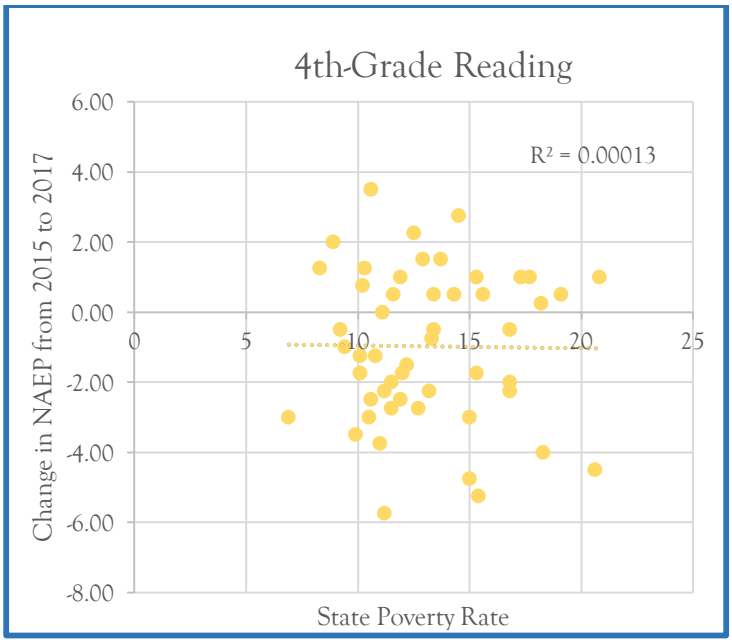


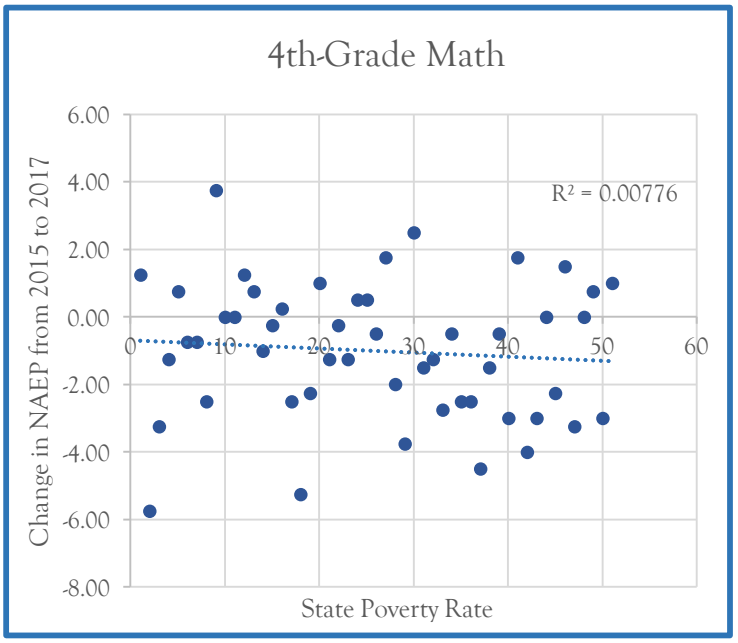


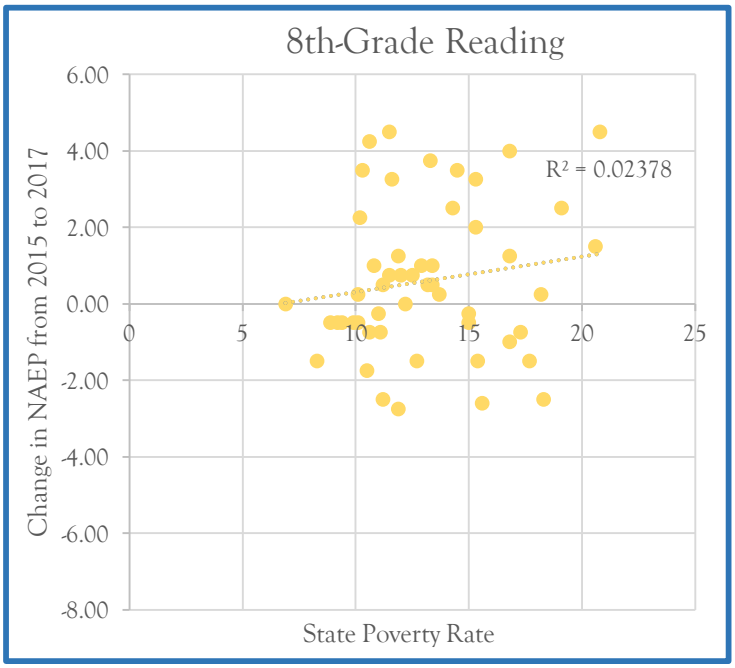


[14] Washington, D.C., is represented in these charts as a state.

Figure 2 Graphs: 2017 NAEP Scores and State Poverty Rates







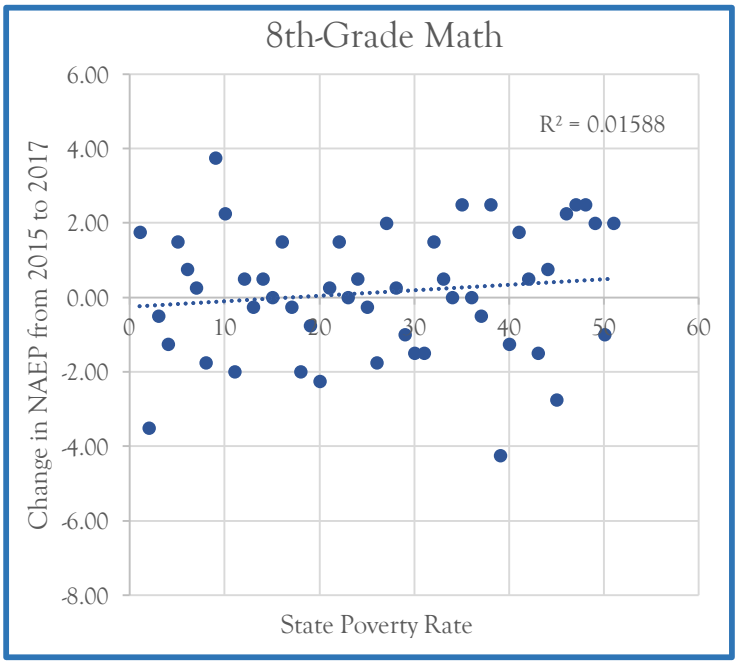


Table 1: States in which the majority (70% or more) of students experienced only paper-and-pencil-based assessments in 2016	
Elementary	**Middle School**
Iowa	Iowa
Kentucky	Kentucky
Louisiana	Louisiana
New York	New York
North Carolina	Pennsylvania
Oklahoma	South Carolina (Reading only)
Pennsylvania	Tennessee
South Carolina	Texas
Tennessee	Wyoming
Texas	
Wyoming	