Deconstructing the 2015 NAEP Results
Summary of a Symposium

The 2015 results of the National Assessment of Educational Progress (NAEP) for grades 4 and 8 were greeted with both disappointment and frustration. Despite years of reform strategies and millions of dollars in public and private investment, average student performance in math slumped at both grades for the first time in 25 years. In reading, the average score was stable at grade 4 but lower at grade 8 than the last time NAEP was given in 2013.

Although the 2015 scores for both grades and subjects were still higher than in the early 1990s, the declines since 2013 stirred immediate speculation about causes. How worried should people be about one year’s results? How has implementation of the Common Core State Standards (CCSS) affected NAEP performance? What other factors in the educational context could have influenced these results? Why did a few states or jurisdictions do better than most? What tentative conclusions or lessons are appropriate?

To help policymakers and the public better understand what the 2015 NAEP results do—and don’t—mean, CEP has put together this brief analysis of the NAEP results. The intent is to offer a reasoned perspective on the 2015 results and the historical trends that NAEP is uniquely able to provide.

The analysis was informed by the views of experts participating in a November 2015 symposium sponsored by CEP and the Graduate School of Education and Human Development (GSEHD) at George Washington University; these experts included Peggy Carr, acting commissioner, National Center for Education Statistics; Michael J. Feuer, GSEHD dean; James Pelligrino, liberal arts and sciences distinguished professor and distinguished professor of education, University of Illinois at Chicago; Fran Stancavage, managing research scientist, American Institutes for Research; Gerunda Hughes, director, Office of Institutional Assessment and Evaluation, Howard University; and Robert Rothman, senior fellow, Alliance for Excellent Education.

How the Unique Characteristics of NAEP Affect Interpretations of Results

NAEP was created in the mid-1960s to be a national barometer of US student achievement. Its assessments are administered to a representative sample of students in grades 4, 8, and 12. Reading and math are tested every two years, and other subjects less frequently. The results from the 2015 administration for grades 4 and 8 were released in October 2015; the grade 12 results will come out in 2016.

Over time, NAEP has been revised to produce data for individual states, in order to create a common measure of student achievement that can be compared across states. In 2001, Congress appropriated funds for a district-level NAEP assessment known as the Trial Urban District Assessment, or TUDA.
Twenty-one large urban districts voluntarily participate in TUDA but they must meet the selection criteria for district size, percentage of African American or Hispanic students, and percentage of students eligible for free or reduced-price school lunch.

“Learning from NAEP requires learning about NAEP.”
—Michael Feuer, dean, GSEHD at GWU

To accomplish NAEP's purpose as the “nation's report card,” several unique characteristics have been built into its design:

- **Only a sample of students and schools take NAEP**, in contrast to the more familiar state tests taken by virtually all students in key grades. Each student participating in NAEP takes only a subset of all the questions on the test. Therefore, NAEP does not produce scores for individual students, classrooms, or schools.

- **NAEP takes a relatively short time to administer**. Students only need about an hour to do a NAEP test, much shorter than the hours spent taking state tests. In light of the value of the data it yields, NAEP is not a major contributor to testing burden.

- **NAEP has low stakes**. Unlike state tests, the results are not used for decisions with important consequences for individuals or schools, such as determining whether students will graduate or pass a course, or evaluating teachers, principals, or schools.

- **The knowledge and skills tested on NAEP are not aligned to any particular state’s curriculum**. The frameworks for the content to be tested on NAEP are established by the National Assessment Governing Board and are updated infrequently.

These features have helped make NAEP the most trusted source of long-term national data on what students know and can do and a model program for other countries. They also affect the interpretation of results. When considering what NAEP results mean for student performance, it is important to remember these points:

- **Students may be less motivated to do their best on NAEP** because of the low stakes and lack of individual scores. Teachers are unlikely to provide the same degree of encouragement to students taking NAEP as they would for tests that “count.” These motivational factors could affect overall NAEP scores.

- **Students may not have had an opportunity to learn the knowledge and skills tested on NAEP**. Because NAEP is not aligned to a particular curriculum, teachers cannot prepare students to master the content on NAEP in the same way they prepare students for state tests. The fact that NAEP is relatively immune to teaching to the test is considered one of its strengths—state NAEP scores have sometimes been used as a check of whether state test scores are inflated. At the same time, this lack of alignment with many states’ curricula may
mean that students have not been exposed to particular areas of knowledge or skills at the grade they are tested on NAEP. (This issue is explored more in the next section on the Common Core.)

- **NAEP provides a historical view of student performance, and it’s important not to make judgments based on one year’s results.** NAEP’s greatest value comes from the long view. As with any test, NAEP scores are estimates; they are subject to year-to-year fluctuations for reasons unrelated to student learning. While the 2015 results warrant deeper analysis, there is reason to be cautious against rushing to conclusions about the US education system or using a single-year’s snapshot for political gain.

### Highlights of the 2015 NAEP Results

- **In math, the national average scores decreased at grades 4 and 8, ending a period of consistent gains that began in 1990.** In 30 states, average math scores dropped at one or both grades. Only three states or jurisdictions—Mississippi, the District of Columbia, and the Department of Defense Overseas Schools—had higher 4th grade scores in 2015 than in 2013, and no state had an increase at 8th grade.

- **Reading results were mixed.** In reading, which has seen less consistent growth over the years than in math, the national average score did not change significantly at grade 4; however, the average score slipped at grade 8 compared with 2013. State data for grade 4 reading revealed an increase in 13 states/jurisdictions, a decrease in one state (Minnesota), and no significant change in the remaining states. But at grade 8, only one state (West Virginia) posted a gain in reading.

- **Most student groups experienced declines in average scores at grade 8 in both reading and math.** Scores in math were lower for white, black, and Hispanic students, and males and females; there was no significant change for Asian/Pacific Islander students. Scores in reading remained stable for all racial/ethnic and gender groups at grade 4 but dropped for female and white students at grade 8.

- **Since the early 1990s, large city schools have made greater gains in reading and math than the nation as a whole.** The average 2015 scores for large city districts were not significantly different from 2013 in either subject or grade.
How Well Is NAEP Aligned to the Common Core, and Should It Be?

To interpret what the NAEP results mean and what constitutes a reasonable reaction, researchers, educators, and policymakers must look at the broader educational and policy context. Perhaps the most conspicuous change in context in recent years is the adoption of the Common Core State Standards in reading and math by 42 states and D.C. When NAEP scores improved in 2013, then-Secretary of Education Arne Duncan and other leaders were pleased to attribute gains to implementation of the CCSS, but with the 2015 release they found themselves backpedaling to revise their narrative. On the other end of the political spectrum, commentators hastened to blame the Common Core for the dispiriting 2015 NAEP results. While it is ill advised and unproductive to put any kind of political spin on NAEP results, it is perfectly appropriate to explore the relationship of the CCSS to students’ performance on NAEP.

To do that first means determining how well NAEP assesses the curriculum being taught in classrooms. This question has become easier to study now that a large majority of the nation’s students attend school in states that have adopted the CCSS. Even though states have implemented these standards at different rates and with varying degrees of individualization, the CCSS have brought more consistency to curriculum across states than existed in the past.

The American Institutes for Research (AIR) studied the alignment between the mathematics test items on the 2015 NAEP at grades 4 and 8 and the content of the CCSS. A panel of mathematicians, classroom teachers, math supervisors, and math educators reviewed all of math test items in the 2015 NAEP item pool for these grades and attempted to match each item to an appropriate standard in the CCSS. The items were classified as matched, not matched, or uncertain, according to the majority of the review panel. Presumably, students in CCSS-adopting states would have a reasonable likelihood of being taught the content in the NAEP items that were included in the Common Core for the tested grade or a lower grade. If a particular NAEP item was not represented in the CCSS or was included in the standards for a higher grade, then students most likely would not have received instruction in that content.

The AIR alignment study reported several interesting findings:

- **Overall, the alignment between NAEP and the CCSS was reasonable at grade 4 and strong at grade 8.** The percentage of NAEP items represented in the CCSS for the same grade or a lower grade was 79% at grade 4 and 87% at grade 8.

- **However, the alignment between NAEP items and the CCSS was much lower for certain math content areas.** NAEP categorizes the content of its math test by “subscales,” including number properties and operations; measurement; geometry; data analysis, statistics, and probability; and algebra. The lowest rate of alignment was in the area of data analysis,
statistics, and probability—47% of the items in this area were matched at grade 4 and 74% at grade 8. The match rates at grade 4 were also low for algebra (62%) and geometry (68%).

- Most of the NAEP items that were not matched to the CCSS were covered by the Common Core standards at a higher grade than the tested grade in NAEP. The exception was in data analysis, statistics, and probability at grade 8, where a substantial majority of the items did not match the CCSS at all.

The AIR study also looked at alignment from the other direction—what percentage of the Common Core standards in math through grade 4 or grade 8 can be matched to at least one NAEP item? Since the CCSS are intended to guide instruction, students may not have an opportunity to learn the content assessed by NAEP if it is not represented in the standards for their grade or a lower grade.

- The extent to which standards in the Common Core are represented on NAEP items was reasonably high at grade 4 but much lower at grade 8. The percentage of Common Core standards through grade 4 that could be matched to at least one NAEP item was 77%; the comparable figure through grade 8 was just 58%.

- Some content areas of the Common Core are not well represented in NAEP. At grade 4, the Common Core content area with the lowest percentage match (57%) between the standards and NAEP items was operations and algebraic thinking—an area of knowledge that is central to the progression from arithmetic to algebra, according to the AIR authors. At grade 8, the areas with the lowest match were ratios and proportional relationships (50%) and statistics and probability (41%).

- A closer look at NAEP math results by subscales reveals a reasonably consistent pattern of greater score declines in content areas that are less well aligned with the Common Core. Of the four content area/grade level combinations that had match rates lower than 75% in the AIR analysis, three showed a drop in average NAEP scores of 4 or 5 points from 2013 to 2015. These included data analysis, probability and statistics at grades 4 and 8, and geometry at grade 8; the exception was algebra at grade 4.

These findings about alignment between NAEP and the Common Core in math suggest that many students may not have had an opportunity to learn important subsets of knowledge and skills before taking the NAEP test. This could help to account for the lower scores on some subscales. No similar current study has been done for reading and English language arts, however.
The findings from the AIR alignment study suggest that the NAEP content frameworks should be revised to better reflect the curriculum being taught to students in the many states that have adopted the CCSS, but there are good reasons to approach any revisions to NAEP cautiously.

On one hand, the long-term stability of the NAEP frameworks has been crucial to maintaining trend data. The math frameworks have been consistent since 1990; the reading frameworks were updated in 2009, but special analyses were done to ensure trends are comparable going back to 1992.

In addition, NAEP has value as an independent indicator precisely because its content is separate from state testing programs. Due to the high stakes attached to state tests, teachers and principals pay more attention to those tests and may engage in certain kinds of test-specific preparation that could inflate scores without improving students’ learning. NAEP has been designed so that it is not susceptible to these kinds of pressures.

Disparities between state-level NAEP results and state test results have often been seen as evidence that some states have established less rigorous expectations for student learning, as Bob Rothman of the Alliance for Excellent Education has noted. The Common Core was developed in part to address the lack of rigor in some states’ standards. But we shouldn’t expect a perfect match between NAEP and state tests because NAEP’s purpose is unique, and should remain so.

On the other hand, NAEP could become a less useful indicator if its content strays too far from the curricula being taught to students. The NAEP math results could be a signal that NAEP is not assessing what students are being taught, and it may be time to strike a better balance between NAEP and the Common Core. If NAEP is to provide accurate data of students’ learning, then the alignment issue will have to be addressed. In the past, NAEP frameworks have been tweaked without breaking the trend line, and this issue will likely be explored in the future.

**Is “Implementation Chaos” a Factor in Declining NAEP Scores?**

Within a relatively short time, the implementation of the Common Core has led to major changes in many aspects of the educational systems of adopting states. States have revamped their curriculum, professional development, instructional strategies, teaching materials, and state testing programs. However, the pace of implementation has varied across and within states, and so have the challenges and complexity of the process.

Even states that have not adopted the Common Core have revised their standards, curriculum, testing programs, and other aspects of education in recent years. Some of these changes have been spurred by
federal policies, particularly the Race to the Top program and waivers of the No Child Left Behind requirements, while others have been state-initiated.

These changes have brought some turmoil. It takes time for students, teachers, principals, and parents to adapt to new standards and curricula and the 2015 NAEP tests were administered in the midst of the transition. It is important to consider the following points about the impact of the Common Core on NAEP:

- Many students and teachers are still adapting to new standards and curricula, which are often more rigorous than those they replaced. Instruction in some content areas may be in flux. For example, teachers may not have had sufficient professional development about teaching data analysis skills, an area where NAEP subscale scores were lower.

- States and school districts are taking different approaches to such issues as the sequence of content in the Common Core. States that have adopted the CCSS, and districts within those states, have made different choices about the sequence for teaching topics in the standards. Not only are certain topics taught at different grade levels, but there is also variation in the order for presenting topics in the same grade. Students may have less time to learn content that is taught later in the year, and may not do as well on test items that assess that content.

- Shifts in instructional practices could affect student performance on NAEP, even when content is similar. The same content may be taught in different ways. Math instruction is in transition in many classrooms, as teachers introduce new approaches to help students develop the problem-solving skills and conceptual understanding embodied in the CCSS. Even teachers with a good grasp of the new standards may still be experimenting with the best ways to teach them. Students also need time to adapt to teaching methods that may differ radically from how they were previously taught. This instructional flux could affect NAEP performance. In addition, some teaching approaches may better prepare students for the kinds of problems found on NAEP.

- Differences in the pace and substance of implementation of new standards make it difficult to tease out the impact of “chaos” on NAEP performance. Implementation varies not only across states, but also across different components of the educational system (for example, designing curriculum materials versus providing professional development to all teachers). Common Core implementation has been tumultuous in some states for the past few years, but
at some point the system will become more stable and chaos will no longer be a possible explanation for volatility in test scores. The question is, when do we reach that point?

What Issues Should Be Explored Further?

The underlying reasons for NAEP trends are typically not apparent without deeper study, and even then may remain elusive, as NCES Acting Commissioner Peggy Carr has emphasized. Take, for example, the NAEP “reading anomaly,” which showed a steep drop in performance among some students on the 1986 reading assessment. An expert panel assembled to probe the data concluded that the declines may have stemmed from changes in test design and administration, including the color of the ink used, rather than from an actual decline in achievement. A contrasting example occurred in 2002 in Delaware, where the average NAEP reading score rose by 17 points compared with the state’s previous score from 1998. NAEP did a special study of the surprising gains and found no technical or analytical problems. However, Delaware had systematically revised its curriculum and staff development during the period in question, which could have contributed to the improved NAEP results.

As we look ahead to future NAEP assessments, some additional questions about the 2015 results are worth considering:

- Did NAEP’s paper/pencil format affect performance? The majority of state tests are now administered by computer, according to recent data from EdTech Strategies. Most notably, the state consortia-developed tests aligned to the CCSS are computer-based. By contrast, NAEP is administered in a paper/pencil format, which may have an impact on results, especially in grades where students have little familiarity with paper/pencil tests. Findings from research comparing computer-based and paper/pencil tests have been mixed; moreover, the impact of computer-based testing appears to vary by student characteristics and content area. Future NAEP tests will be computer-based, which will provide additional data for research.

- Do the math results show a plateau effect? Some analysts have speculated that the 2015 results may indicate that the two decades of uninterrupted progress in math have reached a plateau. The percentage of students scoring at the lowest end of the achievement continuum—“below basic” according to the achievement levels set by NAEP—has decreased substantially in math since 1990 at both grades. The earlier gains in math could have come largely from improvements among lower-achieving students. Once that’s been accomplished, it is harder to raise the performance of students closer to the top of the range. Research on whether and how this plays out in NAEP could be useful.

“We’ve picked the low-hanging fruit. Now the challenge is getting to the next level.”
—Bob Rothman, Alliance for Excellent Education
• **What can be gleaned from the NAEP data by student subgroups?** Achievement gaps by race/ethnicity and gender did not change significantly between 2013 and 2015, with one notable exception. In grade 4 math, the black-white achievement gap was smaller, but that was because white students did worse—not a desirable way to close gaps. Although racial/ethnic gaps have narrowed for some grades and subjects since the early 1990s, the 2015 results show persistent disparities. Reducing these gaps will become an even more urgent priority as the nation, and the NAEP test-taking population, continues to become more diverse. A deep dive into the NAEP data by subgroup for the nation and states is essential to understanding how to raise achievement for lower-performing groups and narrow achievement gaps.

• **What lessons can be learned from in-depth studies of states that showed progress on NAEP?** A closer analysis of the NAEP data and the policies, curriculum, and instruction in these states might shed light on best practices that could be helpful to others. One topic for further research is the math curriculum sequence and instructional practices in states that did better than average. For example, states with the smallest declines on the geometry subscale include some that have not adopted the Common Core, but they also include Kentucky, which was an early implementer of the CCSS. Research on these states might help explain math performance and contribute to our understanding of the best ways to teach math.

• **What factors are present in urban districts that did better on NAEP?** Three TUDA districts—Chicago, D.C. Public Schools, and Miami-Dade—had 2015 score gains in more than one subject and grade. There is an opportunity to probe the more detailed NAEP data for these and other districts that did better than average, and to simultaneously examine their policy changes, curriculum, instructional practices, and larger educational contexts.

In closing, it is important to remember that NAEP exists to provide educators and policymakers with a historical timeline of student performance. The findings and fluctuations of each administration of NAEP should stimulate a deeper conversation about education before and after the test. The 2015 results have generated national introspection, but making instant inferences based on the results is not prudent or productive. With NAEP, the best strategy is to make the most of its unique qualities and play the long game.