# TESTS ARE TURNING OUR KIDS INTO ZEROES: A FOCUS ON FAILING

Fred Smith
Robin Jacobowitz

THE BENJAMIN CENTER
SUNY NEW PALTZ
FOR PUBLIC POLICY INITIATIVES

DISCUSSION BRIEF #20 | SUMMER 2018

> Totally inaccurate. Unintelligible. Indecipherable.

Beginning in 2010, the New York State Education Department (SED) specified and established Common Core Learning Standards (CCLS) in response to a major federal financial incentive, *Race to the Top*. The department awarded a five-year, $38.8 million contract to NCS Pearson, Inc. for the development of tests in English Language Arts and Mathematics to measure the standards' impact.[1] This report, the first in a series, examines the New York State English Language Arts Test (ELA) developed under this contract, and used during the 2013–2016 period.

## I. THE DEBATE

From the outset of their use in 2013, there were parents and educators who questioned the character, value, and impact of Core-based assessments. These tests were reported to be more rigorous than those previously used, often based upon texts that exceeded grade-level.[2] Skeptics warned that the greater emphasis on testing came with a heavy cost—taking resources away from instruction; sacrificing time that should have been spent on non-tested subjects and activities; shifting the balance of classroom hours to test preparation; and bringing undue stress to children who were now subjected to testing over a longer six-day period.

Additional criticisms were voiced about the ambiguity and inappropriateness of the test questions, the frustrating experiences English Language Learners and students with disabilities had with the exams, and the lack of transparency that thwarted scrutiny of the program. There was particular concern about the developmental appropriateness of the reading passages and items used to assess eight- and nine-year-old students in grades 3 and 4.

These initial complaints were dismissed by officials as not empirically grounded—the scattered griping of overprotective parents, a sign of low expectations for children, or reflective of teachers' reluctance to be evaluated using these scores. As the annual testing program continued, however, the criticism amplified, leading to a groundswell of opposition which saw more than 20 percent of the potential test population refuse to take the tests in 2015.[3] Despite the growing grassroots unhappiness with the examination system, SED adhered to the position that the CCLS were beneficial and that the aligned testing was measuring progress toward the implementation of more rigorous standards.

## II. HOW DID THE TESTS DO?

Now, after several years, it is fair to investigate the efficacy of this ongoing testing program, which targets 1.2 million students each year and costs taxpayers millions of dollars. Student performance on these instruments is widely reported and commented on. It is now time to ask not only "How did the students do?" but also "How did the tests do?"

To answer this question, we examine the test questions themselves. These are the bricks of the exams—irreducible elements on whose reliability the value of all results and

We conclude that testing instruments that put children in a virtual stupor cannot be defended as sound testing practice, nor as a way to raise standards or serve as a foundation for high-stakes decisions and statistical models to evaluate teachers, rate principals, or close schools.

usage depends. We are specifically interested here in how the Constructed Response Questions (CRQs) performed because they are the ones intended to assess the higher-level thinking and critical reasoning ability that the CCLS sought to foster.

We analyze the scores students received on these open-ended questions which, unlike multiple-choice items, require students to write their answers rather than select the correct answer. Specifically, we look at the overall number and percentage of students who got zeroes on the CRQs and how the percentage changed over time. We also analyze how some subgroups fared: English Language Learners (ELLs), students with disabilities (SWD), and members of racial and ethnic groups.

Why zeroes? Because a zero score on a CRQ reflects a student's complete inability to cope with the test material. According to the test scoring rubrics, a zero is given to an answer that is "totally inaccurate," "unintelligible," or "indecipherable."

Our data reveal high percentages of zeroes in the scoring distributions generated by the ELA at all grade levels, after CCLS-alignment. We conclude that testing instruments that put children in a virtual stupor cannot be defended as sound testing practice, nor as a way to raise standards or serve as a foundation for high-stakes decisions and statistical models to evaluate teachers, rate principals, or close schools.

Overall, our findings validate skeptics' misgivings. The data we obtained reveal:

- a steep sweeping increase in the percentage of students receiving zeroes on the CRQs in 2013 when the CC-aligned tests debuted;
- particularly sharp increases, sustained over time, in the percentage of zeroes for students in grades 3 and 4 and for English Language Learners and students with disabilities;
- a substantial gap in the zero scores between black and Hispanic students and white and Asian students.
- a substantial number of students getting zeroes on at least half of the CRQs.

Following we detail these and other findings.

## III. COMMON CORE LEARNING STANDARDS AND GRADES 3–8 ELA TESTS

In 2010, New York State sought and secured additional federal education funding through the *Race to the Top* initiative. As a condition of this success, the SED embraced the Common Core State Standards. Supporters hailed these standards as challenging, claiming they would create higher expectations for NYS students, better preparing them for college or careers in the 21st century. Specifically, proponents claimed that the learning standards would foster the development of students' critical thinking skills, their capacity to process complex material, their skills in analyzing information, and their ability to use facts and present evidence logically. For use in New York the SED made slight adjustments to the generic Common Core, incorporated State educational and learning preferences, and renamed them the Common Core Learning Standards (CCLS). These were adopted by the Regents in 2011. To assist in launching these new standards, in January of 2011 the SED offered guidance to school districts, including curriculum guides and materials, anticipating that they would come into use in classrooms shortly thereafter.[4]

As a final step in the march toward more demanding standards and higher expectations, the annual Grade 3–8 ELA and math assessments were constructed under contract by NCS Pearson Inc. to align with CCLS. These exams were first administered to students in 2013.[5]

The SED heralded 2013's exams as "rigorous" and establishing a baseline against which students' progress toward the CCLS standards—and college and career readiness—could be measured. The *Educator's Guide*s that accompanied their introduction stated that the assessments' content will be "more advanced and complex" than in previous years.[6] The changeover would not be easy, the SED warned. In administering these exams, the agency foresaw a precipitous decrease in the percentage of students shown to be "proficient." This outcome was indeed realized.[7]

It is important to note that although the content of the assessments changed significantly with alignment to the CCLS, the format did not. The CCLS-aligned ELA, as with previous assessments, are comprised of multiple choice (MC), short response (SR), and extended response (ER) questions.

MC items measure a range of language arts skills and require students to discern between the correct and "plausible" answers. SR questions require students to make an inference, supported by two pieces of evidence, from a text provided them. And ER questions aim to measure how well students express themselves in writing.[8] SR and ER questions therefore require what is called a "constructed response." They are the CRQs. On the ELA each year, seven or eight questions are SRs; another two are ERs.

From 2013–2016, the ELA assessment was administered over three days. On the first day, the test contained only MC items. Day 2 presented a mix of MC and SR questions plus one ER. Day 3 consisted entirely of CRQs; most of them were SR, with one an ER. *Table 1* presents the number of CRQs, and the time allotted for them, on days 2 and 3 of ELA testing.

**Table 1.** Organization of the Constructed Response Questions (CRQs) on the ELA tests, 2012–2016[9]

| | | DAY TWO | | | DAY THREE | | | TOTAL |
|---|---|---|---|---|---|---|---|---|
| | Grades | SR | ER | Time | SR | ER | Time | # of CRQs |
| 2012 | 3 and 4 | 3 | 1 | 90 | 4 | 1 | 90 | 9 |
| | 5 to 8 | 3 | 1 | 90 | 4 | 1 | 90 | 9 |
| | Grades | SR | ER | Time | SR | ER | Time | # of CRQs |
| 2013 to 2015 | 3 and 4 | 3 | 1 | 70 | 5 | 1 | 70 | 10 |
| | 5 to 8 | 3 | 1 | 90 | 5 | 1 | 90 | 10 |
| | Grades | SR | ER | Time | SR | ER | Time | # of CRQs |
| 2016 | 3 and 4 | 2 | 1 | | 5 | 1 | | 9 |
| | 5 to 8 | 2 | 1 | | 5 | 1 | | 9 |

Short Responses are worth from 0–2 points. Extended Responses are worth from 0–4 points
Multiple-choice items are included on Day 2. Day 3 consists entirely of CRQs.

**Table 2.** Weight of the CRQs on the New York State ELA Test Results, 2012–2016

| POINTS STUDENTS COULD ATTAIN ON CRQS IN RELATION TO TOTAL POINTS | | | | | | | |
|---|---|---|---|---|---|---|---|
| YEAR | GR 3 | GR 4 | GR 5 | GR 6 | GR 7 | GR 8 | |
| 2012 | 20/67 | 22/73 | 22/73 | 22/73 | 22/73 | 22/67 | |
| 2013 | 24/55 | 24/55 | 24/66 | 24/66 | 24/66 | 24/66 | |
| 2014 | 20/49* | 24/55 | 24/66 | 24/66 | 24/66 | 24/66 | |
| 2015 | 24/55 | 24/55 | 24/66 | 24/66 | 24/66 | 24/66 | |
| 2016 | 22/47 | 22/47 | 22/57 | 22/57 | 22/57 | 22/57 | |
| **WEIGHT OF CRQS ON ELA TEST RESULTS** | | | | | | | |
| YEAR | GR 3 | GR 4 | GR 5 | GR 6 | GR 7 | GR 8 | Average |
| 2012 | 30% | 30% | 30% | 30% | 30% | 33% | 30.5% |
| 2013 | 44% | 44% | 36% | 37% | 36% | 36% | 38.9% |
| 2014 | 41% | 44% | 36% | 36% | 37% | 36% | 38.4% |
| 2015 | 44% | 44% | 36% | 36% | 36% | 36% | 38.8% |
| 2016 | 47% | 47% | 39% | ***39%*** | ***39%*** | 39% | 41.3% |

* In 2014, grade 3, SED eliminated the last question on the test after it was administered. It was a 4-point CRQ. This reduced the number of points 3rd graders could earn from 24 to 20, lowering the weight of the CRQs to 41%.

## IV. DATA AND METHOD

This report relies on two sources of data. The SED provided aggregate information about the scores students received on each CRQ for the statewide testing population, including NYC, grades 3–8, from 2012–2016. Data were obtained for the six grade levels over the five years, including for nine CRQs per grade in 2012 and 2016 and ten per grade in 2013, 2014, and 2015. The collective result was 288 CRQs, yielding 30 (6x5) scoring distributions for all CRQs.

Data were obtained from SED via a protracted FOIL process and are inclusive of approximately 200,000 children per grade—or about 1.2 million children each year.

New York City's Department of Education (DOE) responded to Requests for Information promptly and facilitated the transfer of data files for the City's approximately 440,000 children—more than 70,000 students per grade. These data were reported individually and permitted an examination of how English Language Learners (ELLs), Students with Disabilities (SWDs), and students from different racial and ethnic groups fared on the ELA.[10] Grade 4 data were not available for 2016.

We focus our analyses in this paper on the CRQs for several reasons. First, the writing and application of knowledge required to address these questions are central to the Common Core's aspirational goals.[11] Second, since the alignment with the CCLS, CRQs have weighed more heavily in student outcomes (*Table 2*). Inability to do well on the CRQs had a substantial negative effect on student test performance, particularly in the early grades. SR questions could be scored 0, 1, or 2 points depending on the quality of the answer. ERs had a scoring range that went from 0 to 4.

Moreover, our interest in CRQs centers on the zero scores, which reflect a student's inability to cope with the test material. A zero is given to an answer that is "totally inaccurate," "unintelligible," or "indecipherable" according to testing documentation. Pearson's 2013 training package sets forth the guidelines for scoring the SR and ER questions:[12]

- On 2-Point (SR) questions, a zero score is appropriate when the response is **totally inaccurate**, or is **unintelligible, not written in English,** or **indecipherable.**

- The rubric for evaluating 4-Point (ER) "expository writing" essays defines a zero-level response as **demonstrating a lack of comprehension of the text(s) or task involved; providing no evidence or completely irrelevant evidence; being minimal with respect to the conventions of standard English; exhibiting no evidence of organization, or of a concluding statement, or using language that is predominantly incoherent or copied directly from the text.** (Emphasis added.)[13]

It is noteworthy that children can receive one point on an SR for an answer that addresses only part of the question being asked or that contains incomplete sentences. In contrast, a zero, received from trained scorers, indicates an inability to write answers that reach even a partial level of understanding of the test material. Therefore, we are interested in knowing whether there has been an increase in the number of zero scores received prior to and after alignment with the CCLS. Investigating the way the zero scores were distributed across the nine or ten CRQs that were posed each year is one way to study the difficulty and impact of the ELA.

We determined the percentage of zeroes that were received by children statewide for each grade level and each year. These percentages are an average over the nine or ten CRQs on a given ELA test. Further analyses probed the distribution of zeroes each test generated and the percentage of students who received 5 or more zeroes on the CRQs.

We replicated these analyses for the New York City test population and analyzed zero scores by subgroup for ELLs, SWDs and racial and ethnic groups.

Limitations in this study on the interpretation of data and findings:

- The opt-out movement took hold in New York State in 2015, with 20 percent of the test-eligible population refusing to be tested. This figure went to 21 percent in 2016, with the level of resistance remaining much higher in schools and districts outside of New York City. There is disagreement about how the selection factors in this movement might have influenced overall student outcomes. Research at the national level suggests that parents who refuse the test for their children are highly educated, white, and with a median income that is above average.[14] If this finding holds for New York, it would skew some of the data upon which we rely downward.

- In 2016, SED relaxed the time limits on the tests. January and March memoranda to superintendents and principals announcing the transition to untimed tests for the spring of 2016 stated that SED had received "extensive feedback from educators from throughout the State about the inability of students to work at their own pace… In general, this will mean that as long as students are productively working they will be allowed as much time as they need to complete the tests. Additionally, this change in policy may help alleviate the pressures that some students may experience as a result of taking an assessment they must complete during a limited amount of time."[15] Allowing schools discretion to create their own approach meant the tests would be administered under conditions that were no longer uniform. The change in procedures was another reason the test results were not comparable with those of prior years. Variations in the composition of the test population and the shift to untimed testing confounded efforts to study trends. Nevertheless, we report the 2015 and 2016 data because we think there is information to be gleaned.

**Graph 1.** New York State–Percentage of Zero Scores on CRQs by Grade, 2012–2016

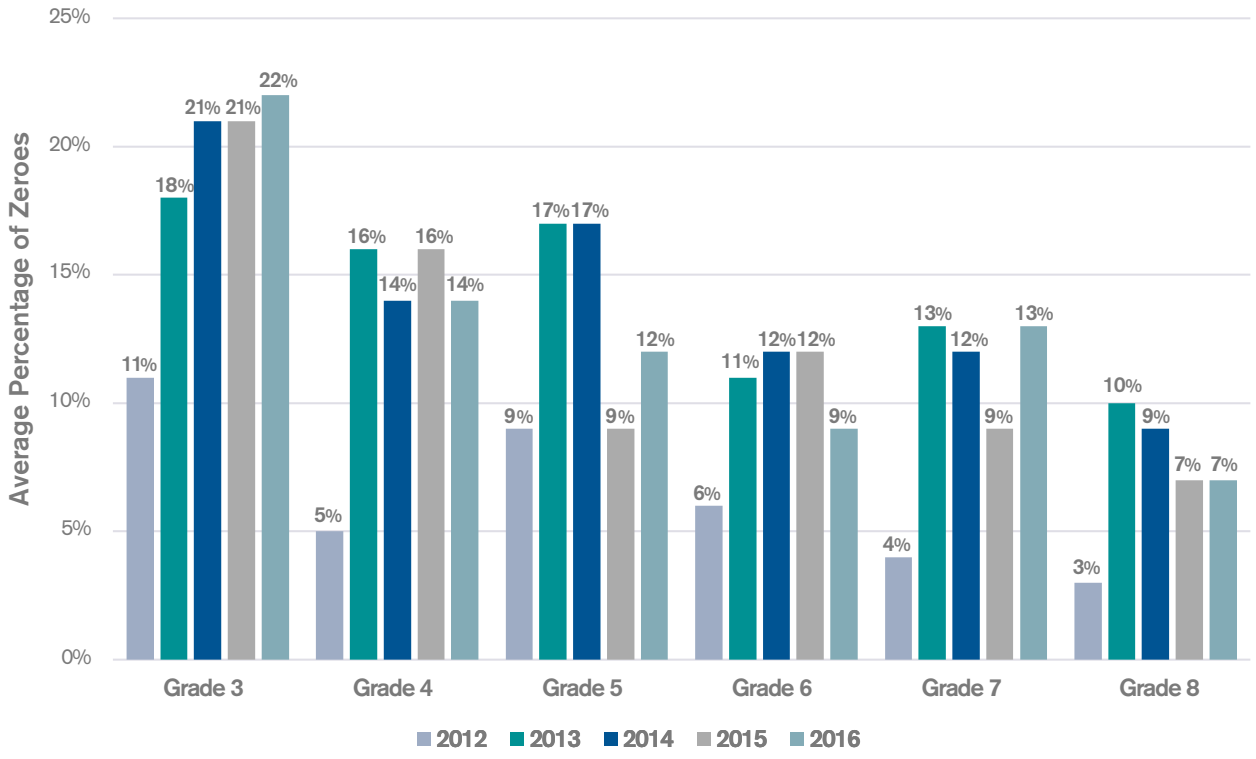



**Table 3:** New York State–Number of Students Receiving at Least One Zero on CRQs, 2012–2016

| GRADE | 2012 | 2013 | 2014 | 2015 | 2016 |
|---|---|---|---|---|---|
| 3 | 22,746 | 35,723 | 41,108 | 31,574 | 37,579 |
| 4 | 10,044 | 32,844 | 27,559 | 24,411 | 22,931 |
| 5 | 18,213 | 32,624 | 33,374 | 13,033 | 18,971 |
| 6 | 13,453 | 22,367 | 25,981 | 17,152 | 14,834 |
| 7 | 7,203 | 25,247 | 23,430 | 11,562 | 18,815 |
| 8 | 5,692 | 19,997 | 18,336 | 8,783 | 9,480 |
| TOTAL | 77,351 | 168,802 | 169,788 | 106,515 | 122,610 |

Fundamental to this study, we sought a frame of reference by which to judge what an acceptable percentage of zeroes might be. That is, how many zeroes should be routinely expected in a distribution of scores generated by the CRQs? There is no definitive answer to this question. Still, we probed the data at our disposal and performed a number of analyses in an effort to gauge the extent of the zero-score phenomenon in the CCLS testing years.

What follows is an endeavor to apply standards of logic and reasonableness to the statistics we compiled in order to gauge their significance. We believe that our findings show the number of zeroes students got on the ELA is evidence that the massive New York State Testing Program was built on poorly developed exams that have returned little educational value and have been particularly problematic for the youngest students.

## V. ANALYSIS AND FINDINGS: NYS AND NYC—ALL STUDENTS

### A. NYS Percentage of Zeroes by Grade

*Graph 1* shows the average percentage of zero scores NYS students in grades 3–8 received on the CRQs from 2012–2016. There were approximately 1.2 million test takers annually. The height of the bars represents the percentage of zeroes received on each question, averaged over the number of CRQs given in that year. For example, in 2013, students in grade 4 wrote answers to ten CRQs; 16 percent of these, across the ten CRQs, received a zero score.

Sharp increases in zeroes occur in all grades from 2012 (pre-CCLS alignment) to 2013 (CCLS alignment). Grade 3 shows a jump from 11 percent in 2012 to 18 percent in 2013. This increase is sustained in 2014–2016, reaching up to 22 percent in 2016. The steepest rise occurs in grade 4, where the zeroes go from 5 percent in 2012 to 16 percent in 2015. The increase is also sustained, generally, through 2016.

The percentage of zeroes fluctuates for grades 5–7 and declines gradually for grade 8 from the initial 2013 surge. But for all grades in all years, the percentage of zero scores post-CCLS alignment (2013–2016) remains well above the 2012 percentage.

When we look at the actual numbers of students receiving zeroes (*Table 3*), we express the percentages in human terms. In 2014, for example, 169,788 students received at least one zero score on the CRQs. That's 169,788 children, aged 8 to 14, who were befuddled enough by a question that their response was "inaccurate, unintelligible, incoherent"—more than enough children to fill the seats in Madison Square Garden eight times!

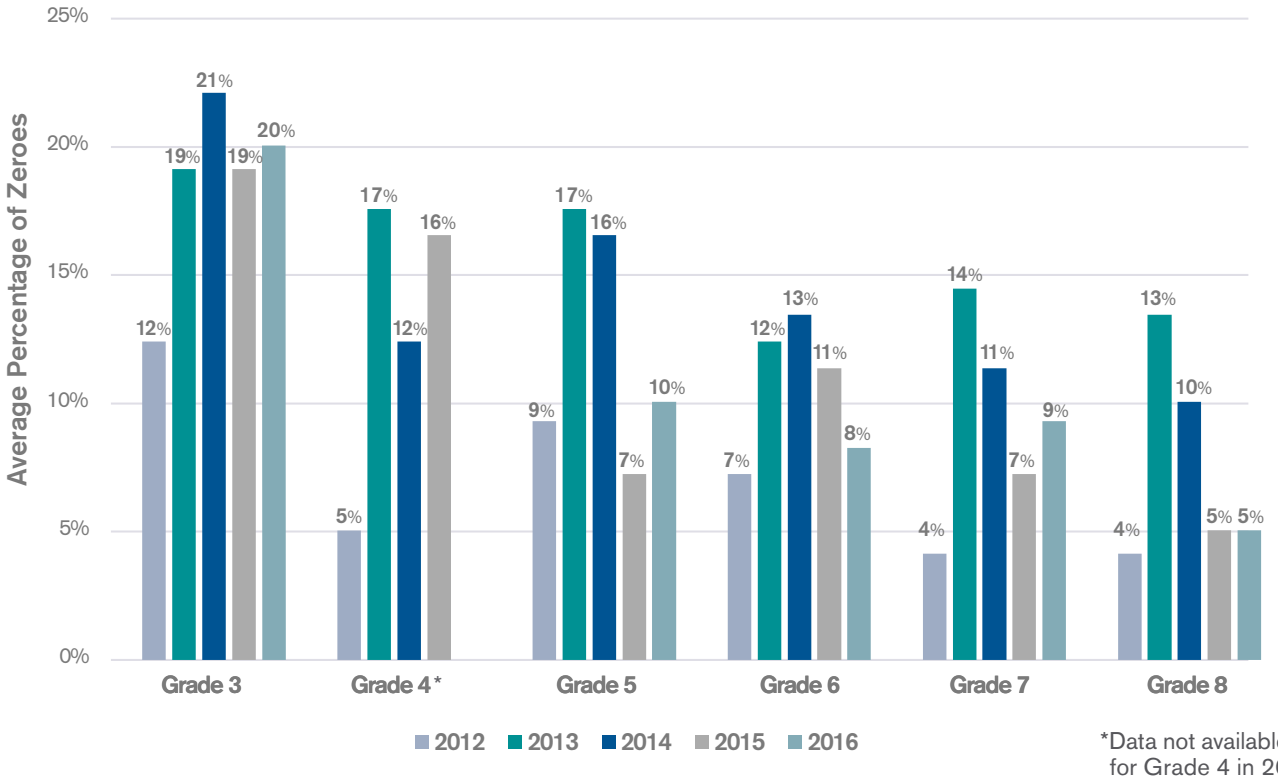
### B. NYC Percentage of Zeroes by Grade

*Graph 2* shows the average percentage of zeroes received by students in NYC on the CRQs. NYC's test population exceeded 440,000 students in grade 3 through 8. As in the analysis of NYS data, each percentage is an average taken over a test's nine or ten CRQs.

It is not surprising to find a parallel between the results that are exhibited for the State and City, as NYC comprises 37 percent of the state's entire test population; its results are included in the statewide numbers. Here again we see a steep increase in the percentage of zeroes between 2012 and 2013 in all grades. Grades 3 and 4 stand out from the rest, with average zeroes that climb in 2013 and then remain high for all exams post CCLS-alignment.[16] The percentage of zeroes begins a decline in 2014 for grades 5–8, but nevertheless remains above the percentages in 2012.

As with the NYS data, in order to understand the magnitude of these percentages it is helpful to see them in terms of the number of children they represent, not just as statistical abstractions. The 66,319 zeroes in NYC 2013 are more than three times the seating capacity of Madison Square Garden.

**Graph 2.** New York City—Percentage of Zero Scores on CRQs by Grade, 2012–2016




*Data not available for NYC for Grade 4 in 2016.

**Graph 3.** New York State—Percentage of Unanswered CRQs by Grade, 2012–2016

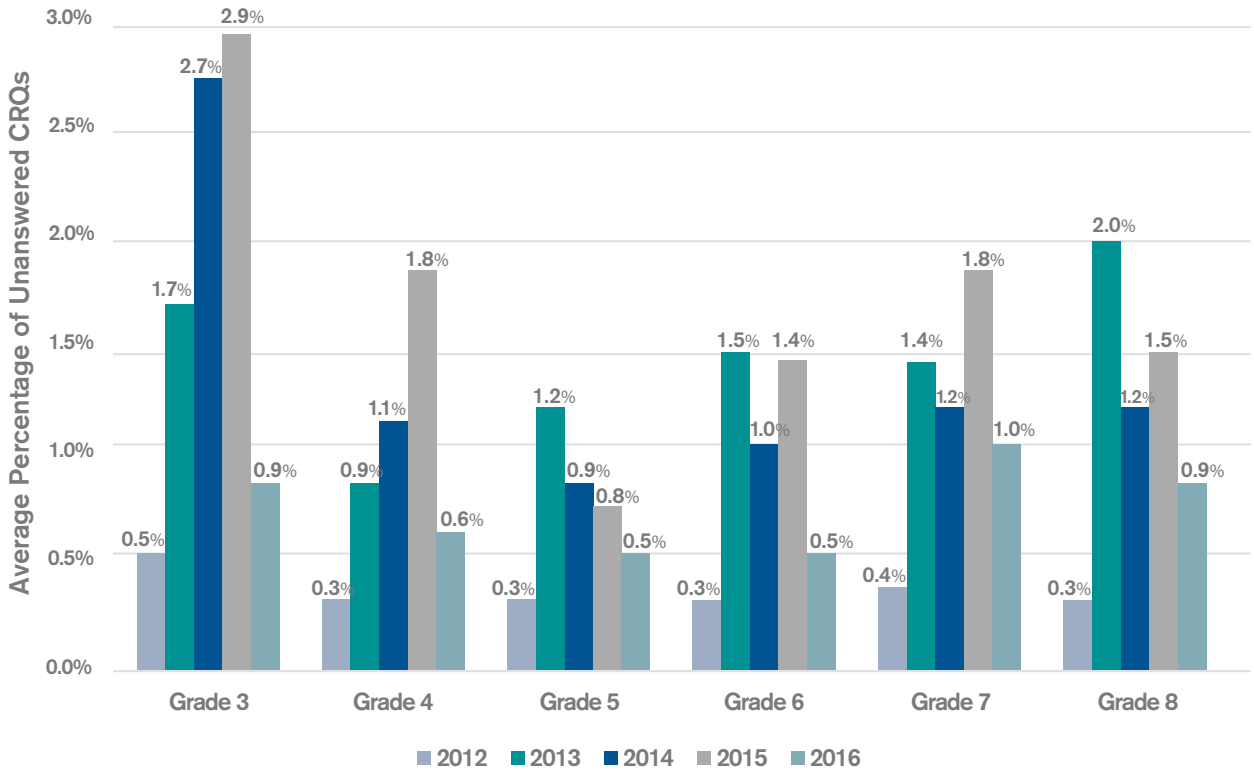


**Table 4:** New York City–Number of Students Receiving at Least One Zero on CRQs by Grade, 2012–2016

| GRADE | 2012 | 2013 | 2014 | 2015 | 2016 |
|---|---|---|---|---|---|
| 3 | 9,250 | 14,295 | 15,850 | 14,795 | 15,622 |
| 4 | 3,562 | 12,385 | 8,883 | 11,979 | — |
| 5 | 6,632 | 12,080 | 11,631 | 5,406 | 7,081 |
| 6 | 5,435 | 8,678 | 9,456 | 7,707 | 5,754 |
| 7 | 2,882 | 9,954 | 8,049 | 4,617 | 6,634 |
| 8 | 2,572 | 8,928 | 7,399 | 3,766 | 3,465 |
| TOTAL | 30,333 | 66,319 | 61,267 | 48,271 | — |

### C. NYS Impact of Unanswered Questions

Most students who take the ELA attempt to answer all CRQs. At times, however, they do not respond to a question or do not reach it. The latter could point to insufficient time given to complete the test. Whatever the reason, when students do not put an answer on paper, they receive no points for that question. This limits the scores students receive and thereby lowers the ceiling they can reach on the ELA. Unanswered questions are different than responses that are scored zero. They form a separate residual category.

*Graph 3* shows that the percentage of unanswered CRQs increases sharply in 2013. The percentage in grade 3 stands apart from the pattern observed in all other grades. Even grade 4 traces a path comparable to the upper grades. Perhaps, eight-year-olds struggle to complete the test within the allotted timeframe or are less likely to attempt a question they find confusing or they are less "test wise" than older children, who have been taught not to skip questions. Regardless of cause, this finding is consistent with the charges that parents made about the exams' difficulty and inadequate time limits, particularly for young children, when CCLS-aligned tests were introduced.[17]

The decrease in unanswered questions in every grade in 2016 was a predictable consequence of the shift to untimed tests. That year, SED also reduced the number of CRQs from ten to nine per grade. This gave students more time to respond to fewer questions, but it took uniformity out of the testing process. With less pressure, students could complete more CRQS.

While zeroes are valid scores, blanks also tell a part of the story. Unanswered questions offer insight into the experiences of test takers. Logic suggests that to the extent they occur, they signal poor test development and testing procedures that spell difficulty and discouragement for test takers.

### D. Analysis of Percentages of Zeroes—NYS and NYC

Before moving on to additional analyses, we provide an overview of our primary findings. The data show that there has been an increase in the percentage of zero scores since the administration of exams aligned with the Common Core. Although fluctuations occur, a clear pattern emerges: Children in grades 3 and 4 have struggled with the CRQs during the Core-based testing years. Of note, there are substantially more zeroes in grade 3 than in the other grades and the percentage—approximately 21—remains consistent for four years.

This raises general questions about the efficacy of this kind of test for such young students. It also opens the floor to a discussion of the quality of Pearson's work, the worth of its product, and SED's management of the test program. With a four-year average of 21 percent of responses rated as "incomprehensible, incoherent, or irrelevant," we must ask to what degree, if at all, such an assessment yields valuable information about our

> The data show that there has been an increase in the percentage of zero scores since the administration of exams aligned with the Common Core.

third-grade students (*Graphs 1* and *2*). The failure here may be in the questions themselves or the belief that it is developmentally acceptable to ask eight-year-olds to sit for extended periods of time to take these exams over several days.

A similar argument could be made for grade 4. While there were fewer zeroes in grade 4 than grade 3, the post-CCLS jump is higher (5 percent in 2012 to 16 percent in 2013). And for all grades in all years after CCLS-alignment, the percentage of zero scores remains above pre-CCLS levels.

We anticipate that officials will claim this outcome to be the consequence or evidence of increased rigor in the standards. We maintain that rigor is not the chief narrative here. Recall that a zero score marks an unintelligible answer, a total lack of ability on the part of a student to engage with the test material. (A score of 1 is given to a response that only partially addresses the question being asked or that contains incomplete sentences.)

Certainly, some zero scores were likely, but 21 percent in grade 3, 15 percent grade 4 and 14 percent in grade 5 over four testing rounds seem inordinately high and beyond reasonable expectations. Zero score levels exceeded pre-CCLS results even as students and teachers became acclimated to the CCLS and hoped to attain the Regents' goal of college and career readiness as gauged by these tests.

Further evidence of faulty testing can be seen in the decline of zeroes and unanswered CRQs in 2016 when SED removed time limits. After three years of CCLS-aligned testing, given to 1.2 million students annually, including multiple field tests administered each year, the SED acknowledged that the duration of the exams was problematic. This, in itself, is an after-the-fact admission that the tests were poorly devised.

Matters of administration, including timing, are normally resolved in the planning and research phases of test development, before exams become operational. Further, in this case, the elimination of time limits meant the exams were no longer given under standard conditions. In 2016, districts and schools set their own limits. SED had recognized a fundamental problem with its CCLS-aligned exams. But SED's solution was late and flawed.[18]

## VI. ANALYSIS AND FINDINGS: NYC SUBGROUPS

The New York City ELA results could be analyzed in greater depth because item level data was provided by the DOE. Analyses below explore zero scores based on English Language Learner (ELL), and student with disability (SWD) status, and ethnicity.

### A. Distribution of Zeroes

*Table 5* shows the average number of zero scores students received annually by grade level. To get this average, the number of zeroes was tallied for every child on each CRQ and the mean was taken.[19]

From 2012 to 2013, there is a sweeping increase in the averages. In grade 4 for example, the mean number of zeroes rose from less than half (0.4) to 1.7 for the City's test population. The average for ELLs rose from 1.0 to 3.3 zeroes and from 1.1 to 3.4 for students with disabilities. The average for black and Hispanic students went from 0.5 to 1.9 and from 0.6 to 2.0 (an increase of 1.4 in both groups). The zero scores for their white and Asian counterparts increased from 0.3 to 1.0 respectively.

**Table 5:** New York City—Average Number of Zeroes on CRQs by Grade, Year and Subgroup

| Grade 3 | Total | ELL | SWD | Asian | Black | Hispanic | White |
|---|---|---|---|---|---|---|---|
| 2012 | 1.1 | 2.0 | 2.2 | 0.7 | 1.24 | 1.31 | 0.67 |
| 2013 | 1.9 | 3.0 | 3.3 | 1.2 | 2.1 | 2.2 | 1.2 |
| 2014 | 1.8 | 3.0 | 3.2 | 1.2 | 2.1 | 2.1 | 1.2 |
| 2015 | 1.9 | 3.3 | 3.6 | 1.2 | 2.22 | 2.26 | 1.22 |
| 2016 | 1.8 | 3.5 | 3.6 | 1.1 | 2.0 | 2.2 | 1.0 |
| **Grade 4** | **Total** | **ELL** | **SWD** | **Asian** | **Black** | **Hispanic** | **White** |
| 2012 | 0.4 | 1.0 | 1.1 | 0.3 | 0.48 | 0.56 | 0.25 |
| 2013 | 1.7 | 3.3 | 3.4 | 1.0 | 1.9 | 2.0 | 1.0 |
| 2014 | 1.2 | 2.4 | 2.6 | 0.6 | 1.4 | 1.5 | 0.6 |
| 2015 | 1.6 | 3.2 | 3.1 | 0.9 | 1.84 | 1.89 | 0.94 |
| 2016 | — | — | — | — | — | — | — |
| **Grade 5** | **Total** | **ELL** | **SWD** | **Asian** | **Black** | **Hispanic** | **White** |
| 2012 | 0.8 | 1.8 | 1.7 | 0.5 | 1.0 | 1.0 | 0.5 |
| 2013 | 1.7 | 3.2 | 3.2 | 0.9 | 1.9 | 2.0 | 1.0 |
| 2014 | 1.6 | 3.1 | 3.0 | 1.0 | 1.8 | 1.9 | 1.0 |
| 2015 | 0.7 | 2.1 | 1.8 | 0.4 | 0.9 | 0.9 | 0.4 |
| 2016 | 0.9 | 2.2 | 2.0 | 0.5 | 1.0 | 1.0 | 0.5 |
| **Grade 6** | **Total** | **ELL** | **SWD** | **Asian** | **Black** | **Hispanic** | **White** |
| 2012 | 0.7 | 1.8 | 1.4 | 0.4 | 0.8 | 0.8 | 0.4 |
| 2013 | 1.2 | 2.9 | 2.5 | 0.7 | 1.4 | 1.4 | 0.7 |
| 2014 | 1.3 | 2.7 | 2.7 | 0.7 | 1.5 | 1.6 | 0.7 |
| 2015 | 1.1 | 2.7 | 2.3 | 0.6 | 1.3 | 1.3 | 0.5 |
| 2016 | 0.7 | 2.0 | 1.7 | 0.4 | 0.9 | 0.9 | 0.4 |
| **Grade 7** | **Total** | **ELL** | **SWD** | **Asian** | **Black** | **Hispanic** | **White** |
| 2012 | 0.4 | 1.2 | 0.9 | 0.2 | 0.4 | 0.5 | 0.2 |
| 2013 | 1.4 | 3.3 | 2.8 | 0.8 | 1.6 | 1.6 | 0.8 |
| 2014 | 1.1 | 2.6 | 2.3 | 0.6 | 1.3 | 1.3 | 0.6 |
| 2015 | 0.6 | 1.8 | 1.4 | 0.3 | 0.8 | 0.8 | 0.3 |
| 2016 | 0.8 | 2.6 | 2.0 | 0.5 | 1.0 | 1.0 | 0.4 |
| **Grade 8** | **Total** | **ELL** | **SWD** | **Asian** | **Black** | **Hispanic** | **White** |
| 2012 | 0.3 | 1.1 | 0.7 | 0.2 | 0.3 | 0.4 | 0.2 |
| 2013 | 1.2 | 3.0 | 2.5 | 0.7 | 1.4 | 1.5 | 0.7 |
| 2014 | 1.0 | 2.6 | 2.1 | 0.6 | 1.2 | 1.2 | 0.6 |
| 2015 | 0.5 | 1.7 | 1.3 | 0.3 | 0.6 | 0.7 | 0.3 |
| 2016 | 0.4 | 1.5 | 1.1 | 0.2 | 0.5 | 0.6 | 0.2 |

Using 2013 as the baseline, the averages appeared to stabilize in 2014, at a relatively high level, most notably in grades 3 and 4, with some variation within grade 4 and for the subgroups. Overall, in grade 3, there was an average of almost two zeroes on the CRQs from 2013 to 2016. For grade 4, from 2013 to 2015 (2016 data were not available) the average was steady with 1.6 zeroes registered in 2015. In grades 6 and 8 the averages are lower in 2016 than 2013 and appear to taper off. For grades 5 and 7 the zeroes decrease but there is an uptick in 2016.

Although the averages appear to be small, there were only nine or ten CRQs per exam. In that context, three zeroes means students could not respond to at least 30 percent of the questions in a way scorers deemed to be on task and comprehensible.

As with NYS data, NYC data show that children in grades 3 and 4 are overwhelmed by the CRQs. Notably too, black and Hispanic students averaged more zeroes than their Asian and white peers, indicating a continued—and in some instances increasing—achievement gap after CCLS-alignment.[21] Average zero scores are highest for ELLs and students with disabilities. Again, the sharp differences observed during the transition to the CCLS-aligned tests could be expected, because the new tests were intended to be more stringent than the preceding exams. But even after this changeover, zeroes remained high, particularly for the youngest students, students of color, students with disabilities, and ELLs.

### B. Zeroes and the Achievement Gap

The findings regarding the average number of zeroes led this study to focus on grades 3 and 4, where the overall percentage is highest and the gap among subgroups most extreme. These two grade levels saw eight- and nine-year-olds consistently falter on the ELA exams. The data in *Table 5* show how the ELA reflects the ethnic divide early on and how the divide increases with the advent of CCLS-based testing.

**Grade 3:** In 2012, as shaded in orange in *Table 5*, there was a difference of .5 zeroes (1.2 minus .7) between black and white 3rd graders and a difference of .6 (1.3 minus .7) between Hispanic and white children. After the 2015[21] ELA, the distance between the groups was 1.0 (2.2-1.2) and 1.1 (2.3-1.2) respectively. So, after the third

**Table 6.** New York City—Percentage of Students Receiving 5 or More Zeroes on CRQs, 2012–2016[22]

| GRADE 3 | Total | ELL | SWD | Asian | Black | Hispanic | White |
|---|---|---|---|---|---|---|---|
| 2012 | 4.5% | 13.1% | 15.0% | 2.1% | 4.9% | 6.1% | 2.0% |
| 2013 | 11.3% | 23.7% | 29.9% | 5.1% | 13.9% | 14.5% | 5.2% |
| 2014 | 10.2% | 23.5% | 27.0% | 4.8% | 12.6% | 12.5% | 5.4% |
| 2015 | 13.2% | 30.7% | 34.4% | 6.0% | 15.8% | 16.8% | 6.6% |
| 2016 | 12.7% | 33.5% | 35.5% | 6.0% | 14.7% | 16.7% | 6.1% |
| **GRADE 4** | **Total** | **ELL** | **SWD** | **Asian** | **Black** | **Hispanic** | **White** |
| 2012 | 1.6% | 5.8% | 5.6% | 1.0% | 1.4% | 2.2% | 0.8% |
| 2013 | 9.9% | 29.9% | 32.2% | 4.6% | 11.5% | 13.0% | 4.4% |
| 2014 | 7.3% | 19.8% | 22.8% | 2.9% | 8.7% | 9.7% | 3.2% |
| 2015 | 9.6% | 28.1% | 26.7% | 4.8% | 11.4% | 12.2% | 4.5% |
| 2016 | – | – | – | – | – | – | – |

administration of the CCLS-aligned ELA exam, the gap in the percentage of zero scores grew by an average of .4 zeroes.

**Grade 4:** Similarly, in 2012, the black/white gap was an average of .2 zeroes, and .3 for Hispanic and white students. The gap between both groups and whites was .9 in 2015. It had expanded by an average of .7 and .6 zeroes for minority group students.

Though the focus of this analysis is narrow, it provides quantitative evidence that the gap that exists between minority group and white youngsters in 2012 appears to have widened over the four CCLS-aligned testing cycles, ending in 2015.

### C. Five or More Zeroes—Depth of Difficulty

Students receiving five or more zeroes—those who were not awarded any points on half or more questions designed to test their reading and writing ability and capacity to think critically—presumably had very limited competency in meeting the standards tapped by this set of questions. Using NYC data, and again with a focus on grades 3 and 4, *Table 6* shows the percentages of students in each subgroup who received 5 or more zeroes on the CRQs.

The "Total" column shows that in 2012, fewer than five percent of students received five or more zeroes on the CRQs in either 3rd or 4th grade. In 2013, the percentage increased for all groups, reaching one out of ten children, and then remained relatively high from 2014 through 2016.

The data indicate that ELLs and SWDs were hit the hardest. In many instances 25 percent score at least five zeroes on the questions. A similar picture is unveiled for Black and Hispanic children. In grade 3, from 13 percent to 17 percent of Black and Hispanic students, taken together, received no points on at least half of the questions, after exams were aligned with the Common Core. Grade 4 outcomes are only slightly better at 9 to 13 percent.

**Table 7:** New York City—Number of Grade 3 and 4 Students Receiving 5 or More Zeroes, 2012–2015[23]

| 2012 | 2013 | 2014 | 2015 | 2016 |
|---|---|---|---|---|
| 4,560 | 16,396 | 13,443 | 17,494 | 10,050 (grade 3 only) |

It is important to step back from the impersonal nature of statistics and show the actual number of children who faced many questions that they couldn't answer. *Table 7* conveys the information.

The peak number came in 2015—17,494 NYC children in grades 3 and 4, approximately 12 percent of the grade 3 and 4 students in that year, so baffled by the CRQs that they received zeroes on half, or more, of the questions.

## VII. DISCUSSION AND IMPLICATIONS

Even before the first administration of the CCLS-aligned assessments in NYS, parents and teachers expressed concern about the content of these exams. When these criticisms were dismissed by the SED, parents took matters into their own hands; in 2013 a handful refused to let their children take the 2013 exam. Following that, the number of resisters increased, reaching 20 percent of eligible test takers in 2015, a stunning mobilization of opposition in such a short time.

And that led to this: Governor Cuomo, after having strongly endorsed swift imposition of the Common Core and accompanying testing, including the use of test scores in teacher evaluations, saw a need to adjust his position. He assembled the Common Core Task Force; the charge was to gather information and feedback from the public in a series of meetings around the State concerning the "Common Core" and to bring recommendations back to him about its implementation. The first session of this task force was convened on October 29, 2015 at the College of New Rochelle.

Tensions were high. One outspoken public school parent pointed to a lack of research about the merits of the Common Core and the accompanying testing.

> The CRQs that SED and Pearson approved for use on these tests appear patently inappropriate for the youngest test-takers, ELLs, and students with disabilities.

She denounced developmentally inappropriate tests with reading passages that were above grade level and ambiguously-worded questions that caused undue stress in children as they struggled to complete the tests within an unrealistic timeframe. Task force members were wary of these assertions. Said one member: "I keep hearing all the complaints about the tests—that things like age-inappropriate, too long, too frequent, too hard, obtuse vocabulary, trick questions, misaligned—all of those sort of things. And I'm asking… I just want to know what research evidence are people relying on when they can say these things about the test?" The meeting was closed when another task force member called for "some analytical work that underlies the expressions, the assertions that were raised today."[24]

Now, after an unnecessarily long and winding road to secure the necessary data from SED,[25] we finally have evidence to address these questions. Our findings vindicate those who protested the testing as unnecessarily difficult, especially for our youngest students, ELLs, and students with disabilities. And our work permits, for the first time, an overdue, independent postmortem examination of the tests that does not rely on assertions by SED and the publisher that the tests were properly assembled.

Two avenues of discussion follow. The first proceeds from this study's analyses and findings. The second addresses of the importance of transparency in SED processes.

Recall that our research focuses on the constructed response questions on the ELA exam, specifically the zero scores, which are given when a student's response to a question is "totally inaccurate," "unintelligible," or "indecipherable." Our research reveals a sharp increase in the percentage of students receiving zeroes on the CRQs in the ELA tests from 2012 to 2013 when the CCLS-aligned tests debuted. Most striking, the percentage of zero scores was highest—and remained consistently so over the years of this study—for grade 3 students, the youngest test takers. Students in grade 4 clearly struggled with the material as well. NYC data allow further analysis of the results by subgroups. ELL students and students with disabilities had a substantially higher percentage of zero scores than their general education peers; likewise black and Hispanic students fared worse than white and Asian students. It is important to note, here, that SED is seeking a waiver from the USDOE that would allow ELLs to be tested at their level of language proficiency, rather than their age level; a similar waiver is being sought for students with disabilities. This is a commendable, reasonable first step. Nevertheless, we remain concerned about this type of testing for these groupings and for our youngest students.

The important thing is this: The CRQs that SED and Pearson approved for use on these tests appear patently inappropriate for the youngest test-takers, ELLs, and students with disabilities. Questions that yield so many zeroes do not return much substantive, diagnostic information about test takers.

Moreover, the elimination of time limits in 2016, rather than a benign concession, is a glaring admission that the test development was fundamentally flawed. Why did it take SED and Pearson three years to make this correction; issues of timing should be resolved by publishers at the time of test design and piloting, before the operational tests are given. This is an acknowledgement of failure by Pearson to perform its contractual obligations responsibly. The decline in zeroes in 2016 further reinforces this consequential oversight. And, to further complicate matters, the solution proposed—to give students unlimited time

on the tests—meant that tests results could no longer be compared from one year to the next.

Questions that do not yield meaningful information for our youngest students and a tacit admission of flawed test development? The public deserves some accounting for this. And yet, SED has not been forthcoming with data needed for scrutiny by independent reviewers. In addition to not releasing test questions, as was done in pre-Pearson years, SED stopped releasing a full range of item-level statistics in its technical reports.

Moreover, Pearson and SED need to explain how the CRQs were field tested: When and where did the field testing take place, who was sampled and how big was the sample size, what statistics were generated by children trying out these questions? That is, did Pearson know how many zeroes were likely when the CRQs became operational? How good was Pearson's "intelligence" about how the questions would work when posed, for instance, to 200,000 children in grade 3?

The response from SED will surely be that they replaced Pearson with a new testing company, Questar Assessment. This is true. While the Pearson contract ended in 2016, the SED continued to use Pearson-created questions for testing beyond that year. Nevertheless, what has been lacking since the advent of CCLS-aligned testing is a way to hold the testers themselves accountable. It may be that there is a new testing company in town, but without understanding the nature and depth of Pearson's flawed work over five years, we may well repeat those mistakes with Questar.

## VIII. CONCLUSION

Our boldest conclusions tie together important aspects of the testing story: children upset and dumbstruck by the exams, especially the youngest ones; unhappy parents whose views were disparaged; SED's suppression of data needed by the public, especially parents to stay informed and make intelligent decisions about their children's education; the surge in zero scores and omissions that this study uncovered; ill-conceived tests and their perpetuation; the strong case parents have for opting out; the overriding need for transparency, timely data and unfettered review by analysts. These rest most solidly on findings for grades 3 and 4, and for ELLs, students with disabilities, and minority students.

In the final analysis, we are dealing with children here at a formative time in their lives, when education matters most. For every discussion and news story about the increase or decrease in test scores, we must remember that behind each statistic is a child—a young child—who lives each day with the decisions that we make about testing. The 3rd graders who took the first CCLS-linked test in 2013 are taking the 8th grade test this spring. Everything that has been wrong with the core-aligned tests has framed the education of these young people.

It's time to create a legitimate assessment process, unified with standards and curricula that work in harmony to foster the development of every child's intellect, abilities, and dreams. Federal education law, the Every Student Succeeds Act (ESSA), dictates that we test our young students in math and ELA each year. We must determine how to do that in a way that serves children and the educational goals we value.

## References


[1] The agreement with Pearson began in January 2011 and ended in June 2016. The total cost was $38.8 million. Source: New York State Comptroller's online Open Book New York—Contract# C010713. Questar Assessment, Inc. succeeded Pearson with a five-year contract in the amount of $44.7 million—Contract# C012427 to develop the assessments in elementary and intermediate schools in ELA and mathematics from August 2015 until November 2020. For 2016, SED used questions developed by Pearson. Questar's role was to create "test forms and guidance material" for the exams. http://www.p12.nysed.gov/assessment/ei/2016/changes2016grades3-8ela-math-tests.pdf

[2] Educator Guide to the 2013 Grade 3 Common Core English Language Arts Test, EngageNY, 2013, http://www.p12.nysed.gov/assessment/ei/2013/guides/grade-3-ela-guide.pdf; language is the same in educators' guides for other grades and years and can be found on the EngageNY website.

[3] https://www.nytimes.com/2015/08/13/nyregion/new-york-state-students-standardized-tests.html

[4] Field Memo, Implementation of the Common Core Learning Standards, pg. 3, https://www.engageny.org/resource/field-memo-transition-to-common-core-assessments

[5] Accountability requirements of Elementary and Secondary Schools Act was reauthorized in 2002 (called No Child Left Behind Act), this included annual testing of ELA and math in grades 3 through 8. Such exams were first administered in NYS in 2006.

[6] Educator Guide to the 2013 Grade 3 Common Core English Language Arts Test, EngageNY, 2013, page 5. http://www.p12.nysed.gov/assessment/ei/2013/guides/grade-3-ela-guide.pdf. This language is reflected in all Educators' Guides for CCLS-aligned testing for all grades, multiple years.

[7] NYSED, State Education Department Releases Grade 3-8 Assessment Results, August 7, 2013. http://www.nysed.gov/news/2017/state-education-department-releases-grades-3-8-assessment-results

[8] NYS Testing Program, Grade 3, Common Core, ELA, Released Questions with Annotations, August 2013, EngageNY, p. iii.

[9] Educators' Guides to ELA Testing, Grades 3-8, 2013-2016. http://www.p12.nysed.gov/assessment/ei/eiguides-13.html; https://www.engageny.org/resource/test-guides-english-language-arts-and-mathematics

[10] Data do not contain student identifiers.

[11] See ELA standards and standards-setting criteria, http://www.corestandards.org/

[12] These criteria remained constant in the years that the CCLS-aligned tests have been administered (2013-current), as well as in 2012, the year preceding CCLS-alignment.

[13] See Appendix A for a copy of the 2012, 2013, 2016 Grade 3 rubric. Available at www.newpaltz.edu/benjamincenter

[14] Pizmony-Levy, O. and Green Saraisky, N. (2016). Who opts out and why? Results from a national survey on opting out of standardized tests. Research Report. New York: Teachers College, Columbia University.

[15] Infante, A., Swerdzewski, A. (January 2016). Memo: Changes for the 2016 Grades 3-8 English Language Arts and Mathematics Tests. New York State Education Department, Office of Instructional Support, Office of Assessment, page 3. http://www.p12.nysed.gov/assessment/ei/2016/changes2016grades3-8ela-math-tests.pdf

[16] Even in the absence of grade 4 data for 2016, it is reasonable to assume they chart a path similar to the one observed in Graph 1 consistent with the congruent results in the other grades. It is likely, therefore, that the average for grade 4 is near 13 percent.

[17] Appendix B provides charts that show the combined effect of zeroes and unanswered questions. Available at www.newpaltz.edu/benjamincenter

[18] For a discussion of this, and other changes that the SED espoused would improve testing, see: Smith, F "CityViews: As Standardized Tests Loom, Improvements are Illusory," City Limits, May 24, 2016. https://citylimits.org/2016/03/24/cityviews-as-standardized-tests-loom-improvements-are-illusory/

[19] Appendix C shows the number and percentage of minority group students in the NYC test population. Available at www.newpaltz.edu/benjamincenter

[20] The averages for minority group students were close to the percentages for the entire New York City test population—a statistical concomitant, since black and Hispanic students made up 69% of the test takers (see Appendix C).

[21] 2015 is used as an endpoint because it allows New York City's grade 4 to be included. The influence that opting out of the exams had on the zero scores in 2015 remains unknown, although the number of opt outs in NYC that year was relatively small.

[22] Appendix D shows the NYC distribution of zeroes on CRQs, grades 3 to 8, 2013–2016. Available at www.newpaltz.edu/benjamincenter

[23] 2016 data are grade 3 only; we do not have 2016 data for grade 4.

[24] New York Common Core Task Force, New Rochelle, October, 19, 2015.

[25] NYCDOE was forthcoming with the data.

## Biographies

**FRED SMITH** served as a researcher for the New York City Board of Education from 1969–2001, specializing in test-related assignments. These included test development projects, the analysis and reporting of results, and the conduct of federal program evaluations. His first year, 1969, coincided with the inception of the citywide testing program mandated by New York State's school decentralization law. Many aspects of the current New York State testing derive from that program.

Since retiring from the New York City Board of Education, Mr. Smith has consulted on testing issues, pursued self-initiated studies of state and city exams, written critical Op-Eds on the subject, and strongly advocated for a rational and transparent assessment system. He holds a Professional Diploma in Measurement and Evaluation from Teachers College, Columbia University. Mr. Smith is a graduate of the City's public schools, as are his daughter and son. He lives in the Bronx. He uses his skills to keep official statistics for the NY Jets and the NFL.

**ROBIN JACOBOWITZ, PhD,** is the director of education projects at the Benjamin Center for Public Policy Initiatives at SUNY New Paltz. Previously, Robin worked with Janice Hirota and Associates on an evaluation of school reform initiatives in New Orleans, Washington DC, New York City, and Dallas. She also worked at New York University's Institute for Education and Social Policy, where her research centered on charter schools, New York City small high schools, and leadership transitions in new schools in New York City. She worked with the University of Chicago's Chapin Hall Center for Children, where her research focused on the relationship between constituency building and policy work in effecting systemic school reform in New York State.

Robin holds a MEd in education policy from the Harvard University Graduate School of Education, and a Ph.D. from the Robert F. Wagner Graduate School of Public Service at New York University. She is currently a trustee on the Kingston City School District Board of Education and serves on the executive committee of the Ulster County School Boards Association.

**Editorial staff**
Robin Jacobowitz
Gerald Benjamin
Janis Benincasa

*Database queries available on request*

THE BENJAMIN CENTER
SUNY NEW PALTZ

910350-99

The Benjamin Center for Public Policy Initiatives
State University of New York at New Paltz
1 Hawk Drive
New Paltz, NY 12561-2443

ADDRESS SERVICE REQUESTED

Nonprofit Organization
U.S. Postage
P A I D
Permit #6127
Newburgh, New York

## THE BENJAMIN CENTER for Public Policy Initiatives

Independently and in collaboration with local governments, businesses, and not-for-profits in the Hudson Valley, The Benjamin Center (formerly CRREO):

- **conducts studies on topics of regional and statewide importance;**
- **brings visibility and focus to these matters;**
- **fosters communities working together to better serve our citizenry;**
- **and advances the public interest in our region.**

The Benjamin Center connects our region with the expertise of SUNY New Paltz faculty. We assist in all aspects of applied research, evaluation, and policy analysis. We provide agencies and businesses with the opportunity to obtain competitive grants, achieve efficiencies and identify implementable areas for success.

**www.newpaltz.edu/benjamincenter**