



U.S. DEPARTMENT OF EDUCATION
INSTITUTE OF EDUCATION SCIENCES

NATIONAL CENTER FOR EDUCATION STATISTICS

April 9, 2018

John C. White
Louisiana State Superintendent of Education
PO Box 94064
Baton Rouge, LA 70804-9064

Dear Superintendent White,

Thank you for your letter regarding the upcoming release of mathematics and reading results from the 2017 administration of the National Assessment of Educational Progress (NAEP). I also want to thank you for our phone conversation on March 29 that I trust provided helpful background information. The intent of this letter is to more formally respond to your letter of March 23 with answers to the specific questions you raised.

As you indicated in your letter, 2017 was indeed an important transition year for NAEP, as it marked the transition of the mathematics and reading instruments from paper-based assessments (PBA) to digitally based assessments (DBA). The National Center for Education Statistics (NCES) is always careful about any changes made to the assessments. In the case of the PBA to DBA transition, it was assumed in designing the 2017 administration that a mode effect would be observed—that is, without proper linking, scores from the two versions of the assessments would not be comparable. This assumption was based, in part, on the experiences of other programs and past research, some of which you cited in your letter.

It is also clear from your letter and our phone conversation that you have a firm understanding of the approach taken by NCES while designing this important transition. In order to provide some context for the additional information you requested, however, I will first highlight some important aspects of the analyses that led to the trend reporting decisions we made for the nation, states, and districts. It is my hope that in reviewing the additional information provided here, you will agree that this transition was conducted with integrity and scientific rigor. Of course, NCES stands ready to provide whatever additional support we can as you prepare for the public release of the assessment results.

I would like to confirm that we are confident that the data we will release on April 10 will provide the fair, meaningful, and honest comparisons of student performance over time for the nation, states, and districts that are the hallmarks of the NAEP enterprise.

Overview of the DBA Transition Analysis and Trend Reporting Decisions

NAEP is probably one of few assessment programs that place a premium on maintaining trends while transitioning to DBA. This is due to many factors, including our legislated mandate to report on student performance trends, as well as the value placed on these trends by educators, policymakers, and the general public.

The value of these trend data to so many of our public is why NCES designed the 2017 administration to collect the data needed to determine if meaningful trends could be reported across paper and digital modes of the assessment. The PBA and DBA versions of each assessment were administered to randomly equivalent groups of students so that we could evaluate the mode effect both within and among states. Randomization was done within schools to optimize the comparability of samples of students who took paper versus digital versions of the assessments. Approximately 500 PBA and 2,200 DBA students were sampled in each state, resulting in a total national sample of roughly 40,000 PBA students and 150,000 DBA students within each subject and grade.

Of primary concern to us in this transition was the impact of changing from a paper-based mode to a digitally based mode of administration for all types of students across the nation and for all the states and districts that rely on NAEP to provide valid performance trends. As you correctly observed in your letter, the national linking of PBA and DBA scores provided for an overall adjustment of the observed mode effect. Once this was completed, NCES proceeded to evaluate the degree to which the linkage established for national PBA and DBA samples held for national subgroups, as well as for states and districts.

In a series of carefully planned empirical studies and evaluations to examine the impact at the state and district levels, and in consultation with several expert advisory panels, we concluded that the small number of mode effects observed were inconsistent within a state across grades and subjects and mostly not statistically significant. In fact, the amount of variation we observed was largely consistent with what might be expected from sampling variability. Given the lack of any overwhelming evidence to the contrary, NCES determined that it was feasible and scientifically defensible to maintain the existing trend lines that now extend back for more than 25 years.

Additional Information Requested in March 23 Letter

As I mentioned earlier, I hope the additional information provided here will be helpful and responsive to your request—but, of course, we look forward to providing any additional information, clarification, or support you might need. You are an important consumer of the valuable data we collect in NAEP, and it is our goal to make them as useful and meaningful to you as possible. I have structured this part of the letter around the four specific types of information you requested.

1. The mode effect adjustment applied to each grade and subject nationally

The rigorous experimental design and the large size of the national PBA and DBA samples provided for a robust linking of the two assessments, which allows for

differences in difficulty between the PBA and DBA at the national level. PBA scale scores were estimated by first establishing a link between the 2015 and 2017 PBA results, as NAEP typically does through common sets of items administered in both years. DBA scale scores were then estimated through randomly equivalent-groups linking of the 2017 PBA and DBA results based on them being randomly drawn from the same population.

The best measure of the difficulty of the two assessments is probably a statistic we call “the average p+” (this is essentially the average percentage of students answering each question correctly, or mean score in the case of constructed response items). As had been anticipated by previous research and other programs that have made this transition, the average p+ of the PBA items were higher than that of the DBA items in each grade and subject. This means that, on average, a higher percentage of students answered the paper-based questions correctly than the digitally based questions. In other words, the paper version appeared to be easier for most students on most items.

The table below provides a summary of the differences between PBA and DBA average p+ for each grade and subject. In essence, this can be interpreted as the “mode effect” expressed as the difference between the difficulties of the two assessments.

| P+ differences between PBA and DBA versions of mathematics and reading assessments (paper minus digital) | | |
|--|---------|---------|
| | Grade 4 | Grade 8 |
| Mathematics | 3.7 | 2.9 |
| Reading | 5.3 | 1.9 |

The table above does indicate a couple of interesting patterns that we examined in great detail as we conducted our analyses. First, the differences were smaller at grade 8 than they were at grade 4, which indicates a smaller mode effect at the older grade level. There was also little difference between the two subjects at grade 8. At grade 4, however, the mode effect appeared larger for reading than for mathematics. These data are provided here as an indication of the mode effect observed, but as indicated earlier, **the design of the administration and the linking of the PBA and DBA versions made it possible to account for these differences in aligning the two scales at the national level.**

2. The average mean scores for students taking the paper-based test and for students taking the tablet-based test, at the state level and at the national level, in each grade, subject, and subgroup

In the following table, you will find the average DBA scores for Louisiana, as well as for the nation and two national subgroups (gender and race) after performing the national alignment. In addition, you will see an estimate of average scores after applying a state-specific adjustment for the assessment mode (see the column in the table labeled *PBA Estimate*). The state-specific mode effect is the difference in the estimated means from DBA and PBA in 2017. Due to the size of the PBA sample in each state, we are not able to estimate scale scores for state-level subgroups. I intend to make the same information available upon request for the other states and districts. NCES will also publish a technical white paper that will include all of this information in one place soon after the public release of the results.

2017 DBA and PBA score estimates for Louisiana, the nation, and national gender and race/ethnicity subgroups

| | Grade 4 | | | Grade 8 | | |
|----------------------|--------------|--------------|------------|--------------|--------------|------------|
| | DBA Reported | PBA Estimate | Difference | DBA Reported | PBA Estimate | Difference |
| Louisiana | | | | | | |
| Mathematics | 229 | 228 | +0.5 | 266 | 269 | -2.5 |
| Reading | 211 | 214 | -3.0 | 256 | 255 | +1.5 |
| Nation | | | | | | |
| Overall | | | | | | |
| Mathematics | 240 | 240 | 0.0 | 283 | 283 | 0.0 |
| Reading | 222 | 222 | 0.0 | 267 | 267 | 0.0 |
| Male | | | | | | |
| Mathematics | 241 | 241 | -0.2 | 283 | 283 | +0.5 |
| Reading | 219 | 218 | +0.5 | 262 | 261 | +0.7 |
| Female | | | | | | |
| Mathematics | 239 | 238 | +0.2 | 282 | 283 | -0.5 |
| Reading | 225 | 226 | -0.5 | 272 | 272 | -0.7 |
| White | | | | | | |
| Mathematics | 249 | 249 | -0.1 | 293 | 291 | -0.1 |
| Reading | 232 | 233 | -0.9 | 275 | 276 | -1.4 |
| Black | | | | | | |
| Mathematics | 222 | 223 | -0.5 | 260 | 259 | +0.6 |
| Reading | 205 | 205 | +0.2 | 249 | 248 | +1.5 |
| Hispanic | | | | | | |
| Mathematics | 229 | 229 | 0.0 | 269 | 269 | 0.0 |
| Reading | 209 | 208 | +0.9 | 255 | 253 | +1.7 |
| Asian/Pacific | | | | | | |
| Mathematics | 261 | 259 | +1.8 | 313 | 313 | +0.4 |
| Reading | 242 | 239 | +2.7 | 284 | 284 | -0.2 |

For Louisiana, as well as for the nation and national subgroups (with the exception of fourth-graders with disabilities on the mathematics assessment) we found no statistically significant differences between the DBA and PBA results. Moreover, we found no consistent pattern in the apparent (non-significant) differences between DBA and PBA results. For example, the fourth-grade mathematics DBA average score in Louisiana was 0.5 points higher than the PBA average, while the reading DBA average score was 3 points lower than the PBA average. At grade 8, the apparent differences were reversed. None of these differences, however, were statistically significant.

It is worth noting that we performed similar tests of significance between overall DBA and PBA results across all the 52 states and jurisdictions. This produced 208 comparisons for the two subjects at the two grade levels.

We found the following results for those 208 observations:

- Only 21 out of 208 comparisons were statistically significant.
- The 21 statistically significant differences were spread across 19 states.
- Only two states had more than one statistically significant PBA vs. DBA difference across the four subject/grade comparisons.

We detected the above differences without statistically adjusting for the fact that running many comparisons such as this at once tends to produce “false positives”—cases in which we detect a difference when there really is no difference. When we did control for multiple comparisons, we found just two differences across the 52 jurisdictions, two subjects, and two grades (out of 208 total comparisons).

Despite finding so few statistically significant differences between PBA and DBA scores, we continued to dig into the data to see if, for any of these 21 states, there might be a connection between the mode difference and the scale score trends to be reported for that state. Our investigations concluded the following:

- In 13 of the 21 cases, there were no differences in the “trend story” between the DBA and PBA results.
- Of the eight cases in which the trend result would have been different for PBA than it is for DBA, seven of them resulted in the state having a more positive trend outcome with the DBA results.

3. Evidence of the random equivalence of the groups of students taking the paper-based test and the students taking the tablet-based tests, at the state level and at the national level

Attached are two charts showing demographic comparisons of the national and Louisiana PBA and DBA samples. I hope you will agree that the two samples taking the paper-based and digitally based assessments are nearly identical in terms of these major demographic characteristics. We believe this provides strong support for the random equivalence of the groups of students taking the assessments in each mode.

4. National and subgroup performance trends, reported by performance quintile, quartile, or decile

Attached is an Excel file that contains national and subgroup performance trends at the 10th, 25th, 50th, 75th, and 90th percentiles. The file contains four spreadsheets—one for each subject/grade. At the national level, we do see an overall tendency for scores to have lowered since 2015 among lower-performing fourth-graders, and scores to have risen since 2015 among higher-performing eighth-graders.

Conclusion

I certainly hope that you find this additional information and data helpful in preparing for the release of the NAEP 2017 mathematics and reading results. I was deeply grateful for your thoughtful letter about the embargoed data that was shared with you and your staff prior to the release, and the care with which you are approaching the use and interpretation of these important data. I reiterate that my team and I stand ready to support you in any additional ways that we can.

We at NCES are very confident in the scientific rigor of the approach we have taken to design the NAEP digital transition and to conduct the analyses that led to the results that will soon be released. We believe that the full extent of the evidence we reviewed confirms that we can and should maintain NAEP's extremely valuable trend lines (spanning more than a quarter of a century) through this digital transition. Given the importance and critical impact of this decision on many of our key constituents (like yourself), we maintained close and constant contact throughout the process with several panels of national experts, including the NAEP Quality Assurance Technical Panel, the NAEP Design and Analysis Committee, the NAEP Validity Studies Panel, and the National Assessment Governing Board's Committee on Standards, Design, and Methodology.

As a reminder, all information shared in this letter, its attachments, and all other forms of communication about these data is embargoed, and the Louisiana Department of Education has agreed to honor the terms of the embargo set by NCES. As such, these data may not be copied, published, announced, or in any other way made public—including email, voicemail, and other electronic forms of communication—prior to their official release on April 10, 2018.

Again, thank you for being a valued partner in the NAEP program and a key constituent who certainly has much at stake in the reliability and validity of NAEP results. Please let me know how I can be of any additional service to you.

Sincerely,



Peggy G. Carr, Ph.D.
Associate Commissioner
National Center for Education Statistics

Enclosures

cc: Jessica Baghain, Louisiana Department of Education
James Blew, U. S. Department of Education
Bill Bushaw, National Assessment Governing Board
Carissa Miller, Council of Chief State School Officers