2017 NAEP Transition to Digitally Based Assessments in Mathematics and Reading at Grades 4 and 8: Mode Evaluation Study

White Paper

Paul Jewsbury, Robert Finnegan, Nuo Xi, Yue Jia, ETS
Keith Rust, Westat
Samantha Burg, NCES

In Collaboration With

Patricia Donahue, John Mazzeo, ETS
Brian Cramer, Optimal Solutions Group
Amy Lin, Natalia Weil, Westat

March 2020

Contents

Tables

Figures

Executive Summary

On April 10, 2018, the National Center for Education Statistics (NCES) released results from the main National Assessment of Educational Progress (NAEP) 2017 mathematics and reading assessments at grades 4 and 8 for the nation, states, and 27 districts that are part of NAEP's Trial Urban District Assessment (TUDA). The NAEP program officially transitioned from paper-based assessments (PBAs) to digitally based assessments (DBAs) in mathematics and reading in 2017. This white paper details the purpose, technical methodology, and results of evaluations conducted as part of the transition to DBA.

To adapt to an increasingly digital world, NAEP began studying a transition to digitally based assessment in the late 1990s. Through a series of pilot studies comparing paper and digitally based assessments, NAEP concluded that taking an assessment on computer had a small, but meaningful, effect on performance on the test. Given the central mandate of NAEP to monitor achievement changes across time, transitioning to DBA presented a potential challenge to the validity of continued trend comparisons; however, in not transitioning, NAEP would risk not measuring students' skills in the most relevant mode for an increasingly digital society.

When NAEP has been faced with necessary instrument or administration changes in the past, carefully designed bridge studies were conducted to allow for comparisons across years. As part of the 2017 mathematics and reading assessments, NAEP conducted its most ambitious bridge study to date to evaluate the effect of the mode of administration on performance and to try to allow for comparisons of the 2017 results to later assessments administered digitally, as well as to the earlier assessments administered on paper.

The 2017 bridge study involved randomly equivalent samples receiving the NAEP mathematics or the NAEP reading assessment in either the digital or the paper format. The digital instruments were designed to resemble the paper instruments as closely as reasonable to minimize content differences and facilitate the mode comparisons. The 2017 bridge study confirmed small but meaningful differences in performance depending on mode, as found in earlier NAEP studies. To link the digital and paper assessment scores in a way that accounts for instrument differences, NAEP followed generally accepted best practice in psychometrics and employed common population linking, a technique applied by NAEP in the past.

To evaluate the validity and appropriateness of the common population link, NAEP systematically evaluated the alignment of the linked scores from the digital and paper assessments across the proficiency range, major reporting subgroups, and states and participating urban districts. Few statistically significant differences were observed after applying the linking transformation. Most statistically significant differences were not associated with different reporting outcomes on the 2015 and 2017 score changes, and all had small effect sizes.

In summary, based on an exhaustive number of analyses conducted by the authors of this paper and vetted by numerous external advisory panels, there was no clear evidence of consistent bias in the linked results. Consequently, NCES decided to report all the state/district trends based on the national PBA-DBA linking.

1. Introduction

A central mission of the National Assessment of Educational Progress (NAEP), as the largest nationally representative and continuing assessment of what America's students know and can do in various subject areas, is to measure and report academic progress over time, or trends. As the educational landscape has evolved to place greater value on technology-related skills and outcomes, and as technology has played a larger role in classroom instruction and assessment, the NAEP program has begun transitioning to digitally based assessments (DBAs). Excluding pilots and field trials, the first operational DBA was the NAEP writing assessment in 2011 (National Center for Education Statistics [NCES] 2012), followed by a new assessment for technology and engineering literacy (TEL) in 2014 (NCES, 2016). In 2017, the operational mathematics assessment and the reading assessments were administered digitally for the first time to students at grades 4 and 8.

As a measure of academic proficiency dating back decades, the design of NAEP has always involved balancing two competing pressures (Beaton, 1990a): (1) the pressure to minimize the changes in the instrument across years, so that differences in proficiency estimates can be unambiguously interpreted as changes in the performance of the population on fixed constructs (Beaton, 1990b); and, (2) the pressure to adapt to changes in the educational landscape to ensure that NAEP measures the knowledge and skills that are valued and relevant to the policy goals of NAEP (Zwick, 1992). The 2017 NAEP mathematics and reading assessments were carefully designed to balance these two competing pressures and to allow NAEP to study extensively the risk to continued trend reporting.

The potential risk to continued trend reporting was the possibility that changes in the assessment instrument and administration would result in reporting large proficiency changes for a student group or jurisdiction that do not accurately reflect growth or decline in the NAEP subject matter proficiency; or, conversely, that meaningfully large changes that do reflect actual changes in subject matter proficiency are not reported. To discontinue reporting trends, however, would be to lose comparisons of results over decades for the main NAEP mathematics and reading assessments. Retaining continued trend reporting was considered a priority to NAEP, if scientifically defensible (National Assessment Governing Board [NAGB], 2015).

This paper describes the research conducted to evaluate the impact of the digital transition on the main NAEP mathematics and reading instruments and assessment results. In this chapter, the motivation and context for the digital transition, prior NAEP studies, and research questions of this study are described. Chapter 2 provides a description of the design of the 2017 assessments and digital transition study, chapter 3 the analysis methods, and chapters 4-6 the analysis results. Finally, chapter 7 provides concluding comments about the 2017 digital transition.

1.1 Motivation and Context for the Digital Transition

Throughout the process of planning the transition to DBAs, NAEP program leaders at the National Center for Education Statistics (NCES) and the Governing Board considered the changing context of learning in the United States. Growing evidence in policy, research, and practice suggests digital technologies are transforming instruction and assessment across the nation (Pearson, 2015; SRI, 2018). Innovative curricula and digital tools have already shifted the way K-12 students learn to read, write, and do mathematics. Additionally, federal education legislation emphasizes digital learning (SRI

International, 2018). The NAEP survey results from 2017 showed that between 95 and 99 percent of the sampled students, regardless of school location, district, or state, had access to a computer in their classroom (NCES, 2018).

Surveys have shown that students are using computers more frequently in their classrooms and that most teachers see digital technologies as essential to teaching and learning (Common Sense Media, 2013). According to the NAEP survey results in 2015, the use of computers for certain tasks and activities for classroom learning and instruction in mathematics and reading increased in 2015 when compared to 2013 and 2009. As part of gathering contextual information to support the reporting of 2017 NAEP results, SRI reviewed the current status and use of digital technologies in U.S. classrooms. For mathematics, SRI found that the use of technology in classrooms tended to focus primarily on personalization of instruction and for practice. In addition, SRI identified the use of emerging technological features to support collaboration, exploration, and problem solving in mathematics classrooms (SRI, 2018). Similarly, for reading and writing, SRI found that digital tools were used primarily to create opportunities for targeted instruction, including supporting close reading skills and engagement in collaborative and multimedia-supported writing processes. There is growing evidence that schools are using or exploring ways to provide instructional content through digital means (SRI, 2018).

Use of Computers for Mathematics and ELA State Assessments at Grades 4 and 8, 2016-17

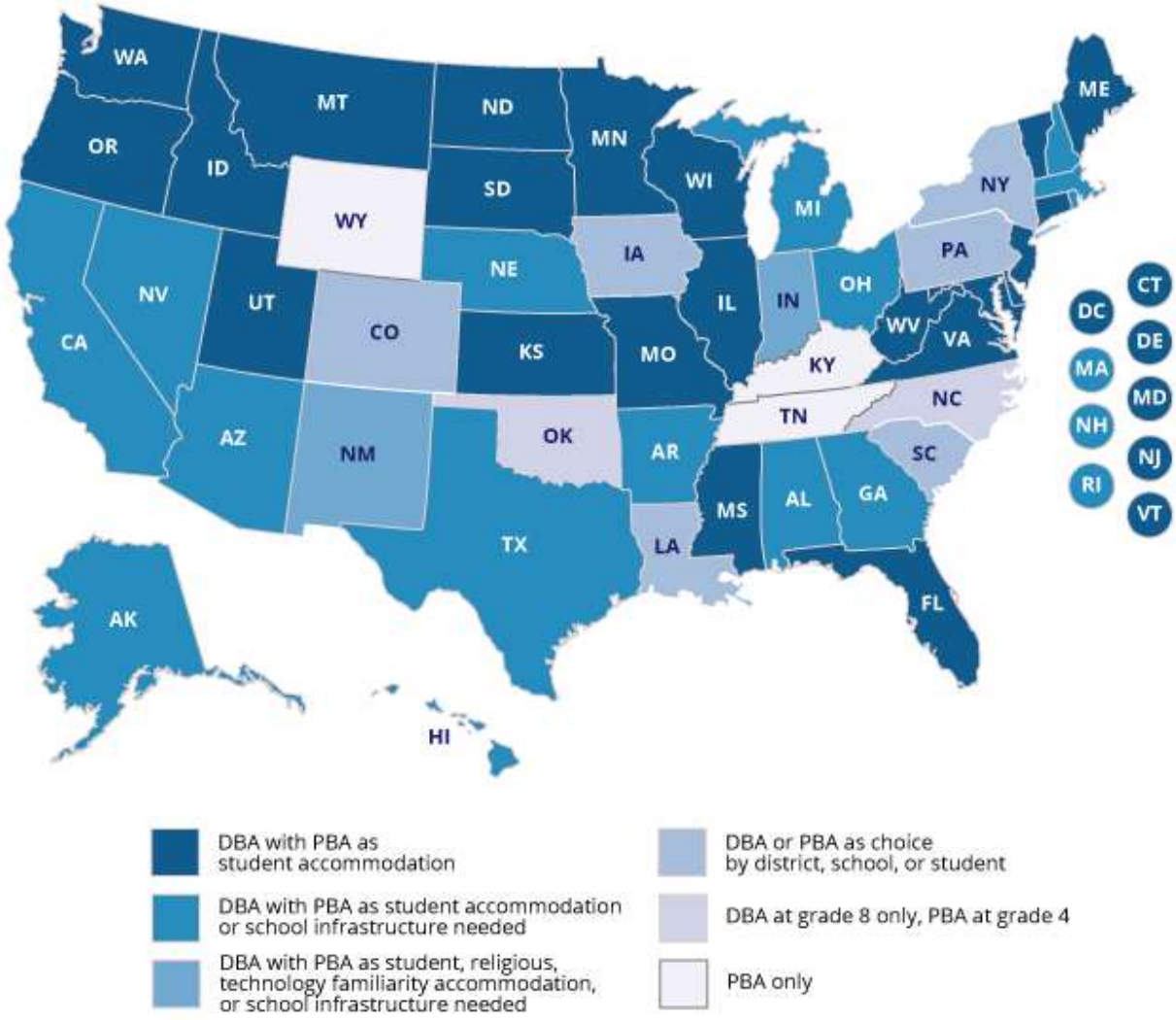
In addition to a general shift toward learning and teaching with digital technologies, the administration of summative assessments has been shifting toward digital delivery. In recent years, states have started to transition away from the traditional paper administration of their end-of-year mathematics and English language arts (ELA) assessments and toward administering those same assessments in a digital format.

In figure 1.1a, based on the state assessment's administration mode at grades 4 and 8, states are grouped into six categories by their use of digitally based assessment (DBA) and paper-based assessment (PBA):

(1) DBA at both grades 4 and 8 with PBA as an accommodation for students with disabilities;
(2) DBA at both grades 4 and 8 with PBA as an accommodation for students with disabilities, or when the school lacked the infrastructure to support DBA administration;
(3) DBA at both grades 4 and 8 with PBA as an accommodation for students with disabilities, students with religious reasons, students lacking necessary computer skills, or when the school lacked the infrastructure to support DBA administration;
(4) DBA or PBA by choice of districts, schools, or students;
(5) DBA offered at grade 8 only, and PBA available as an accommodation for students with disabilities or when the school lacked the infrastructure; and
(6) PBA only.

Forty-five U.S. states and the District of Columbia used some form of digitally based mathematics and ELA state assessments at grades 4 and 8 in the 2016-17 school year, though many of these jurisdictions also provided PBA as an option.

Figure 1.1a. Administration of state assessments in English language arts (ELA) and mathematics in grades 4 and 8, by type of assessment mode and state: school year 2016-17



NOTE: PBA = paper-based assessment. DBA = digitally based assessment.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, https://nces.ed.gov/programs/statereform/tab2_22.asp.

In addition to state assessments, several international assessments are moving or have moved to DBA. The Program for International Student Assessment (PISA) began transitioning to a digital platform in 2006 by experimenting with digitally based science assessments (Beller, 2013). As part of its main operations, PISA has administered optional DBAs in reading literacy since 2009, mathematics literacy since 2012, and science literacy since 2006, as well as in problem solving in 2012, collaborative problem solving in 2015, and global competency in 2018. As of 2018, 70 countries were participating in PISA DBAs. Other international assessments, such as the International Computer and Information Literacy

Study (ICILS), the Progress in International Reading Literacy Study (PIRLS), and the Trends in International Mathematics and Science Study (TIMSS) have either already offered DBAs or are in the process of transitioning to DBA.

1.2 Prior NAEP Studies on the Digital Transition

NAEP has been exploring the transition to digital administration since the late 1990s. In 1999, NCES commissioned three early studies evaluating the impact of digital administration on assessments for mathematics, writing, and problem solving with technology. Three field investigations were conducted: (1) the 2001 mathematics online (MOL) study (Sandene et al., 2005; Bennett et al., 2008); (2) the 2002 writing online (WOL) study (Sandene et al., 2005); and (3) the 2003 problem solving in technology-rich environments (TRE) study (Bennett, Persky, Weiss, & Jenkins, 2007).[1]

All three studies provided key insights for the development of digitally based NAEP assessments, but the results from the 2001 MOL study were particularly relevant to later developments in the digital transition on mathematics. In the spring of 2001, NAEP conducted a study to compare eighth-grade student performance on paper and digital formats of the mathematics assessment. An existing NAEP block from the operational NAEP assessment comprised of 26 questions was converted to the digital format. Approximately 1,000 students took the digital version of the block, and approximately 1,000 students took the original paper version. Prior to receiving the paper or digital block, both groups of students received another block composed of 20 NAEP mathematics items, but this block was administered to both student groups in paper format.

The MOL study revealed two important aspects of the properties of the mathematics items converted to digital format (Bennett et al., 2008). First, the properties of the mathematics items in digital format were highly correlated to the properties of the items in the original paper versions. In terms of Item Response Theory (IRT) item parameters, the discrimination parameters were correlated at $r = .86$ and the difficulty parameters were correlated at $r = .96$. Second, the difficulty parameters were generally slightly higher for the mathematics items in digital format. This suggests that the conversion to digital format did not strongly alter the psychometric properties of the converted mathematics items, but it did slightly increase the difficulty of the items in a way that was relatively consistent across all of the items.

At the item level, there was a mean difficulty difference of .05 on the proportion-correct scale between the digitally based and paper-based assessments (Sandene et al., 2005), and this difficulty difference translated to about 4 scale score points on the NAEP mathematics scale (Bennett et al., 2008). This suggests that if the differences in the difficulty between the two formats were not taken into account, students taking a digitally based mathematics assessment may be expected to be disadvantaged by about 4 scale score points. For this assessment, 4 scale score points was .14 of a standard deviation. Performance on the items in digital format was also more variable (e.g., proficiency estimates were more spread out), suggesting that if the differences were not taken into account, students taking a digitally based mathematics assessment may appear to be more variable in ability.

[1] NAEP has also conducted other digitally based national assessments (e.g., the 2014 and 2018 Technology and Engineering Literacy [TEL] assessments) and studies (e.g., the 2011 Mathematics Computer Based Study, which investigated use of adaptive testing); however, those studies were not directly used in the design of the 2017 transition study design.

While the digital format was more difficult than the paper format, the magnitude of the increase in difficulty was not statistically different between subgroups based on gender, race/ethnicity, school urban locale (city, urban, rural), region (northeast, southeast, central, west), and school type (public, nonpublic). The only exception found was for students reporting that at least one parent had graduated college, although the difficulty difference for this subgroup was only slightly larger than that for the overall sample (about 2 points greater; Bennett et al., 2008).

Finally, Bennett and colleagues (2008) also found that self-reported prior computer experience did not predict performance on the items in digital format, after controlling for performance on the block of 20 mathematics items in paper format provided to the students before they took the studied items.

The MOL study compared digitally based and paper-based testing in terms of issues related to measurement, equity, efficiency, and operations (Sandene et al., 2005). One important lesson learned from the MOL study is that although the digital transition did not strongly alter the psychometric properties of the mathematics instrument, taking the assessment in a digital format had a negative effect on performance (Bennett et al., 2008). While it is known that a true equivalence between digitally based and paper-based testing is impossible to achieve because two different presentation and response modes are being used (Noyes & Garland, 2008), the difficulty difference between the two modes can be properly accounted for in the analysis process so that it will not translate into the final scale score estimates.

1.3 Development Work Leading to the 2017 Assessment

The successful launch of the 2017 digitally based mathematics and reading operational assessments was the result of a multi-year transition process, the key steps of which included

- confirming that most of the existing mathematics and reading items were suitable for digital delivery;
- converting the existing item pool to digital and developing new DBA content in 2013;
- field testing the converted DBA content in 2015; and
- pilot testing the new DBA content in 2016.

To minimize the risk to continued trend reporting, the transition process began with the migration of existing paper-and-pencil items to digital delivery with the intention that new item types would be introduced in subsequent assessment years. For the first operational mathematics and reading digitally based instruments in 2017, the development focus was on trans-adapting items; that is, migrating paper-and-pencil items into a digital environment in a way that measures the constructs of interest and maintains correspondence to past PBAs for the preservation of the trend lines. Assessment specialists, cognitive scientists, and usability experts evaluated the PBA items for formal features that might preclude digital administration using the initially small number of response formats, primarily multiple-choice and text entry. While most items were deemed suitable for adaptation—since they were presented in the PBA in the same formats that would be initially available in the DBA—some could not

be easily migrated. More details on the process of creating the digitally based instrument and the process of trans-adapting items from paper to digital format are outlined in section 2.2.

In addition to the item trans-adaptation work, preparation for the 2017 transition included the development of a NAEP testing system which included a delivery application/platform as well as tablets for students to use while taking the assessment. As part of the effort to create a standardized administration environment, the NAEP program developed a secure web application called eNAEP to deliver DBAs. The eNAEP delivery system was built to provide a consistent user experience for all students. A multi-year, progressive rollout of eNAEP functionalities was also planned: an initial version with some item types and tools in 2014; an enhanced version with more item types, tools, and features in 2015 and 2016; and the 2017 operational version with a capacity to support a wider range of item types in future assessments.

One of the primary design goals for the eNAEP system was to ensure that as many students as possible could participate in the assessment. To achieve that goal, the program pursued a two-part strategy:

- Provide a user interface (UI) containing features and tools that are usable and accessible to all test takers regardless of their physical, perceptual, or cognitive characteristics. These features and tools, which are grounded in the principles of Universal Design, include
    - zoom;
    - text-to-speech (mathematics only);
    - color/contrast adjustment; and
    - highlighter.
- Provide accommodations and tools that are essential for people with various kinds of disabilities or other needs. These tools include
    - translation to Spanish;
    - magnification;
    - extra time; and
    - presentation in braille.

In addition to system tools, the 2017 assessment included additional tools or features unique to the reading or mathematics assessments. For example, mathematics students used a digital calculator for selected test forms. In reading, some items included links back to the referenced part of the reading material. To view sample released tasks, readers can visit the reading and mathematics reports on the Nation's Report Card website.

To ensure a consistent test experience across schools and to relieve schools from having to free up their own devices to use during NAEP testing, the program planned on using Microsoft Surface Pro tablets to administer the assessments. NCES decided to use the Surface Pro in the 2017 operational administration after proof-of-concept testing in 2014 as well as the successful field trial in 2015. In addition to meeting the specifications required for the eNAEP system and NAEP administration, the Surface Pro also included several built-in security features that were beneficial for the delivery of the assessment in the field. Trained administrators brought the devices to the schools, set up and managed a local network of devices in the testing locations, oversaw the testing process, and took back the devices when the testing period ended. The NAEP-provided Surface Pro tablets were equipped with an external keyboard, a

stylus, and a pair of ear buds for audio. The tablets ran the eNAEP test delivery system in which students engaged with stimuli and input their answers. Before beginning the assessment, students watched a tutorial to become familiar with the device and the system. Tutorials and other details about the digital transition are available on the NAEP website at https://nces.ed.gov/nationsreportcard/dba/.

As part of the development process for the 2015 field trial, the Surface Pro tablets were thoroughly tested for functionality and the NAEP program performed usability testing of the eNAEP interface. Usability testing explored and documented students' interactions with the tablets and with features of the eNAEP tablet-based assessment environment. Surface Pro hardware elements such as the touch screen, stylus, trackpad, and keyboard as well as a number of different eNAEP control elements such as the scrollbar, tabs, and Next/Back buttons were tested to determine how easy they were for students to use. For example, in the eNAEP assessment system, accessing all the content requires the use of tabs, scrollbars, and the Next and Back buttons on the toolbar. Usability testing found the majority of fourth- , eighth- , and twelfth-graders used the Next button in the toolbar rather than the tab and that navigation between screens and among tabs was shown to be highly intuitive for students at all three grades. Data gathered from usability tests served as the basis for changes made to improve the overall clarity and usability of the interface. For example, usability testing resulted in a recommendation for a more prominent scrolling indicator to emphasize to students that they need to scroll to see additional assessment content.

The DBA field trial in 2015

In 2015, NAEP administered extensive field trials of the complete reading, mathematics, and science assessments to nationally representative samples in the digital format at grades 4, 8, and 12. Only mathematics and reading at grades 4 and 8 will be considered here, as only these subjects and grades were transitioned to DBA in 2017. Each of the four DBAs (two subjects times two grades) involved about 12,000 to 15,000 students. In 2015, the DBA field trials were administered in the same window of time as the 2015 operational NAEP assessments, which were administered on paper.

The 2015 operational NAEP results provided a paper-based comparison to the field trial results to evaluate differences between the paper-based and digitally based administrations. The four operational NAEP samples that have corresponding field trial samples considered here (i.e., mathematics and reading, grades 4 and 8) had sample sizes of approximately 140,000 students. While no student and no school participated in both the (operational) paper-based and (field trial) digitally based assessments, the PBA and DBA samples were both nationally representative and randomly equivalent.

The 2015 NAEP digital field trials were much larger in scope than previous research studies, including the early NAEP studies summarized above (Bennett et al., 2007, 2008; Sandene et al., 2005). Sampling and weighting were carefully conducted following the operational NAEP standards for both PBA and DBA samples (Johnson & Rust, 1992). The administration of both instruments was carefully designed and administered to minimize potential confounds, and almost the entire operational paper-based item pools were converted to the digital format for the comparison.

The results from the 2015 DBA field trials confirmed what was learned from the work of the MOL study (Sandene et al., 2005; Bennett et al., 2008). The 2015 field trial studies found that the digital versions of both mathematics and reading items were harder than the original paper format items. For grade 4, the

difference of the mean item score[2] on a set of items presented in both DBA and PBA administrations was -6.7 percent for reading and -5.2 percent for mathematics. For grade 8, the difference was -2.0 percent for reading and -2.5 percent for mathematics.

Consistent with earlier research (e.g., Bennett et al., 2008), item properties (e.g., discrimination and difficulty parameters as defined by IRT) were highly correlated between the digital and paper formats. Also, the difficulty of each item relative to the entire item pool was largely consistent between digital and paper formats. Although items in the digital format were more difficult than their corresponding original paper versions, this increase in difficulty was largely consistent across all items. One slight exception was found for the constructed-response items in reading grade 4. These items had a greater increase in difficulty than was found in the reading grade 4 multiple-choice items. Most importantly, the large sample sizes in the 2015 field trials and the use of a complete NAEP instrument enabled a much more powerful test of whether the increased difficulty of the DBA administration was consistent across major student groups than was possible with the earlier NAEP studies conducted with smaller sample sizes and with a subset of the NAEP instrument.

A linear transformation was applied to the DBA scale scores to align the overall distribution of DBA scale scores to the overall distribution of PBA scale scores. As a similar linear transformation was also used for the 2017 analysis, details on the transformation are available in chapter 3. The transformation, or *linking procedure* (see chapter 3), accounts for the overall or average differences in difficulty of the paper-based and digitally based instruments across all student groups and states/jurisdictions.

After the linking procedure, residual differences in scale score estimates between the transformed DBA scores and the PBA scores for the major NAEP subgroups were examined. Residual differences, or mode residuals, are defined as the difference between a DBA statistic (e.g., a subgroup mean, a state mean, a percentile) and the corresponding PBA statistic on the reporting metric, *after* the linking transformation (see chapter 4). Across gender, race/ethnicity, English language learner status, disability status, and socioeconomic status, no subgroup performed significantly better or worse on the DBA than the PBA for the grades 4 and 8 mathematics and reading assessments.

Designing the 2017 digital transition

In designing the 2017 transition from paper to digital format in mathematics and reading at grades 4 and 8, NAEP program leaders looked to other successful transitions in NAEP's past. During the history of NAEP, the program has made a number of changes to the administration of NAEP or the NAEP instruments. One example is the 2004 redesign of the NAEP long-term trend assessments. Compared to previous assessments, there were a number of administrative changes in 2004, such as reorganizing and standardizing the assessment booklet design to a common design used in other NAEP assessments (Perie, Moran, & Lutkus, 2005). Another example is the administrative change made for the 1988 NAEP assessments to assess each student in only one subject area (Johnson & Zwick, 1990). These changes, like the transition to DBA, required special methodology in order to estimate results that may be meaningfully compared to results of previous assessments.

---

[2] The mean item score for an item is calculated as $\frac{\sum_{i=i}^{N} w_i x_{ij}}{\sum_{i=1}^{N} w_i M}$ where w_i is the weight for student i, x_{ij} indicates the score obtained by student i, and M represents the maximum possible score of this item. Incorrect responses are scored as zero. Note that students who did not reach this item are excluded from the computation.

A typical NAEP assessment with no major administrative changes comprises most, but not all, items in common with the previous assessment. NAEP uses concurrent calibration with Multiple Group IRT to link the assessment results to a common scale to enable comparisons of results between years (Yamamoto & Mazzeo, 1992). Multiple Group IRT accommodates the differences in the item composition of instruments across different assessments. The approach requires that a subset of items appears in both assessments and has the same psychometric properties in both years, including the same difficulty.

However, when substantial changes are made to the administration of the assessment, the requirement of a subset of items with invariant psychometric properties may be tenuous. The prior research described above found that paper-and-pencil items tend to become systematically harder when delivered in a digital format (e.g., Bennett et al., 2008). As such, the assumption that digital and paper versions of the same items could be treated as common items, as required by concurrent calibrations, was not deemed appropriate for the digital transition.[3]

Historically, when NAEP has been faced with the challenge that assuming common items was either not possible or needed careful evaluation, the program has employed a special design, called a bridge study, in which the assessment is given with and without the administrative changes to two randomly equivalent samples (Haertel, 2016; Johnson & Zwick, 1990). This was the case, for example, with the 2004 NAEP long-term trend assessments and the 1988 NAEP assessments referenced above (Johnson & Zwick, 1990; Perie, Moran, & Lutkus, 2005). The linking of the scores from the new administration design to the existing NAEP scale is accomplished by aligning the proficiency distributions of the two samples, considering that the randomly equivalent samples can be reasonably assumed to have equivalent proficiency distributions (within the margin of sampling error).

Such bridge studies have been successful in the history of NAEP. They share the assumption that the new assessment does not introduce construct-irrelevant confounds—that is, the new assessment does not change the subject-matter construct measured by the assessment in a meaningful way. Consequentially, data collection of NAEP bridge studies is designed so that the impact of the administration change on subgroup scores and trends can be psychometrically evaluated. The overarching benefit of bridge studies is that they enable the linking of the results of an assessment to prior assessment results, in a way that accounts for differences in the assessment instrument and administration, provided that construct-irrelevant confounds are not introduced (Dorans, Pommerich, & Holland, 2007). Similarly, the 2017 bridge study was designed to enable the linking of the 2017 DBA mathematics and reading results to the previous PBA results, through the bridge of the 2017 PBA.

[3] Note that while NAEP routinely evaluates the assumption of common items and relaxes the assumption on an item-by-item basis where necessary in every assessment that is linked to a previous assessment (NCES, 2000-2018), a systematic change across all items, such as all items becoming more difficult by a similar amount, presents a model identification issue that cannot be resolved with concurrent calibration and requires a stronger study design (de Ayala, 2009).

The 2017 assessments were designed to replicate and extend research findings from prior NAEP assessments. Psychometric evaluations concentrated on three primary research questions:

1. **What was the impact of the mode transition on the measurement instrument?** The findings of prior research with subsets of NAEP instruments (Bennett et al., 2007, 2008; Sandene et al., 2005), or with the majority of NAEP instruments (2015 NAEP digital transition study, unpublished), were evaluated in a more powerful design involving the full 2017 instruments and larger samples. Notably, the findings that NAEP items retain most of their original properties (such as their difficulty relative to the item pool) when trans-adapted (moved to digital with minimal changes), yet become systematically harder, was studied. The 2017 study also re-evaluated the conclusion of prior research that assuming common items between PBA and DBA was untenable.
2. **After linking the scale scores, what was the impact of the mode transition on state/jurisdiction and district estimates?** Prior research was limited in scope and did not collect the necessary data to evaluate this research question. However, estimating and reporting changes across time for not just the nation but also for states/jurisdictions and districts is a priority for NAEP. As a result, this issue was studied for the first time with the 2017 data. This research question has two aspects: (1) to understand the differences in the instruments as reflected in performance across different jurisdictions, and (2) to evaluate the degree to which the bridge study design accounted for differences in the paper-based and digitally based instruments across all states/jurisdictions and districts.
3. **After linking the scale scores, what was the impact of the mode transition on subgroup estimates?** Again, the findings of prior research (Bennett et al., 2007, 2008; Sandene et al., 2005; 2015 NAEP digital transition studies, unpublished), were evaluated in a more powerful design involving the full 2017 instruments and very large samples. An important observation in earlier work, the conclusion that the increased difficulty of the digitally based assessment was consistent across all major subgroups, was re-evaluated. This research question also had two main components. The first was to provide insight into the nature of the differences between the instruments as reflected in performance across different subgroups. The second was to evaluate the degree to which the bridge study design accounted for differences in the paper-based and digitally based instruments across major student subgroups.

2. 2017 Bridge Study Design

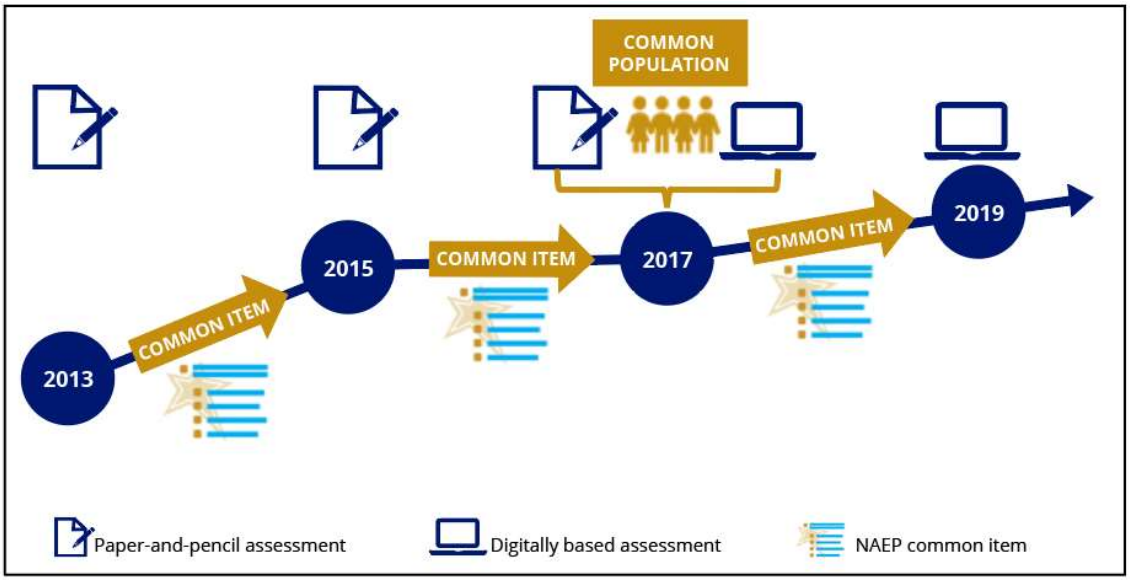
2.1 Summary of the Design

In accordance with the principles of bridge studies historically employed by NAEP (e.g., Beaton, 1990a, p. 5), the 2015 operational NAEP mathematics and reading assessments were re-administered in 2017. The re-administered assessments are referred to as the 2017 paper-based assessments (PBAs). The 2017 PBAs were identical to the 2015 operational assessments in terms of the instrument, including all of the items, with the same design of the test booklets divided cognitive testing time into two 25-minute sections, and the same administration procedures as 2015. For each of the four grade and subject combinations (mathematics and reading, grades 4 and 8), each PBA was administered to approximately 40,000 students.

In addition to the 2017 PBAs, a digitally based assessment (DBA) was also administered for each subject and grade combination, which divided cognitive testing time into two 30-minute sections. Intended as reporting samples, each of the four DBAs was administered to approximately 150,000 students. The sampling and weighting procedures for the PBAs and DBAs are detailed below in the current chapter.

The 2017 PBA served three distinct, but related, purposes. First, the 2017 PBA measured the trend for the overall national sample in a non-interruptive manner using the typical common item linking approach. As illustrated in figure 2.1a, the 2015 assessment results were linked to the trend line through a set of common items between 2013 and 2015. Similarly, in the transition year of 2017, the 2017 PBA results were placed onto the reporting scale through the same approach, providing a performance measure at the national level without any impact from the DBA transition. Second, the 2017 PBA provided a bridging sample to link the DBA to prior assessments. Since the 2017 PBA and 2017 DBA were administered to randomly equivalent samples drawn from a common population, placing the DBA results onto the reporting scale by assuming a similar proficiency distribution between the PBA and DBA samples accounted for differences in the administration and assessment instruments between DBA and PBA. This common population linking is also illustrated in figure 2.1a, above the year 2017. And third, the 2017 PBA component provided data to critically evaluate the psychometric consequences of the transition to DBA, and to inform the decision of whether or not to continue reporting trend comparisons between DBA results (current and future) and previous PBA results. Once the assessment is transitioned to digital, the common item linking approach will be used to connect the 2017 DBA with future DBAs.

Figure 2.1a. How common item linking and common population linking were used on NAEP



2.2 Digitally Based Assessment Instruments

A primary goal in the design of the 2017 DBA was to ensure that the digitally based instruments measure the same content and skills as the paper-based instruments, while still being appropriately adapted for digitally based administration.

Items in the 2017 digitally based instruments for mathematics and reading fell into two categories:

1. Trans-adapted items
    a. Direct transfer to digital format. These items were essentially unchanged from the paper-and-pencil version. This includes most multiple-choice items as well as most constructed-response questions.
    b. Transfer to digital format with new tools. These were items that were modified from the paper format to accommodate administration in a DBA environment, but they have the same measurement target as the original paper version. For example, some reading items that referred to a specific phrase or page in the passage were enhanced with a look-back button to take students to the relevant phrase and page. Similarly, in mathematics, questions were enhanced with a drag-and-drop function so that students could respond by moving a source, such as a number or shape, to a target.
2. New DBA items. In 2016, a pilot was conducted to evaluate items that were newly developed for the 2017 DBA. Some of the items piloted in 2016 were included in the digitally based instrument, and they had no corresponding paper version in the 2017 PBA, which was based on the 2015 assessment.

The paper-based versions of the NAEP assessment traditionally included both multiple-choice and constructed-response questions to which students responded either by marking a response option or by providing a written response in their test booklet. The transition to a digital version of the assessment allowed item developers to both enhance traditional question formats and to consider new question types. In the 2017 reading and mathematics assessments, students were still assessed with both

multiple-choice and constructed-response questions; for the multiple-choice, students clicked on their answer choice, and for constructed-response, students typed their answers in the provided space. Students continued to construct their own responses to questions that required a written explanation, interpretation, justification, or a description of steps for solving of a given problem. In addition, the 2017 assessments included new selected-response questions where students were asked to match elements by dragging and dropping, to make a selection in a stimulus, grid, or text, or to use other interactive components.

At both grades (4 and 8), 67 percent of the mathematics items and approximately 92 percent of the reading items in the 2017 digitally based instruments were trans-adapted from paper (see tables 2.2a and 2.2b). Note that students participating in NAEP do not receive all items in the assessment. Instead, items are assembled into separately timed sections termed "blocks," and students receive a random selection of two blocks of items (Johnson, 1992; NCES, 2000-2018).

For the mathematics assessment, each grade had a total of ten blocks of items. Six of the ten blocks administered at each grade comprised only trans-adapted items. The other four blocks were newly developed for the 2017 DBA. Some of the paper-based mathematics items were not appropriate for DBA administration. For example, items that required a student to manipulate artifacts (such as geometric shapes) or use tools were not transitioned; however, similar content was assessed using appropriate digital tools (such as a digital ruler). Thus, while an effort was made to preserve the order of items, the mathematics trans-adapted blocks were not exact conversions of the original paper blocks. Instead, a small number of items from other paper blocks were trans-adapted and used in place of the items that could not be trans-adapted.

For the reading assessment, nine of the ten blocks of passages and items administered at grade 4, and 12 of the 13 blocks administered at grade 8, were trans-adapted. One new block was developed at each grade for the digitally based reading assessment. Unlike mathematics, the reading trans-adapted blocks consisted of trans-adapted versions of all the passages and items in the original paper blocks, in the same context and position. In line with the framework's principle of authenticity, all trans-adapted passages were presented as closely as possible to their original publication. If a passage source had been digital, the passage was presented as scrolled in the assessment; and if the passage was originally paged, it was presented as paged. Digital delivery allowed for the color and styling of the original passage source to be restored.

It should be noted that students typed responses into specified areas in the interface. Students could not write freehand in answer spaces. For mathematics items that might have previously involved drawing a picture, students answered the questions by selecting answers or using system tools to provide their answers. To view sample released tasks, visit the reading and mathematics reports on the Nation's Report Card website.

Table 2.2a. Percentage of trans-adapted and new items and number of items in NAEP mathematics at grades 4 and 8: 2017

Grade	Trans-adapted items	New DBA items	Total items
4	118 (67%)	59 (33%)	177
8	120 (67%)	59 (33%)	179

NOTE: DBA = digitally based assessment.

Table 2.2b. Percentage of trans-adapted and new items and number of items in NAEP reading at grades 4 and 8: 2017

Grade	Trans-adapted items	New DBA items	Total items
4	88 (91%)	9 (9%)	97
8	120 (92%)	10 (8%)	130

NOTE: DBA = digitally based assessment.

Both the paper-based and digitally based instruments were designed to assess content and skills delineated in the current NAEP frameworks developed by the Governing Board.[4] Essentially, the DBA instruments were designed to measure the same content as assessed in PBA except in a new digital platform. It should be noted that scoring procedures were not altered between the DBAs and PBAs; trans-adapted items were scored using the same criteria in both DBA and PBA. Nevertheless, NCES designed the bridge study described in this paper to allow for trend reporting despite known differences in percent correct statistics across DBA and PBA item formats. Keeping the cognitive content as close as possible between the two formats allows the program to continue to test critical cognitive abilities and skills despite the differences in modality that were linked statistically through the bridge study.

For both subjects, the content distribution was similar for the paper-based and corresponding digitally based instruments. Details for mathematics are provided in tables 2.2c, d, and e, while details for reading are provided in tables 2.2f, g, and h.

For mathematics, the content and subscale distributions were very similar across the DBA and PBA (table 2.2c). Similarly, the distribution of the level of complexity of items was also very similar across the two modes (table 2.2d). The item type distribution was similar across the DBA and PBA, although a higher proportion of the new DBA items were constructed-response items, resulting in a slightly higher proportion of constructed-response items in DBA (table 2.2e).

For reading, the vast majority of the PBA items were trans-adapted to digital format, and thus comprised the vast majority of the reading DBA items. Consequently, reading had almost identical content distribution in terms of subscale distribution, cognitive target, and item type (tables 2.2f, 2.2g, and 2.2h).

Table 2.2c. Percentage of items by content area for mathematics, for the target, paper-based instrument, and digitally based instrument at grades 4 and 8: 2017

	Grade 4			Grade 8		
Content area	Target	PBA	DBA	Target	PBA	DBA
Number properties and operations	40%	41%	37%	20%	20%	22%
Measurement	20%	18%	17%	15%	15%	15%
Geometry	15%	15%	17%	20%	20%	16%
Data analysis	10%	13%	14%	15%	15%	17%
Algebra	15%	14%	15%	30%	30%	30%

NOTE: PBA = paper-based assessment. DBA = digitally based assessment. Target refers to the targeted distribution specified in the NAEP mathematics framework. Detail may not sum to totals because of rounding.

[4] See https://www.nagb.gov/naep-frameworks/frameworks-overview.html

Table 2.2d. Percentage of assessment time by item complexity for mathematics, for the target, paper-based instrument, and digitally based instrument at grades 4 and 8: 2017

	Grade 4			Grade 8		
Level of complexity	Target	PBA	DBA	Target	PBA	DBA
Low	25%	55%	48%	25%	50%	50%
Moderate	50%	33%	37%	50%	42%	40%
High	25%	12%	15%	25%	8%	10%

NOTE: PBA = paper-based assessment. DBA = digitally based assessment. Target refers to the targeted distribution specified in the NAEP mathematics framework. Detail may not sum to totals because of rounding.

Table 2.2e. Percentage of assessment time by item type for mathematics, for the target, paper-based-instrument, and digitally based instrument at grades 4 and 8: 2017

	Grade 4			Grade 8		
Item type	Target	PBA	DBA	Target	PBA	DBA
Selected response	50%	46%	39%	50%	50%	39%
Constructed response	50%	54%	61%	50%	50%	61%

NOTE: PBA = paper-based assessment. DBA = digitally based assessment. Target refers to the targeted distribution specified in the NAEP mathematics framework. Detail may not sum to totals because of rounding.

Table 2.2f. Percentage of blocks by subscale for reading, for the target, paper-based instrument, and digitally based instrument at grades 4 and 8: 2017

	Grade 4			Grade 8		
Subscale	Target	PBA	DBA	Target	PBA	DBA
Informational	50%	50%	50%	55%	46%	46%
Literary	50%	50%	50%	45%	54%	54%

NOTE: PBA = paper-based assessment. DBA = digitally based assessment. Target refers to the targeted distribution specified in the NAEP reading framework. Detail may not sum to totals because of rounding.

Table 2.2g. Percentage of items by cognitive target for reading, for the target, paper-based instrument, and digitally based instrument at grades 4 and 8: 2017

	Grade 4			Grade 8		
Cognitive target	Target	PBA	DBA	Target	PBA	DBA
Locate/Recall	30%	23%	24%	20%	17%	19%
Integrate/Interpret	50%	61%	58%	50%	60%	58%
Critique/Evaluate	20%	16%	19%	30%	23%	23%

NOTE: PBA = paper-based assessment. DBA = digitally based assessment. Target refers to the targeted distribution specified in the NAEP reading framework. Detail may not sum to totals because of rounding.

Table 2.2h. Percentage of items by item type for reading, for the target, paper-based instrument, and digitally based instrument at grades 4 and 8: 2017

	Grade 4			Grade 8		
Item type	Target	PBA	DBA	Target	PBA	DBA
Selected response	50%	68%	67%	40%	60%	59%
Short constructed response	40%	22%	23%	45%	30%	31%
Extended constructed response	10%	10%	10%	15%	10%	10%

NOTE: PBA = paper-based assessment. DBA = digitally based assessment. Target refers to the targeted distribution specified in the NAEP reading framework. Detail may not sum to totals because of rounding.

Tutorial for Digitally Based Assessments

For the transition to DBA, the NAEP program created a stable and consistent administration environment by bringing tablet equipment to schools. At the beginning of each assessment session, students viewed an interactive tutorial that provided information about how to take the assessment on the tablet. Some of the information was pertinent to both mathematics and reading, such as how to indicate answers for multiple-choice questions and how to proceed from one item to the next. Other information was subject specific, such as how to use online tools (e.g., the equation editor or scratch-work tool) when performing mathematical computations, or how to toggle between the reading text and questions. The interactive nature of the tutorial allowed students to familiarize themselves with the digital delivery system before beginning the actual assessment. Each touchscreen tablet had an attached keyboard for entering answers to constructed-response questions.

2.3 Sample Design and Weighting

The 2017 NAEP sample followed the usual NAEP school and student sampling procedures with a few additional features to accommodate the dual mode design. Below, the special features of the 2017 NAEP sample are detailed.

2017 Target Sample Sizes

To enhance the statistical power of comparisons between modes, a key design component for 2017 was to sample some students for DBA and others for PBA within each sampled school, where practical. At the jurisdiction level, the target sample size of the operational DBA was comparable to the sample sizes of the 2015 operational assessments, which allowed for reporting at the usual NAEP level of reliability for each jurisdiction. A smaller sample was drawn for PBA. This provided sufficient power to detect differences between the PBA and DBA results and the ability to estimate national paper-based trends.

Target sample sizes varied by type of school and type of jurisdiction. The target sample sizes by mode type for both subjects combined, within a single grade, are provided in Table 2.3a, along with the target sample size for the 2015 operational PBA. The sample size targets were the same for each grade (4 and 8).

Table 2.3a. Target student sample sizes for operational assessments, by sample type and assessment mode for mathematics and reading combined, for each of grades 4 and 8: 2015 and 2017

	2015	2017		
Sample type	PBA	DBA	PBA	Total
Public schools				
State samples	4,400	4,400	1,000	5,400
Large TUDA district	3,300	3,300	1,000	4,300
Small TUDA district	2,200	2,200	1,000	3,200
Puerto Rico	5,000	3,000	3,000	6,000
Private schools	6,000	6,000	1,200	7,200
Aggregate national	277,000	291,000	79,200	370,200

NOTE: PBA = paper-based assessment. DBA = digitally based assessment. TUDA = Trial Urban District Assessment. Large TUDA districts are those with enrollment of over 17,000 per grade; the remaining districts are labelled as "Small."

Student Sample Selection

The student sampling process for NAEP was a systematic procedure, consisting of two parts. The first part was the within-school student sample selection, and the second part was the assignment to the assessment mode of the selected students. Within each sampled school, a sample of students was drawn from a list of students enrolled in the targeted grade, such that every student had an equal chance of selection. For schools where student demographic data were available, the student lists were sorted by *gender* and *race/ethnicity* before the sample was selected to implicitly stratify the sample.

Once the overall student sample was obtained, sampled students generally were assigned to assessment modes with predefined subsampling rates. Again, students were sorted by *gender* and *race/ethnicity* to ensure the two samples were sufficiently balanced. Any potential imbalance was adjusted in the weighting procedure.

In small schools it was not practicable to conduct both DBA and PBA sessions. Thus, small schools[5] were randomly assigned to have all students assigned to a single mode. In each case, the assignment was conducted using a systematic sampling procedure, so that the schools assigned to each mode were as comparable as possible with regard to the school stratification variables.

Table 2.3b provides the within-school target sample size per sampled grade, the mode size, and mode rates for each sample type. The “Take-all cutoff” column indicates the school size below which all students from a sampled school were selected. For example, for a public school sampled from a large district, if this school had fewer than or equal to 75 students, then all students from this school were selected. If this school had more than 75 students, then 66 students were randomly selected. These targets were the same for each grade.

[5] The definition of a small school was a school with fewer than 21 grade-eligible students for public schools, fewer than 20 for Puerto Rico, or fewer than 18 for private schools.

Table 2.3b. Within-school target sample size, within-school assessment mode size, and assessment mode rates, by sample type: 2017

	Schools where Enrollment > Target sample size				Other Schools	
	Sample	Take-all	Mode		Mode rate	
Sample type	size	cutoff	DBA	PBA	DBA	PBA
Public schools						
State samples	62	75	50	12–25	50/62	12/62
Large TUDA district	66	75	50	16–25	50/66	16/66
Small TUDA district	74	80	50	24–30	50/74	24/74
Puerto Rico	50	55	25	25–30	25/50	25/50
Private schools	60	75	50	10–25	50/60	50/60

NOTE: Small schools were assigned to only paper-based or digitally based assessment. PBA = paper-based assessment. DBA = digitally based assessment.

DBA Session Size

The assignment of students to assessment modes generally followed the guidelines described above. However, the assignment to mode within a school had to balance two considerations: (1) the need to achieve the target sample size for each mode within each jurisdiction, and (2) the fact that DBA sessions could not be larger than 25 students each, and that, in general no more than two such sessions were to be conducted in a school.

For the mode assignment, sampled students were assigned to DBA with the subsampling rates provided above. This would, in general, result in 50 students in every non-certainty school receiving DBA (i.e., schools selected with a probability of less than one). However, in schools with fewer than the target sample size of students, applying the designated rate of digital sample assignment (the DBA mode rate) would, in some cases, result in DBA sessions with slightly more or slightly fewer than 25 students. These situations were deemed operationally inefficient, and therefore, to satisfy the second condition, adjustments were made to the DBA sample assignment process to prevent these occurrences. In schools where slightly more than 25 students would have been assigned to a DBA session, that number was decreased to 25 and the remainder of students were assigned to a PBA session. Similarly, in order to maintain balance in the overall proportions of DBA and PBA students, in schools where slightly fewer than 25 students would have been assigned to a DBA session, the number assigned to the DBA session was increased to 25. In certainty schools with a larger student sample size, the same adjustments were made based on multiples of 25 instead of exactly 25. Thus, the proportion of students assigned to each mode varied somewhat across the schools within a jurisdiction, for reasons of operational efficiency. This meant that the PBA sample size in a school varied somewhat with the size of the school. For example, in a state school with 62 students enrolled in the sampled grade, 50 students were assigned to DBA and 12 to PBA. In a state school with enrollment between 63 and 75 inclusive in the sampled grade, 50 students were assigned to DBA, and the remainder to PBA, leading to PBA sizes in the range of 13 to 25. In a state school with more than 75 students in the sampled grade, only 62 students were assessed, 50 with DBA and 12 with PBA. In a state school with fewer than 62 students in the sampled grade, 25/31

of the students were assigned to DBA and 6/31 to PBA. All of these variations in the allocation of students to DBA and PBA modes were accounted for in the calculation of sampling weights.

Sample Weights

As in all NAEP assessments, sample weights are applied at the student level to ensure that the samples are representative of the jurisdictions from which they were selected. The goals of NAEP weighting are to provide accurate and approximately unbiased estimates of student achievement for the nation and its various subpopulations. Sample weights are created separately for each NAEP sample defined by grade (4 and 8) and assessment subject (mathematics and reading).

Sample weights were derived for each mode, treating each mode as a separate sample, for each grade and subject. In general, the methods used to conduct the weighting were the same as in all past NAEP assessments, going back to 2002 when national and state assessments were combined and the TUDA program was introduced (NCES, 2000-2018). However, an additional step was used to equilibrate the DBA and PBA samples in each jurisdiction (see Equilibration of Different Mode Samples via Raking (Iterative Proportional Fitting)). This step was taken to reduce the sampling variance of the DBA and PBA samples, and in particular the variances of estimates of the differences in student achievement between these modes.

Replicate Weights

In addition to the final full-sample weight, a set of replicate weights was provided for each student. Replicate weights are used to calculate the variances of survey estimates using the jackknife repeated replication method. The methods used to derive these weights aim at reflecting the features of the sample design, so that when the jackknife variance estimation procedure is implemented, approximately unbiased estimates of sampling variance are obtained. In particular, estimates of sampling variance for comparisons by mode appropriately reflect the fact that the samples for each mode generally were drawn from the same schools, thus reducing such variance.

In addition, the various weighting procedures were repeated on each set of replicate weights to appropriately reflect the impact of the weighting adjustments on the sampling variance of a survey estimate. A finite population correction (fpc; Lohr, 2010) factor was incorporated into the replication scheme so that it could be reflected in the variance estimates for the mathematics and reading assessments.

The procedures used in 2017 were the same as in 2015 and other recent assessments, and used the same principles that NAEP has used since 1984 for estimating sampling variance.

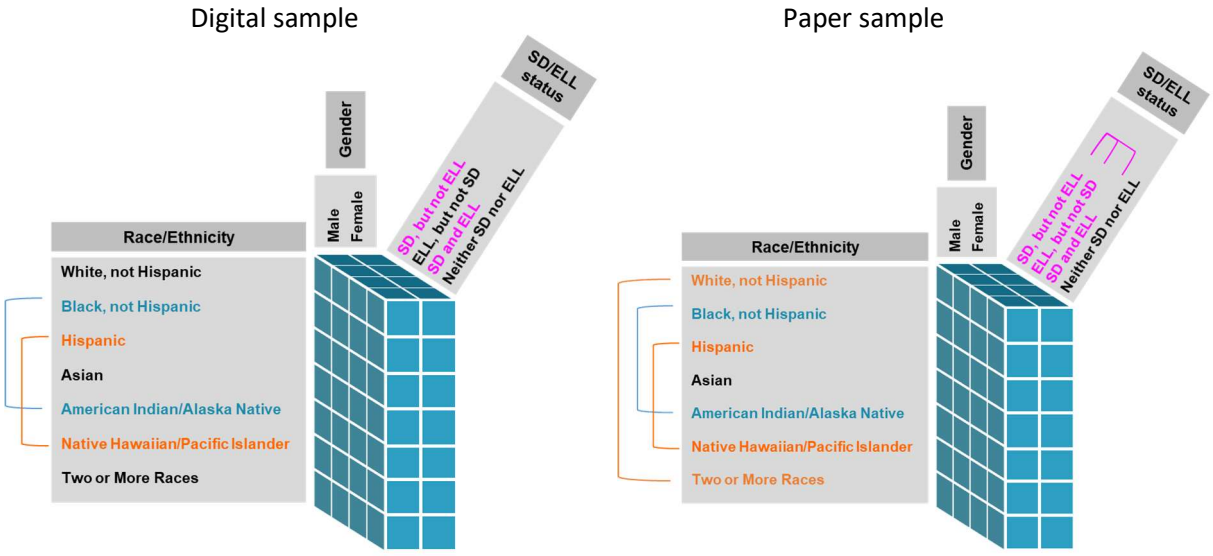
Equilibration of Different Mode Samples via Raking (Iterative Proportional Fitting)

Weighted estimates of population totals for student-level subgroups for a given grade will vary across subjects even though the student samples for each subject generally come from the same schools. These differences result from sampling error associated with the random assignment of subjects to students through a process known as spiraling. For state assessments in particular, any difference in demographic estimates between subjects, no matter how small, may raise concerns about data quality. To remove these random differences and potential data quality concerns, a step was added to the NAEP weighting procedure in 2009. This step adjusted the student weights in such a way that the weighted sums of population totals for specific student groups were the same across all subjects. It was implemented using a raking procedure and applied only to public school assessments.

Raking (also known as iterative proportional fitting) is a weighting procedure that involves simultaneous ratio adjustments to two or more marginal distributions of population totals. In 2017, the student raking adjustment was carried out independently by mode in each state for the mathematics and reading public school samples at grades 4 and 8. The dimensions used in the raking process were race/ethnicity, gender, and SD/ELL status. The control totals for the raking dimensions for all four samples (i.e., mathematics and reading, each for DBA, and PBA) per grade (4 and 8) were obtained from the NAEP student sample weights of the aggregated mathematics and reading, DBA and PBA, public school samples.

Figure 2.3a graphically illustrates the dimensions of the raking procedure. This example shows the levels of the raking dimensions for the New Jersey grade 4 public school DBA sample (left) and PBA sample (right) that had to be collapsed prior to raking because of small marginal cells. The PBA sample required more collapsing, due to its substantially smaller size. For example, in the PBA case the four SD/ELL categories had to be collapsed into two categories: SD or ELL; Neither SD nor ELL. This was because of the relatively small numbers of SD and ELL students. In the case of the PBA sample, the racial/ethnic groups of American Indian/Alaska Native, Native Hawaiian/Pacific Islander, and Two or More Races were each too small to comprise a cell in their own right. Each was collapsed with another racial/ethnic group that has traditionally had similar performance on NAEP assessments (so that, for example, the Two or More Races category was collapsed with the White, not Hispanic category).

Figure 2.3a. Raking cube showing the levels of the raking dimensions for New Jersey grade 4 public school digital sample (left) and paper sample (right): 2017



Comparison Across the Two Samples on Major Demographics

Table 2.3c demonstrates the effect of student raking on the distribution of student characteristics used in the raking by mode and subject, for the state of New Jersey at grade 4. The table shows the student distributions for the raking dimensions before and after raking by mode (PBA and DBA) and subject (mathematics and reading). Estimates before and after raking compared to control totals show that after raking, the differences in demographic estimates between subjects were largely eliminated. This result for New Jersey grade 4 is typical of other states, districts, subjects, and grades. Detailed tables for each jurisdiction are available upon request.

Table 2.3c. Effect of student raking on the distribution of student characteristics used in the raking by assessment mode and subject for New Jersey grade 4 public school digital and paper samples: 2017

Characteristic		Before raking				After raking				Control totals ****
		PBA		DBA		PBA		DBA		
		Math	Reading	Math	Reading	Math	Reading	Math	Reading	
SD/ELL*	SD, not ELL	17.2%	16.1%	15.1%	17.0%	17.1%	16.3%	16.0%	16.3%	16.2%
	ELL, not SD	3.3%	3.4%	4.0%	3.8%	3.3%	3.5%	3.8%	3.8%	3.8%
	SD and ELL	0.2%	0.7%	0.7%	0.4%	0.2%	0.7%	0.7%	0.4%	0.5%
	Not SD or ELL	79.4%	79.7%	80.2%	78.7%	79.4%	79.4%	79.4%	79.4%	79.4%
	Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Race/ethnicity	White, not Hispanic	49.4%	48.8%	48.1%	46.4%	47.3%	47.9%	47.5%	47.5%	47.5%
	Black, not Hispanic	11.6%	12.0%	12.5%	13.4%	12.9%	12.9%	12.8%	12.7%	12.8%
	Hispanic	27.9%	27.5%	28.0%	28.7%	27.9%	28.3%	28.3%	28.2%	28.2%
	Asian	8.1%	9.7%	9.1%	8.8%	8.9%	8.9%	8.9%	8.9%	8.9%
	AI/AN**	-	-	0.1%	0.2%	-	-	0.1%	0.2%	0.1%
	NH/OPI***	0.6%	0.2%	0.2%	0.3%	0.6%	0.2%	0.2%	0.3%	0.3%
	Two or More Races	2.5%	1.8%	2.0%	2.2%	2.4%	1.8%	2.1%	2.1%	2.1%
	Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Gender	Male	50.1%	50.7%	51.6%	51.2%	51.2%	51.2%	51.2%	51.2%	51.2%
	Female	49.9%	49.3%	48.4%	48.8%	48.8%	48.8%	48.8%	48.8%	48.8%
	Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

*SD = students with disabilities, ELL = students with English Language Learner status.

**AI/AN = American Indian and Alaska Native.

***NH/OPI = Native Hawaiian or Other Pacific Islander.

****The control totals for the raking dimensions for all six samples (i.e., mathematics and reading, each for DBA, PBA, and combined) per grade (4 and 8) were obtained from the NAEP student sample weights of the mathematics and reading public samples combined.

3. Analysis Procedures

3.1 The NAEP Model

Analysis of NAEP data is optimized for two primary design features of NAEP (von Davier et al., 2006). First, each student receives only a randomly selected subset of all the items in the assessment, following a matrix sampling design. This is done to balance competing desires of limiting testing time and maintaining an item pool that is comprehensive of a broad framework. Second, NAEP is required to report results only for aggregated groups of students, such as subgroups (e.g., race/ethnicity, gender), states, and districts; NAEP is not allowed to report results for individual students (Johnson, 1992).

Among other benefits, the use of item response theory (IRT) enables NAEP to model the responses of students who received different, but overlapping, sets of items. IRT assumes that the response to item *j* for student *i* is dependent on the *p* latent proficiencies of the student, $\boldsymbol{\theta}_i = (\theta_1,\ldots, \theta_p)$, and the properties of the item, $\boldsymbol{\beta}_j$. The responses to individual items are assumed to be conditionally independent such that,

$$\Pr(\mathbf{Y}_i = \mathbf{y}_i \mid \boldsymbol{\theta}_i, \boldsymbol{\beta}) = \prod_{j=1}^{J} \Pr(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\beta}_j),$$

where $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{iJ})$ is a vector comprising the response variables for all *J* items. The dependency of the responses on the latent proficiencies and the item response functions are modelled with either the two-parameter logistic model (2PL), three-parameter logistic model (3PL), or the general partial credit model (GPCM; Muraki, 1992), depending on the type of item. The dimensionality of the latent proficiencies, *p*, represents the number of subscales in a NAEP subject. For reading, *p* is equal to 2, and for mathematics, *p* is equal to 5. However, in the current NAEP operational analyses, only items from the same subscale are calibrated together and the correlations among the subscales are not directly considered in the IRT model. Therefore, the likelihood function $\Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{\theta}_i, \boldsymbol{\beta})$ can be further decomposed as,

$$\Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{\theta}_i, \boldsymbol{\beta}) = \prod_{k=1}^{p} \prod_{j\in I_k} \Pr(Y_{ij} = y_{ij}|\theta_{ki}, \boldsymbol{\beta}_j),$$

where $j \in I_k$ indicates that the item *j* is from the k^{th} subscale, $k=1,\ldots,p$, and θ_{ki} represents the student *i*'s proficiency on the k^{th} subscale.

To estimate results for different student groups, a regression is employed where the NAEP proficiencies are the dependent variables and group indicators are the independent variables. Because the NAEP proficiencies are latent variables, the regression is a latent regression (Mislevy, 1984, 1985). Conditional on $\mathbf{x}_i$, the vector of contextual variables and group indicators for student *i*, $\boldsymbol{\theta}_i$ is assumed to follow a multivariate normal distribution. That is, $\boldsymbol{\theta}_i|\mathbf{x}_i \sim N(\boldsymbol{\Gamma}^T\mathbf{x}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\Gamma}$ is the vector of regression parameters, and $\boldsymbol{\Sigma}$ is a common variance-covariance matrix.

The complete NAEP model combining IRT and latent regression is obtained by assuming $\mathbf{Y}_i$ and $\mathbf{x}_i$ are independent conditional on $\boldsymbol{\theta}_i$, leading to the marginal likelihood,

$$\prod_{i=1}^{N} [\int \Pr(\mathbf{Y}_i = \mathbf{y}_i \mid \boldsymbol{\theta}_i, \boldsymbol{\beta}) \phi(\boldsymbol{\theta}_i; \boldsymbol{\Gamma}^T \mathbf{x}_i, \boldsymbol{\Sigma}) d\boldsymbol{\theta}_i]^{w_i},$$

where w_i is the sampling weight for student *i*, *N* is the total number of students, and $\phi(\cdot|\cdot,\cdot)$ is the multivariate normal density function.

3.2 NAEP Linking Procedures

IRT models have a well-known indeterminacy that the latent scale is arbitrary under linear transformation (Lord, 1980). Therefore, a *linking procedure* is required to place newly estimated IRT-based scores (*scale scores*) on the common metric of existing reported scores, defined as the *reporting metric*. The scale scores from typical NAEP assessments are linked to the reporting metric following a *common item linking* approach, which assumes that a subset of items that were presented in both the current and previous administrations have common properties across the two administrations. A typical operational NAEP administration involves 70–80 percent of the items from the previous operational NAEP administration for the relevant subject area, while the remaining items are newly developed for the current operational administration. NAEP uses the two-population concurrent IRT calibration method, with each year defining a population. The item parameters of the common items across the two administrations are constrained to be equal (Bock & Zimowski, 1997), and the two populations combined are constrained to have an overall mean of 0 and standard deviation of 1 to make the model identified.

2017 bridge study linking procedures

Using the standard NAEP common linking methodology to link between the previous PBA and the current DBA was untenable based on previous research (Bennett et al., 2007, 2008; Sandene et al., 2005), as items converted to digital may appear more difficult and therefore violate the assumption of items being common. Instead, linking of the DBA scores to the reporting metric was accomplished through the combination of common item linking using the 2015 PBA results and a *bridge* PBA sample in 2017 and common population linking within 2017 to link the DBA results to 2017 PBA results on the reporting metric. This linking procedure was made possible by readministering the 2015 PBA instrument in 2017 and sampling both the DBA and the PBA samples within the same population.

Types of Assessment Linking Methodologies

- Unchanged test questions used to measure performance across years

- Randomly equivalent samples from a common population used to bridge across modes

In general, when common items cannot be assumed, common population linking methods are the standard alternative both within NAEP and the education field (e.g., Yamamoto & Mazzeo, 1992), and are often used when linking scores from assessments with different modes (Eignor, 2007). For example, bridge samples were used when focused balanced incomplete block matrix sampling was introduced in the 1988 NAEP assessments (Johnson & Zwick, 1990), when updates were made to the 2004 NAEP long-term trend assessments (Perie, Moran, & Lutkus, 2005), and when the reading item design framework was changed for the 2009 NAEP assessments (NCES, 2009). When common population linking is used in NAEP, the analysis to conduct the link and evaluate the effectiveness and fairness of the link is sometimes referred to as a *bridge study*, due to the requirement of a bridge sample.

In the 2017 linking procedures, the results from the 2017 PBA were analyzed separately from the DBA data but in combination with the 2015 PBA following standard NAEP procedures, based on the common item linking method. These standard procedures to estimate and link scores to the reporting metric have been comprehensively described previously (e.g., in Mislevy, Johnson, & Muraki, 1992; Yamamoto & Mazzeo, 1992).

The DBA data were also analyzed separately, resulting in DBA scale scores on an arbitrary metric referred to as a *theta metric*. The common population linking procedure involved equating the first two moments of the DBA scores on the theta metric to the first two moments of the PBA scale scores on the reporting metric. That is, DBA scale scores on the theta metric, θ_{DT}, were transformed to DBA scores on the reporting metric, θ_{DR}, by,

$$\theta_{DR} = A\theta_{DT} + B,$$

where the transformation coefficients are calculated as,

$$A = \frac{\hat{\sigma}_{PR}}{\hat{\sigma}_{DT}}; B = \hat{\mu}_{PR} - A\hat{\mu}_{DT},$$

and where $\hat{\mu}_{PR}$ and $\hat{\sigma}_{PR}$ are the estimated national mean and standard deviation on the reporting metric based on the PBA results, and $\hat{\mu}_{DT}$ and $\hat{\sigma}_{DT}$ are the estimated national mean and standard deviation on a theta metric based on the DBA results. Note that linking transformations were applied for each subscale, but the subscale subscripts on all variables were suppressed to simplify notation.

3.3 Error Variance Estimation

For any scale score-based statistic in NAEP, two distinct components of error variance are estimated: (1) *sampling error* and, (2) *error because of latency of the scale scores*. For the 2017 NAEP mathematics and reading DBAs, a third distinct component of error variance was estimated—linking error.

Sampling error is complicated by the complex sampling design used by NAEP (Johnson & Rust, 1992), and NAEP uses a jackknife-based approach to estimate the sampling error component (NCES, 2000-2018). For each sample, 62 replicate datasets are produced. This is accomplished through the use of a set of replicate weights which, when paired with test-taker response data, effectively generates the jackknife replicate datasets. Each set of weights involves deleting a portion of the dataset by setting the weights for students within 1 of 62 primary sampling units to 0, and reweighting the remaining students so that each replicate dataset remains representative. The jackknife procedure uses a function of the variance of the statistic across the replicate datasets to measure the component of variance resulting from the sampling of schools and students.

Because the true proficiencies are latent and not directly observed, using the NAEP scale scores as an estimate of the latent proficiencies may introduce some error variance. The type of error is estimated based on random draws from the posterior distribution of proficiency (i.e., $\boldsymbol{\theta}$) for each test taker conditional on the test taker's responses to the assessment questions and a full array of demographic and contextual variables. These random draws are referred to as *plausible values* and have other utility (see, e.g., Mazzeo, 2018 or von Davier et al., 2007). Adapting the work of Rubin (1987) to the context of latent variable models, NAEP uses a function of variance of the statistic across the multiple sets of the random draws as an estimate of the component of error variance due to latency.

As noted above, the linking transformation for the 2017 NAEP mathematics and reading DBAs involved an additional common population linking between the PBA and DBA samples within 2017. This additional transformation introduced an additional source of error variance, specifically associated with the common population linking, which needed to be taken into account for comparisons with prior PBA results as well as for comparisons of DBA and PBA results within 2017.

A procedure was developed to estimate the linking error component. The central idea behind the procedure was the use of a variant of the general jackknife approach as described above to estimate the sampling and latency error associated with the linking coefficients *A* and *B*. These estimated errors in the linking constants were, in turn, used to quantify the linking error to add to the typical two error components for all NAEP DBA-based reporting group statistics. Two distinct approaches to estimating the error components were developed and evaluated: (1) formula-based estimates based on substituting sample moments for the required population quantities (essentially a method-of-moments approach; Mood, Graybill, & Boes, 1974, pp. 74-76), and (2) Monte Carlo estimates modeled after an empirical Bayes approach (Carlin & Lewis, 1996). Both approaches produced essentially equivalent results. Operationally, an approximation to the Monte Carlo estimates was used. A more detailed description of the development and evaluation of the linking error estimation and the final operational procedure can be found in Mazzeo et al. (2018).

The 2017 operational assessments in mathematics and reading were designed to be administered in both modes: paper and digital. When comparing results between the 2017 DBA and PBA samples, it was necessary to ensure that the jackknife procedure for standard errors of the differences between results for the two modes was implemented appropriately for this purpose. In particular, it was essential to reflect the fact that, for the most part, some students in each sampled school took DBA while others in the same school took PBA. Thus, the two student samples were not independent. This design provided considerable statistical power for making comparisons between the samples assessed in each mode, compared to a design using independent samples. The method of implementing the jackknife procedure used in this study ensured that the dependency between the DBA and PBA samples was reflected in all estimates of sampling variance associated with comparisons by mode. In brief, this was accomplished by forming the jackknife replicate groups of schools without regard to mode. That is, a single procedure was used to assign schools to replicate groups, reflecting the overall school sample design. The replicates were not created by treating the samples for each mode as if they were distinct samples of schools. However, the part of the replication procedure that involved the assignment of students to replicate groups did respect the fact that the student samples were of distinct groups of students, as no student was assessed using both modes.

4. Impact of the Transition on the Measurement Instrument

An important first step in evaluating the transition from PBA to DBA was to consider the possible *mode difference* in item characteristics. A mode difference is defined as the difference between a statistic for a PBA item and the same statistic for the corresponding trans-adapted digital version, prior to the linking transformation. Evaluating the item-level mode difference corresponds to the first research question (see the end of chapter 1). This chapter compares mode differences between items in paper and digital formats only for items classified as *comparable* items. Items in paper and digital formats were classified as comparable only when the digital conversion of the item was associated with no or little context change due to the trans-adaptation and DBA assembly process (see 2.2 for a description of trans-adapted items).

The number of comparable items varies by subject. For reading, the PBA blocks were kept intact when trans-adapted to DBA. Therefore, all reading trans-adapted items at grades 4 and 8 were classified as comparable and included in the following item-level comparisons. The assembly process for DBA trans-adapted blocks is a little more complicated for mathematics. At both grades, six out of the ten mathematics DBA blocks were trans-adapted blocks and the items were selected from six designated PBA parent blocks. Because some mathematics items from these six PBA blocks could not be trans-adapted, items from other PBA blocks were substituted. The NAEP program has observed that even small changes to context or position of an item within a block can affect its psychometric properties (Zwick, 1991). To exclude potential confounding of context or position effects, only those trans-adapted mathematics items that were from the six designated PBA blocks and had no or only one position shift from their PBA parent blocks were classified as comparable and used in the following item-level comparisons.

Table 4.1a summarizes the distribution of the comparable items for each subject and grade combination. For reading, 88 items were included in the mode comparison at grade 4 and 120 items at grade 8. For mathematics, 78 and 77 items were included at grades 4 and 8, respectively. Table 4.1a also shows the breakdown of the comparable items by content area (or subscale) and by item type. The NAEP reading framework (NAGB, 2017a) specifies two types of texts on the assessment: Literary texts and Informational texts. The NAEP mathematics framework (NAGB, 2017b) requires items to measure one of the five major content areas: number properties and operations; measurement; geometry; data analysis, statistics, and probability; and algebra.

As table 4.1a shows, the 2017 NAEP assessments utilized two types of items: selected-response (SR) items and constructed-response (CR) items. For reading, SR items were traditional single-selection multiple-choice (SSMC) items (i.e., multiple response options were provided for each item, only one response option can be selected, and there is only one correct response option). For mathematics, there are several different SR formats—SSMC items, multiple-selection multiple-choice (MSMC) items, matching items, in-line choice items, and grid items (for example, see NCES 2018). CR items in both mathematics and reading require students to write/type short responses. Note that apart from three matching items at grade 4 and two matching items at grade 8, the SR items included in the mode comparisons are all SSMC items. Items with the other SR formats were only present in the digital assessment.

The impact on the measurement instrument was evaluated by analyzing the item-level mode difference on the items listed in table 4.1a, from the perspective of classical test theory (CTT), and IRT. For CTT,

mean item scores of the comparable items were compared between PBA and DBA. For IRT, item difficulty and discrimination parameter estimates were compared. Note that, because the samples of students taking the PBA and DBA were randomly equivalent samples selected from a common population, differences in these statistics reflect differences in the instrument and sampling error, but not population differences.

Table 4.1a. Item counts within content area and item type for the paper and digital instruments on a comparable set of trans-adapted items: 2017

	Grade 4			Grade 8		
Subject and content area	SR	CR	Total	SR	CR	Total
Reading						
Literary	34	16	50	34	26	60
Information	26	12	38	37	23	60
Total	60	28	88	71	49	120
Mathematics						
Numbers and operations	29	4	33	17	3	20
Measurement	11	2	13	11	1	12
Geometry	10	2	12	11	1	12
Data analysis	5	3	8	9	3	12
Algebra	8	4	12	13	8	21
Total	63	15	78	61	16	77

NOTE: SR = selected response. CR = constructed response.

4.1. Classical Test Theory Item Performance Across Modes

Examining item property differences within the CTT framework is optimal for directly evaluating performance differences on the items, without assumptions about the item response process. To evaluate the impact of the paper to digital transition on the assessment instrument, the mean item scores[6] for the comparable items were compared between the two formats.

[6] For multiple-choice and dichotomous constructed-response items, the mean item score, or weighted percent correct, is the percentage of examinees who received a correct score on the item. For polytomous items, weighted percent correct is the sum of percentage proportion of examinees in each score category weighted by the magnitude of each score category and standardized with a maximum credit of 1. For example, if there are 3 scoring categories (0, 1, and 2) for an item and percentage distribution for the item across three score categories is 20%, 40%, and 40%, respectively, then the weighted percent correct will be: 20 (percent)* 0 (point)/2 (maximum score) + 40 (percent)* 1 (point)/2 (maximum score) + 40* (percent)*2 (point) /2 (maximum score) = 60 (percent). Average weighted percent correct refers to an average of weighted percent correct across items.

Table 4.1b shows the overall mean item score averaged across the comparable items within each assessment and the corresponding mode difference for both subjects and both grades. For each subject and grade combination, the mode difference was significantly different from zero ($p < .05$) and negative. As the samples taking PBA and DBA were randomly equivalent, these significant differences suggest that in general the mathematics and reading items became more difficult when trans-adapted to the digital format. The mode difference was greater at grade 4 than at grade 8. The largest difference was observed at grade 4 reading, while the smallest difference was observed at grade 8 reading.

Table 4.1b. Overall weighted average item score for the paper and digital instruments on a comparable set of trans-adapted items: 2017

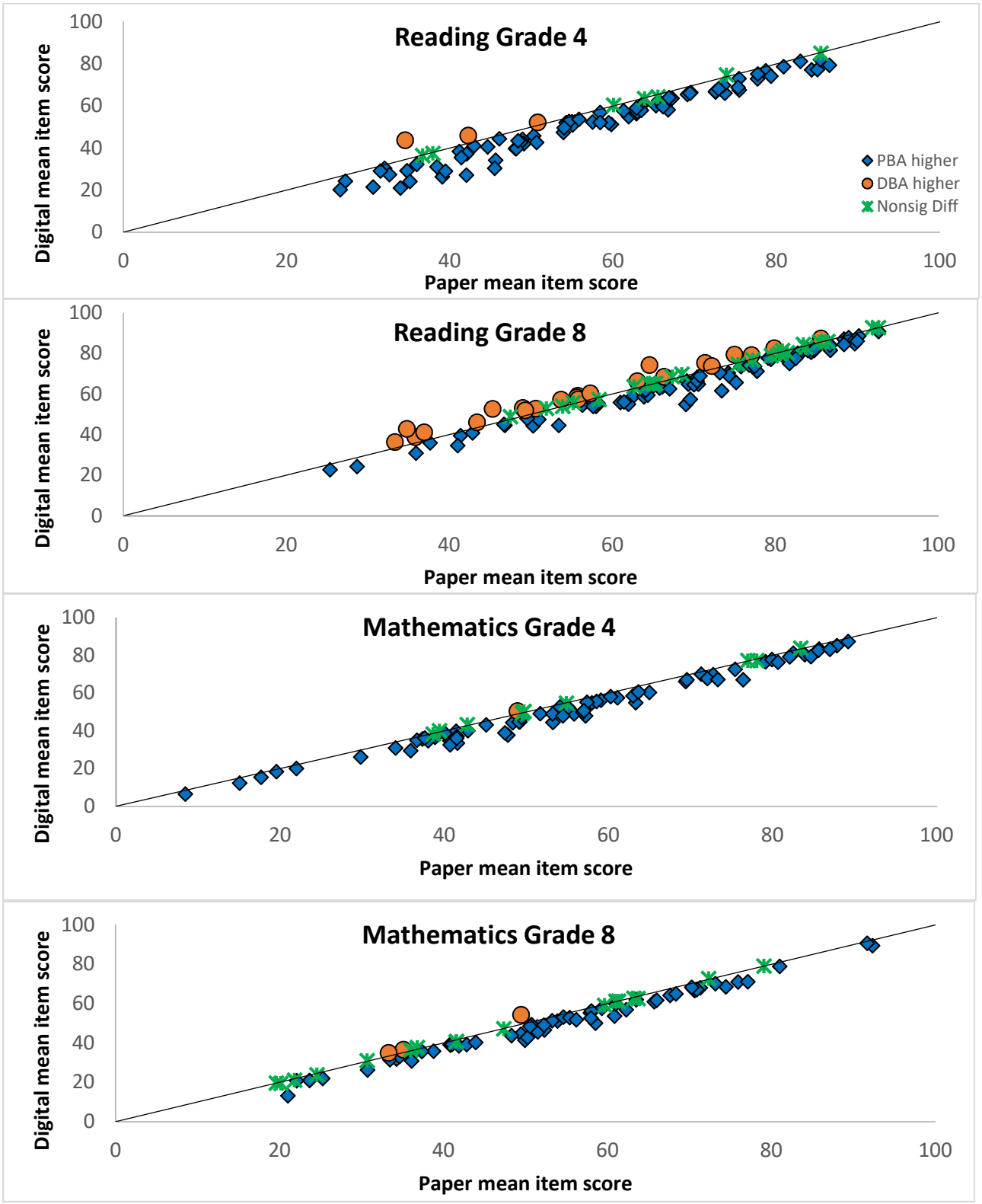
	Grade 4			Grade 8		
Subject	DBA	PBA	DBA-PBA (SE)	DBA	PBA	DBA-PBA (SE)
2017 reading	47%	53%	-5.3% (.18)	61%	63%	-1.9% (.20)
2017 mathematics	52%	55%	-3.7% (.23)	49%	52%	-2.9% (.25)

NOTE: All differences were significant ($p < .05$). DBA = digitally based assessment. PBA = paper-based assessment. Detail may not sum to totals because of rounding.

Figure 4.1a shows scatterplots comparing the mean item scores for the paper and digital formats of each comparable item with a line of equivalence at a 45-degree angle. The vast majority of the items for grade 4 reading, grade 4 mathematics, and grade 8 mathematics had significantly lower mean item scores in the digital format than in the paper format (the blue squares). For grade 8 reading, which had the smallest overall mean item score difference (table 4.1b), most items had significantly different mean item scores, but many items had significantly higher mean item scores for the digital format (orange circles). As the samples assessed on DBA and PBA were randomly equivalent, these differences in mean item scores can be attributed to differences in the difficulty of the items in the paper and digital formats.

The correlations between the mean item scores in the two formats were very high. For reading, the correlations were .97 for both grades. For mathematics, the correlations were .99 for both grades. For each grade and subject combination, the plotted points scatter around the 45-degree diagonal line, albeit more are below the line, suggesting that the mean item score differences observed in table 4.1b were relatively consistent across all items. In other words, the items generally retained their difficulty ranking relative to other items within the instrument, but all items for a given subject and grade combination became more difficult by roughly the same amount.

Figure 4.1a. Paper versus digital mean item scores on comparable items for each subject and grade combination: 2017



NOTE: DBA = digitally based assessment. PBA = paper-based assessment.

A small number of items within each subject and grade combination were observed as having outlier mode differences, either in a positive or negative direction, and these items were closely examined. For example, one item in reading grade 4 had a mean item score that was 9 percentage points higher in the digital format. Examination of the actual item suggested that the relevant text for answering this question may be more prominent in the digital format due to a change in how image and text aligned in the digital format. The reasons for most other items showing outlier mode differences were not clear. These items were still used in the following analyses and comparisons because the typical common item linking was not used in 2017, and it was therefore not required to maintain a set of items that function exactly the same between the two modes. See chapters 1 and 2 for a detailed description of the linking design employed in 2017.

Table 4.1c shows the mean item score mode difference for SR items and CR items. For both subject areas in grade 4, the CR items show a noticeably larger mode difference than the SR items. In contrast, the mode differences across the two item types were more similar for grade 8. Note that the reading CR items were more difficult than the SR items, meaning that the larger mode differences on the CR items for reading grade 4 result in greater mode differences for harder items (see the lower range of the scale in the first panel in figure 4.1a).

CR items for both mathematics and reading at grade 4 show the largest mode difference, of almost 7 percent, which was significantly different from the mode difference for SR items for both subjects. This suggests that some of the greater overall mode differences observed for reading grade 4 compared to mathematics grade 4 may be because of a higher proportion of CR items in the reading grade 4 assessment. For example, within the comparable item pool, 28 out of the 88 grade 4 reading comparable items or 32 percent (see table 4.1a) were CR, compared to 19 percent and 21 percent of the mathematics items at grades 4 and 8. Taken together, these results suggest that the CR items involving typing short answer responses (for reading) or the use of equation editor (for mathematics) may be more demanding in the digital format—particularly for grade 4 students. Furthermore, though the relative difficulty ordering of items was highly similar across modes—as indicated by the strength and slope of the relationships between the paper and digital formats of the item (described earlier in this subsection)—the item-type results showed there were some items that had larger differences than others.

Table 4.1d shows the mean item score mode difference for the two content areas of reading and the five content areas of mathematics. For reading, the two content areas have mean item score mode differences that are similar in magnitude to the overall mean item score difference. For mathematics, there is noticeable variation in the mean item score mode difference across the five content areas at grade 4. Note that the distribution of the content area coverage for mathematics is relatively uneven at grade 4, with the most represented content area (Numbers and Operations) covering around 40 percent of the testing time and the least represented content area (Data Analysis) covering only around 10 percent of the testing time.

Table 4.1c. Mean item score within item type for the paper and digital instruments on a comparable set of items: 2017

	Grade 4			Grade 8		
Subject and item type	DBA	PBA	DBA-PBA (SE)	DBA	PBA	DBA-PBA (SE)
Reading						
Selected response	60%	64%	-3.8% (.22)	74%	76%	-1.6% (.19)
Constructed response	35%	42%	-6.8% (.18)	53%	55%	-2.0% (.24)
Mathematics						
Selected response	54%	56%	-2.4% (.24)	51%	53%	-2.5% (.26)
Constructed response	46%	52%	-6.9% (.31)	44%	48%	-3.5% (.30)

NOTE: All differences were significant ($p < .05$). DBA = digitally based assessment. PBA = paper-based assessment. SE = standard error. All reading selected-response items were single-select, multiple-choice items. Mathematics selected-response items included the following item types: single select (SS) multiple choice (MC), multiple select (MS) MC, MS matching, MS grid, MS zone, in-line choice, and composite-selected, as described in the text. Detail may not sum to totals because of rounding.

Table 4.1d. Item distribution by content area and mean item score within content area for the paper and digital instruments on a comparable set of trans-adapted items: 2017

Subject and content area	Grade 4				Grade 8			
	Item distribution	DBA	PBA	DBA-PBA (SE)	Item distribution	DBA	PBA	DBA-PBA (SE)
Reading								
Literary	57%	48%	54%	-5.6% (.24)	50%	62%	64%	-2.4% (.26)
Information	43%	46%	51%	-5.0% (.23)	50%	61%	62%	-1.3% (.25)
Mathematics								
Numbers and operations	42%	52%	55%	-2.5% (.27)	26%	56%	58%	-2.6% (.31)
Measurement	17%	54%	60%	-5.9% (.29)	16%	53%	55%	-2.1% (.39)
Geometry	15%	52%	56%	-3.6% (.35)	16%	41%	44%	-2.8% (.39)
Data analysis	10%	39%	43%	-4.3% (.36)	16%	35%	39%	-3.8% (.37)
Algebra	15%	55%	60%	-4.6% (.30)	27%	51%	54%	-2.8% (.26)

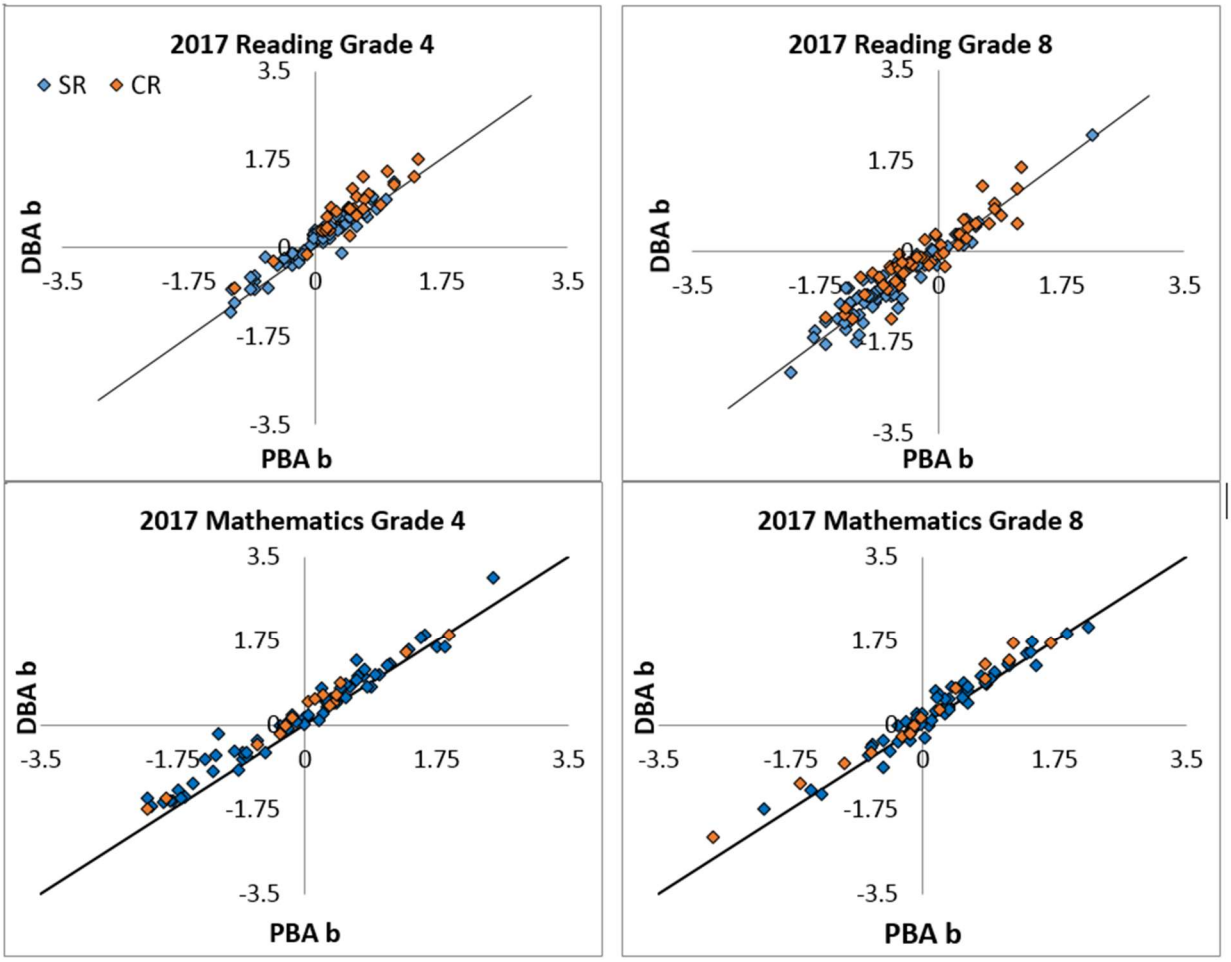
NOTE: All differences were significant ($p < .05$). Detail may not sum to totals because of rounding. DBA = digitally based assessment. PBA = paper-based assessment. SE = standard error.

4.2 Item Response Theory Parameter Comparisons

Examining item property differences within the IRT framework is optimal for directly evaluating item differences in relation to the scale score generating model described in chapter 3, which is based on the IRT. The paper-format item parameters were estimated with a concurrent calibration of 2017 PBA and 2015 PBA following standard NAEP procedures. The DBA item parameters were estimated with a 2017 DBA-only one-population model, and then transformed to a metric comparable to that of 2017 PBA item parameters by equating the mean and standard deviation of the 2017 DBA theta estimates to that of the 2017 PBA theta estimates.

Figure 4.2a shows comparisons of the IRT difficulty parameter (or the *b* parameter) estimates for comparable items between PBA and DBA for each grade and subject combination. The difficulty parameters are conceptually similar to the mean item score, and unsurprisingly, the difficulty parameter estimates scatterplots led to similar conclusions as the mean item score scatterplot. While an almost perfect correlation was found between PBA and DBA item parameter estimates, the DBA item difficulty estimates were higher on average, and the mode differences were larger at grade 4. For reading grade 4, CR items were harder on average, and also showed greater mode differences on average. These conclusions are consistent with the conclusions described above in relation to the CTT average item score differences.

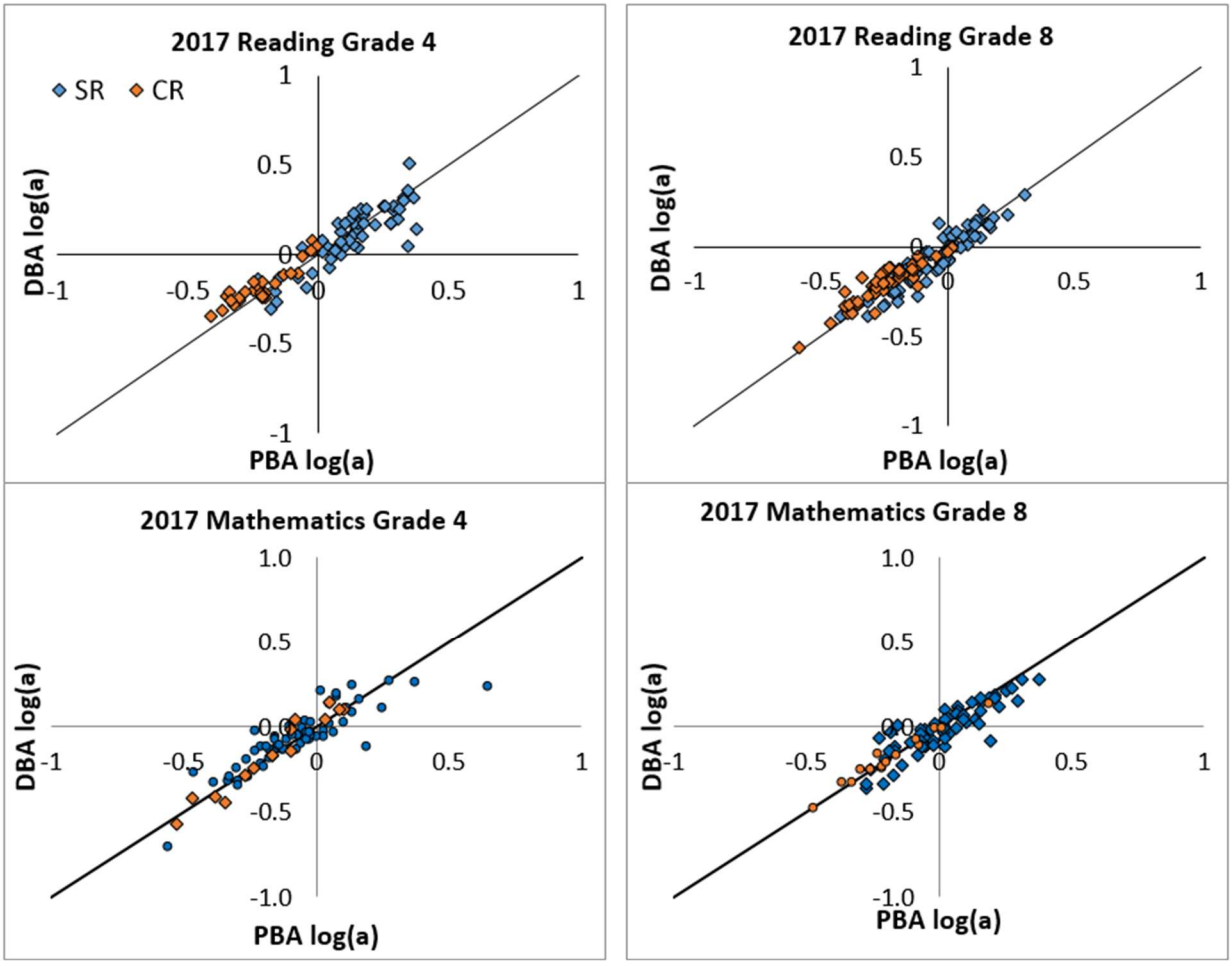
Figure 4.2a. Difficulty parameter estimate comparisons, by item type: 2017



NOTE: DBA = digitally based assessment. PBA = paper-based assessment. CR = constructed response. SR = selected response.

Figure 4.2b shows comparisons of the log of the IRT discrimination parameter (or the *a* parameter) estimates for each comparable item by subject and grade. The IRT discrimination parameter represents how well the item discriminates between test takers with different latent ability. A log transformation helps conform the discrimination parameter's distribution to normality and reduce the variability of data. Similar to the mean item score and difficulty parameter estimates scatterplots, the discrimination parameter estimates also fall closely around the PBA=DBA line. For reading, CR items have slightly higher discrimination parameters in DBA. For mathematics grade 8, items tended to have lower discrimination parameters in DBA.

Figure 4.2b. Log of discrimination parameter estimate comparisons, by item type: 2017



NOTE: DBA = digitally based assessment. PBA = paper-based assessment. CR = constructed response. SR = selected response.

Discussion

Three general conclusions can be inferred from the results. First, the most salient impact of the mode transition on item-level statistics was an increase in difficulty for the digital versions of items, as expressed by lower-mean item scores and higher-difficulty parameter estimates for the digital versions with respect to the original paper versions. This finding provides strong evidence that the transition to DBA had systematic impact on the difficulty of the assessment instrument, and it implies that the method used to model and link the DBA results to results from PBA must not assume the digital and paper instruments were equivalent in terms of difficulty.

Second, apart from the shift in difficulty, items generally retained the properties of the original paper versions, including most notably the rank-order of the item difficulty with respect to the other items. This observation, in addition to subgroup, state/jurisdiction, and district results as described in later chapters, provides support for construct similarity underlying PBA and DBA, despite instrument differences.

Third, there was no theoretically demarcated group of items that had zero difference in item properties across the two modes. As noted in chapter 3, typical NAEP assessment results are linked to the metric of prior results via the assumption that a subset of items have common properties across the current and previous administration of the NAEP subject. However, the lack of evidence for a subset of items with common properties suggests the usual linking procedure is not appropriate. Consequently, the findings support the use of the common population linking method, as was used in the 2017 NAEP digital transition bridge study.

5. Evaluation of Mode Transition on State/Jurisdiction and District Estimates

While chapter 4 provides contextual information about the mode differences on the measurement instrument to confirm prior findings that informed the methodological approach (see chapter 2), this chapter explores the DBA and PBA scale score differences after linking. This chapter addresses the second research question described at the end of chapter 1. Specifically, was the linear, national-level linking transformation effective at aligning the DBA scale scores with the PBA scale scores for all states/jurisdictions and districts participating in the Trial Urban District Assessment (TUDA)?

This chapter focuses on *mode residuals*. Mode residuals are defined as the difference between a DBA statistic and a PBA statistic, *after* the linking transformation (e.g., a subgroup mean, a state mean, or a percentile). In contrast to the mode differences of chapter 4 concerning item-level differences prior to linking that may be non-zero, mode residuals are calculated after linking and are expected to be close to zero, except resulting from estimation error, if minimal construct differences are introduced by the digital transition.

The first section describes the estimation and evaluation of the mode residuals for states/jurisdictions and TUDA districts within the 2017 assessment results. The focus was on whether the mode residuals were significantly different from zero. NAEP is primarily concerned with whether any state/jurisdiction or district had significant mode residuals, rather than concerned with any one state/jurisdiction or district in isolation. Therefore, correction for multiple comparisons was made to evaluate this hypothesis. However, for the purpose of evaluating the mode residuals for a particular state/jurisdiction or district in isolation, uncorrected significance test outcomes are also provided.

The second section provides scatterplots and correlations of the mode residuals across subjects and across grades. The purpose of this analysis is twofold: (1) to examine whether there is a hypothetical common cause of the mode residuals across subjects and across grades, such as computer access and familiarity, which could cause correlations between the mode residuals, and (2) to examine whether there is any evidence that any states/jurisdictions or TUDA districts have been systematically advantaged or disadvantaged across the assessments.

The final section examines whether there is any evidence that the mode transition impacted the reported trend results or changes in proficiency across the previous paper-based NAEP administration in 2015 and the current digitally based NAEP administration in 2017.

5.1. Mode Residuals Across States/Jurisdictions and Districts

Table 5.1a lists the significant mode residuals between the DBA and PBA scale scores for the states/jurisdictions that participated in the NAEP 2017 mathematics and reading assessments. All 50 U.S. states participate in NAEP mathematics and reading at grades 4 and 8. In addition, NAEP provides state-like estimates for two additional jurisdictions: Washington, D.C. public schools and Department of Defense Education Activity (DoDEA) schools. In total, there were 52 states/jurisdictions that participated in the NAEP 2017 operational mathematics and reading assessments. Although the NAEP 2017 mathematics administration for grades 4 and 8 also included assessments for Puerto Rico, a separate instrument was used for the Puerto Rico assessments and the analysis was completed separately from the 52 states/jurisdictions; therefore, Puerto Rico is not included in the evaluation.

As more statistical tests are considered simultaneously, the expected occurrence of false positives increases. This is known as the "multiple comparison" problem. Within a subject and grade combination, the standard NAEP reporting procedure when considering changes in proficiency across years for all states simultaneously is the Benjamini-Hochberg procedure which controls the false discovery rate (FDR; Benjamini & Hochberg, 1995). While the focus on the NAEP mode evaluation study was the entire set of states/jurisdictions and not any particular state/jurisdiction, users of NAEP may be interested in only one particular state/jurisdiction. To account for both purposes, the state/jurisdiction mode residual statistical tests were conducted with and without any adjustment for multiple comparisons.

Table 5.1a lists the states/jurisdictions for which the mode residuals are statistically significant from zero with $p < .05$ and no FDR procedure applied. Differences for the two states with names in bold are still significant with the FDR procedure within that subject/grade combination, for a family size of 52 states/jurisdictions. The table shows that in three of the four subject and grade combinations, no significant mode residuals were observed after controlling for multiple comparisons with the FDR procedure. Only grade 8 reading had significant mode residuals with the FDR procedure, where two significant mode residuals were observed. These were for Georgia with a higher mean scale score estimate on DBA and for Kentucky with a higher scale score estimate on PBA.

Without adjusting significance tests for multiple comparisons, table 5.1a shows that 8 percent to 15 percent of the state/jurisdiction mode residuals were significant within each subject and grade combination, for a total of 21 significant residuals. Slightly more significant residuals were associated with higher scores on DBA than on PBA. Within states, the results across the subject and grade combinations were very inconsistent. Only two states showed significant mode residuals in more than one subject and grade combination: Georgia and Hawaii. For both of these states, the estimate based on DBA was higher than the estimate based on PBA for both mathematics and reading at grade 8.

Table 5.1a. States/jurisdictions with significant mode residuals: 2017

Grade and subject	DBA > PBA	DBA < PBA	Nonsignificant
Grade 4 reading	AZ, CT, RI	ND	48
Grade 8 reading	CA, **GA**, HI, IN	AL, IA, **KY**, NH	44
Grade 4 mathematics	OR, SD, VA	AR, PA	47
Grade 8 mathematics	GA, HI, TN	WI	48

NOTE: Bolded states are significant with the FDR procedure within that subject and grade family size of 52 states/jurisdictions. All listed states/jurisdictions are significant with $p < .05$ and no multiple comparison adjustment. DBA = scale score estimate based on the digitally based assessment. PBA = scale score estimate based on the paper-based assessment.

Table 5.1b lists the districts for which the mode residuals are statistically significant from zero with $p < .05$ and no FDR procedure applied. A total of 27 districts participated in the NAEP 2017 operational mathematics and reading assessments.

Table 5.1b shows that, without adjusting for multiple comparisons, between 0 percent and 11 percent of the district mode residuals were significant within each subject and grade combination. Among those with significant mode residuals, more districts had higher proficiency scores in PBA. All four of the subject and grade combinations had no significant mode residuals with the FDR control for multiple comparisons.

Table 5.1b. TUDA districts with significant mode residuals: 2017

Grade and subject	DBA > PBA	DBA < PBA	Nonsignificant
Grade 4 reading	-	-	27
Grade 8 reading	-	Denver, Philadelphia, Hillsborough County	24
Grade 4 mathematics	Fresno	Atlanta, Houston	24
Grade 8 mathematics	-	-	27

NOTE: All listed districts are significant with $p < .05$ and no multiple comparison adjustment. DBA = scale score estimate based on the digitally based assessment. PBA = scale score estimate based on the paper-based assessment.

5.2 State and District Mode Residuals Across Subjects and Grades

To examine the consistency of the state and district mode residual estimates across subject and grade, effect sizes for the mode residual estimates for the states and TUDA districts were compared. The effect size was calculated as the mode residual divided by the 2017 reported (i.e., DBA scale scores) standard deviation for the relevant state/jurisdiction. The use of effect sizes enabled the comparison of the mode residuals across subjects and grades that differ in standard deviations, and they are expressed in a metric that is meaningful in relation to the reported results of the state/jurisdiction.

Figure 5.2a shows scatterplots of the effect sizes for states/jurisdictions, comparing across either subject or grade. As stated before, the purpose of this analysis is to examine whether there is any common cause of the mode residuals across subjects and across grades that could cause correlations between the mode residuals. The top two panels in the figure compare the effect sizes across the two subjects within a grade. The correlations of the state/jurisdiction mode residuals between subjects were .15 and .32 for grade 4 and grade 8, respectively. The bottom two panels in the figure compare the effect sizes across the two grades within a subject. The correlations between grades were .10 and .12 for mathematics and reading, respectively. Across the four correlations, only the correlation between grade 8 mathematics and reading was significantly different from zero ($p < .05$). Note that the majority of schools had students tested in both mathematics and reading, so a small correlation between subjects may be expected because of common school-level sampling. The scatterplots, however, do not indicate any noticeable pattern across the states/jurisdictions.

Figure 5.2a. Effect sizes for state/jurisdiction mode residuals, across subject and across grade: 2017

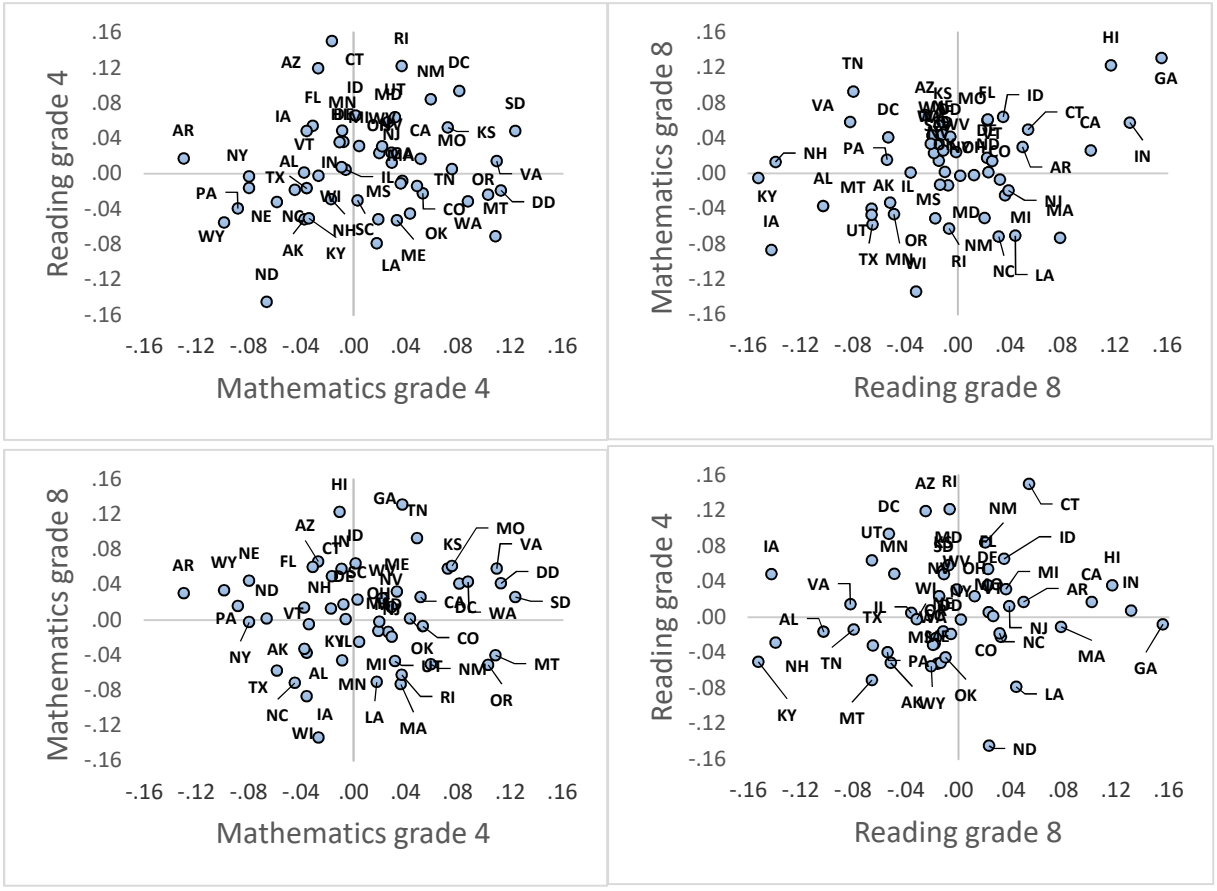
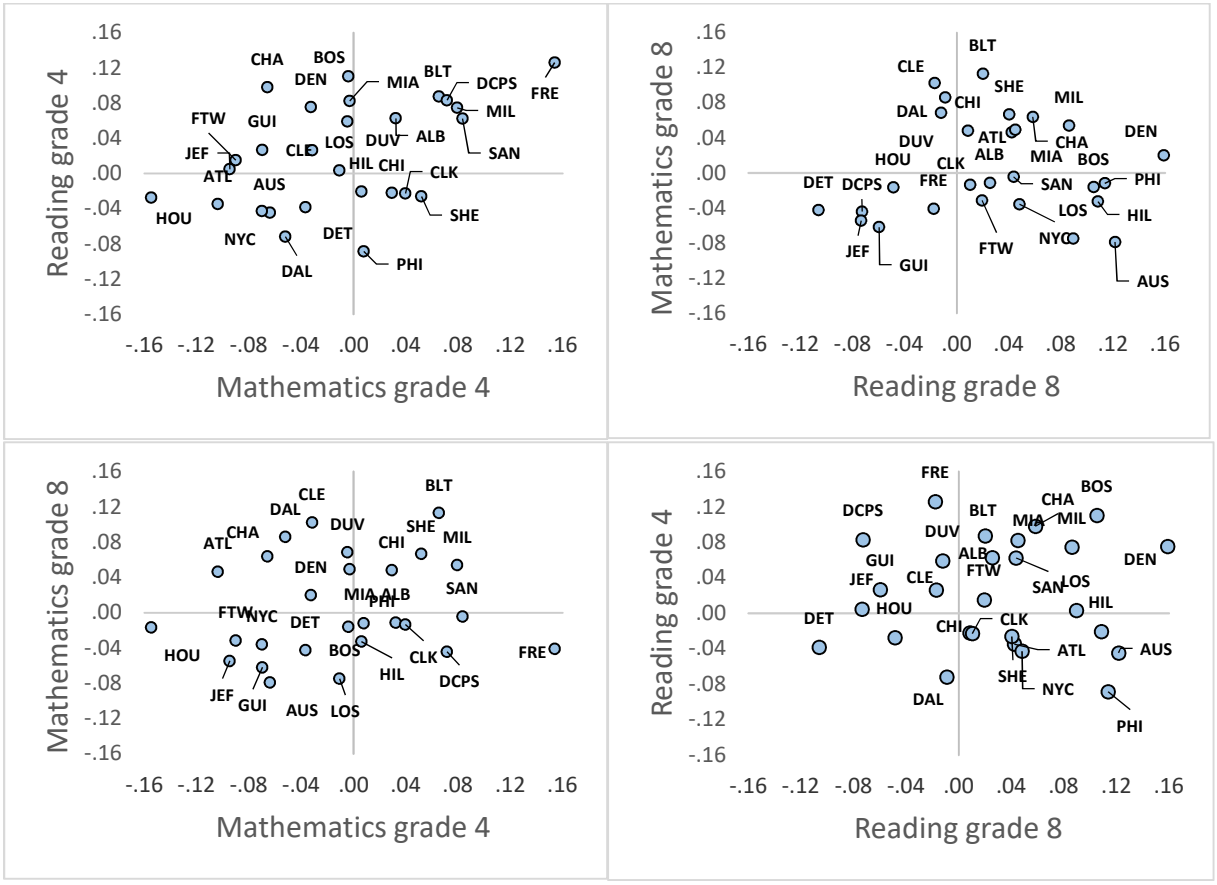


Figure 5.2b shows scatterplots of the effect sizes for TUDA districts, comparing across either subject or grade. The top two panels in the figure compare the effect sizes across the two subjects within a grade. The correlations between subjects were .47 and .05 for grade 4 and grade 8, respectively. The bottom two panels in the figure compare the mode residuals across the two grades within a subject. The correlations between grades were .13 and -.02 for mathematics and reading, respectively.

Across the four correlations between district mode residuals, only the correlation between grade 4 mathematics and reading was significantly different from zero ($p < .05$). The magnitude of the correlation between grade 4 mathematics and reading, .47, may appear moderate; however, it is partially driven by an extreme case on the upper right corner, or Fresno Unified School District (FRE). The correlation between the two subjects without FRE is .36, which suggests only about 10% of the mode residual variance in one subject could be explained by the other subject. Note that the majority of schools sampled had students tested in both mathematics and reading, so a small correlation between subjects is expected due to common school-level sampling error.

Figure 5.2b. Scatterplots of TUDA district mode residuals for comparing subjects within a grade and for comparing grades within a subject: 2017



5.3 Impact of Mode Residuals on Trend Results

To examine the impact of mode residuals on state/jurisdiction and district average proficiency from 2015 to 2017, or *trend*, contingency tables (tables 5.3a, 5.3b, 5.3c, and 5.3d) were produced comparing reporting outcomes based on DBA (*the reported trend*) and a trend adjusted for the mode residual specific to that state/jurisdiction (*the adjusted trend*). Specifically, the adjusted trend is equal to the reported trend minus the mode residual. In tables 5.3a, 5.3b, 5.3c, and 5.3d, both reported trend and adjusted trend represent the difference between the 2017 and 2015 proficiency estimates, where the outcome could be either a significantly positive trend (higher in 2017), a significantly negative trend (higher in 2015), or no significant trend. Note that for each state the reported trend was based on the DBA results in a larger sample (approximate 2,200 or greater per jurisdiction) and the adjusted trend was based on the PBA results in a smaller sample (approximately 500 or greater per state), so a higher frequency of significant reported trends was expected than adjusted trends, due to increased power associated with larger sample size.

Table 5.3a shows the contingency tables between the adjusted trend and the reported trend for all states/jurisdictions. In no case were both the reported and adjusted trends significant and in different directions (i.e., the off-diagonal cases where one is “<” and the other is “>”). For each subject and grade

combination, between 41 and 46 of the 52 states/jurisdictions had the same reporting outcome for both trends (i.e., the diagonal cases). Between 6 and 11 of the states/jurisdictions within each subject and grade combination had one significant and one nonsignificant trend (i.e., the off-diagonal cases where one is "n.s." and the other is ">" or "<"). In the majority of these cases, the reported trend was significant while the adjusted trend was nonsignificant. Again, because PBA and DBA were given to different samples, some degree of inconsistency in the trend outcomes is expected due to sampling errors.

Table 5.3a. Contingency tables for reported trends versus adjusted trends, for all states/jurisdictions: 2015 and 2017

Grade 4 reading				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	9	0
	n.s.	2	41	0
	>	0	0	0

Grade 8 reading				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	1	0
	n.s.	0	41	0
	>	0	9	1

Grade 4 mathematics				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	4	6	0
	n.s.	1	38	2
	>	0	0	1

Grade 8 mathematics				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	3	0
	n.s.	1	46	0
	>	0	2	0

NOTE: "<" means a significant negative trend, ">" means a significant positive trend, and "n.s." means a nonsignificant trend. All trends are estimates of the change in proficiency across 2017 and 2015.

Table 5.3b is similar to table 5.3a, but counting only cases where a significant mode residual was detected (see table 5.1a). In no case were both the reported and adjusted trends significant and in different directions. A total of 8 cases across the two subjects and two grades had one significant and one nonsignificant trend (i.e., the off-diagonal cases where one is "n.s." and the other is ">" or "<"), 7 of which had a more positive outcome on the reported trend (a positive reported trend but a nonsignificant adjusted trend, or a nonsignificant reported trend but a negative adjusted trend). Only one case, for reading grade 4, had a more negative outcome on the reporting trend than the adjusted trend. In this case, the reporting trend was a significant decrease while the adjusted trend was nonsignificant.

Table 5.3b. Contingency tables for reported trends versus adjusted trends, for states/jurisdictions where the associated mode residual was significantly different from zero: 2015 and 2017

Grade 4 reading				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	1	0
	n.s.	2	1	0
	>	0	0	0

Grade 8 reading				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	0	0
	n.s.	0	4	0
	>	0	4	0

Grade 4 mathematics				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	1	0	0
	n.s.	0	4	0
	>	0	0	0

Grade 8 mathematics				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	0	0
	n.s.	1	3	0
	>	0	0	0

NOTE: "<" means a significant negative trend, ">" means a significant positive trend, and "n.s." means a nonsignificant trend. All trends are estimates of the change in proficiency across 2017 and 2015.

Table 5.3c shows the contingency tables between the adjusted trend and the reported trend for all TUDA districts that participated in both 2017 and 2015 operational mathematics and reading assessments and therefore for which trend estimates are possible. A total of 21 districts (out of 27) were included. As was the case for state results, in no instance were the reported and adjusted trends statistically significant and in different directions. For each subject and grade combination, between 12 and 19 of the 21 districts had the same reporting outcome for both trends. Between 2 and 9 of the districts within each subject and grade combination had one significant and one nonsignificant trend. In the majority of these cases, the reported trend was significant while the adjusted trend was nonsignificant.

Table 5.3c. Contingency tables for reported trends versus adjusted trends, for all TUDA districts: 2015 and 2017

Grade 4 reading				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	0	0
	n.s.	1	19	0
	>	0	1	0

Grade 8 reading				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	0	0
	n.s.	0	19	0
	>	0	2	0

Grade 4 mathematics				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	4	0
	n.s.	0	12	1
	>	0	4	0

Grade 8 mathematics				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	1	0
	n.s.	1	18	1
	>	0	0	0

NOTE: "<" means a significant negative trend, ">" means a significant positive trend, and "n.s." means a nonsignificant trend. All trends are estimates of the change in proficiency across 2017 and 2015. Only 21 of the 27 districts that participated in the 2017 NAEP operational assessment also participated in the 2015 NAEP operational assessment and are represented here.

Table 5.3d is similar to table 5.3c, but counting only cases where a significant mode residual was detected (see table 5.1b). Again, in no case were the reported and adjusted trends significant and in different directions. In other words, the differences between the reported and adjusted trends did not change a score gain to a score decline (or vice versa). A total of two cases across the two subjects and two grades had one significant and one nonsignificant trend, both in mathematics grade 4: one had a more positive outcome on the reported trend (a positive reported trend but a nonsignificant adjusted trend, or a nonsignificant reported trend but a negative adjusted trend) and the other had a more negative outcome on the reporting trend.

Table 5.3d. Contingency tables for reported trends versus adjusted trends, for all TUDA districts where the associated mode residual was significantly different from zero: 2015 and 2017

Grade 4 reading				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	0	0
	n.s.	0	0	0
	>	0	0	0

Grade 8 reading				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	0	0
	n.s.	0	2	0
	>	0	0	0

Grade 4 mathematics				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	0	0
	n.s.	0	1	1
	>	0	1	0

Grade 8 mathematics				
		Adjusted trend		
		<	n.s.	>
Reported trend	<	0	0	0
	n.s.	0	0	0
	>	0	0	0

NOTE: "<" means a significant negative trend, ">" means a significant positive trend, and "n.s." means a nonsignificant trend. All trends are estimates of the change in proficiency across 2017 and 2015. Only 21 of the 27 districts that participated in the NAEP 2017 operational assessment also participated in the NAEP 2015 operational assessment, and they are represented here.

Discussion

As an overall test of whether any state/jurisdiction or district had significant mode residuals, significance tests adjusted for multiple comparisons using the FDR procedure were conducted. With multiple comparison adjustment, only one of the four subject and grade combinations had instances of significant mode residuals for the state/jurisdiction comparisons and no subject and grade combination had instances of significant mode residuals for the district comparisons.

Taking a less conservative approach and evaluating jurisdiction-by-jurisdiction, a total of 316 significance tests were conducted across subjects and grades (as summarized in this chapter), and 8 percent were statistically significant. In only two cases was a state/jurisdiction or district significant in more than one subject and grade combination. Specifically, Georgia and Hawaii had significantly better DBA scale scores for both subjects at grade 8.

Mode residuals were mostly inconsistent across subjects and grades. For both states/jurisdictions and districts, the correlation between mode residuals across subjects or across grades was nonsignificant in three of the four subject/grade pairings. The corresponding scatterplots did not suggest any noticeable patterns across the four subject and grade combinations, again suggesting that the mode residuals were randomly distributed across the states/jurisdictions and districts.

Finally, examining the trend implications of the digital transition found that the majority of states had consistent trend results regardless of whether the trend was based on the DBA scores or the PBA scores. Because the statistical tests employed have some possibility of false positive and false negative results,

some degree of inconsistency is expected between the trend results derived from the two samples even if there were no actual differences. Consequently, focusing only on the cases where the mode residuals were significant provides a more direct test where differences in trend reporting due to the mode transition may be more reasonably expected. When focusing only on instances where the mode residuals were significant, only a small number of cases had different trend outcomes: eight for states/jurisdictions (table 5.3b) and two for districts (table 5.3d). Furthermore, only one state/jurisdiction and one TUDA district had a more negative reporting trend outcome than adjusted trend outcome.

6. Evaluation of the Mode Transition on Subgroup Estimates

This chapter explores the third and final research question described in chapter 1—after linking the scale scores, what was the impact of the mode transition on the subgroup estimates? Included in this chapter are comparisons on the student subgroup performance between PBA and DBA.

The first section is an analysis of *differential item functioning* (DIF) between PBA and DBA. The purpose of DIF analysis is to flag items that should be evaluated for possible bias across subgroups (AERA, APA, & NCME, 2014).[7] Here, the digitally based and paper-based assessments are compared on the number of items flagged for DIF.

The remainder of the chapter is an analysis of mode residuals for subgroups. As defined in chapter 5, mode residuals are the difference between a DBA statistic and a PBA statistic on the reporting metric, *after* the linking transformation. Mode residuals are expected to be close to zero, except when resulting from estimation error, if minimal construct differences are introduced by the digital transition.

In section 6.2, we consider whether the DBA scale scores and PBA scale scores align across the entire proficiency range at every percentile of performance. This analysis tests whether students may have been advantaged or disadvantaged by the digital transition based on their proficiency level, for example, whether low performing students and certain student groups may have been disadvantaged.

Next in section 6.3, we report the mode residuals for the major reporting subgroups, as well as students with more or less computer access and familiarity at the national level. Finally, we look within states/jurisdictions and TUDA districts at mode residuals for the major reporting subgroups and students with more or less computer access and familiarity.

6.1 Differential Item Functioning

During operational analyses of NAEP data, items are evaluated for potential biases across major reporting subgroups using differential item functioning (DIF) methods to decide if the items will be included in the following score estimation process. An item is flagged for DIF when the item response data suggest that individuals that have the same underlying ability have different probabilities of getting a correct response on the item depending on a group membership (e.g., race/ethnicity). As sample size allows, the items are usually tested for DIF across the comparisons of male versus female, White versus Black, White versus Hispanic, White versus Asian, and White versus American Indian.

Items flagged for DIF are not necessarily biased or unfair (AERA, APA, & NCME, 2014; Zieky, 1993). Following recommended best practices in educational testing, all items flagged for DIF are reviewed by a diverse expert panel to bring a variety of perspectives and background experiences to understand whether the item is unfair. More detail is available on the NAEP Technical Documentation on the Web (TDW; NCES, 2000-2018).

While the DIF procedures are part of the standard analysis procedure for all NAEP assessments, the procedures were especially relevant to the digital transition as to evaluate whether the digital instrument measures the same knowledge and skills as required by the NAEP frameworks. Under the DIF context, this evaluation can be approached from two directions. First, the introduction of the digital

[7] For more information on DIF analyses, see https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_checks_dif_proced.aspx

testing environment and the various digital-only tools available for some items (e.g., drag-and-drop) could potentially introduce construct-irrelevant performance differences across student subgroups on the item level, which may show up as more DBA items being flagged by the DIF analysis. To evaluate the hypothesis of whether there was a greater number of DBA items flagged for DIF, the frequency of DIF-flagged items on DBA was compared to the frequency of DIF flagged items in PBA, based on the usual NAEP DIF analyses conducted on the gender and ethnicity subgroups. Second, if certain items were more technically demanding because of special demands such as typing, students less familiar with computers or other digital devices may have been disadvantaged on those items. Again, a DIF procedure could be applied to help identify such items. To evaluate this in DBA, DIF was tested across a special comparison of students who reported having a desktop, laptop, or tablet that they could use in their home versus students who reported they did not have such a device[8] (COMPACC). Approximately 90% of public school students in the nation reported having access to such devices at home.

Table 6.1a shows the counts of items flagged for DIF in DBA and PBA for the three[9] main DIF comparisons (male – female, White – Black, and White – Hispanic) as well as for the student-reported computer access variable. As Table 6.1a indicates, most of the items, either in DBA or PBA, were not flagged for the considered DIF comparisons.

For each subject and grade combination, the number of DIF flagged items was greater in PBA than DBA. Reading grade 4 had no flagged items in DBA versus one in PBA. Reading grade 8 had one flagged item in DBA versus five in PBA. Mathematics grade 4 had two in DBA, compared with five in PBA. Finally, mathematics grade 8 had no flagged items in DBA, while there were three in PBA. No item was classified as biased for any subgroup after review by a diverse expert panel.

For the comparison of students who reported having a computer at home versus students who reported not having a desktop, laptop, or tablet at home, there were no items flagged for DIF in any of the four subject and grade combinations.

Table 6.1a. Counts of items flagged for DIF across all items in the digitally based and paper-based assessments: 2017

	DBA				PBA		
Subject and grade	MF	WB	WH	COMPACC	MF	WB	WH
Reading grade 4	0	0	0	0	0	0	1
Reading grade 8	1	0	0	0	1	2	2
Mathematics grade 4	1	0	1	0	2	1	1
Mathematics grade 8	0	0	0	0	1	2	0

NOTE: DBA = digitally based assessment. PBA = paper-based assessment. MF = male – female comparison. WB = White – Black comparison, WH = White – Hispanic comparison. COMPACC = Student-reported computer access at home, versus no computer access at home comparison.

[8] See, for example, items 9e and 9f in https://nces.ed.gov/nationsreportcard/subject/about/pdf/bgq/student/2017_sq_student_read_g4.pdf

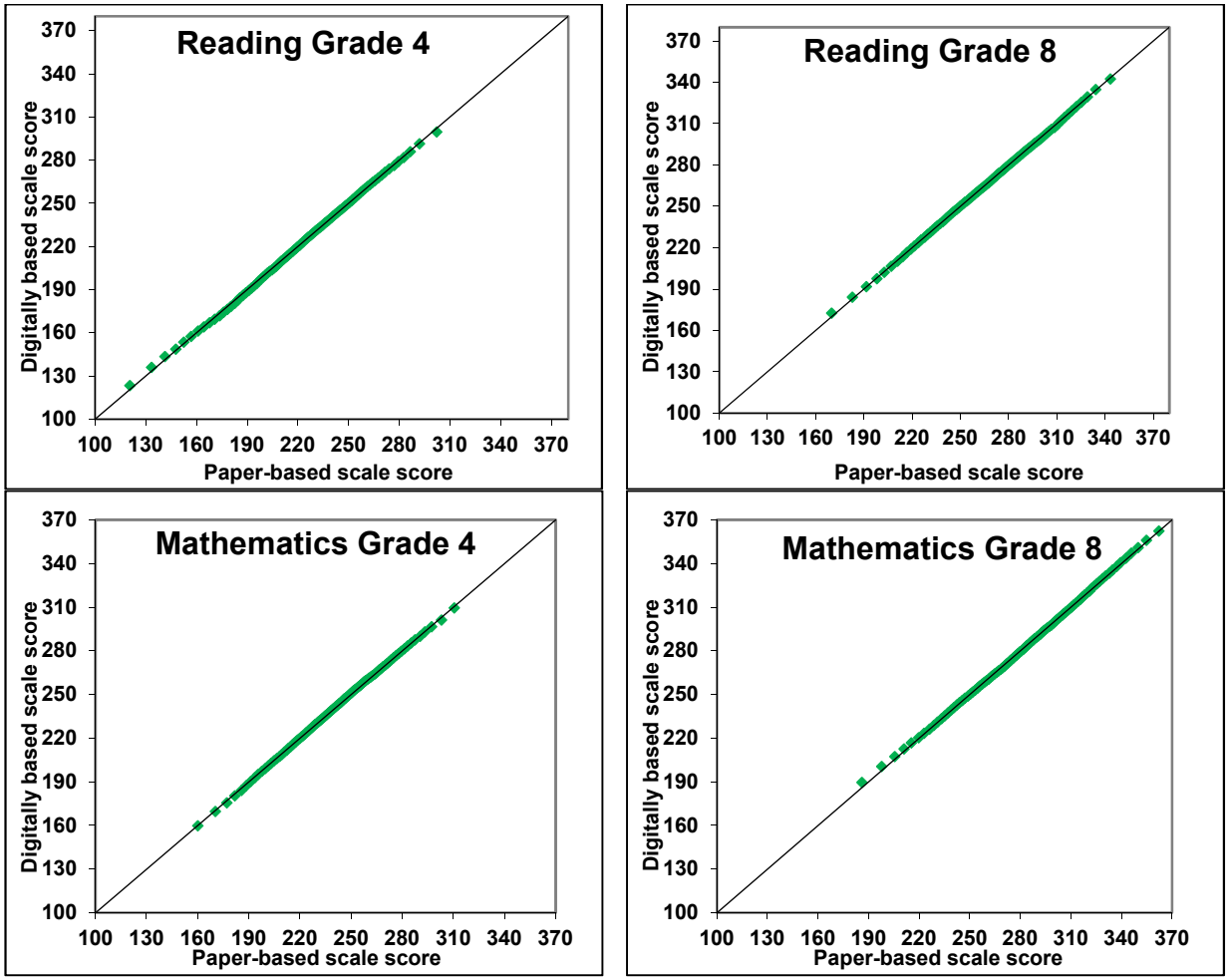
[9] The White versus Asian and White versus American Indian comparisons were not included due to the focal groups' sample sizes being not large enough on the PBA side to provide stable analysis results.

6.2 Mode Residuals Across the Proficiency Range

The alignment of the PBA and DBA scale scores across the proficiency range was evaluated with the use of Q-Q (quantile-quantile) plots. The Q-Q plot is a graphical tool for visually comparing the shape of two distributions. Following the standard NAEP procedure, the scale score estimate at every corresponding percentile from the PBA and DBA scale scores was graphed.

The Q-Q plots for each subject and grade combination for the composite scale are shown in the panels of figure 6.2a below. The line in the plots is the DBA=PBA reference line where the PBA and DBA estimates are equal. Figure 6.2a suggests that the PBA and DBA proficiency distributions line up very closely across the entire proficiency range. This has two major implications. First, the use of linear transformation (as opposed to a higher-order nonlinear transformation such as an equipercentile method) was supported. Second, little evidence was found that portions of the proficiency range were differentially impacted by the mode transition. For example, a reasonable concern may be that lower-performing students or student groups may be more disadvantaged, or advantaged, by the digital instrument than higher-performing students. However, both lower and higher percentiles were associated with similar scale scores across PBA and DBA.

Figure 6.2a. Paper-based versus digitally based scale scores at every percentile: 2017



6.3 Mode Residuals Across the Major Student Groups

Table 6.3a lists the mean scale score difference between composite scale scores based on PBA and DBA for the major reporting student groups and student-reported computer access at home (based on the COMPACC variable). While computer access is not a major reporting group, the variable was included in this analysis as a likely source of difference in performance on the digitally based and paper-based assessments.

The corresponding standard errors are given in parentheses. These major reporting student groups are defined by the five main contextual variables NAEP is federally mandated to measure: race/ethnicity, gender, socioeconomic status (i.e., student eligibility for the National School Lunch Program [NSLP]), disability status, and English language learner status (Elementary and Secondary Education Act reauthorized in 2001). The computer access variable, although not a major reporting variable in NAEP, was included as a way to measure a source of difference between DBA and PBA performance based on whether or not students reported having a computer at home.

To test for significant differences across modes after linking, the standard NAEP procedure was used involving *t*-tests at a significance level of .05, and the Benjamini-Hochberg procedure where all categories within a variable (e.g., male and female within gender) form a family, as comprehensively described in the NAEP Technical Documentation on the Web (NCES, 2000–2018).

Table 6.3a. Major reporting student groups and computer access mode residuals: 2017

	Reading		Mathematics	
Subgroup	Grade 4	Grade 8	Grade 4	Grade 8
White	-0.9 (0.6)	-1.4 (0.6)	-0.1 (0.6)	-0.1 (0.6)
Black	0.2 (0.9)	1.5 (1.0)	-0.5 (0.9)	0.6 (1.0)
Hispanic	0.9 (0.9)	1.7 (0.9)	0.0 (0.8)	0.0 (1.1)
Male	0.5 (0.7)	0.7 (0.6)	-0.2 (0.6)	0.5 (0.7)
Female	-0.5 (0.6)	-0.7 (0.6)	0.2 (0.5)	-0.5 (0.7)
NSLP	0.1 (0.6)	1.1 (0.6)	0.0 (0.5)	0.0 (0.8)
No NSLP	0.2 (0.6)	-0.5 (0.6)	0.6 (0.6)	0.1 (0.7)
SD	-0.5 (1.5)	-0.5 (1.2)	-3.6 (1.0)*	1.8 (1.2)
Non-SD	0.0 (0.5)	0.1 (0.5)	0.4 (0.5)	-0.2 (0.6)
ELL	2.5 (1.5)	3.6 (1.8)	0.0 (1.0)	0.6 (1.8)
Non-ELL	-0.3 (0.5)	-0.3 (0.5)	-0.1 (0.5)	0.1 (0.6)
Computer access	1.1 (0.5)*	0.2 (0.5)	0.9 (0.5)	0.3 (0.6)
No computer access	-1.3 (2.1)	-0.1 (1.9)	0.4 (1.5)	2.5 (1.9)

* Significantly different from zero ($p < .05$).

NOTE: SD = students with an Individualized Education Program or on a Section 504 Plan. ELL = English language learner. NSLP = students eligible for the National School Lunch Program. Students with no information available about their status in the National School Lunch Program were not included in either the NSLP or No NSLP categories. Standard errors in parentheses.

As defined in chapter 5, mode residuals are the difference between a DBA statistic and a PBA statistic on the reporting metric, *after* the linking transformation. A positive mode residual indicates that the DBA statistic is higher than the corresponding PBA statistic, and vice versa. Table 6.3a shows that in three of the four subject and grade combinations, no significant mode residuals across the major reporting student groups were observed. For grade 4 mathematics, the SD student group (students with an Individualized Education Program or on a Section 504 Plan) was the only major reporting student group with a significant mode residual. The grade 4 SD student group's mathematics scale score estimate was significantly higher by 3.6 points when based on PBA.

The last row of table 6.3a shows mode residuals for the categories of the computer access variable COMPACC. Both categories of the computer access variable show no significant difference between modes in three of the four subject and grade combinations. For grade 4 reading, the mean scale score estimate for the group of students reporting access to at least one type of digital device at home was significantly higher when based on DBA by 1.1 points.

6.4 Subgroup Mode Residuals Aggregated to the State/Jurisdiction and TUDA District Levels

While few mode residuals were significant, the remaining text of this subsection explores potential causes of the observed mode residuals and examines whether states/jurisdictions were disadvantaged as a function of demographics (e.g., minorities, wealth) or computer access. To consider these questions, available contextual information on student demographics and computer access was aggregated to the state/jurisdiction level, and the relationships between these state-level contextual variables and the state/jurisdiction mode residual within each subject and grade combination were estimated. All contextual variables aggregated to the state/jurisdiction level were dichotomous, meaning that the state-aggregated variable represents the proportion of students in the relevant category within each state (e.g., proportion non-White). The state/jurisdiction mode residuals represent the mean scale score difference between DBA and PBA, after linking, for the corresponding state/jurisdiction.

Table 6.4a shows correlations between the mode residuals for states/jurisdictions and major state/jurisdiction demographics. A positive correlation indicates that as the proportion of the considered demographic subgroup increases, the state/jurisdiction mode residuals tend to increase (i.e., the difference between DBA and PBA becomes more positive), and a negative correlation indicates the opposite. The percentage of non-White, NSLP, ELL, and SD students within each state/jurisdiction did not show a statistically significant correlation with the mode residuals in any subject and grade combination. For grade 8, the percentage of students within states attending schools in suburban areas was the only demographic variable significantly correlated with state/jurisdiction mode residuals in more than one subject and grade combination, while three of the four urban locale categories were significantly correlated with the mode residuals for reading grade 4. In both grades for reading, states/jurisdictions with a greater proportion of students in suburban schools had higher scale scores on DBA. In contrast, states/jurisdictions with a greater proportion of students in town and rural schools had higher scale scores on PBA for reading grade 4.

Table 6.4a. Correlations between state/jurisdiction mode residuals and demographics: 2017

Grade and subject	Non-White	SD	ELL	NSLP	City	Suburb	Town	Rural
Grade 4 reading	.26 (.14)	-.15 (.14)	.17 (.14)	.08 (.14)	.23 (.14)	.33 (.13)*	-.32 (.13)*	-.47 (.12)*
Grade 8 reading	.27 (.14)	.05 (.14)	.16 (.14)	.00 (.14)	-.09 (.14)	.30 (.13)*	-.17 (.14)	-.21 (.14)
Grade 4 mathematics	.14 (.14)	-.14 (.14)	.13 (.14)	.09 (.14)	.16 (.14)	.08 (.14)	-.13 (.14)	-.19 (.14)
Grade 8 mathematics	.15 (.14)	-.15 (.14)	-.10 (.14)	.06 (.14)	.09 (.14)	-.04 (.14)	.04 (.14)	-.08 (.14)

* Significantly different from zero ($p < .05$).
NOTE: $N = 52$, except for NSLP, where $N = 51$ (excluding DoDEA schools, where NSLP information is not available). SD = students with an Individualized Education Program or on a Section 504 Plan. ELL = English language learners. NSLP = students eligible for the National School Lunch Program. The demographic variables are the percentage of students within each state/jurisdiction identified as belonging to that demographic group. Standard errors in parentheses.

Table 6.4b shows correlations between the mode residuals for TUDA districts and major district demographics. The percentage of non-White, SD, ELL, and NSLP students within each district did not significantly correlate with the mode residuals for any subject and grade combination. As all schools within many of the district samples are exclusively within cities, the urban locale columns from the equivalent state/jurisdiction table (table 6.4a) are not shown here.

Table 6.4b. Correlations between TUDA district mode residuals and demographics: 2017

Grade and subject	Non-White	SD	ELL	NSLP
Grade 4 reading	-.08 (.20)	.10 (.20)	-.03 (.20)	-.14 (.20)
Grade 8 reading	-.03 (.20)	.23 (.19)	.19 (.20)	-.17 (.20)
Grade 4 mathematics	.22 (.20)	.09 (.20)	-.15 (.20)	.20 (.20)
Grade 8 mathematics	.31 (.19)	.09 (.20)	-.05 (.20)	.31 (.19)

NOTE: $N = 27$. SD = students with an Individualized Education Program or on a Section 504 Plan. ELL = English language learners. NSLP = students eligible for the National School Lunch Program. The demographic variables are the percentage of students within each TUDA district identified as belonging to that demographic group. Standard errors in parentheses.

Tables 6.4d and 6.4e show correlations between the mode residuals for states/jurisdictions (table 6.4d) or TUDA districts (table 6.4e) and contextual data collected by NAEP based on teacher and school responses to survey questions about within-school tablet access. A total of four survey questions are considered and aggregated to the state level. The percentage of students that have tablets available for their use within classrooms was based on a question given to teachers. The percentage of students from schools with fewer than ten students per tablet, the percentage of students from schools with an average tablet age of less than 4 years old, and the percentage of students who have tablets in all classrooms within their school, were based on responses from school administrators. Missing data were excluded when calculating the percentages within each state/TUDA district. Tablet access and familiarity were considered separately from other digital devices because DBAs were given on tablets. To provide context for the correlations, table 6.4c provides the overall percentage of public school students in the nation for each variable.

Table 6.4c. Percentage of public school students in the nation by teacher- and school-reported tablet familiarity variables: 2017

Grade and subject	Tablet available for student use	<=10 stud/tablet	Avg tablet <4 years old	Tablets in all classrooms
Grade 4 reading	57%	50%	78%	19%
Grade 8 reading	45%	33%	77%	7%
Grade 4 mathematics	57%	50%	78%	19%
Grade 8 mathematics	45%	33%	77%	7%

At the state/jurisdiction level, significant correlations were only observed with the grade 4 mathematics mode residual. There, greater tablet availability within schools and fewer students per tablet were related to higher performance on PBA, which is indicated by negative correlations between the state/jurisdiction mode residuals and the corresponding variables.

Table 6.4d. Correlations between state/jurisdiction mode residuals and teacher- and school-reported tablet familiarity variables: 2017

Grade and subject	Tablet available for student use	<=10 stud/tablet	Avg table <4 years old	Tablets in all classrooms
Grade 4 reading	-.23 (.14)	-.11 (.14)	.08 (.14)	-.15 (.14)
Grade 8 reading	-.25 (.14)	-.01 (.14)	.27 (.14)	-.04 (.14)
Grade 4 mathematics	-.34 (.13)*	-.33 (.13)*	.24 (.14)	-.23 (.14)
Grade 8 mathematics	.02 (.14)	.11 (.14)	.24 (.14)	.07 (.14)

* Significantly different from zero ($p < .05$).

NOTE: $N = 52$. *Tablet available for student use* is the percentage of students within each state/jurisdiction who have tablets available for student use within classrooms. *<= 10 stud/tablet* is the percentage of students within each state/jurisdiction who go to a school with less than 10 students per tablet. *Avg tablet <4 years old* is the percentage of students within each state/jurisdiction who go to a school with tablets of average age less than 4 years old. *Tablets in all classrooms* is the percentage of students within each state/jurisdiction who have tablets in all classrooms within their school. Standard errors in parentheses.

Among TUDA districts, significant correlations were observed only between the percentage of students who have tablets available for their use within classrooms and the reading grade 4 mode residuals, where greater tablet availability was related to higher performance on PBA.

Table 6.4e. Correlations between TUDA district mode residuals and teacher- and school-reported tablet familiarity variables: 2017

Grade and subject	Tablet available for student use	<=10 stud/tablet	Avg tablet <4 years old	Tablets in all classrooms
Grade 4 reading	-.40 (.18)*	.14 (.20)	-.11 (.20)	.03 (.20)
Grade 8 reading	-.37 (.19)	.08 (.20)	-.07 (.20)	.10 (.20)
Grade 4 mathematics	-.19 (.20)	-.01 (.20)	-.13 (.20)	-.09 (.20)
Grade 8 mathematics	-.21 (.20)	-.28 (.19)	.32 (.19)	.00 (.20)

* Significantly different from zero ($p < .05$).

NOTE: $N = 27$. *Tablet available for student use* is the percentage of students within each district who have tablets available for student use within classrooms. *<= 10 stud/tablet* is the percentage of students within each district who go to a school with less than 10 students per tablet. *Avg tablet <4 years old* is the percentage of students within each district who go to a school with tablets of average age less than 4 years old. *Tablets in all classrooms* is the percentage of students within each district who have tablets in all classrooms. Standard errors in parentheses.

Tables 6.4f and 6.4g show correlations between the mode residuals for states/jurisdictions (table 6.4f) or TUDA districts (table 6.4g) and contextual data collected in NAEP survey questionnaires relating to non-tablet digital device access. Four of the aggregated state-level variables, percentage of students from schools with no more than 3 students per laptop, percentage of students from schools with no more than 3 students per device (laptop or desktop), percentage of students from schools with an average desktop age of less than 4 years old, and percentage of students from schools with an average laptop age of less than 4 years old were based on responses from school administrators. The fifth variable, percentage of students reporting more frequent use of computers for the relevant subject in classes, was based on responses from students. Missing data were excluded when calculating the percentages within each state/TUDA district.

At the state/jurisdiction level (table 6.4f), the only significant correlation observed was between the percentage of students reporting more frequent use of computers for the relevant subject in classes and the reading grade 4 mode residuals, where more frequent computer use was associated with higher performance on DBA.

Table 6.4f. Correlations between state/jurisdiction mode residuals and other computer access and familiarity variables: 2017

Grade and subject	<=3 stud/laptop	<=3 stud/device	Avg desktop <4 years old	Avg laptop <4 years old	Computer use more frequent
Grade 4 reading	.13 (.14)	.11 (.14)	.06 (.14)	.13 (.14)	.39 (.14)*
Grade 8 reading	.13 (.14)	.16 (.14)	-.03 (.14)	.21 (.14)	.22 (.15)
Grade 4 mathematics	.09 (.14)	.01 (.14)	.00 (.14)	-.01 (.14)	-.01 (.15)
Grade 8 mathematics	.05 (.14)	.09 (.14)	.07 (.14)	-.14 (.14)	-.18 (.15)

* Significantly different from zero ($p < .05$).

NOTE: $N = 52$ except for *Computer use more frequent*, where $N = 47$, due to five states opting out of the 2017 subject-specific student survey questionnaire. *<= 3 stud/laptop* is the percentage of students within each state/jurisdiction who go to a school with less than 3 students per laptop. *<= 3 stud/device* is the percentage of students within each state/jurisdiction who go to a school with less than 3 students per device (laptop or desktop). *Avg desktop <4 years old* is the percentage of students within each state/jurisdiction who go to a school with desktops of average age less than 4 years old. *Avg laptop <4 years old* is the percentage of students within each state/jurisdiction who go to a school with laptops of average age less than 4 years old. For mathematics, *Computer use more frequent* is the percentage of students who report use of a computer or other digital device once per week or more for mathematics at school. For reading, *Computer use more frequent* is the percentage of students who report use of a computer or other digital device 30 minutes or more for ELA schoolwork and homework. Standard errors in parentheses.

Among TUDA districts (table 6.4g), the only significant correlations observed were between the percentage of students from schools with no greater than 3 students per device (laptop or desktop) and the reading grade 4 mode residuals, and the percentage of students from schools with an average desktop age of less than 4 years old and the grade 8 mathematics mode residuals. Higher performance on DBA was observed for both conditions.

Table 6.4g. Correlations between TUDA district mode residuals and other computer access and familiarity variables: 2017

Grade and subject	<=3 stud/laptop	<=3 stud/device	Avg desktop <4 years old	Avg laptop <4 years old	Computer use more frequent
Grade 4 reading	.32 (.19)	.42 (.18)*	-.30 (.19)	.18 (.20)	.18 (.23)
Grade 8 reading	.35 (.19)	.33 (.19)	-.01 (.20)	.32 (.19)	.15 (.23)
Grade 4 mathematics	.15 (.20)	.33 (.19)	-.21 (.20)	-.08 (.20)	.09 (.23)
Grade 8 mathematics	.13 (.20)	-.02 (.20)	.22 (.19)	.38 (.18)*	.34 (.22)

* Significantly different from zero ($p < .05$).

NOTE: $N = 27$, except for *Computer use more frequent*, where $N = 26$, due to one district opting out of the 2017 subject-specific student survey questionnaire. *<= 3 stud/laptop* is the percentage of students within each district who go to a school with less than 3 students per laptop. *<= 3 stud/device* is the percentage of students within each district who go to a school with less than 3 students per device (laptop or desktop). *Avg desktop <4 years old* is the percentage of students within each district who go to a school with desktops of average age less than 4 years old. *Avg laptop <4 years old* is the percentage of students within each district who go to a school with laptops of average age less than 4 years old. For mathematics, *Computer use more frequent* is the percentage of students who report use of a computer or other digital device once per week or more for mathematics at school. For reading, *Computer use more frequent* is the percentage of students who report use of a computer or other digital device 30 minutes or more for ELA schoolwork and homework. Standard errors in parentheses.

Discussion

This chapter examined the issue of the mode transition on subgroup scale scores in multiple ways. In nearly all cases, the comparisons did not find evidence that the DBA scale scores were significantly different from the PBA scale scores.

In every grade and subject combination, the digital instrument showed less evidence of DIF than the paper-based instrument. Notably, very few items in either assessment were flagged for DIF and no item was classified as biased for any subgroup after review by a diverse expert panel.

After the linking transformation, the scale scores estimated from DBA and PBA fell on the line of equivalence. Little evidence was found that low proficiency students were more impacted than higher proficiency students.

Across the major reporting groups, no student group had a significant mode residual in three of the four subject and grade combinations. In the remaining subject and grade combination (i.e., grade 4 mathematics), only one subgroup (i.e., SD students) had a significant mode residual. For the computer access variable, a significant mode residual was found only for one subject and grade combination (i.e., grade 4 reading).

Finally, when the mode residuals, demographics, and technology-related contextual variables were aggregated to the state/jurisdiction and district levels, the majority of correlations with the mode residuals were non-significant. No demographic measure, except school urban locale, was significantly correlated with state/jurisdiction or district mode residuals—in reading, some significant correlations between school urban locale and the mode residuals were observed for states/jurisdictions. For technology-related contextual variables, some significant correlations between computer access and the state/jurisdiction and district mode residuals were observed. However, significant correlations were in

inconsistent directions. Specifically, of the 72 tested correlations (see tables 6.4c to 6.4f), 6 were significant. Three were in the direction that greater computer access related to better PBA scale score estimates, and three were in the direction that greater computer access related to better DBA scale score estimates.

Taken as a whole, the analyses show little evidence of any disadvantage or advantage for student subgroups, states/jurisdictions or TUDA districts from the transition to the digital format. There were few significant mode residuals for major student groups, and differences in proportions of students by demographics or access to technology were largely uncorrelated with state/jurisdiction and district scores. This suggests that the mode residuals are random with little noticeable patterns across the demographics and digital access variables considered.

7. Discussion

In some ways, the NAEP 2017 transition to digitally based assessment (DBA) resembled several challenges NAEP has faced in the past. In reference to the first NAEP bridge study in 1984, Beaton (1990a, p. 5) wrote,

> There is a clear tension between the need to maintain constant measurement procedures in order to estimate changes in performance and the desire to continue to improve the assessment by using the most modern, best available technology. The new design introduced in 1984 responded to this tension by assessing student achievement in two ways: in one set of samples using the methods of past assessments and in another set using the best available methodology. The samples using the methods of the past were called "bridge" samples, since they provided bridges to the performance of students in past assessments. The result was parallel assessments, using different technologies, that could be compared and for some purposes, perhaps, equated [or linked]. In this way, innovations could be introduced without losing comparability with the past.

In other ways, the digital transition presented novel challenges, affecting not just one aspect of the NAEP administration but virtually all aspects, and was the most substantial change to the instruments and administration that NAEP has undergone since its inception. Correspondingly, the digital transition study was the most comprehensive bridge study NAEP has conducted to date.

NAEP began studying a transition to DBA around the turn of the century (as later published in Bennett et al., 2007, 2008; Sandene et al., 2005), and conducted field trials of the digitally based mathematics and reading assessments in 2015. The NAEP 2017 transition studies described in this paper replicated the prior findings. Notably, the studies found that mathematics and reading items generally became more difficult in a digital format, but otherwise retained similar item properties such as their difficulty relative to other items. In other words, the increased difficulty was mostly consistent across items; however, there was a notable exception of the constructed-response items at grade 4 which tended to become more difficult by a larger amount than did selected-response items.

After linking the DBA scale scores to the PBA scale scores, the DBA scale scores were not significantly different from the PBA scale scores for all major student groups for three of the four subject and grade combinations. For mathematics grade 4, the scale scores for PBA and DBA were significantly different for students with disabilities, where performance in DBA was lower than in PBA.

In terms of differential item functioning, DBA showed less evidence of subgroup DIF than PBA. Examining performance with reference to students' self-reported access to digital devices produced inconsistent results. While access to digital devices was sometimes significantly related to relative performance on DBA, the direction was not consistent and the strength of the relationship observed was quite modest.

For states/jurisdictions, three of the four subject and grade combinations showed no significant evidence of inconsistency between the linked DBA scale scores and the PBA scale scores when using a significance testing procedure that controlled for multiple comparisons. One of four subject and grade combinations (i.e., grade 8 reading), two states had significantly different scale scores after controlling for multiple comparisons. Of these two states for grade 8 reading, one had a more positive outcome on

the reported trend (a positive reported trend but a nonsignificant adjusted trend) and the other had the same outcome on both the reported and adjusted trends (nonsignificant trend difference). Differences in the DBA scale scores and the PBA scale scores were small in terms of the effect size, and the differences for each state/jurisdiction were mostly inconsistent between subjects and between grades.

For districts, all four subject and grade combinations showed no significant evidence of inconsistency between the linked DBA scale scores and the PBA scale scores when using a significance test that controlled for multiple comparisons. Similarly, differences in the DBA scale scores and the PBA scale scores were small in terms of effect sizes, and the differences for each district were mostly inconsistent between subjects and between grades.

The NAEP 2017 digital transition bridge study was the largest and most comprehensive evaluation of the impact of paper-based versus digitally based assessments among NAEP studies to date. While some previous studies did not find a significant difference in difficulty between the DBA and PBA instruments (e.g., Wang et al., 2007), the smaller sample sizes of those studies may not have had the statistical power to detect the difficulty difference between the two testing modes. The finding in the more powerful 2017 NAEP study that the PBA and DBA instruments were not equivalent in difficulty suggests that linking approaches that do not assume items have common properties across the two assessment modes, such as common population linking, are more appropriate for digital transitions of PBAs.

Studies evaluating the impact of the DBA versus PBA instruments on proficiency estimates do not always use a consistent methodology. Relatedly, DBAs are likely to involve administrative changes, so studies from one assessment may not generalize to others that do not entail the same administrative changes (e.g., the timing difference between the two modes, whether schools are providing the required digital devices, etc.). Therefore, generalizing results between different assessments should take into account differences in the instruments, administrations, and methodology of the assessments.

For example, Bennett and colleagues (2008) examined the impact when items across DBA and PBA were erroneously assumed to have equivalent properties, including difficulty. They found differences in estimated proficiency when no linking procedure was applied that accounts for differences in digital and paper instruments, and the differences were highly consistent across student groups. Other studies involving linking methods that differ from the method NAEP uses, also found differences in estimated proficiency between samples, but generally the differences were highly consistent across student groups (e.g., Backes & Cowan, 2018).

In contrast, the research summarized in chapters 5 and 6 examines scale score estimate differences between PBA and DBA after using common population linking to align scale scores from the two assessments. Common population linking does not require the assumption that items have equivalent properties across modes, thereby accounting for the greater difficulty of the digital instruments, as well as the greater spread of performance on the digital instruments. Furthermore, the bridge study design used in this paper accounts for the finding that not all items were impacted in the same manner by the conversion to digital format: for example, the design accounts for the observation that constructed-response items had a greater increase in difficulty than selected-response items. However, the transformation as part of the common population linking does not guarantee scores will be aligned for all states/jurisdictions, districts, and subgroups. Consequently, the alignment of scores for states/jurisdictions, districts, and subgroups were carefully evaluated in chapters 5 and 6.

The authors and NCES are confident in the scientific rigor of the approach taken to design the NAEP digital transition and to conduct the analyses that led to the results. Overall, based on the analyses conducted, no clear evidence of consistent bias in the linked results was detected. Therefore, NAEP's extremely valuable trend lines can be preserved through this digital transition.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Backes, B., & Cowan, J. (2018). *Is the pen mightier than the keyboard? The effect of online testing on measured student achievement* (CALDER Working Paper No. 190). Washington, DC: CALDER, American Institutes for Research.

Beaton, A. E. (1990a). Introduction. In E. G. Johnson and R. Zwick (Eds.), *Focusing the new design: the NAEP 1988 technical report* (Report No. 19-TR-20, pp. 3–9). Princeton, NJ: Educational Testing Service.

Beaton, A. E. (1990b). Introduction. In A. E. Beaton and R. Zwick (Eds.), *The effect of changes in the national assessment: Disentangling the NAEP 1985-86 reading anomaly* (Report No. 17-TR-21, pp. 1–13). Princeton, NJ: Educational Testing Service.

Beller, M. (2013). Technologies in Large-scale Assessments: New Directions, Challenges, and Opportunities. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 25–45). Netherlands: Springer.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289–300.

Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). Problem solving in technology-rich environments: A report from the NAEP technology-based assessment project (NCES 2007–466). Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, *6*, 1–38.

Bock R. D., & Zimowski M. F. (1997). Multiple Group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433-448). New York, NY: Springer-Verlag.

Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis*. London: Chapman & Hall.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Common Sense Media. (2013). Teaching with technology: The promise and reality of the U.S. digital classroom. Retrieved from https://www.commonsense.org/education/sites/default/files/tlr-blog/teachingwithtechnology-thepromiseandrealityoftheusdigitalclassroom.pdf.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales*. New York, NY: Springer.

Eignor, D. R. (2007). Linking scores derived under different modes of test administration. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 135-159). New York, NY: Springer.

Haertel, E. H. (2016). *Future of NAEP long-term trend assessments* (Paper commissioned for the National Assessment Governing Board). Retrieved from https://www.nagb.gov/content/nagb/assets/documents/newsroom/naep-releases/naep-long-term-trend-symposium/long-term-trends.pdf.

IEA. (2017). *TIMSS & PIRLS – TIMSS 2019.* Retrieved from https://timssandpirls.bc.edu/timss2019/.

IEAa. (n.d.). *IEA International Computer and Information Literacy Study 2018: How well are students prepared for study, work, and life in the digital age?* Retrieved from https://www.iea.nl/fileadmin/user_upload/Studies/ICILS_2018/IEA_ICILS_2018_Leaflet.pdf.

IEAb. (n.d.). *ICCS: Join ICCS 2022!*. Retrieved from https://iccs.iea.nl/cycles/join-2022.html.

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement*, *29*, 95-110.

Johnson, E. G., & Rust, K. F. (1992). Population inference and variance estimation for NAEP. *Journal of Educational Statistics*, *17*, 175-190.

Johnson, E. G., & Zwick, R. (1990). *Focusing the new design: the NAEP 1988 technical report* (Report No. 19-TR-20, pp. 3-9). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mazzeo, J., Donoghue, J. R., Liu, B., & Xu, X. (2018). *Estimating standard errors for NAEP that Incorporate random-groups linking error for the transition from paper-based to digital-based assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Mislevy, R. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359-381.

Mislevy, R. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, *80*, 993-997.

Mislevy, R., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133-161.

Mislevy, R., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 131-154.

Mood, A., Graybill, F., & Boes, D. (1974). *Introduction to the theory of statistics* (3rd ed.). New York, NY: McGraw-Hill Companies.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *ePIRLS 2016 international results in online informational reading*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/pirls2016/international-results/.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

National Assessment Governing Board. (2015). *Resolution on maintaining NAEP trends with the transition to digital-based assessments (DBA)*. Retrieved from https://www.nagb.gov/content/nagb/assets/documents/policies/resolution-on-trend-and-dba.pdf.

National Assessment Governing Board. (2017a). *Reading framework for the 2017 National Assessment of Educational Progress*. Retrieved from https://www.nagb.gov/content/nagb/assets/documents/publications/frameworks/reading/2017-reading-framework.pdf.

National Assessment Governing Board (2017b). *Mathematics framework for the 2017 National Assessment of Educational Progress*. Retrieved from https://www.nagb.gov/content/nagb/assets/documents/publications/frameworks/mathematics/2017-math-framework.pdf.

National Center for Education Statistics. (2000-2018). *NAEP technical documentation on the web*. Retrieved from https://nces.ed.gov/nationsreportcard/tdw/.

National Center for Education Statistics. (2009). *The nation's report card: Reading 2009* (NCES 2010–458). Institute of Education Sciences, U.S. Department of Education, Washington, DC. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/main2009/2010458.pdf.

National Center for Education Statistics. (2012). *The nation's report card: Writing 2011* (NCES 2012–470). Institute of Education Sciences, U.S. Department of Education, Washington, DC. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf.

National Center for Education Statistics. (2016). *The nation's report card: Technology and engineering literacy (TEL) 2014*. Institute of Education Sciences, U.S. Department of Education, Washington, DC. Retrieved from https://www.nationsreportcard.gov/tel_2014/.

National Center for Education Statistics. (2017). *Facts for parents: About the eTIMSS 2017 pilot*. Retrieved from https://nces.ed.gov/timss/parents/pdf/etimss_2017_parent_faq.pdf.

National Center for Education Statistics. (2018). *NAEP mathematics 2017 report card: Sample questions*. Institute of Education Sciences, U.S. Department of Education, Washington, DC. Retrieved from https://www.nationsreportcard.gov/math_2017/#sample-questions.

National Center for Education Statistics. (2018). *Students' computer access and use*. National Center for Education Statistics. Washington, DC. Retrieved from www.nationsreportcard.gov/sq_computer/.

NCESa. (n.d.). *International computer and information literacy study.* Retrieved from https://nces.ed.gov/surveys/icils/.

NCESb. (n.d.). *International computer and information literacy study – ICILS participating countries.* Retrieved from https://nces.ed.gov/surveys/icils/countries.asp.

NCESc. (n.d.). *Trends in international mathematics and science study (TIMSS) – TIMSS participating countries*. Retrieved from https://nces.ed.gov/timss/countries.asp.

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics, 51*, 1352-1375.

OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en.

OECD. (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems (Volume V).* PISA. OECD Publishing. http://dx.doi.org/10.1787/9789264208070-en.

OECDa. (n.d.). *PISA 2015 Integrated design*. Retrieved from https://www.oecd.org/pisa/pisaproducts/2015-Integrated-Design.pdf.

OECDb. (n.d.). *Beyond PISA 2015: A longer-term strategy of PISA*. Retrieved from https://www.oecd.org/pisa/pisaproducts/Longer-term-strategy-of-PISA.pdf.

Pearson. (2015). *Student mobile device survey 2015, National report: Students in grades 4–12.* Retrieved from https://www.pearsoned.com/wp-content/uploads/2015-Pearson-Student-Mobile-Device-Survey-Grades-4-12.pdf.

Perie, M., Moran, R., & Lutkus, A. D. (2005). *NAEP 2004 trends in academic progress: Three decades of student performance in reading and mathematics* (NCES 2005–464). Washington, DC: U.S. Government Printing Office.

Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project* (NCES 2005-457). Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457.

SRI International. (2018). *An overview of current digital curricula and instructional tools in mathematics, reading, and writing* (Unpublished commissioned paper). Prepared by Everson, H., Kim, H. J., Yarnall, L., and Ziker, C. under contract ED-IES-13-C-0017. Menlo Park, CA: SRI International.

Tennessee Department of Education. (n.d.). *2017–18 assessment FAQ*. Retrieved from https://www.tn.gov/education/assessment/tnready/assessment-faq.html.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1039–1055). Amsterdam: Elsevier.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K–12 mathematics tests. *Educational and Psychological Measurement*, *67*, 219-238.

Wyoming Department of Education. (n.d.). *WY-TOPP accommodations FAQ*. Retrieved from https://edu.wyoming.gov/downloads/assessments/2018/WY-TOPP-Accommodations-Guide.pdf.

Wyoming Department of Education. (August 25, 2018). *FAQ WY-TOPP for parents*. Retrieved from https://edu.wyoming.gov/downloads/communications/2017/Wy-TOPP-FAQ-Parents.pdf.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics*, *17*, 155-173.

Zieky, M. (1993). Practical questions in the use of DIF statistics. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, *10*, 10-16.

Zwick, R. (1992). Chapter 7: Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. Journal of Educational Statistics, 17, 205-218.