# 2019 NAEP Transition to DBA and Mode Evaluation for the Mathematics and Reading Assessments at Grade 12

The NAEP mathematics and reading assessments at grade 12 were last administered in 2015 as paper-based assessments (PBAs) to national samples. In 2019, these two NAEP assessments were transitioned to digitally based assessments (DBAs), with 2019 being the first year for operational DBAs for grade 12. As with the grades 4 and 8 mathematics and reading DBA transition in 2017 and grade 8 social sciences DBA transition in 2018, bridge studies were designed and implemented for evaluating the effects of the change in administration mode from paper-and-pencil to digital. Bridge studies document and evaluate how trends on the core NAEP scales may be interpreted in reference to previously reported PBA results.

## Bridge study design

For each subject, the bridge study incorporated two components, i.e., a PBA component and a DBA component. For the PBA component, the 2019 paper instrument was exactly the same as that used in 2015 (but with updated survey questionnaire items), making direct comparisons of PBA results between 2019 and 2015 possible. On the other hand, the digital instrument largely drew upon the existing "legacy" item pool content established for PBA but represented these items (referred to as trans-adapted items) on tablet devices. The digital instruments in both math and reading also included several blocks of items that were specifically developed for DBA. Based on previous digital transition experience, the trans-adapted DBA items were not expected to function exactly the same as their paper-version counterparts and therefore could not be linked to the existing paper scales through NAEP's usual common item linking approach. Instead, the DBA to PBA linking process relied on the random equivalency between the two samples taking the corresponding instrument, or the *common population assumption*. In this linking process, the bridge PBA component served three purposes: 1) to link the DBA component results to the existing scale through common population linking; 2) to evaluate the validity and fairness of the linking results across the range of student proficiency for major subgroups; and 3- to serve as part of the 2019 reporting sample.

## Analysis procedures

### Common population linking

Typically, NAEP relies on the common item linking method to place the proficiency estimates from the current assessment to the trend line. The current assessment would share between 70 and 80 percent of the items with the previous assessment. By assuming these common items would maintain their psychometric properties across assessments, a two-group concurrent IRT calibration is used to scale all the items while holding the IRT parameters of the common items equal between the two assessments.

However, it was not appropriate to assume that the trans-adapted items would function exactly the same between DBAs and PBAs. Previous research on psychological and educational assessments has shown that it is difficult to achieve equivalence in a digital transition as two different presentation and response modes are being used (Bennett et al., 2008). The 2015 DBA transition field trial and the 2017 DBA transition on mathematics and reading at grades 4 and 8 added empirical evidence to that the trans-adapted digital items appeared more difficult than their paper parent counterparts on average while both were administered to randomly equivalent groups. In addition, the actual difference in mean

item score differed by subject and by grade. Thus, the DBA results were bridged to the existing trend line through the common population linking method. Sampled students were randomly assigned to take either mode to ensure that the DBA sample and the PBA sample would be randomly equivalent to one another. Demographic composition of the two samples were carefully compared and the results indicated strong comparability between the two samples.

To facilitate the common population linking, data collected from the DBA component and the PBA component were analyzed separately. Through the usual NAEP procedure of common item linking, the 2019 PBA scores were placed onto the NAEP reporting scale. The mean and standard deviation of the 2019 DBA scores were then set to those of the 2019 PBA scores through common population linking.

### Error variance estimation

Similar to the 2017 NAEP mathematics and reading digital transition at grades 4 and 8 and the 2018 NAEP social sciences transition, placing the 2018 social science DBA scores onto the existing trend line through common population linking required calculating an additional source of error variance associated with the linking transformation (i.e., "linking variance"), in addition to the usual error variances due to sampling and measurement error. The total jackknife procedure that was developed and used during the 2018 NAEP social sciences transition to account for the linking variance was also used in the 2019 reading and mathematics digital transition at grade 12.

## Impact of the transition on item-level properties

To evaluate the impact of the paper-to-digital transition on the item-level properties, multiple item-level statistics from both a classical test theory (CTT) and item response theory (IRT) framework were compared. Because students taking the DBA and PBA were randomly equivalent samples selected from a common population, any difference observed on the statistics that were compared reflected differences in the instrument and sampling error rather than population differences. Below, the comparison of the mean item score[1] between the paper and digital formats is shown for the trans-adapted items in both the reading and mathematics assessments at grade 12. For both the reading and mathematics assessments at the composite and subscale levels, a negative mode difference was observed, indicating that on average the trans-adapted DBA assessment items were more difficult than their PBA counterparts.

[1] 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 1. Overall weighted mean item score comparison between digitally based assessment (DBA) and paper-based assessment (PBA) for the grade 12 mathematics and reading composite scales and the corresponding subscales: 2019

| Subject and content area | Number of Items | 2019 DBA | 2019 PBA | DBA-PBA (SE) |
|---|---|---|---|---|
| Mathematics | 113 | 41.1% | 43.7% | -2.6 (0.29)* |
| Number properties and operations | 14 | 50.0% | 52.5% | -2.5 (0.45)* |
| Measurement and geometry | 36 | 35.6% | 38.4% | -2.8 (0.35)* |
| Data analysis, statistics, and probability | 27 | 44.5% | 48.2% | -3.7 (0.37)* |
| Algebra | 36 | 41.8% | 43.3% | -1.4 (0.39)* |
| Reading | 110 | 57.2% | 59.1% | -1.9 (0.33)* |
| Literary | 39 | 58.4% | 61.0% | -2.6 (0.46)* |
| Informational | 71 | 56.5% | 58.0% | -1.6 (0.38)* |

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

Table 1 compares the overall mean item score averaged across the trans-adapted items within each subject and the corresponding mode difference for the 2019 mathematics and reading assessments at grade 12. The difference between the two mean item scores is also listed under a separate column named "DBA-PBA (SE)", with the standard error (SE) of the difference enclosed in the parentheses. Results followed by an asterisk (*) under the "DBA-PBA (SE)" column indicate that the difference is significantly different from zero. Table 1 also lists the comparisons for each content area under mathematics and reading at grade 12.

## Evaluation of the mode transition at grade 12 on subgroup estimates

The 2019 mathematics and reading DBA and PBA components were analyzed separately following the standard NAEP operational analysis procedures. The DBA and PBA results were compared at various analysis steps to determine to what extent the two operational components function similarly at the national level. After the 2019 PBA results were placed onto the reporting scale, the mean and standard deviation of the DBA results were made equal to those of the PBA scale scores, using the transformation procedure described above under Common population linking. The next evaluation step was to see whether this mean-SD transformation could effectively and successfully adjust the mode difference across the entire proficiency range and whether there were any meaningful mode-by-subgroup interactions.

The alignment of the DBA and PBA scale scores across the proficiency range was evaluated with the use of quantile-quantile plots (i.e., Q-Q plots). The Q-Q plot is a graphical tool for visually comparing the shapes of two distributions. The scale score estimate at every corresponding percentile from the PBA and DBA scale scores was graphed to compare the distributions of the PBA and DBA scale scores. For both reading and mathematics at each subscale, the DBA and PBA scale scores showed close alignment.

Mode-by-subgroup interactions were evaluated by calculating the mode residuals, or mean composite scale score differences, between DBA and PBA. Table 2 lists these mode residuals for the main reporting

subgroups with the corresponding standard errors given in the parentheses. These main reporting subgroups are defined by the five main contextual variables NAEP is federally mandated to measure: gender, race/ethnicity, student disability, English learner status, and socioeconomic status (No Child Left Behind Act of 2001, 2002).

Table 2. Mode residuals for major reporting subgroups at grade 12 in mathematics and reading: 2019

| Subgroup | Mathematics | Reading |
|---|---|---|
| **Male** | 0.7 (0.8) | -1.0 (0.9) |
| **Female** | -0.7 (0.6) | 1.0 (0.7) |
| **White** | -0.1 (0.6) | 0.8 (1.0) |
| **Black** | 0.5 (1.2) | 0.8 (1.4) |
| **Hispanic** | 0.0 (1.0) | -1.5 (1.3) |
| **Asian** | -0.1 (2.3) | -1.4 (2.2) |
| **American Indian/Alaska Native** | -5.4 (3.0) | 0.5 (4.6) |
| **SD** | 3.5 (1.8) | -2.4 (2.3) |
| **Non-SD** | -0.3 (0.6) | 0.2 (0.6) |
| **EL** | 0.8 (2.2) | -4.6 (2.1) |
| **Non-EL** | -0.1 (0.5) | 0.3 (0.6) |
| **Eligible for NSLP** | 0.5 (0.7) | -1.2 (0.8) |
| **Not eligible for NSLP** | -0.3 (0.7) | 0.5 (0.9) |

* Significantly different from zero ($p < .05$).
NOTE: SD = students with an Individualized Education Program or on a Section 504 Plan. EL = English learner. NSLP = students eligible for 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. The standard error variance for mode residual is the sum of two components: sampling variance and measurement variance. The sampling variance accounts for dependency between the PBA and DBA samples. The measurement variance is the sum of measurement variances for the DBA and PBA subgroup averages, respectively.

Table 2 shows that for mathematics and reading, no significant mode residuals were detected for any of the considered major reporting subgroups.

Taking the score distribution comparison and the subgroup performance comparison into consideration, the evidence supported strong comparability between the DBA and PBA scale scores. The analyses showed little evidence of any disadvantage for student subgroups from the transition to the digital format.

## Summary

In 2019, the NAEP mathematics and reading assessments at grade 12 transitioned from paper-based assessments (PBA) to digitally based assessments (DBA). Following the example of the digital transitions of both the 2017 reading and mathematics assessments at grades 4 and 8 and the 2018 social sciences

assessments at grade 8, the analysis of the mathematics and reading assessments at grade 12 included a mode evaluation study to examine the impact of the transition and provide evidence to support the continuation of trend reporting. To ensure the feasibility of the proposed linking methodology, the DBA and PBA instruments were administered to randomly equivalent samples of students drawn from a common population. The PBA results were placed onto the trend line through usual common item linking by concurrently calibrating the 2019 and 2015 PBA data, while the DBA results were put onto the existing trend line by lining up the mean and SD of the DBA scores to those of the PBA scale scores.

After linking the DBA results to the PBA scales, the differences between the DBA scale scores and PBA scale scores were not statistically significant for any major reporting subgroups for either mathematics or reading. The QQ plots between the DBA quartiles and PBA quartiles confirmed the consistency between the DBA and PBA scale score results for both. The results of the mode evaluation study supported the decision to report on the combined DBA and PBA results.

Reference

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.