**2019 NAEP Transition to DBA and Mode Evaluation for the Science Assessments at Grades 4, 8, and 12**

The NAEP science assessments at grades 4, 8, and 12 were last administered in 2015 as paper-based assessments (PBAs) to national and state samples, for grades 4 and 8, and to a national sample at grade 12. In 2019, the NAEP science assessments were transitioned to digitally based assessments (DBAs) for all three grades. As with the grade 12 mathematics and reading DBA transition in 2019, the grades 4 and 8 mathematics and reading DBA transition in 2017, and the 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. The white paper by Jewsbury et al. (2020) details the purpose, technical methodology, and results of the mode evaluations conducted as part of the 2017 NAEP reading and math transitions to DBA. The bridge study for 2019 science was designed and analyzed in a fashion very similar to that of the 2017 reading and math bridge studies.

**Bridge study design**

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 presented these items (referred to as trans-adapted items) on tablet devices. The digital instruments also included several blocks of items that were specifically developed for DBA. Additionally, the 2019 DBA assessments at all three grades included new block types, called scenario-based task (SBT) blocks. 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 were not used in linking to the existing paper assessment-based 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 onto the trend line. The current assessment would share between 75 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 for science added empirical evidence so that, for grades 4 and 8, the trans-adapted digital items appeared more difficult than their paper parent counterparts on average while both were administered to randomly equivalent groups. Thus, the results from the 2019 DBAs were bridged to the existing trend line through the common population linking method. Within each grade, sampled students were randomly assigned to take either mode. Demographic composition of the two samples was 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**

As with subjects that have transitioned from PBA to DBA, placing the DBA scores onto the existing trend line through common population linking required calculating an additional source of error variance associated with the linking transformation (linking variance). Linking variance is calculated 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 science digital transition, as in all other previous NAEP transitions.

**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 an 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 item-level statistics that were compared reflected differences in the instrument and sampling error rather than population differences. Table 1 compares the overall mean item score averaged across the trans-adapted items within each grade between the paper and digital formats.[[1]](#footnote-2) The difference between the two mean item scores is also listed under a separate column named “DBA-PBA,” with the standard error (SE) of the difference in the last column. Results followed by an asterisk under the “DBA-PBA” column indicate that the difference is significantly different from zero. At grades 4 and 8 at the overall 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. This same pattern is observed at grade 12 for only the Earth and Space Sciences subscale. For the remaining subscales and at the overall level, there was no significant mode effect.

Table 1. Overall weighted mean item score comparison between digitally based assessment (DBA) and paper-based assessment (PBA) for the overall science scale and subscales by grade: 2019

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Grade** | **Content area** | **Number of Items** | **2019 DBA** | **2019 PBA** | **DBA-PBA** | **SE** |
| **4** | Science Overall | 123 | 47.8 | 50.1 |  **-2.3\*** | 0.25 |
| Physical Science | 40 | 39.4 | 42.3 |  **-2.8\*** | 0.29 |
| Earth and Space Sciences |  | 50.0 | 51.8 |  **-1.9\*** | 0.31 |
| 37 |
| Life Science | 46 | 52.9 | 55.0 |  **-2.1\*** | 0.28 |
| **8** | Science Overall | 149 | 44.4 | 45.5 |  **-1.1\*** | 0.26 |
| Physical Science | 49 | 43.5 | 45.0 |  **-1.6\*** | 0.30 |
| Earth and Space Sciences | 58 | 41.9 | 42.7 |  **-0.9\*** | 0.29 |
| Life Science | 42 | 48.5 | 49.4 |  **-0.9\*** | 0.32 |
| **12** | Science Overall | 167 | 39.9 | 39.9 | **0.0** | 0.27 |
| Physical Science | 60 | 39.8 | 39.8 | **-0.1** | 0.27 |
| Earth and Space Sciences | 37 | 36.3 | 37.1 |  **-0.8\*** | 0.36 |
| Life Science | 70 | 42.3 | 41.8 | **0.5** | 0.31 |
| \* Significantly different from zero (*p* < .05). |  |  |

Note: SE = standard error.

**Evaluation of the science mode transition on subgroup estimates**

The 2019 science 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 earlier under common population linking. The next evaluation step was to consider 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 (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 all three grades at the overall level and for 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 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 in science by grade: 2019[[2]](#footnote-3)

|  |  |  |  |
| --- | --- | --- | --- |
| **Subgroup** |  **Grade 4** |  **Grade 8** |  **Grade 12** |
| **Male** |  0.3 (0.6) |  -1.7 (0.5)\* |  -0.1 (0.7) |
| **Female** | -0.3 (0.6) |  1.8 (0.5)\* |  0.1 (0.7) |
| **White** | -1.1 (0.5) |  -0.9 (0.5) |  -0.7 (0.5) |
| **Black** |  1.6 (1.3) |  -0.3 (1.2) |  -1.8 (1.4) |
| **Hispanic** |  1.3 (0.8) |  1.4 (0.8) |  2.6 (0.8)\* |
| **Asian** |  1.4 (2.2) |  2.3 (2.0) |  1.1 (2.6) |
| **American Indian/Alaska Native** |  0.3 (3.4) |  2.7 (3.9) |  1.1 (4.7) |
| **SD** | -4.8 (1.2)\* |  -3.6 (1.2)\* |  -1.0 (1.9) |
| **Non-SD** |  0.8 (0.3)\* |  0.6 (0.3)\* |  0.1 (0.3) |
| **EL** |  1.6 (1.3) |  1.9 (1.6) |  7.2 (2.4)\* |
| **Non-EL** | -0.2 (0.3) |  -0.1 (0.3) |  -0.2 (0.2) |
| **Eligible for NSLP** | -0.5 (0.6) |  0.4 (0.6) |  0.8 (0.7) |
| **Not eligible for NSLP** | -0.5 (0.6) |  0.2 (0.5) |  -1.0 (0.6) |
| \* Significantly different from zero (*p* < .05). |

NOTE: SD = Students identified as students with disabilities. EL = English learners. NSLP = National School Lunch Program. Students with no information available about their status in the National School Lunch Program were not included in either the eligible for NSLP or not eligible for NSLP categories. Standard errors are in parentheses.

Across all three grades, the overall pattern of subgroup results yielded few instances of statistically significant differences in mode effects. One exception pertained to students identified as students with disabilities (SD). SD students were found to show a negative mode effect at grades 4 and 8. Students who were not identified as students with disabilities (Non-SD) showed a positive mode effect at grades 4 and 8.

The intended reporting sample for 2019 was the combined 2019 DBA and PBA samples (referred to as combined sample). Trend inferences between 2019 and 2015 were evaluated among three different samples: 2019 PBA, 2019 DBA, and the 2019 combined sample. As discussed earlier, the 2019 PBA results were placed onto the reporting scale through common item linking, and the 2019 DBA results were put onto the same reporting scale through common population linking by aligning the mean and standard deviation of the 2019 DBA results with those of the 2019 PBA scale scores. The same common population linking approach was used in linking the 2019 combined sample results to the same reporting scale. Table 3 contains trend results for each of the three grades and for each of the three 2019 samples. The overall trend inferences between 2019 and 2015 on subgroups are consistent when using the three 2019 samples, with few exceptions. Note that SD students at grade 4 performed significantly lower in 2019 PBA than in 2015. This demonstrates that, even without transitioning to DBA, SD students at grade 4 show a decline in performance over time. That decline over time is magnified in value using the 2019 DBA sample or the 2019 combined sample. There was a significant mode effect for SD students at grade 8 as well. The trend for this group of students would show a significant increase in performance if the 2019 PBA sample were used. That significant gain over time is reduced to a nonsignificant increase when the 2019 combined sample is used.

Table 3. Trend results for major reporting subgroups for all grades in science for three possible 2019 reporting samples

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Grade 4** | **Grade 8** | **Grade 12** |
|   | **Across years (2019 – 2015)** | **Across years (2019 – 2015)** | **Across years (2019 – 2015)** |
| **Group** | **PBA trend (SE)** | **DBA trend (SE)a** | **Combined trend** | **PBA trend (SE)** | **DBA trend (SE)a** | **Combined trend** | **PBA trend (SE)** | **DBA trend (SE)a** | **Combined trend** |
|  **(SE)a** |  **(SE)a** |  **(SE)a** |
| **Overall** |  -2.4(0.6)\* |  -2.4(0.6)\* |  -2.4(0.6)\* | -0.2(0.6) | -0.2(0.6) | -0.2(0.6) | -0.7(0.7) | -0.7(0.7) | -0.7(0.7) |
| **Male** |  -2.9(0.8)\* |  -2.6(0.8)\* |  -2.7(0.7)\* | -0.1(0.7) |  -1.9(0.7)\* | -1.3(0.7) | -1.3(1.0) | -1.4(0.9) | -1.4(0.9) |
| **Female** |  -1.9(0.8)\* |  -2.3(0.7)\* |  -2.1(0.7)\* | -0.3(0.8) |  1.5(0.7)\* | 0.9(0.7) | 0.0(1.0) | 0.1(0.9) | 0.1(0.9) |
| **White** |  -2.0(0.8)\* |  -3.1(0.7)\* |  -2.7(0.7)\* | -0.3(0.7) | -1.1(0.6) | -0.8(0.6) | 0.6(1.0) | -0.1(0.9) | 0.1(0.9) |
| **Black** |  -3.9(1.4)\* | -2.2(1.1) |  -2.8(1.0)\* | 1.2(1.1) | 0.9(1.1) | 1.0(0.9) | 1.8(1.9) | -0.1(1.7) | 0.6(1.7) |
| **Hispanic** | -1.5(1.1) | -0.2(1.0) | -0.6(1.0) | 0.1(1.0) | 1.5(1.0) | 1.1(1.0) | -2.0(1.3) | 0.5(1.2) | -0.4(1.1) |
| **Asian** | -1.7(3.4) | -0.3(2.7) | -0.7(2.8) | 0.4(2.1) | 2.7(1.6) | 1.9(1.5) | -1.7(3.3) | -0.5(3.4) | -0.9(3.1) |
| **American Indian/Alaska Native** | 2.0(3.7) | 2.3(3.0) | 2.2(2.8) | 2.9(3.5) | 5.6(3.2) | 4.7(2.7) | 6.3(6.7) | 7.4(5.9) | 7.0(5.8) |
| **SD** |  -4.3(1.4)\* |  -9.1(1.2)\* |  -7.5(1.1)\* |  3.5(1.3)\* | -0.1(1.2) | 1.2(1.1) | -1.4(2.5) | -2.5(2.2) | -2.1(2.1) |
| **Non-SD** |  -1.8(0.6)\* | -1.0(0.6) | -1.3(0.6) | -0.3(0.6) | 0.3(0.6) | 0.1(0.6) | -0.2(0.7) | -0.2(0.7) | -0.2(0.7) |
| **EL** | -0.9(1.7) | 0.7(1.4) | 0.2(1.4) | -0.5(1.9) | 1.3(1.6) | 0.7(1.6) |  -10.6(3.3)\* | -3.4(3.1) | -5.8(3.0) |
| **Non-EL** |  -2.1(0.6)\* |  -2.3(0.6)\* |  -2.2(0.6)\* | 0.3(0.6) | 0.3(0.6) | 0.3(0.6) | 0.3(0.7) | 0.0(0.7) | 0.1(0.7) |
| **Eligible for NSLP** |  -2.7(0.8)\* |  -3.2(0.7)\* |  -3.0(0.7)\* | 0.2(0.7) | 0.6(0.7) | 0.5(0.7) | 0.2(1.2) | 1.0(1.1) | 0.7(1.1) |
| **Not eligible for NSLP** |  -2.3(0.9)\* |  -2.8(0.8)\* |  -2.6(0.8)\* | -1.3(0.7) | -1.1(0.7) | -1.1(0.6) | 0.5(0.9) | -0.5(0.9) | -0.2(0.8) |

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

a Includes linking error component.

NOTE: SD = Students identified as students with disabilities. EL = English learners. NSLP = National School Lunch Program. Students with no information available about their status in the National School Lunch Program were not included in either the eligible for NSLP or not eligible for NSLP categories. Standard errors are in parentheses.

**Summary**

In 2019, the NAEP science assessment at all three grades transitioned from paper-based assessments (PBA) to digitally based assessments (DBA). Following the example of the digital transitions of the 2017 reading and mathematics assessments at grades 4 and 8, the 2018 social sciences assessments at grade 8, and the 2019 reading and mathematics assessments at grade 12, the analysis of the science assessment 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 standard deviation of the DBA scores with 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 most major reporting subgroups. The Q-Q plots between the DBA quantiles and PBA quantiles confirmed the consistency between the DBA and PBA scale score results for both. The results of the mode evaluation study (within-year and trend inferences) supported the decision to report on the 2019 NAEP results using the combined DBA and PBA sample.

**References**

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.

Jewsbury, P., Finnegan, R., Xi, N., Jia, Y., Rust, K., & Burg, S. (2020). 2017 NAEP transition to digitally based assessments in mathematics and reading at grades 4 and 8: Mode evaluation study [White paper]. <https://nces.ed.gov/nationsreportcard/subject/publications/main2020/pdf/transitional_whitepaper.pdf>

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. [↑](#footnote-ref-2)
2. 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. [↑](#footnote-ref-3)