**2018 NAEP Transition to DBA and Model Evaluation for the U.S. History, Geography, and Civics Assessments at Grade 8**

The NAEP social science assessments (U.S. history, geography, and civics) were last administered in 2014 as paper-based assessments (PBAs) to national samples of grade 8 students only. In 2018, the three NAEP social science subjects were transitioned to digitally based assessments (DBAs), with 2018 being the first operational social science DBAs for grade 8. As with the reading and mathematics DBA transition in 2017, bridge studies were designed and implemented for evaluating potential effects of the change in administration mode from paper-and-pencil to digital, and to understand how trends on the core NAEP scales in U.S. history, geography, and civics may be interpreted in reference to the PBA results. In support of these goals, along with the operational DBAs, the 2014 PBA instruments in these three subjects were re-administered in 2018 at the national level.

# Bridge study design

For each social science subject, the bridge study incorporated two components: a PBA component and a DBA component. For the PBA component, the 2018 paper instrument was the same as that used in 2014 (but with updated survey questionnaire items), making direct, error-free comparisons of PBA results between 2018 and 2014 possible. On the other hand, the digital instrument largely made use of the existing “legacy” item pool established for PBA but represented these items on tablet devices (referred to as trans-adapted items). The digital instrument also included a small portion 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 the usual common item linking approach. Instead, the DBA to PBA linking process relied on the common population assumption (i.e., the random equivalency between the two samples taking the corresponding instrument). 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 2018 reporting sample.

# Analysis procedures

### Common population linking

Typically, NAEP analyses employ 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 to be equal between the two assessments. However, it was not appropriate to assume 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 of reading and mathematics at grades 4 and 8 added empirical evidence to the previous research as the trans-adapted digital items appeared more difficult in comparison to their paper parent counterparts while both were administered to randomly equivalent groups. Thus, the results from the 2018 social science DBAs were bridged to the existing trend line through the common population linking method. Within each subject, sampled students were randomly assigned to take either mode. 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 2018 PBA scores were placed onto the NAEP reporting scales for each subject. The mean and standard deviation (SD) of the 2018 DBA scores were then set to those of the 2018 PBA scores through common population linking.

### Error variance estimation

Similar to the 2017 NAEP reading and mathematics digital transition, 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 (i.e., “linking variance”). Linking variance is calculated in addition to the usual error variances due to sampling and measurement error. The total jackknife procedure was developed for the 2018 social sciences to account for the additional error variance, or linking error variance, in the total error variance.

# 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 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 in the “DBA-PBA” column, 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.

Table 1. Overall weighted mean item score comparison between digitally based assessment (DBA) and paper-based assessment (PBA) for the 2018 social sciences subjects for the overall scale and corresponding subscales.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Number of Items** | **2018 DBA** | **2018 PBA** | **DBA-PBA** | **SE** |
| **U.S. History** | 132 | 44.2% | 44.5% | -0.3  | 0.34 |
| **Democracy** | 30 | 42.5% | 43.3% | -0.8 | 0.42 |
| **Cultures** | 42 | 43.6% | 43.1% | 0.6  | 0.42 |
| **Technology** | 33 | 45.7% | 47.3% | -1.6\* | 0.45 |
| **World Role** | 27 | 46.0% | 45.3% | 0.7 | 0.43 |
| **Geography** | 100 | 47.3% | 46.2% | 1.1\* | 0.35 |
| **Space & Place** | 44 | 54.0% | 53.9% | 0.0 | 0.44 |
| **Environment & Society** | 26 | 41.9% | 38.8% | 3.2\* | 0.41 |
| **Spatial Dynamics** | 30 | 44.9% | 45.2% | -0.3 | 0.45 |
| **Civics** | 128 | 49.0% | 49.9% | -0.9 \* | 0.41 |

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

The mode difference based on the overall mean item score statistics is not consistent across the three subjects. As Table 1 shows, among the 132 U.S. history items considered, on average the DBA sample answered 44.2% of the digital content correctly, while the PBA sample answered an average of 44.5% of the paper content correctly, and the difference between the two modes (0.3%) is not significantly different from zero. For geography and civics, this difference is 1.1% and -0.9% respectively, which are significantly different from zero but in opposite directions. For the subscales under U.S. history and geography, the comparison results are also not consistent. Although the overall U.S. history average percent correct difference is insignificant, the subscale “Economic and Technological Changes and Their Relationship to Society, Ideas, and the Environment” (Technology) has a negative mode difference that is significantly different from zero. The other three U.S. history subscales have both positive and negative differences. For geography, the only subscale having a significant mode difference is “Environment and Society”, and the average percent correct difference is greater than 3%, which seems to explain why geography has an overall significant mode difference, as the other two geography subscales have close to zero mode differences. Note that although the NAEP civics framework specifies five content areas, civics is treated as a univariate subject in the operational analysis and therefore the subscale information is not listed in this and the following comparisons.

# Evaluation of the mode transition on subgroup estimates

The 2018 social sciences DBA and PBA components for each subject 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 functioned similarly at the national level. After the 2018 PBA results were placed onto the reporting scales for each subject, 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 (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 all three subjects, 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 the parentheses. These main reporting subgroups are defined by the five main contextual variables the NAEP program 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

|  |  |  |  |
| --- | --- | --- | --- |
| **Subgroup** | **U.S. History** | **Geography** | **Civics** |
| **Male** | 0.3 (0.9) | 0.7 (0.9) | 0.0 (1.0) |
| **Female** | -0.3 (0.8) | -0.7 (0.9) | 0.0 (1.0) |
| **White** | -0.6 (0.8) | 1.0 (1.1) | -2.0 (1.1) |
| **Black** | 2.5 (1.5) | -2.4 (1.6) | 0.0 (2.0) |
| **Hispanic** | -0.1 (1.1) | 0.1 (1.1) | 3.0 (1.3) |
| **Asian** | 1.1 (2.3) | -3.7 (3.0) | 3.1 (2.6) |
| **American Indian** | 3.3 (5.6) | -5.5 (9.3) | 2.3 (7.7) |
| **SD** | -2.7 (1.8) | -1.4 (1.8) | 0.4 (2.1) |
| **Non-SD** | 0.3 (0.6) | 0.1 (0.7) | -0.1 (0.9) |
| **ELL** | -1.9 (2.1) | -3.0 (2.4) | 4.9 (2.1)\* |
| **Non-EL** | 0.1 (0.6) | 0.3 (0.8) | -0.4 (0.9) |
| **Eligible for NSLP** | 0.1 (0.7) | -0.4 (1.0) | 1.6 (1.2) |
| **Not eligible for NSLP** | 0.3 (0.9) | 0.4 (1.0) | -1.5 (1.0) |

*\*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 in parentheses.*

Table 2 shows that for U.S. history and geography, no significant mode residuals were detected for any of the considered major reporting subgroups. For civics, the English learner (EL) subgroup was the only major reporting subgroup with a significant mode residual.

The intended reporting sample for 2018 was the combined 2018 DBA and PBA samples (referred to as combined sample). Trend inferences between 2018 and 2014 were evaluated among three different samples: 2018 PBA, 2018 DBA, and the 2018 combined sample. As discussed earlier, the 2018 PBA results were placed onto the reporting scale through common item linking, and the 2018 DBA results were put onto the same reporting scale through common population linking by aligning the mean and standard deviation of the 2018 DBA results those of the 2018 PBA scale scores. The same common population linking approach was used in linking the 2018 combined sample results to the same reporting scale. Tables 3, 4 and 5 contain trend results for U.S. history, geography, and civics for each of the three 2018 samples. As these three tables show, the trend results were very consistent among the three trend lines based on different 2018 samples, with somewhat smaller standard errors (SE) for the 2018 combined sample. No difference was observed between the DBA trend and the combined trend for civics, for any of these considered subgroups. For U.S. history and geography, only three subgroups and one subgroup, respectively, had different trend results between the two trend lines, and the actual differences were usually small. Given the consistency between the DBA trend and the combined trend and the considerations listed above, the combined sample was chosen as the final reporting sample. This reporting option fully utilized the collected data and provided a more comprehensive evaluation of the student performance on social sciences to the public and secondary users.

Table 3. Three sets of trend results for major reporting subgroups, 2018 U.S. history

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Subgroup** | **Mode effect (SE)** | **PBA trend (SE)** | **DBA trend (SE)a** | **Combined trend (SE)a** |
| **Overall** | 0.0 (0.6) | -4 (1.2)\* | -4 (1.2)\* | -4 (1.2)\* |
| **Male** | 0.3 (0.9) | -5 (1.4)\* | -5 (1.5)\* | -5 (1.4)\* |
| **Female** | -0.3 (0.8) | -4 (1.2)\* | -4 (1.2)\* | -4 (1.2)\* |
| **White** | -0.6 (0.8) | -4 (1.3)\* | -5 (1.3)\* | -5 (1.2)\* |
| **Black** | 2.5 (1.5) | -6 (2.0)\* | -3 (1.8) | -5 (1.8)\* |
| **Hispanic** | -0.1 (1.1) | -4 (1.4)\* | -4 (1.3)\* | -4 (1.3)\* |
| **Asian** | 1.1 (2.3) | -5 (3.2) | -4 (3.0) | -5 (3.0) |
| **Am. Indian** | 3.3 (5.6) | --- | --- | --- |
| **SD** | -2.7 (1.8) | -2 (2.1) | -5 (2.1)\* | -4 (1.9) |
| **Non-SD** | 0.3 (0.6) | -5 (1.3)\* | -4 (1.2)\* | -4 (1.3)\* |
| **EL** | -1.9 (2.1) | -3(2.6) | -5 (2.4)\* | -4 (2.3) |
| **Non-EL** | 0.1 (0.6) | -4 (1.2)\* | -4 (1.2)\* | -4 (1.2)\* |
| **Eligible for NSLP** | 0.1 (0.7) | -4 (1.2)\* | -4 (1.2)\* | -4 (1.2)\* |
| **Not eligible for NSLP** | 0.3 (0.9) | -5 (1.4)\* | -5 (1.3)\* | -5 (1.3)\* |

\* 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 in parentheses.*

Table 4. Three sets of trend results for major reporting subgroups, 2018 geography

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Subgroup** | **Mode effect (SE)** | **PBA trend (SE)** | **DBA trend (SE)a** | **Combined trend (SE)a** |
| **Overall** | 0.0 (0.7) | -3 (1.3)\* | -3 (1.3)\* | -3 (1.3)\* |
| **Male** | 0.7 (0.9) | -4 (1.5)\* | -3 (1.4)\* | -3 (1.4)\* |
| **Female** | -0.7 (0.9) | -2 (1.4) | -3 (1.4)\* | -3 (1.3)\* |
| **White** | 1.0 (1.1) | -4 (1.5) | -3 (1.4) | -4 (1.4)\* |
| **Black** | -2.4 (1.6) | -4 (2.0) | -6 (1.9)\* | -5 (1.9)\* |
| **Hispanic** | 0.1 (1.1) | -2 (1.5) | -2 (1.5) | -2 (1.5) |
| **Asian** | -3.7 (3.0) | -3 (3.5) | -7 (3.4) | -5 (3.1) |
| **Am. Indian** | -5.5 (9.3) | --- | --- | --- |
| **SD** | -1.4 (1.8) | -2 (2.3) | -4 (2.2) | -3 (2.1) |
| **Non-SD** | 0.1 (0.7) | -3 (1.4) | -3 (1.4) | -3 (1.4) |
| **EL** | -3.0 (2.4) | -1 (3.1) | -4 (3.1) | -3 (2.9) |
| **Non-EL** | 0.3 (0.8) | -3 (1.4) | -3 (1.4) | -3 (1.4) |
| **Eligible for NSLP** | -0.4 (1.0) | -3 (1.4) | -3 (1.4) | -3 (1.3) |
| **Not eligible for NSLP** | 0.4 (1.0) | -3 (1.6) | -2 (1.5) | -3 (1.5) |

\* 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 in parentheses.*

Table 5. Three sets of trend results for major reporting subgroups, 2018 civics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Subgroup** | **Mode effect (SE)** | **PBA trend (SE)** | **DBA trend (SE)a** | **Combined trend (SE)a** |
| **Overall** | 0.0 (0.8) | -1 (1.4) | -1 (1.4) | -1 (1.4) |
| **Male** | 0.0 (1.0) | -2 (1.6) | -2 (1.6) | -2 (1.6) |
| **Female** | 0.0 (1.0) | 0 (1.5) | 0 (1.5) | 0 (1.4) |
| **White** | -2.0 (1.1) | -1 (1.5) | -3 (1.4) | -2 (1.4) |
| **Black** | 0.0 (2.0) | -2 (2.7) | -2 (2.6) | -2 (2.5) |
| **Hispanic** | 3.0 (1.3) | -1 (2.0) | 2 (1.9) | 0 (1.8) |
| **Asian** | 3.1 (2.6) | 3 (4.0) | 6 (3.7) | 5 (3.6) |
| **Am. Indian** | 2.3 (7.7) | --- | --- | --- |
| **SD** | 0.4 (2.1) | -3 (2.8) | -3 (2.6) | -3 (2.6) |
| **Non-SD** | -0.1 (0.9) | 0 (1.3) | 0 (1.3) | 0 (1.3) |
| **EL**  | 4.9 (2.1)\* | 0 (2.8) | 5 (2.8) | 2 (2.6) |
| **Non-EL** | -0.4 (0.9) | 0 (1.4) | -1 (1.4) | -1 (1.4) |
| **Eligible for NSLP** | 1.6 (1.2) | 0 (1.6) | 1 (1.6) | 0 (1.5) |
| **Not eligible for NSLP** | -1.5 (1.0) | -1 (1.4) | -2 (1.5) | -2 (1.4) |

\* 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 in parentheses.*

# Summary

In 2018, the NAEP U.S. history, geography, and civics assessments at grade 8 transitioned from PBA to DBA. Following the example of digital transitions of the 2017 reading and mathematics assessments at grades 4 and 8, the 2018 social sciences administration 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 2018 and 2014 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 DBA scale scores were not significantly different from the PBA scale scores for any major reporting subgroups for U.S. history and geography, and for civics only one subgroup (English learners) had a significant mode residual after linking. The Q-Q plots between the DBA quantiles and PBA quantiles confirmed the consistency between the DBA and PBA scale score results for all three subjects. The results of the mode evaluation study (within-year and trend inferences) supported the decision to report on the 2018 NAEP results using the combined DBA and PBA samples.

# 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.

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)