**2024 NAEP Digital Device Bridge Study for Mathematics and Reading at Grade 12**

**John R. Donoghue, Xueli Xu**

**Educational Testing Services**

Beginning in 2019, the NAEP Mathematics and Reading assessments were administered digitally on NAEP-supplied devices (Microsoft Surface Pros [SP] tables with attached keyboards). In 2024, sampled students were assessed using either a NAEP-supplied SP or a NAEP-supplied Chromebook (CB). As with the previous NAEP assessment transitions from paper-based to digitally based assessments (see Jewsbury et al., 2020), a bridge study was designed and implemented for evaluating the effects of the change in assessment device from SP to CB. Bridge studies document and evaluate how trends on the core NAEP scales relate to previously reported results. This report details the 2024 NAEP SP to CB bridge study design and device evaluation for the 2024 NAEP Mathematics and Reading grade 12 assessments.

**Bridge study design**

The 2024 digital device bridge studies were designed in accordance with the principles historically employed by NAEP (e.g., Jewsbury et al., 2020). The 2024 NAEP Mathematics and Reading grade 12 samples followed the usual NAEP school and student sampling procedure with a few additional features to accommodate the bridge study design. One key design component for the 2024 NAEP Mathematics and Reading assessments at grade 12 was that SP and CB were systematically assigned to the sampled students within selected schools in a ratio of 13:12, where practical. This design ensured comparability of the two random samples assessed using SP versus CB. Under this design, just over half of the 2024 Mathematics and Reading samples took the assessments on the SP. This would allow the continuation of trend reporting if score comparability between SP and CB devices was not supported after linking the scale scores. The administration on both devices enables evaluation of the device transition impact at the national level, and to potentially adjust the reporting metric for device transition effects. Including both devices also enables the potential of a combined SP and CB sample to be used as the basis for reporting on student performance and trends in 2024.

The bridge study was designed to (1) evaluate the comparability of item psychometric properties between devices; (2) allow for linking scores between device conditions within and across years through the common item linking approach (if feasible); (3) enable alternative score linking (i.e., common population linking) through a random groups design and account for differences in device type if a device effect was present; and (4) evaluate the validity and fairness of the linking results across the range of student proficiency for major subgroups.

The assessment was identical in terms of items, timing, and administration procedures regardless of device. However, there were several differences in technical specifications for the two device types (i.e., input modality, screen size, and aspect ratio; see Table 1).

Table 1. Device Specifications for NAEP 2024 Assessments

|  |  |  |
| --- | --- | --- |
| **Specification** | **Surface Pros (SP)** | **Chromebooks (CB)** |
| Input Modalities | Touchscreen, Trackpad, Keyboard, Stylus | Touchscreen, Trackpad, Keyboard |
| Screen size[[1]](#footnote-2) | 12" | 11.6" |
| 12.3" |
| Aspect ratio | 3:2 | 16:9 |

**Analysis procedures**

**Within-2024 Device Evaluation Summary**

To place the proficiency estimates from the 2024 assessments onto the existing trend line, the bridge study design allows for the typical NAEP procedure of common item linking in which the current assessment shares between 70 and 80 percent of the items in common with the previous assessment. By assuming these common items would generally maintain their psychometric properties across assessments in adjacent administrations, a two-group concurrent Item Response Theory (IRT) calibration is used to scale all the items while holding the IRT parameters of the common items equal between the two assessments. However, prior to the adjacent-administration common item linking, it was first necessary to determine the within-2024 linking procedure in which the CB sample was linked to the SP sample. For the Reading and Mathematics assessments at grade 12, an evaluation based on observed statistics and IRT modeling showed little practical evidence of device differences in psychometric properties.

**Error variance estimation**

When evaluating grade 12 trend between the 2024 Surface Pro and Chromebook combined sample and previous paper-based assessments (PBA) in 2015, 2013, 2009, and 2005, as well as comparisons to the most recent combined 2019 PBA + digitally-based assessment (DBA) assessment, a linking error correction term was included to compute the variance estimates for the trend comparisons. The overall error accounts for measurement error, sampling error and linking error. Measurement and sampling error variance estimates follow the NAEP conventional two-part variance estimation procedure.

**Impact of device on item-level psychometric properties**

To evaluate the impact of device type on the item-level psychometric properties, multiple item-level statistics from both a classical test theory (CTT) and an IRT framework were compared. Because students assessed on SP and CB were randomly equivalent samples selected from a common population and the identical assessments were given to both samples, differences on the item-level statistics that were compared would reflect differences due to device types and/or margin of error due to uncertainty in sampling and latency of proficiency, rather than population differences. Tables 2 and 3 compare the overall mean item score[[2]](#footnote-3) averaged across all items within each content area for Mathematics and Reading between the SP and CB device types. The difference between the two mean item scores is also listed under a separate column named “SP-CB,” with the standard error (SE) of the difference in the last column. None of the mean item differences were statistically significant.

Table 2. Overall weighted mean item score comparison between Surface Pro (SP) and Chromebook (CB) for the Mathematics composite scale and subscales: 2024

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade | Content area | Number of items | SP | CB | Difference (SP-CB) | SE |
| 12 | Mathematics overall scale | 182 | 37.8 | 37.7 | 0.1 | 0.29 |
| Numbers and Operations | 23 | 48.3 | 47.6 | 0.7 | 0.46 |
| Geometry and Measurement | 55 | 34.4 | 34.6 | -0.2 | 0.35 |
| Data Analysis and Statistics | 44 | 41.2 | 41.0 | 0.2 | 0.34 |
|  | Algebra and Functions | 60 | 34.3 | 34.3 | 0.0 | 0.35 |

Note: SE = standard error.

Table 3. Overall weighted mean item score comparison between Surface Pro (SP) and Chromebook (CB) for the Reading composite scale and subscales: 2024

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Grade | Content area | Number of items | SP | CB | Difference (SP-CB) | SE |
| 12 | Reading overall scale | 138[[3]](#footnote-4) | 55.2 | 54.9 | 0.2 | 0.37 |
| Literary | 38 | 54.1 | 53.7 | 0.4 | 0.51 |
|  | Informational | 100 | 55.6 | 55.4 | 0.1 | 0.43 |

**Evaluation of the device transition on subgroup performance estimates**

To evaluate the effect of the device transition on performance for subgroups, SP and CB results were compared at various analysis steps to determine to what extent the two components function similarly. The next evaluation steps were to consider whether the within-year common item linking approach was effective to link scores between device conditions and whether there were any meaningful device-by-subgroup interactions. The alignment of the SP and CB 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 SP and CB scale scores was graphed to compare the distributions of the SP and CB scale scores.

Mathematics

For the overall mathematics scale and for each subscale, the SP and CB scale scores showed close alignment (see Figure 1 for Q-Q plots).

A graph of a graph

AI-generated content may be incorrect.

A graph with a line

AI-generated content may be incorrect.A graph with a line

AI-generated content may be incorrect.

A graph with a line

AI-generated content may be incorrect.A graph with a line

AI-generated content may be incorrect.

Figure 1. Q-Q plots of SP and CB scale scores for the overall Mathematics scale and subscales at grade 12: 2024

Device-by-subgroup interactions were evaluated by calculating the difference in average scale scores between SP and CB, for the overall Mathematics scale. Table 4 lists these average overall scale score differences for the overall national sample and main reporting subgroups. The corresponding standard errors are in parentheses. These main reporting subgroups are defined by the five contextual variables NAEP is federally mandated to measure: gender, race/ethnicity, student disability status, English learner status, and economically disadvantaged status. None of the device differences across subgroups are statistically significant.

Table 4. Differences in subgroup averages (Surface Pro minus Chromebook) by device in Mathematics grade 12: 2024

|  |  |
| --- | --- |
|  | **Overall Mathematics Scale** |
| **Subgroup** | **Grade 12** |
| **Overall** | 0.4 (0.6) |
| **Male** | 0.1 (0.9) |
| **Female** | 0.7 (0.8) |
| **White** | 1.3 (1.0) |
| **Black** | -1.1 (1.4) |
| **Hispanic** | 1.1 (0.9) |
| **Asian** | -1.8 (2.6) |
| **American Indian/ Alaska Native** | 5.6 (5.8) |
| **SD** | 3.6 (2.0) |
| **Non-SD** | 0.0 (0.6) |
| **EL** | 0.3 (1.8) |
| **Non-EL** | 0.4 (0.6) |
| **ED** | 0.7 (0.8) |
| **Non-ED** | 0.1 (0.8) |

NOTE: SD = Students identified as students with disabilities. EL = English learners. ED = Economically Disadvantaged. Standard errors are in parentheses.

Reading

For the overall Reading scale and for each subscale, the SP and CB scale scores showed close alignment (see Figure 2 for Q-Q plots).

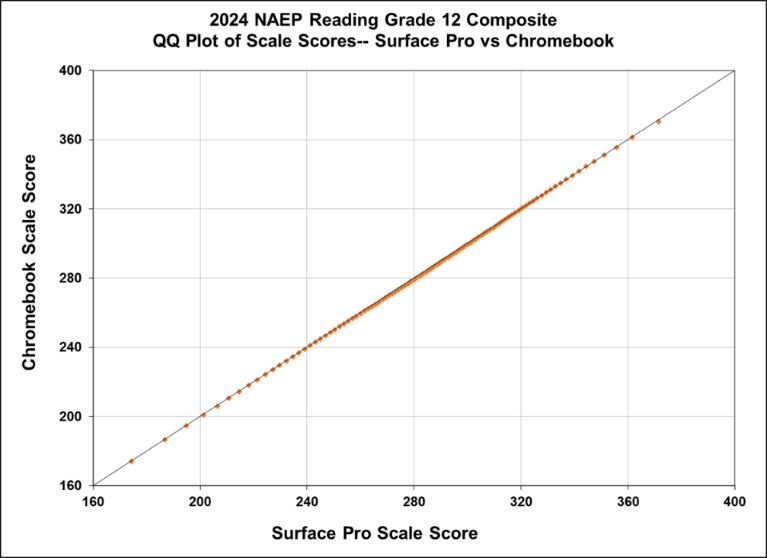


Figure 2. Q-Q plots of SP and CB scale scores for the overall Reading scale and subscales at grade 12: 2024

Device-by-subgroup interactions were evaluated by calculating the difference in average scale scores between SP and CB, for the overall Reading scale. Table 5 lists these average overall scale score differences for the overall national sample and main reporting subgroups. The corresponding standard errors are in parentheses. These main reporting subgroups are defined by the five contextual variables NAEP is federally mandated to measure: gender, race/ethnicity, student disability status, English learner status, and economically disadvantaged status. None of the device differences across subgroups are statistically significant.

Table 5. Differences in subgroup averages (Surface Pro minus Chromebook) by device in Reading grade 12, 2024

|  |  |
| --- | --- |
|  | **Overall Reading Scale** |
| **Subgroup** | **Grade 12** |
| **Overall** | 0.5 (0.7) |
| **Male** | 0.5 (1.0) |
| **Female** | 0.3 (1.0) |
| **White** | -1.0 (0.9) |
| **Black** | 2.6 (1.4) |
| **Hispanic** | 1.6 (1.1) |
| **Asian** | 1.8 (2.5) |
| **American Indian/ Alaska Native** | 1.7 (6.0) |
| **SD** | -3.6 (2.2) |
| **Non-SD** | 1.0 (0.7) |
| **EL** | 0.8 (1.9) |
| **Non-EL** | 0.5 (0.7) |
| **ED** | 1.7 (0.9) |
| **Non-ED** | 0.3 (0.8) |

NOTE: SD = Students identified as students with disabilities. EL = English learners. ED = Economically Disadvantaged. Standard errors are in parentheses.

**Summary**

In 2024, the NAEP Mathematics and Reading assessments at grade 12 were administered using Surface Pro (SP) devices and Chromebook (CB) devices. Similarly to the mode evaluations that were done in previous transitions from paper-based to digitally based assessments, a device evaluation study was conducted to examine any possible impact of the new device on performance and provide evidence to support the continuation of trend reporting. To ensure the feasibility of the proposed linking methodology, the SP and CB instruments were administered to randomly equivalent samples of students drawn from a common population. The 2024 CB sample was linked to the 2024 SP sample using common item linking, and the combined sample results were put onto the existing trend line using the usual approach for adjacent administrations of common item linking. The device evaluation consisting of item-level analyses, Q-Q plots, and subgroup device differences supported the decision to use common item linking to link the 2024 SP and CB samples and to report results based on the combined SP and CB sample.

**References**

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. Two different SP models with different screen sizes were used. [↑](#footnote-ref-2)
2. For multiple-choice and dichotomous constructed-response items, P+ is 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 score 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-3)
3. 3 A total of 139 items were in the assessment. However, one multiple-choice item (Literary) was dropped due to bias concerns since 2019. Statistics in Table 3 are based on performance excluding this item. [↑](#footnote-ref-4)