**Summary of the 2019 NAEP Science Scenario-based Task Integration Evaluation**

For NAEP Science, scenario-based task (SBT) blocks were administered for the first time as part of the 2019 digitally based assessment (DBA) assessments at grades 4, 8, and 12. Science SBTs have been administered and analyzed in some limited fashion in the past in the form of proof of concept, probe, or pilot administrations since 2007. However, 2019 represents the first administration where the NAEP science operational assessments at all three grades were designed to allow the integration of SBTs into the main reporting scales. A study was conducted to evaluate the suitability of including the SBTs together with the traditional discrete items in the analysis and reporting of the 2019 NAEP science results. The 2019 NAEP science assessment included two types of SBTs: interactive computer tasks (ICT) and hybrid hands-on tasks (HHOT).

The key question was whether the unidimensional latent variable models were sufficient for describing the 2019 NAEP science assessment data, which were composed of both discrete items and SBTs. If SBTs measure a different dimension than discrete items, then incorporating the SBTs into the NAEP science assessments would present a confound when comparing results to the previous science assessments, which were composed of only discrete items. The SBT integration evaluation comprised two main components: dimensionality analyses and impact analyses on group-level scale score results. Dimensionality analyses sought to evaluate the extent to which the SBTs formed a coherent trait that differed from the discrete items. Impact analyses on group-level scale score estimation serve a complementary role, addressing the question of whether the estimated scores for key NAEP reporting variables would be changed if the SBTs were to be included.

The 2019 DBA science forms were composed of two separately timed blocks of items. As a result, the forms can be described as discrete-only (made up of two discrete item blocks0F[[1]](#footnote-2)), mixed (with one discrete and one SBT block) and SBT-only (made up of two SBT blocks). All three grades used 30-minute long SBTs1F[[2]](#footnote-3). The preponderance of discrete items/blocks in the assessments resulted in uneven percentages of students receiving each kind of form. Table 1 contains the unweighted rounded sample size and percentage of students by form and grade.

Table 1. Unweighted sample size (N) and percentage of students by form type and grade

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | |  | | | **Grade** | | | |  | | | | |
|  |  | **4** | | |  | | **8** | | | | |  | **12** | |
| **Form type** | **N** | **Percentage** | | | **N** | | | **Percentage** | | | | **N** | **Percentage** | |
| Discrete-only form | 13,560 | | | 67% | 12,840 | | | | 62% | | 11,360 | | | 66% |
| Discrete/SBT (mixed) form | 5,420 | | | 27% | 6,950 | | | | 34% | | 4,430 | | | 25% |
| SBT-only form | 1,250 | | | 6% | 880 | | | | 4% | | 1,500 | | | 9% |
| Total | 20,230 | | | 100% | 20,670 | | | | 100% | | 17,290 | | | 100% |

**Dimensionality Analyses**

In the dimensionality analyses, the goal was to evaluate the dimensionality of the SBT data and the data based on the traditional discrete items (DI) in order to determine whether unidimensional models are sufficient in modeling the DI and SBT data. Both exploratory and confirmatory analyses were conducted. As is usual in NAEP analyses, student sampling weights were used to adjust for the oversampling of certain student groups. Finally, all analyses were conducted at the content area level (Physical Science, Life Science, and Earth and Space Sciences) as well as at the overall science level. Results for content areas are very similar to the overall results. For brevity, only results at the overall science level are presented here.

**Exploratory Analysis**

First, exploratory factor analysis (EFA) was conducted on the full DBA sample of students at each grade using the Mplus 7.2 software (Muthén & Muthén, 1998–2014). The EFA results were evaluated in terms of Bayesian Information Criterion (BIC) model fit statistics (lower values indicate better model data fit), factor loadings, eigenvalues, as well as evidence of overfactoring for multi-factor solutions2F[[3]](#footnote-4). The results suggest that a 1-factor model most appropriately describes the grade 4 and grade 12 data. At grade 8, the BIC for the 1-factor model was 867,125. For the 2-factor model, the BIC was 866,479. While the BIC fit statistics indicate that the 2-factor model fits the data better, a single dominant factor is observed in the scree plot of ordered eigenvalues in Figure 1. Additionally, the factor loadings from the 2-factor solution do not indicate a meaningful pattern that shows that SBT items are dimensionally different from DI items. The factor loadings from the 2-factor solution for grade 8 are shown in the Appendix.

Figure 1. Scree plot of eigenvalues: Grade 8

Out of concern that the EFA results were potentially dominated by the subset of students taking DI forms, additional EFA analyses were conducted on the subset of students who took mixed forms (i.e., forms assembled with one discrete and one SBT block). The conclusion based on these analyses was consistent with the findings from the EFAs based on the full DBA sample of students—that a 1-factor model most appropriately describes the data at grades 4 and 12. At grade 8, despite better fit statistics of the 2-factor model, the pattern of factor loadings does not meaningfully indicate separate factors for SBT and DI.

**Confirmatory Analysis**

To further evaluate if unidimensional models are sufficient in modeling the DBA data with SBT integrated, two types of confirmatory analyses were conducted: multidimensional item response theory (MIRT) model and bifactor model.

Using the full DBA sample, a confirmatory, two-factor simple structure MIRT model was used to examine whether the DI and SBT items show evidence of distinct dimensionality. In this MIRT model, one dimension represented a DI latent factor, and the other dimension represented a correlated SBT latent factor. The estimated disattenuated correlations between DI and SBT dimensions were .88 at grade 4, .82 at grade 8, and .93 at grade 12, suggestive of a single common underlying factor.

For SBTs, because the items within tasks share a common scenario, responses to items within SBTs may be more similar than they would be for discrete items. Unidimensional models ignore this potential clustering of items within SBTs, which may result in violation of the IRT assumptions of item local independence (Lord, 1980). The bifactor models can be considered to test the assumption of item local independence as they explicitly account for these potential testlet effects (Wainer & Kiely, 1987) associated with SBTs. In this study, bifactor models (again using the full DBA sample) were conducted and compared to unidimensional models. At each grade, the bifactor models specified one general “science” factor that all items loaded on. Additional, task-specific factors were specified for each SBT3F[[4]](#footnote-5). As is usual for the bifactor model, all the factors were specified to be orthogonal.

Compared to the unidimensional model, the bifactor model showed improved model fit for all three grades based on the BIC model fit statistic. Table 2 contains the BIC for the unidimensional and bifactor models at each grade.

Table 2.BIC fit statistics for the unidimensional and bifactor models, all grades

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Grade** | | |
|  |  | **4** | **8** | **12** |
| **BIC** | Unidimensional | 722,791 | 860,859 | 731,138 |
| Bifactor SBT | 722,355 | 858,837 | 730,846 |
| Difference | -436 | -2,022 | -292 |

In addition to global model fit statistics, the loading estimates for individual items from the general factor of the bifactor model are compared to those from the unidimensional model. Specifically, differences in item parameter estimates between the two models could be an indication of presence of a testlet effect; that is, potential bias could be introduced into the item parameter estimates of the unidimensional model as a result of neglecting testlet effects. The differences in loading parameters were calculated (as a percentage) by the difference between the unidimensional and the bifactor loading parameter (on the *general* factor), divided by the bifactor loading parameter. Here, the percentage of change in loading less than 15% is considered negligible, following the recommendation by Muthén, Kaplan, and Hollis (1987, p. 446). For discrete items, the differences in loadings between the two models were negligible in all grades. For a few SBT items across the various SBTs, the differences in loading estimates exceeded this criterion. That most loading parameters fall below the 15% criterion suggests that there is little potential systematic error introduced into the unidimensional parameter estimates as a result of ignoring a testlet structure. The percent of change of the loading parameters is presented in the first column in Table 3 for the SBTs at grade 8. Table 3 also contains the loadings and estimated standard error (SE4F[[5]](#footnote-6)) of that loading for all SBT items at grade 8 on both the general and specific factors from the bifactor model and from the unidimensional model. All four items in the hybrid hands-on task (HHOT) are associated with relatively large loadings on both the task-specific and general factors. The first two items in the interactive computer task (ICT) 1 have high loadings on the task-specific factor. This pattern is an indication of potential local item dependence for these two items rather than a testlet effect for the SBT. Upon examination, those items are very similar in content, format, and actions that students must take to respond correctly. The loadings for the remaining items in ICT 1 are moderate. Results are presented only for grade 8, as potential testlet effects are smaller at grades 4 and 12.

Table 3. Bifactor and unidimensional loading results for SBT items in grade 8

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | **Bifactor Model** | | | | **Unidimensional model** | |
|  |  |  |  | **General factor** | | **Specific factor** | |
| **HHOT** | **Item** | **% change of loading** |  | **Loading** | **SE** | **Loading** | **SE** | **Loading** | **SE** |
|  | 1 | -0.50% |  | 0.64 | 0.05 | 0.66 | 0.11 | 0.64 | 0.04 |
|  | 2 | 8.42% |  | 0.81 | 0.05 | 0.47 | 0.08 | 0.88 | 0.05 |
|  | 3 | -11.26% |  | 0.73 | 0.09 | 1.01 | 0.2 | 0.65 | 0.04 |
|  | 4 | -6.56% |  | 0.82 | 0.06 | 0.6 | 0.08 | 0.77 | 0.05 |
| **ICT 1** | 1 | -50.40% |  | 1.89 | 0.18 | 3.6 | 0.31 | 0.94 | 0.05 |
|  | 2 | -41.88% |  | 2.09 | 0.16 | 3.38 | 0.27 | 1.21 | 0.06 |
|  | 3 | -0.43% |  | 0.85 | 0.05 | 0.06 | 0.05 | 0.85 | 0.05 |
|  | 4 | 4.19% |  | 1.1 | 0.07 | 0.28 | 0.06 | 1.15 | 0.07 |
|  | 5 | 2.46% |  | 0.68 | 0.03 | 0.1 | 0.03 | 0.70 | 0.03 |
|  | 6 | -5.27% |  | 0.24 | 0.03 | -0.02 | 0.03 | 0.23 | 0.03 |
|  | 7 | -1.53% |  | 0.81 | 0.05 | 0.06 | 0.05 | 0.80 | 0.05 |
|  | 8 | 1.48% |  | 0.73 | 0.04 | 0.09 | 0.03 | 0.74 | 0.04 |
|  | 9 | 9.12% |  | 0.85 | 0.05 | 0.45 | 0.04 | 0.93 | 0.05 |
|  | 10 | 2.26% |  | 0.73 | 0.04 | 0.09 | 0.03 | 0.75 | 0.04 |
|  | 11 | -3.81% |  | 1.01 | 0.06 | -0.01 | 0.05 | 0.97 | 0.06 |
| **ICT 2** | 1 | -11.01% |  | 0.51 | 0.04 | 0.54 | 0.07 | 0.45 | 0.03 |
|  | 2 | -30.20% |  | 0.81 | 0.11 | 2.12 | 0.34 | 0.57 | 0.04 |
|  | 3 | -4.57% |  | 0.79 | 0.05 | -0.06 | 0.05 | 0.75 | 0.04 |
|  | 4 | -1.73% |  | 0.65 | 0.04 | 0.06 | 0.04 | 0.64 | 0.03 |
|  | 5 | -2.74% |  | 1 | 0.05 | 0.01 | 0.04 | 0.97 | 0.05 |
|  | 6 | -0.82% |  | 1.09 | 0.05 | 0.32 | 0.05 | 1.08 | 0.05 |

To summarize, the bifactor model resulted in slightly better model fit than the unidimensional model at all three grades. In addition, there is evidence of a potential testlet effect for the HHOT at grade 8. Overall, the more parsimonious, unidimensional model is preferred, given that for the most part, little difference was observed in the loading estimates for the SBT items between the bifactor and unidimensional models.

**Impact analysis on student group scale scores**

Student group results are a key part of NAEP’s reporting, serving a vital role in informing educational policymakers. Given the importance of student groups in NAEP reporting, two different sets of scale score comparisons were made to evaluate the impact of SBTs integration on group-level scale scores:

* One set of comparisons evaluated whether there are differential student group effects as a result of integrating SBTs into the science assessment.
* The second set of comparisons evaluated whether there are differential student group effects between students who took a discrete-only form and students who took mixed or SBT-only forms. Because the two sets of sampled students are randomly equivalent, differences can be attributed to differences in science proficiency as measured by discrete versus SBT items.

The first set of scale score comparisons involved two samples: the discrete-only form sample and the full DBA sample. As discussed above, the discrete-only form sample (approximately 65% of the full sample at each grade) was based on the subset of students who took discrete-only forms,5F[[6]](#footnote-7) and the full DBA sample comprised the sample taking all three form types. The set of results based on the discrete-only form sample represents NAEP results if the science assessment continued to use only discrete items, while results based on the full DBA sample incorporate the SBTs. Large differences in the results would indicate that including SBTs has changed what is being measured. This change is confounded with trend differences and so would constitute a threat to validity of one of the program’s key reporting goals.

Table 4 contains mean scale score differences between the full DBA sample and the discrete-only form sample, the standard error of the difference and indication of statistical significance for each grade. Scale score comparisons of these two groups revealed no significant scale score differences for the major student groups. The magnitude of the differences was less than one scale score point except for Asian students (around 1.5 scale score points) on the NAEP science 0 to 500 scale.

Table 4. Student group scale score differences for students in the full sample compared to the discrete-form sample, all grades

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Grade 4** | | | **Grade 8** | | | | | **Grade 12** | | | | | | | | |
| **Student group** | | **Difference (FULL-DI)** | **SE** | |  | | **Difference (FULL-DI)** | **SE** | |  | **Difference (FULL-DI)** | | | **SE** |  | |
| Male | | -0.55 | 0.41 | |  | -0.26 | | 0.46 | |  | | -0.34 | 0.42 | | |  | | |
| Female | | 0.59 | 0.45 | |  | 0.27 | | 0.38 | |  | | 0.32 | 0.47 | | |  | | |
| White | | -0.20 | 0.40 | |  | -0.06 | | 0.44 | |  | | 0.27 | 0.50 | | |  | | |
| Black | | -0.06 | 0.63 | |  | 0.32 | | 0.70 | |  | | 0.53 | 0.88 | | |  | | |
| Hispanic | | -0.21 | 0.57 | |  | 0.02 | | 0.51 | |  | | -0.36 | 0.61 | | |  | | |
| Asian | | 1.45 | 1.15 | |  | -1.29 | | 1.30 | |  | | -1.45 | 1.53 | | |  | | |
| SD | | 0.18 | 0.88 | |  | -0.72 | | 0.89 | |  | | -0.76 | 1.05 | | |  | | |
| Non-SD | | -0.12 | 0.29 | |  | -0.15 | | 0.32 | |  | | -0.06 | 0.32 | | |  | | |
| EL | | -0.18 | 0.83 | |  | 0.12 | | 1.10 | |  | | 0.66 | 1.59 | | |  | | |
| Non-EL | | -0.06 | 0.31 | |  | 0.00 | | 0.33 | |  | | -0.06 | 0.31 | | |  | | |
| Eligible for NSLP | | -0.10 | 0.42 | |  | -0.12 | | 0.45 | |  | | 0.17 | 0.53 | | |  | | |
| Not eligible for NSLP | | 0.07 | 0.39 | |  | 0.03 | | 0.43 | |  | | -0.11 | 0.45 | | |  | | |

NOTE: SD = Students identified as students with disabilities; EL = English learners; NSLP = National School Lunch Program.

Although the discrete-only form sample and full DBA sample comparison analysis reflects what the actual reporting would look like, there was some concern that the discrete-only form sample composed a large proportion of the full DBA sample. Therefore, the second analysis compared two mutually exclusive, randomly equivalent samples of students: a discrete-only form sample and an SBT-only form or mixed-form sample. For the discrete-only form sample, the DI scale was formed based on scaling only the DI items (which are the same results as in Table 3) and linking those results to the science trend reporting scale through a common-population linking to the 2019 paper-based assessment (PBA)6F[[7]](#footnote-8) sample. For the SBT-only form or mixed-form sample, a special set of analyses was conducted to create an SBT scale. The SBT scale was formed using only the SBT items in the IRT scaling, then linking those results to the science trend reporting scale through a common-population linking to the 2019 PBA sample. Table 5 contains mean scale score differences between the SBT scale and the DI scale, the standard error of the difference, and indication of statistical significance for each grade.

Table 5. Student group scale scores differences for students who took a discrete-only form compared to students who took mixed or SBT-only forms, all grades

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Grade 4** | | | **Grade 8** | | | | | **Grade 12** | | |
| **Student group** | **Difference (SBT-DI)** | **SE** |  | **Difference (SBT-DI)** | **SE** |  | **Difference (SBT-DI)** | | | **SE** |  |
| Male | -3.24\* | 1.14 |  | -3.50\* | 0.99 |  | | -2.10 | | 1.06 |  |
| Female | 3.26\* | 0.95 |  | 3.61\* | 0.86 |  | | 2.11 | | 1.17 |  |
| White | -1.50 | 1.16 |  | -2.89\* | 1.06 |  | | -0.21 | | 1.07 |  |
| Black | 0.82 | 1.61 |  | 3.93\* | 1.62 |  | | 3.70 | | 1.76 |  |
| Hispanic | 0.99 | 1.44 |  | 3.86\* | 1.25 |  | | -0.03 | | 1.28 |  |
| Asian | 3.60 | 3.37 |  | -5.39 | 2.99 |  | | -7.03 | | 4.30 |  |
| SD | 3.07 | 2.12 |  | 2.61 | 1.87 |  | | -2.24 | | 2.46 |  |
| Non-SD | -0.73 | 0.80 |  | -1.01 | 0.71 |  | | -0.20 | | 0.87 |  |
| EL | 3.01 | 2.05 |  | 9.43\* | 2.37 |  | | 4.45 | | 3.12 |  |
| Non-EL | -0.62 | 0.85 |  | -0.69 | 0.72 |  | | -0.29 | | 0.84 |  |
| Eligible for NSLP | 0.94 | 1.04 |  | 1.71 | 0.87 |  | | 1.22 | | 1.07 |  |
| Not eligible for NSLP | -0.86 | 1.07 |  | -1.59 | 1.00 |  | | -0.99 | | 1.06 |  |

NOTE: SD = Students identified as students with disabilities; EL = English learners; NSLP = National School Lunch Program.

\*Significant difference (*p* < .05)

**Key results:**

* Significant scale score differences are present for male and female groups at grades 4 and 8. Male student performance on the SBT scale was lower compared to the discrete scale, while female student performance was higher on the SBT scale compared to the discrete scale. The absolute magnitude of the differences for the gender groups are between three and four scale score points.
* Additionally, at grade 8, significant differences in scale scores are observed for White, Black, and Hispanic students as well as for English learners. White student performance was lower on the SBT scale compared to the discrete scale, whereas Black, Hispanic, and English learner students had higher performance on the SBT scale compared to the discrete scale. For those student groups where significant differences in scale scores are present, the absolute magnitude of the differences for the gender groups is three scale score points or higher.
* There are no significant differences observed at grade 12.

**Summary**

Based upon the analyses just described, the following conclusions about the 2019 DBA science data are drawn about the SBT integration evaluation:

* Unidimensional models appear to be sufficient in modeling the 2019 DBA data with SBTs integrated.
* For the NAEP main reporting student groups, no substantive differences in student group means were observed when SBTs were included, compared to when SBTs were excluded from the analysis. These results provide support for including the SBTs in the 2019 science analysis and reporting.
* When directly comparing the student group scores of students who took mixed or SBT-only forms to the student group scores for those who took discrete-only forms, some numerically larger and statistically significant scale score differences between student groups emerged.

The two sets of student group scale score impact analyses are indicative of the impact of including SBTs in the NAEP science assessment. Given the 2019 assessment design, in which the assessment was composed mostly of discrete items with most students consequently taking the discrete-only forms, the impact of including SBTs appears to be moderate, and the integration of SBTs does not appear to have disrupted the reported results for the NAEP science scale. However, differential student group performance was observed when comparing the data based on SBTs only to the data based on discrete items. Significant group differences were restricted to a relatively small number of groups: male and female at grades 4 and 8, as well as English learners and race/ethnicity student groups at grade 8. There were no significant differences at grade 12.

**References**

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

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431-462.

Muthén, L. K., & Muthén, B. O. (1998–2014). *Mplus user’s guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–202.

Appendix

EFA factor loadings from the 2-factor solution: Grade 8

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Item | Factor1 | Factor2 | Item | Factor1 | Factor2 | Item | Factor1 | Factor2 | Item | Factor1 | Factor2 |
| 1 | 0.377\* | -0.166\* | 54 | 0.228\* | 0.089 | 107 | 0. 247\* | 0.001 | 160 | 0.479\* | 0.081 |
| 2 | 0.370\* | -0.062 | 55 | 0.488\* | 0.08 | 108 | 0.191\* | 0.069 | 161 | 0.380\* | -0.073 |
| 3 | 0.457\* | -0.064 | 56 | 0.143\* | 0.023 | 109 | 0.515\* | 0.072 | 162 | 0.283\* | 0.039 |
| 4 | 0.394\* | 0.072 | 57 | 0.570\* | 0.052 | 110 | 0.507\* | 0.173\* | 163 | 0.441\* | 0.051 |
| 5 | 0.402\* | 0.05 | 58 | 0.172\* | 0.035 | 111 | 0.626\* | 0.093 | 164 | 0.559\* | 0.028 |
| 6 | 0.518\* | -0.174\* | 59 | 0.365\* | 0.233\* | 112 | 0.558\* | 0.156\* | 165 | 0.531\* | 0.075 |
| 7 | 0.284\* | 0.047 | 60 | 0.478\* | 0.137\* | 113 | 0.419\* | 0.153\* | 166 | 0.444\* | 0.094 |
| 8 | 0.584\* | -0.035 | 61 | 0.553\* | 0.067 | 114 | 0.470\* | 0.005 | 167 | 0.543\* | -0.135\* |
| 9 | 0.682\* | 0.041 | 62 | 0.540\* | -0.034 | 115 | 0.237\* | 0.187\* | 168 | 0.606\* | 0.041 |
| 10 | 0.542\* | -0.068 | 63 | 0.497\* | 0.171\* | 116 | 0.404\* | 0.127\* | 169 | 0.383\* | 0.001 |
| 11 | 0.343\* | 0.005 | 64 | 0.376\* | 0.056 | 117 | 0.509\* | 0.001 | 170 | 0.561\* | -0.053 |
| 12 | 0.697\* | 0.003 | 65 | 0.546\* | 0.05 | 118 | 0.565\* | -0.145\* | 171 | 0.489\* | -0.111 |
| 13 | 0.501\* | 0.004 | 66 | 0.403\* | 0.165\* | 119 | 0.335\* | 0.02 | 172 | 0.698\* | -0.138 |
| 14 | 0.312\* | -0.013 | 67 | 0.606\* | -0.061 | 120 | 0.238\* | 0.06 | 173 | 0.704\* | -0.212\* |
| 15 | 0.661\* | 0.024 | 68 | 0.604\* | -0.026 | 121 | 0.588\* | 0 | 174 | 0.404\* | 0.05 |
| 16 | 0.543\* | 0.142 | 69 | 0.571\* | 0.099 | 122 | 0.339\* | 0.109 | 175 | 0.243\* | 0.167\* |
| 17 | 0.300\* | 0.064 | 70 | 0.580\* | 0.178\* | 123 | 0.394\* | -0.018 | 176 | 0.615\* | -0.122\* |
| 18 | 0.295\* | 0.033 | 71 | 0.244\* | 0.017 | 124 | 0.536\* | 0.098 | 177 | 0.539\* | 0.003 |
| 19 | 0.232\* | -0.03 | 72 | 0.399\* | 0.155\* | 125 | 0.211\* | 0.123\* | 178 | 0.512\* | 0.081 |
| 20 | 0.526\* | -0.185\* | 73 | 0.254\* | -0.083 | 126 | 0.477\* | 0.135\* | 179 | 0.190\* | 0.119 |
| 21 | 0.444\* | 0.148 | 74 | 0.625\* | 0.059 | 127 | 0.314\* | 0.004 | 180 | 0.374\* | 0.127 |
| 22 | 0.440\* | -0.005 | 75 | 0.431\* | -0.019 | 128 | 0.265\* | 0.138 | 181 | 0.407\* | 0.11 |
| 23 | 0.661\* | -0.103 | 76 | 0.261\* | 0.034 | 129 | 0.603\* | -0.073 | 182 | 0.596\* | -0.131\* |
| 24 | 0.483\* | -0.057 | 77 | 0.616\* | -0.048 | 130 | 0.445\* | -0.016 | 183 | 0.599\* | 0.05 |
| 25 | 0.273\* | 0.188\* | 78 | 0.235\* | 0.061 | 131 | 0.487\* | 0.111 | 184 | 0.577\* | -0.193\* |
| 26 | -0.179\* | 0.076 | 79 | 0.350\* | 0.013 | 132 | 0.529\* | 0.015 | 185 | 0.623\* | -0.151 |
| 27 | 0.093\* | 0.067 | 80 | 0.367\* | 0.002 | 133 | 0.410\* | 0.116 | 186 | 0.194\* | -0.006 |
| 28 | 0.373\* | -0.032 | 81 | 0.669\* | 0.041 | 134 | 0.412\* | -0.047 | 187 | 0.717\* | -0.07 |
| 29 | 0.489\* | -0.013 | 82 | 0.211\* | 0.087 | 135 | 0.611\* | -0.067 | 188 | 0.536\* | 0.057 |
| 30 | 0.361\* | 0.04 | 83 | 0.489\* | 0.105 | 136 | 0.783\* | -0.12 | 189 | 0.292\* | -0.031 |
| 31 | 0.607\* | 0.009 | 84 | 0.370\* | -0.086 | 137 | 0.508\* | -0.087 | 190 | 0.235\* | 0.058 |
| 32 | 0.639\* | 0.06 | 85 | 0.525\* | -0.001 | 138 | 0.259\* | 0.047 | 191 | 0.445\* | -0.017 |
| 33 | 0.285\* | 0.09 | 86 | 0.395\* | 0.043 | 139 | 0.418\* | 0.031 | **192** | 0.556\* | -0.452\* |
| 34 | 0.419\* | -0.036 | 87 | 0.364\* | 0.118 | 140 | 0.565\* | -0.017 | **193** | 0.589\* | -0.300\* |
| 35 | 0.595\* | 0.256\* | 88 | 0.651\* | 0.069 | 141 | 0.525\* | 0.059 | **194** | 0.544\* | -0.374\* |
| 36 | 0.674\* | 0.157\* | 89 | 0.273\* | 0.045 | 142 | 0.528\* | 0.178\* | **195** | 0.656\* | -0.319\* |
| 37 | 0.296\* | 0.038 | 90 | 0.359\* | -0.025 | 143 | 0.576\* | -0.160\* | **196** | 0.298\* | 0.769\* |
| 38 | 0.407\* | 0.009 | 91 | 0.253\* | 0.072 | 144 | 0.478\* | -0.002 | **197** | 0.283\* | 0.910\* |
| 39 | 0.455\* | -0.08 | 92 | 0.335\* | 0.039 | 145 | 0.357\* | -0.025 | **198** | 0.419\* | 0.032 |
| 40 | 0.195\* | 0.042 | 93 | 0.546\* | 0.009 | 146 | 0.417\* | 0.001 | **199** | 0.498\* | 0.112\* |
| 41 | 0.430\* | 0.075 | 94 | 0.581\* | -0.029 | 147 | 0.405\* | -0.181\* | **200** | 0.475\* | 0.063 |
| 42 | 0.305\* | 0.113 | 95 | 0.744\* | -0.002 | 148 | 0.318\* | 0.02 | **201** | 0.194\* | -0.02 |
| 43 | 0.482\* | 0.047 | 96 | 0.527\* | -0.003 | 149 | 0.569\* | -0.084 | **202** | 0.396\* | 0.032 |
| 44 | 0.600\* | 0.201\* | 97 | 0.449\* | 0.450\* | 150 | 0.354\* | -0.019 | **203** | 0.492\* | 0.061 |
| 45 | 0.591\* | -0.098 | 98 | 0.551\* | 0.470\* | 151 | 0.625\* | -0.348\* | **204** | 0.495\* | 0.326\* |
| 46 | 0.324\* | 0.06 | 99 | 0.482\* | 0.345\* | 152 | 0.430\* | -0.136\* | **205** | 0.530\* | 0.063 |
| 47 | 0.571\* | 0.05 | 100 | 0.623\* | 0.024 | 153 | 0.316\* | 0.093 | **206** | 0.486\* | -0.014 |
| 48 | 0.452\* | 0.127\* | 101 | 0.520\* | -0.007 | 154 | 0.423\* | -0.005 | **207** | 0.457\* | -0.161 |
| 49 | 0.391\* | -0.024 | 102 | 0.434\* | 0.013 | 155 | 0.443\* | -0.153\* | **208** | 0.324\* | -0.188 |
| 50 | 0.561\* | -0.119\* | 103 | 0.544\* | -0.059 | 156 | 0.657\* | -0.325\* | **209** | 0.550\* | -0.119 |
| 51 | 0.270\* | 0.152\* | 104 | 0.574\* | -0.039 | 157 | 0.563\* | 0.129\* | **210** | 0.488\* | -0.173\* |
| 52 | 0.264\* | -0.028 | 105 | 0.557\* | -0.055 | 158 | 0.530\* | -0.041 | **211** | 0.636\* | -0.112\* |
| 53 | 0.331\* | 0.216\* | 106 | 0.558\* | 0.049 | 159 | 0.692\* | -0.139\* | **212** | 0.638\* | -0.238\* |

NOTE: Boldface items indicate SBT items

\* *p* < .05.

1. See glossary of NAEP terminology at https://www.nationsreportcard.gov/glossary.aspx#region [↑](#footnote-ref-2)
2. At grade 12, there were two SBTs that were 15 minutes each. These were paired in a single 30-minute unit. [↑](#footnote-ref-3)
3. Note that the estimation does not account for dependence between observations due to clustering in the sample. As a result, estimates may underestimate the true variability. For the purposes of the present analysis, this was deemed acceptable. [↑](#footnote-ref-4)
4. The model at grades 4 and 8 both contained 3 task-specific factors, as there were 3 SBTs in each grade. At grade 12, there were 5 SBTs, so 5 task-specific factors were defined. [↑](#footnote-ref-5)
5. Note that the estimation does not account for dependence between observations due to clustering in the sample. As a result, estimates may underestimate the true variability. For the purposes of the present analysis, this was deemed acceptable. [↑](#footnote-ref-6)
6. 6Results of the full DBA sample were linked to the Science trend reporting scale through common-population linking to the 2019 paper-based assessment (PBA) sample. The same linking approach applies to the results of the discrete-only sample. Consequently, the mean and SD of the full DBA sample are identical to those of the discrete sample. [↑](#footnote-ref-7)
7. The design of the 2019 samples resulted in randomly equivalent groups of students who took PBA and DBA, supporting the assumption of common-population linking. Similarly, within the DBA, the DI sample, mixed-form sample, and SBT sample are randomly equivalent. [↑](#footnote-ref-8)