Committee on Standards, Design and Methodology

August 4, 2023
9:00 am – 11:00 am ET
Hemingway 2, 3

AGENDA

9:00 – 9:05 am	**Welcome and Overview of Agenda** *Suzanne Lane, Chair*	
9:05 – 9:50 am	**NAEP Modernization: Plans for Device Agnostic Administration (CLOSED)** *Jing Chen, NCES* *Helena (Yue) Jia, Educational Testing Service*	Attachment A
9:50 – 10:25 am	**Utility of Effect Sizes** *Suzanne Lane* *Becky Dvorak, Assistant Director, Psychometrics*	Attachment B
10:25 – 11:00 am	**Updates: Achievement Level Communications** *Suzanne Lane*	Attachment C

NAEP Modernization: Plans for Device Agnostic Administration

August 4, 2023

The Governing Board has received high-level updates on modernization plans in recent months, including plans to move towards a device agnostic administration where NAEP would be administered on school-owned devices. There are many technical considerations related to moving to a device agnostic administration such as ensuring that scores remain valid and reliable across devices, and that risks to trend are mitigated. The purpose of this August 2023 session is for members of the Committee on Standards, Design and Methodology (COSDAM) to learn from staff of the National Center for Education Statistics (NCES) regarding steps towards a successful transition.

Background

The NAEP program is considering multiple changes in administration practices to move towards a more efficient and modern assessment program. One such change is to move towards a device agnostic administration. Since NAEP first became a digital assessment, NCES has provided all necessary equipment, currently Microsoft Surface Pro tablets with styluses. The assessment administrators carry large cases, called pelican cases, that hold all equipment required to administer the assessment and to transmit the assessment data without accessing a school's Wi-Fi or using their computers. In 2024 NAEP assessments will be administered on Microsoft Surface Pro tablets and Google Chromebooks. Bridge studies are planned to evaluate score comparability between Surface Pro and Chromebook devices and the feasibility of linking results collected from the hybrid device mode (Surface Pros and Chromebooks) to the existing trendlines. Chromebooks were chosen because of their widespread use in classrooms and their lower cost. Current plans are that in 2026 NAEP will be administered on school-based equipment that meet NAEP's technology requirements, with NAEP-provided Chromebooks as backups. NAEP is also moving towards a fully online-based assessment, which is a necessary step towards administering assessments on school devices.

NCES recently conducted a field test to examine the logistic impacts of moving to an online assessment administered on Chromebooks. Additional proof-of-concept study and field test are planned to examine the impacts of moving towards a hybrid administration mode of school-based equipment and NAEP-provided Chromebooks. The August 2023 COSDAM session will provide an opportunity to learn more about completed and planned studies, and will give COSDAM members an opportunity to ask questions.

Using Effect Size to Understand the Size of Score Differences

August 4, 2023

The purpose of the COSDAM session on effect sizes at the August 2023 meeting will be to further deliberate whether effect size is something COSDAM would like to recommend be considered for inclusion in NAEP reporting. Following the May 2023 COSDAM meeting, the COSDAM chair requested Board staff provide examples of effect sizes with real NAEP data to help conceptualize the value as the committee continues to discuss this issue. This document includes 1) background information, 2) a reminder of recent COSDAM effect size discussions, and 2) example cases using real NAEP data.

Background

The Nation's Report Card provides results from the various NAEP assessments. It includes scale scores and percentages of students that meet each NAEP Achievement Level for the nation, by student subgroups, and sometimes by state or for select districts participating in the Trial Urban District Assessments (TUDA). The Nation's Report Card uses significance testing to highlight score differences. If a score difference is found to be non-significant, it indicates the values should be interpreted as effectively the same. If a score difference is found to be significant, it indicates confidence that the difference represents a real difference in the full population, though it does not inform the size of the difference. NAEP uses a 95% confidence level for determining significance – in other words, it requires 95% confidence that an observed difference in the sample reflects a true difference in the population. Among other factors, statistical significance is dependent on sample size – the larger the sample size, the more likely a difference will be significant.

Alternatively, effect sizes provide practical meaning to score differences. They represent the size of the difference or change in scores. Effect sizes are measured in terms of how many standard deviations away from each other the two numbers are. It is computed based on the size of the difference and the variability of scores across the samples, as measured by standard deviations. Larger differences and smaller variabilities are associated with greater effect sizes – which indicate larger differences. It is not dependent on sample size. There is some debate regarding thresholds for interpreting effect sizes, and Kraft 2019 notes it may depend on the context. However, in general, effect sizes of 0.2 or greater are generally seen as meaningful, and in some contexts, it may be that even smaller effect sizes (e.g., 0.1, 0.15) are considered meaningful.

The Nation's Report Card does not report out standard deviations or effect sizes in its standard reporting; however, one can obtain standard deviations through the NAEP Data Explorer (NDE) tool that allows computation of effect sizes. Effect size computations are not highly sophisticated, and can also be accomplished through

various publicly available online effect size calculators by plugging in the means and standard deviations of two samples of data.

Summary of May 2023 COSDAM Discussion

COSDAM has been exploring the potential utility of incorporating effect sizes with NAEP results to provide additional meaning to score differences. The idea was initially discussed regarding recent reductions in state-level sample sizes. When sample sizes are decreased, it may result in non-significant findings for differences that previously had been identified as statistically significant with larger sample sizes. Thus, some Board members inquired about using effect sizes to provide additional information about the meaningfulness of the difference not tied to significance testing. In May 2023 COSDAM members expressed that effect sizes should be a consideration in general for NAEP, not to address state-level sample sizes specifically.

COSDAM members expressed concern that scale scores and significance testing alone may not provide enough information to fully understand the meaningfulness of score differences on NAEP. Adding the standard deviation more prominently – not requiring one to use the NDE to obtain it – may help by providing additional information about the variability of scores within a sample. Including the effect size could be useful as it provides a standardized measure of the size of the difference. COSDAM members noted effect sizes could make it easier to make quick comparisons about the size of score differences between subject areas or states.

Some concerns have been expressed regarding use of effect size in NAEP reporting. There is a desire to keep the Nation’s Report Card at a level that is accessible to a wide audience and those without a statistics background may not find the information intuitive. Readers may not know how to interpret effect size values and it may be confusing if a non-significant result is associated with a meaningful effect size, or vice versa.

The remainder of this document presents examples of reporting NAEP score differences with significance testing and effect sizes. These examples are intended to aid the August COSDAM follow-up discussion on the utility of incorporating effect size in NAEP reporting.

Examining Differences of NAEP Scores with Effect Size

Figure 1 below illustrates how NAEP scores are presented at the top of the Nation's Report Card Highlights page for 2022 NAEP Mathematics. As shown, the graphic presents the overall score for the nation over time, and an * indicates whether prior years' scores are statistically different from 2022. To the right, the difference in score points since the beginning of trend in 1990 and since 2019 are presented. As shown, the difference between the most recent data, 2022, and each of the three years highlighted – 1990, 2003, and 2019 – show statistically significant differences for grades 4 and 8.

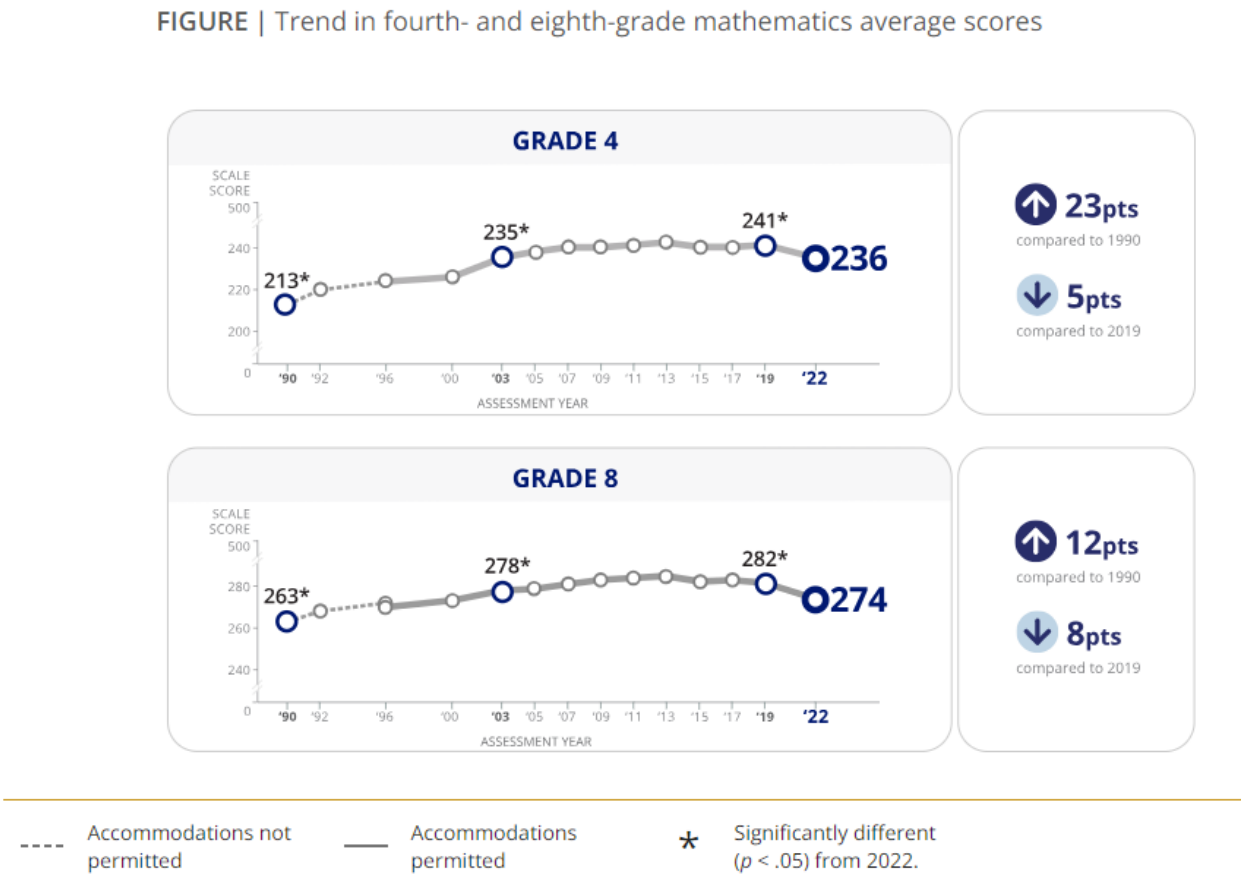

Figure 1. Trend in fourth and eighth grade mathematics average scores.

Table 1 provides the effect sizes of the differences presented in Figure 1. Though all the differences presented are statistically significant, the effect sizes are quite different across the comparisons. For example, the difference between 2022 and 2003 at the national level for grade 4 is only 1 score point, yet was found to be statistically different. The effect size associated with this difference is only 0.03, indicating the difference is only very small, and potentially not meaningful. Whereas the effect size associated with the difference between 1990 and 2022 scores is 0.70, which would be considered at

least a moderate, if not large, difference. These effect sizes illustrate that the effect size is related to the size of the difference – larger differences are typically associated with larger effect sizes. However, the variability of scores (e.g., standard deviations) also impact the effect size. Lower variability in the sample provides greater confidence in the score mean and score differences, and so is associated with larger effect sizes compared to greater variability. For NAEP data, the standard deviations tend to fall between 30 – 40 score points.

Table 1. Examining effect sizes to changes in NAEP Mathematics performance over time.

Comparison	Difference	Statistically Significant?	Effect Size
Grade 4			
2019 – 2022	-5	Yes	-0.14
2003 – 2022	1	Yes	0.03
1990 – 2022	23	Yes	0.70
Grade 8			
2019 – 2022	-8	Yes	-0.20
2003 – 2022	-4	Yes	-0.09
1990 – 2022	12	Yes	0.31

Table 2 provides an example of state-level changes between 2019 and 2022 for NAEP Reading and Mathematics. State-level sample sizes are smaller than the national level, and so it may take larger differences to result in a statistically significant difference.

For Mathematics, all states presented performed statistically significantly worse in 2022 than 2019, and the effect sizes indicate differences of 0.19 or greater (effect sizes should be interpreted based on their absolute value, the sign indicates the direction of the change).

For Reading, two of the states showed statistically decreased performance in 2022 compared to 2019, though the scale score differences were smaller compared to Mathematics, and likewise, so were the effect sizes. New Jersey had less than one point difference between 2019 and 2022 scores, and so the effect size was 0.01.

Table 2. Effect sizes and statistical significance testing for Grade 8 2019 and 2022 NAEP Mathematics and Reading for a selection of states.

State	2022		2019		Difference	Effect Size	Statistically Significant
	Scale Score	SD	Scale Score	SD			
Mathematics							
New Mexico	258.98	34.77	268.77	36.59	-9.8	-0.27	Yes
New Jersey	280.89	41.18	291.82	43.98	-10.9	-0.26	Yes
Florida	271.2	37.47	278.52	39.53	-7.3	-0.19	Yes
Illinois	275.2	38.52	282.56	39.72	-7.4	-0.19	Yes
Reading							
New Mexico	247.8	36.49	251.7	39.26	-3.9	-0.10	Yes
New Jersey	269.78	38.71	270.36	41.89	-0.6	0.01	No
Florida	259.63	36.91	263.35	37.97	-3.7	0.10	Yes
Illinois	261.89	38.54	264.7	38.35	-2.8	0.07	No

Tables 3 and 4 provide comparison between select states and the nation overall for Reading and Mathematics, respectively. These tables were designed to provide examples of state comparisons, and comparing states to the nation. The region is listed across the top row and repeated in the first column. Only half the table includes values to avoid repetition. For each comparison, the differences are presented, followed by whether the difference is significant (yes or no), and then the effect size.

For reading, all differences that are considered statistically significant have effect sizes that exceed 0.2, and so provide support that these are practically meaningful differences. As shown, the effect sizes for differences between New Mexico and other states, and the nation, are larger than others and suggest at least a moderate difference in scores.

Table 3. Comparing National and Select State 2022 NAEP Reading Scale Scores through Statistical Testing and Effect Sizes.[1]

Region	Florida (260, SD = 37)	Illinois (262, SD = 39)	New Jersey (270, SD = 39)	New Mexico (248, SD = 36)
National (260, SD =38)	Diff = 1 Sig - No Effect Size = 0.02	Diff = -1 Sig - No Effect Size = -0.04	Diff = -9 Sig - Yes Effect Size = -0.24	Diff = 13 Sig - Yes Effect Size = 0.34
Florida		Diff = -2 Sig - No Effect Size = -0.06	Diff = -10 Sig - Yes Effect Size = -0.27	Diff = 12 Sig - Yes Effect Size = 0.32
Illinois			Diff = -8 Sig - Yes Effect Size = -0.20	Diff = 14 Sig - Yes Effect Size = 0.38
New Jersey				Diff = 22 Sig - Yes Effect Size = 0.58

For math, all four states have statistically significant differences from one another, and all states are significantly different from the nation except for Illinois. The effect size suggests the difference between Florida and the national scale score averages is small and may not be meaningful. Whereas the effect sizes associated with the differences between New Mexico and the three other states and the nation indicate meaningful differences.

[1] Note that the scale scores, standard deviations, and score differences are presented as rounded values to reduce the reading load of the table; the significance testing and effect sizes were computed using unrounded values.

Table 4. Comparing National and Select State 2022 NAEP Mathematics Scale Scores through Statistical Testing and Effect Sizes. [2]

Region	Florida (271, SD = 37)	Illinois (275, SD = 39)	New Jersey (281, SD = 41)	New Mexico (259, SD = 35)
National (274, SD = 39)	Diff = 3 Sig - Yes Effect Size = 0.08	Diff = -1 Sig - No Effect Size = -0.02	Diff = -7 Sig - Yes Effect Size = -0.17	Diff = 15 Sig - Yes Effect Size = 0.41
Florida		Diff = -4 Sig - Yes Effect Size = -0.11	Diff = -10 Sig - Yes Effect Size = -0.25	Diff = 12 Sig - Yes Effect Size = 0.34
Illinois			Diff = -6 Sig - Yes Effect Size = -0.14	Diff = 16 Sig - Yes Effect Size = 0.44
New Jersey				Diff = 22 Sig - Yes Effect Size = 0.57

[2] Note that the scale scores, standard deviations, and score differences are presented as rounded values to reduce the reading load of the table; the significance testing and effect sizes were computed using unrounded values.

Update on Achievement Level Communications Activities

August 4, 2023

Suzanne Lane, Chair of the Committee on Standards, Design and Methodology (COSDAM), presented a plan for achieving the committee's goals for improving NAEP Achievement Levels communications at the May 2023 COSDAM meeting. This plan was based on feedback from various prior discussions and focused on the need to develop: 1) a series of brief communications documents each focused on specific NAEP content and stakeholder needs, and 2) a validity argument document summarizing appropriate and inappropriate claims based on the evidence, and sources of achievement level validity evidence to support these claims.

The purpose of the August COSDAM session will be to receive an update on communications activities progress, and for COSDAM members to weigh in on expected content of the documents, including an outline of the validity argument document, and appropriate and inappropriate interpretations and uses of achievement levels (AL), for inclusion in the validity argument document and communication briefs.

Achievement Levels Communication Plan

COSDAM members agreed with an approach for developing the two types of communications documents planned described at the May 2023 COSDAM meeting. These plans were discussed in detail at that time, and are briefly summarized below.

Communication briefs:

- Start by drafting documents for NAEP Reading and Mathematics, and for journalists and governors and their staff.
- Board staff will work with Board strategic communications contractor to develop mock-ups that include introductory material, AL policy definitions, example items (and/or item descriptions), and appropriate and inappropriate interpretations and uses.
- COSDAM members and Reporting and Dissemination (R&D) committee members will be given opportunities to review drafts initially and throughout the process.
- COSDAM will consult with R&D on the content of the briefs and strategies for disseminating the information.
- Focus groups will be convened by Board technical services contractor to gather reactions and input and this information will be used to finalize documents.
- COSDAM will work with R&D and the Board strategic communications contractor to disseminate information.

Validity argument document:

- Board staff will draft an outline with the proposed contents of the document for COSDAM feedback.
- COSDAM members will offer feedback on the draft outline and Board staff will adjust the outline accordingly.
- A validity argument document will be developed based on the agreed upon outline. COSDAM members will have the opportunity to weigh in on initial draft, and drafts throughout the process.
- Focus groups with measurement and testing professionals will be convened by the Board technical services contractor to gather reactions and input used to revise and finalize document.

For both document types, we may include additional rounds of focus groups and/or Board feedback depending on the substance of revisions.

Draft Outline of Validity Argument Document

As described above, one of the first steps was for Board staff to develop a draft outline for information to include in the Achievement Levels Validity Argument Document. The validity argument document is intended to compile and summarize validity evidence about NAEP Achievement Levels in one place, and to offer information on the appropriate and inappropriate uses and interpretations of them. This document is most likely to be used by those with some familiarity of assessment and achievement levels but will be publicly available to all.

I. Purpose of NAEP and NAEP Achievement Levels
    a. NAEP history and purpose
        i. Content areas and grades assessed, and frequency
    b. Historical context of Achievement Levels
        i. When and why they were developed
        ii. Policy definitions
    c. Major claims that can be made using achievement levels
II. Achievement Levels development policy and process
    a. Summary of Board achievement level policy (including links to policy documents)
        i. Adherence to field best practices
    b. Achievement Level Descriptions (ALD)
        i. In framework
        ii. Reporting ALDs
III. Validity research
    a. Standard setting process overview (including links to full reports for most recent standard setting for each subject area)
    b. ALD review studies, with focus on alignment ratings
        i. Summarize methodology and alignment ratings for Reading, Mathematics, Science, U.S. History, and Civics

   c. Summary of achievement level evidence from linking studies, state mapping studies

IV. Claims/appropriate and inappropriate uses of ALs based on validity evidence (see Tables 1 and 2)
   a. How achievement levels indicate academic performance and how these differ from state achievement levels and being "on grade level"
   b. Relationship to external measures of achievement and college preparedness
   c. Use of NAEP achievement levels for understanding differences in state achievement levels

Tables 1 and 2 summarize some of the appropriate and inappropriate interpretations and uses of NAEP Achievement Levels, respectively, for inclusion in the validity argument document and communication briefs. This is intended to be more expansive than information presented in The Intended Meaning of NAEP Results adopted by the Board in 2020. These are only draft and not exhaustive. COSDAM members will be asked to weigh in on whether there are other correct or incorrect uses or interpretations they are aware of that should be incorporated.

Table 1. Draft of appropriate interpretations and uses of NAEP Achievement Levels.

Appropriate Uses of NAEP Achievement Levels	**Evidence**
Performance at *NAEP Proficient* represents a solid understanding of subject-matter content	Policy and technical documentation of AL development and the standard setting process
Though not directly related to state achievement levels, NAEP ALs can help inform the comparisons of state achievement level cut-points	State Mapping Studies; information on AL development and the standard setting process; State achievement level documentation
Reporting ALDs provide information on what students performing at each AL can likely do based on assessment data • Include full set of Reporting ALDs for validity argument document; one or two examples for briefs.	ALD Study reports for Reading and Mathematics, and for U.S. History, Civics, and Science
AL performance is related to other academic and college readiness outcomes • NAEP achievement levels associated with greater likelihood of attending a two- or four- year college • Performance in *NAEP Advanced* associated with a greater likelihood	Linking study reports, including: • NAEP linked with High School Longitudinal Study of 2009 (HSLS:09) • NAEP linked with Early Childhood Longitudinal Study (ECLS-K:2011) • Various studies linking NAEP with college entrance exams

of majoring in a STEM field in college compared to other achievement levels • Performance at *NAEP Proficient* or above in grade 4 Reading associated with higher reading trajectories in elementary school	

Table 2. Draft of inappropriate interpretations and uses of NAEP Achievement Levels.

<u>Inappropriate</u> Uses of NAEP Achievement Levels	**Evidence**
Using NAEP Reading ALs to determine the percentage of students that can or cannot read	Policy and technical documentation of AL development and standard setting process, information from framework and Reporting ALDs; State achievement level documentation
The percent *NAEP Proficient* (or *NAEP Basic*, or *NAEP Advanced*) indicates the percentage of students falling at grade level for a given subject	Policy and technical documentation of AL development and standard setting process; note regarding how NAEP achievement levels differ from state achievement levels; State achievement level documentation
Using NAEP AL data as an outcome measure to determine cause and effect impacts of state- or district-level interventions	Policy and technical documentation of AL development; Intended Meaning of NAEP Results; External information on requirements for determining causality