Committee on Standards, Design and Methodology

March 15, 2022

2:00 – 4:00 pm ET (Virtual)

AGENDA

2:00 pm	**Welcome** *Suzanne Lane, Chair*	
2:00 – 2:40 pm	**Update: 2022 NAEP Administration (CLOSED)** *Bill Ward, NCES*	
2:40 – 2:45	**Break**	
2:45 – 3:25 pm	**Briefing and Discussion: NAEP Innovations** *Enis Dogan, NCES* *Eunice Greer, NCES*	Attachment A
3:25 – 4:00 pm	**Discussion: Potential Next Steps for Exploring a NAEP Below Basic Achievement Level** *Suzanne Lane*	Attachment B

NAEP Innovations

NCES has plans to explore various innovations to increase efficiency and precision of the NAEP assessments. During the March 2022 COSDAM meeting, NCES will present on design and scoring changes to be studied in the coming years, including adaptive testing, a two-subject design, and automated scoring. This session will highlight justifications and plans to study the impact of these changes, and will provide an opportunity for COSDAM members to ask questions and consider impacts to NAEP policy. To prepare for this session NCES provided the following background information, included in this attachment:

- NAEP Design Innovations: 2028 and beyond. Provides justifications and plans to study potential changes from a linear, one-subject design to an adaptive, two-subject design. Studies will examine a two-subject design under a linear condition and an adaptive condition.

- NAEP Automated Scoring Challenge. Describes a recently held NCES automated scoring challenge to score constructed response items for the NAEP reading assessment, and presents the winners.

NAEP Design Innovations: 2028 and beyond

Currently, in NAEP assessments each student receives two 30-minute cognitive blocks of items from a single subject (e.g. reading or mathematics). The assessments are linear, not adaptive, meaning that the difficulty of the assessment is not customized to the individual student, and the test booklets are spiraled across the student sample with equal probability.

As part of our long-term innovation research agenda, our plan is to examine potential changes to this current single-subject, linear design. More specifically, we will conduct studies to examine two potential design changes in 2026 with reading and mathematics assessments: 1) two-subject design, where each student takes items from both subjects, and 2) adaptive testing, where the difficulty of the assessment is customized depending on students' performance on an initial set of items. These two design changes have the potential to create significant improvements in measurement and cost efficiency, discussed further below. Two studies will be conducted to study these in 2026: one examining a two-subject design (reading and mathematics) under a linear testing condition, and one examining the same under an adaptive testing condition. Depending on the outcomes of these studies, the design changes will be implemented for the first time, along with a bridge study, in 2028. Below we discuss the rationale behind these potential design changes.

Adaptive testing

There are two major types of adaptive testing: item-adaptive, versus stage-adaptive. In the former, each individual test question the student receives is chosen based on how the student performed in the earlier question. If the student did not correctly answer the earlier question, the student is given an easier question next. If the student correctly answered the earlier question, that question is followed by a relatively more difficult one. In the case of stage-adaptive testing, each student takes two or more sets of items, each set corresponding to a "stage". The set of questions the student receives is chosen based on how the student performed in the earlier set. If the student performs well on the earlier set (based on a preset criteria), the student receives a relatively more challenging set of questions next. If not, the student receives a relatively less challenging set. This design is also known as Multi-Stage Testing (MST).

Each adaptive type design is appropriate for different applications. Item-adaptive testing is particularly attractive when the focus is on individual student scores, as opposed to group level results. MST, on the other hand, is particularly suitable for measuring a construct with a large number of content objectives such as NAEP. This is because MST allows more control over the presentation of items across a finite number of test forms, which means better construct coverage and easier quality control.

Excerpt from advanced materials for COSDAM August 2021 meeting

Would you include the following item in a fourth-grade assessment?

1+1= …

How about this one?

Solve for x, where $\log_x 81 = 4$

Obviously, the answer is no in both cases. Setting aside the fact that these items would not be measuring skills in a fourth-grade assessment framework, the items would not provide any "information" about a (typical) fourth-grader's mathematics "ability." There is not a good alignment between the student ability and these two (hypothetical) items; you already know how the student would perform on these items. This example is to illustrate that items should not be too difficult, nor too easy for the students—they need to be … "just right!" In fact, the level of "information" an assessment provides is proportional to the degree of alignment between student ability and item difficulty. The most efficient way to achieve such alignment is through adaptive testing, where items are selected for the student in a way that their difficulty match the student's "ability."

The basic premise behind adaptive testing is that more psychometric information is gathered when we avoid giving students test questions that are either too easy or too difficult for them. Given that more psychometric information corresponds to smaller measurement error, adaptive testing holds the promise of better measurement, especially for the lowest and highest achieving students. In the case of NAEP, given the relatively large percentage of students performing in the below *NAEP Basic* level, the major reason behind the interest in adaptive testing has been the potential **improvement in measurement at the lower end of the score distribution**. In addition, adaptive tests also elicit **better student motivation** during test-taking, because students are presented items at their most appropriate range of difficulty, eliminating discouragingly difficult items or items that are too easy.

Two-subject design

The main advantage of a two-subject design in mandated reading and mathematics assessments is that this design would lead to substantial reductions in the combined sample size for these two assessments, creating **significant cost savings**. In addition, the design would allow **analysis of the relationships between student performance in these two subjects**. Other large-scale assessments, which administer longer tests to students, test each student in more than one subject. This includes TIMSS (mathematics and science for 72 minutes at grade 4 and 80 minutes at grade 8), PISA (mathematics, reading and science, and additional optional innovative domain, totaling 120 minutes for 15-year-old students), and PIAAC (literacy, numeracy, and digital problem-solving, 60 minutes). NAEP will study this design change in 2026 in conjunction with adaptive testing.

Conclusions

As part of our exploration of these two design innovations, NCES will be conducting two studies in 2026: one examining a two-subject design under a linear testing condition, and one examining the same under an adaptive testing condition (more specifically MST). There are three potential outcomes to the 2026 studies:

1. Two-subject design under adaptive testing is successful in terms of operational feasibility and psychometric quality. In this case, a bridge study will be conducted in 2028, where a portion of the sample takes the assessment under the current linear, single-subject design, whereas the rest takes the assessment under the two-subject adaptive design. The trend reporting will be maintained through this bridge study.
2. Two-subject design is not successful under adaptive testing, but it is successful under linear design. In this case, a bridge study will be conducted in 2028, where a portion of the sample takes the assessment under the current linear, single-subject design, whereas the rest takes the assessment under the two-subject linear design. The trend reporting will be maintained through this bridge study.
3. Two-subject design is not successful under either adaptive or linear design. In this case, there will be no design changes in 2028 assessments.

It is worth noting that the samples for these studies will not be part of the operational sample. 2026 results will be reported entirely based on the current linear, single-subject design.

National Assessment of Educational Progress (NAEP) Automated Scoring Challenge

This past fall, NCES held its first automated scoring challenge to score constructed response items for the National Assessment of Educational Progress's reading assessment. The purpose of the challenge was to help NCES determine the existing capabilities, accuracy metrics, the underlying validity evidence, and costs and efficiencies of using automated scoring with the NAEP reading assessment items. The Challenge required that submissions demonstrate interpretability of models, provide score predictions using these models, analyze models for potential bias based on student demographic characteristics, and provide cost information for putting an automated scoring system into operational use.

The challenge was announced and posted on Challenge.gov.

Start data: 9/16/2021
End date: 11/28/2021

A Request for Information Webinar was held 10/4/2021. Approximately 50 persons attended.

25 teams registered for the challenge, submitted the required non-disclosure agreements and requested data. Teams included commercial entrants, university teams, and independent teams. 17 teams were domestic, 8 were international. While the majority of the teams were comprised of graduate-level data scientists and statisticians, one local team included a high school student.

Description

Automated Scoring using natural language processing is a well-developed application of artificial intelligence in education. Results are consistently demonstrated to be on-par with the inter-rater reliability of human scorers for well-developed items (Shermis, 2014). Currently, the National Assessment of Educational Progress (NAEP) makes extensive use of constructed response items. Annually, contractors assemble teams of human scorers who score millions of student responses to NAEP's assessments.

This challenge sought to ascertain whether a wide array of automated scoring modes could perform well with a representative subset of NAEP Reading, constructed response items administered in 2017 to students in grades 4 and 8. The ultimate goal was to produce reliable and valid score assignments, provide additional information about responses (e.g. length, cohesion, linguistic complexity), and generate scores more quickly while saving money on scoring costs.

There were two components to this challenge; entrants could submit responses to one or both of these components:

1. Component A - Item-Specific Models: Respondents were asked to build a predictive model for each item that could be scored, using current state-of-the-art practices in operational automated scoring deployments. Extensive training data from prior human scoring administrations was provided. The first-place prize for this challenge is $15,000, with up to 4 runner-up prizes of $1,250 each.

2. Component B - Generic Models: Respondents were asked to build a generic scoring model that could score items that were not included in the training dataset, but were from the same administration, subject, and grade level. The prize for this challenge is $5,000, with up to 4 runner-up prizes of $1,250 each.

Participants were provided access to digital files that contain information related the results of human-scored constructed responses to reading assessment items that were administered in 2017, including

item text, passages, scoring rubrics, student responses, and human assigned scores (both single and double scored). The responses correspond to items that accompany two genres of 4th and 8th grade reading passages, literary and informational. Items for this challenge are of two response formats, short and extended.

The data set included 20 items for the item-specific models, and 2 items for the generic models. There was an average of 1,181 double-scored responses per dataset. These were divided into a training dataset (50%), a validation dataset (10%), and a test dataset (40%). The validation dataset was augmented with a much larger number of single-scored responses (average 23,000 per item).

In addition to model accuracy compared to human scorers, successful respondents to this Challenge had to provide documentation of model interpretability through a technical report that was evaluated by NCES's team of scorers for transparency, explainability, and fairness. The documentation was evaluated before respondents' scored submissions were evaluated. Only documentation that met acceptance criteria were considered as valid submissions and evaluated for accuracy of the predicted scores compared to the hold-out test dataset. The Federal Government is particularly interested in submissions that provided accurate results and met these objectives, as they have been absent from a good deal of recent research in automated scoring, particularly for solutions using artificial intelligence (e.g. neural networks, transformer networks) and other complex approaches (Kumar & Boulanger, 2020).

This process is consistent with the operational processes that the Department intends to use as part of the approval process for scoring and reporting; only models that can provide substantive validity evidence would be approved for operational use.

Of the 25 teams that registered, 15 completed the challenge and submitted the required work to be judged. Three submissions did not meet the acceptance criteria and were eliminated from competition.

On January 21l NCES announced that four teams had won top honors in the Challenge. They are *Measurement Incorporated, the University of Massachusetts-Amherst, Cambium Assessment, and the University of Duisburg-Essen.* In addition to awarding the four grand prizes, NCES recognized four runner-up teams, as well. The winners used advanced natural language processing methods that promise to reduce scoring costs while maintaining accuracy similar to human scoring.

Natural language processing uses computer algorithms to identify patterns in language; automated scoring applies these patterns to analyze student responses and assign scores. Those scores are then compared to the scores for each response given by human graders. The most accurate submissions used advanced machine learning approaches based in what are called "transformer network architectures" such as BERT (or "Bidirectional Encoder Representations from Transformers"). These models used NAEP data to fine tune pre-trained language models that were created by analyzing language consistencies and patterns among billions of student writing examples.

This challenge is a key component in modernization efforts to incorporate data science and machine learning into operational activities at NCES. It is the first in a series of challenges that use NAEP data.

Winners

Grand Prizes
Arianto Wibowo, Measurement Incorporated (Item-Specific Model)
Andrew Lan, UMass-Amherst (Item-Specific Model)
Susan Lottridge, Cambium Assessment (Item-Specific Model)
Torsten Zesch, University of Duisburg-Essen (Generic Model)

Runners-up
Fabian Zehner, DIPF | Leibniz Institute for Research and Information in Education, Centre for Technology-Based Assessment (Item-Specific Model)
Scott Crossley, Georgia State University (Item-Specific Model)
Prathic Sundararajan, Georgia Institute of Technology and Suraj Rajendran, Weill Cornell Medical College (Item-Specific Model)
Susan Lottridge, Cambium Assessment (Generic Model)

Potential Next Steps for Exploring a NAEP Below Basic Achievement Level

In recent years, some COSDAM members have expressed an interest in holding discussions to reconsider the Governing Board's current policy on the intentional exclusion of an official below NAEP Basic achievement level. The March 2022 COSDAM discussion will provide an opportunity to consider potential benefits and costs of adding a new achievement level.

The Governing Board's current policy concerning NAEP achievement levels is to *not* include an official achievement level below NAEP Basic. The justification has been guided by NAEP's intent. NAEP is intended to provide a snapshot of student performance in the United States on each of the subject areas assessed at the national, state, and select urban district level. NAEP is intended to be aspirational with the goal of measuring the percentage of students to achieve academically a solid understanding of the content for a given subject and grade. The assessments are designed based on NAEP frameworks, which are intended to reflect current educational requirements across the U.S. Unlike state assessments, NAEP does not provide individual student scores and is not intended to guide classroom instruction. However, in recent years COSDAM and other groups have engaged in discussions questioning whether NAEP would benefit from reconsidering the policy.

Arguments in favor of adding a NAEP Below Basic achievement level include a) the desire to better understand what students at the lowest end of the achievement scale know and can do, and b) the addition might help inform item development at the low end of the scale. Arguments against adding the achievement level include a) it may not be feasible to generate a description at the low-end at a level of detail that would prove useful for NAEP's intended purposes - assessment programs that include a low-end achievement level tend to word them in terms of what students cannot do or what they may be able to do. This may not be useful to NAEP because it is not reported at the individual student level, b) NAEP already provides item maps that show sample items across the entire range – including below NAEP Basic to gauge what these students can do, and c) adding new achievement levels would take time and resources away from other priorities.

To prepare for the March 2022 COSDAM discussion, key findings from recent panel discussions and evaluations regarding considerations for a Below Basic achievement level, or achievement levels in general, are summarized in this document. Links are provided for those wishing to review the complete documentation. Information and links are also provided for information regarding low-end achievement levels in state and international assessments.

2016 Evaluation of the Achievement Levels for Mathematics and Reading on the NAEP

The National Academies of Sciences, Engineering, and Medicine (NAS) conducted an evaluation of NAEP achievement levels in 2016 resulting in seven recommendations related to achievement levels. The National Assessment Governing Board responded with an Achievement Levels Work Plan. The work plan presents specific actions the Board might take in the coming years to address the recommendations – some of which are currently ongoing. The evaluators did

not identify a need to investigate adding a Below Basic achievement level; rather, they focused on the validation and interpretations of existing achievement levels. These recommendations are included for COSDAM members to consider when thinking about how work related to considering a Below Basic achievement level might fit within other priorities.

The key recommendations stemming from the 2016 evaluation are:

RECOMMENDATION 1 Alignment among the frameworks, item pools, achievement-level descriptors (ALDs), and the cut scores is fundamental to the validity of inferences about student achievement. In 2009, alignment was evaluated for all grades in reading and for grade 12 in mathematics, and changes were made to the ALDs, as needed. Similar research is needed to evaluate alignment for the grade-4 and grade-8 mathematics assessments and to revise them as needed to ensure that they represent the knowledge and skills of students at each achievement level. Moreover, additional work to verify alignment for grade-4 reading and grade-12 mathematics is needed.

RECOMMENDATION 2 Once satisfactory alignment among the frameworks, the item pools, the achievement-level descriptors, and the cut scores in National Assessment of Educational Progress mathematics and reading has been demonstrated, their designation as trial should be discontinued. This work should be completed and the results evaluated as stipulated by law: 20 U.S. Code 9622: National Assessment of Educational Progress (https://www.law.cornell.edu/uscode/text/20/9622 [September 2016]).

RECOMMENDATION 3 To maintain the validity and usefulness of achievement levels, there should be regular recurring reviews of the achievement-level descriptors, with updates as needed, to ensure they reflect both the frameworks and the incorporation of those frameworks in National Assessment of Educational Progress assessments.

RECOMMENDATION 4 Research is needed on the relationships between the National Assessment of Educational Progress (NAEP) achievement levels and concurrent or future performance on measures external to NAEP. Like the research that led to setting scale scores that represent academic preparedness for college, new research should focus on other measures of future performance, such as being on track for a college-ready high school diploma for 8th-grade students and readiness for middle school for 4th-grade students.

RECOMMENDATION 5 Research is needed to articulate the intended interpretations and uses of the achievement levels and to collect validity evidence to support these interpretations and uses. In addition, research is needed to identify the actual interpretations and uses commonly made by the National Assessment of Educational Progress's various audiences and to evaluate the validity of each of them. This information should be communicated to users with clear guidance on substantiated and unsubstantiated interpretations.

RECOMMENDATION 6 Guidance is needed to help users determine inferences that are best made with achievement levels and those best made with scale score statistics. Such guidance should be incorporated in every report that includes achievement levels.

RECOMMENDATION 7 The National Assessment of Educational Progress (NAEP) should implement a regular cycle for considering the desirability of conducting a new standard setting. Factors to consider include, but are not limited to, substantive changes in the constructs, item types, or frameworks; innovations in the modality for administering assessments; advances in standard setting methodologies; and changes in the policy environment for using NAEP results. These factors should be weighed against the downsides of interrupting the trend data and information.

2018 Governing Board Achievement Level Panel

In 2018, the Governing Board convened an Achievement Level Panel comprised of eight experts in education measurement and policy. A report summarizing the panel discussions is included with the COSDAM November 2018 Board materials. The purpose of this panel was to update the Board policy on achievement level setting in response to both the evaluation and current advances in the field of standard setting. Though the goal of the panel was not to focus on consideration of adding a NAEP Below Basic achievement level, it did come up in discussion, summarized here:

Experts engaged in extensive debate about whether the Governing Board should consider developing Achievement Level Descriptors (ALDs) for the below Basic category. Many states, but not all, do have descriptions for the lowest category of performance on their assessments. Most Experts felt it was not necessary for NAEP to develop ALDs for below Basic. Given the purpose of NAEP and the lack of individual scores, there are no student or teacher score reports. In addition, it is difficult to describe what students know and can do in the below Basic category because there is no policy definition describing what these students should know and be able to do, and because their performance can range from falling just below the Basic cut score to not answering any items correctly. The NAEP item maps do include items below Basic so this is an alternative way of representing what students at a given score point in the below Basic range are likely able to do.

Two Experts felt strongly that NAEP should contain descriptions for below Basic. Even though there are no individual scores on NAEP, they noted that some jurisdictions do have large numbers of students in the below Basic category. ALDs for this category could be written in terms of what students may be able to do, or could describe the midpoint of the category.

At the March 2018 COSDAM meeting members discussed this report and whether or not the Board should consider developing a policy definition and content ALDs for performance below the Basic achievement level. At that time, members did not see a compelling reason to develop a Below Basic description. They noted that it is difficult to develop an informative description

when the bottom of the category starts at zero; any statements would need to be in terms of what students sometimes or may be able to do. The NAEP item maps do include items below NAEP Basic and therefore provide some information about performance in this range.

2020 NCES Expert Panel Meeting on Performance Below the NAEP Basic Achievement Level

In 2020, the National Center for Education Statistics (NCES) convened a panel of nine experts with backgrounds in educational measurement, research, and policy or in curriculum and teaching in mathematics or reading. A full summary of the findings is presented in the COSDAM August 2021 Board materials. The purpose of this panel was to consider how NAEP might better serve students who perform below the NAEP Basic level.

The panel identified the following major recommendations:

1. *The panel recommended the development of achievement-level descriptors for students who perform in the score range below the NAEP Basic cut point by outlining what students at this level know and can do. (The panel did note that achievement-level descriptions and cut points are set by NAGB.) The panel believes that the NAEP framework needs to carefully describe the construct of measurement and skill progressions required across all of the achievement levels, including what is now described as below NAEP Basic. This recommendation also underscores the need to name the level that is below the NAEP Basic achievement level. Given the large range of scores below NAEP Basic, the panel also suggested giving consideration to including multiple levels below NAEP Basic, as is done in other largescale assessment programs, such as the Program for International Student Assessment (PISA). Naming the below NAEP Basic score range and providing descriptions of what students who perform at this level know and can do would enrich the reporting of NAEP.*
2. *The panel recommended that the distribution of items included in NAEP assessments correspond to the distribution of student ability, especially at the lower range. The current distribution of NAEP item difficulty is right-skewed and, therefore, lower performing students may become discouraged by what they see as inaccessible items. The panel suggested adding more items measuring the lower part of the NAEP scale so that the distribution of item difficulty more closely mirrors the entire distribution of student performance. The items of more appropriate difficulty will allow more precise measures of what students performing below NAEP Basic know and can do and add more insight into the performance of these students.*
3. *The panel recommended that the NAEP reporting emphasis on students who perform below the NAEP Basic achievement level should, at a minimum, match the reporting emphasis for the three current achievement levels (NAEP Basic, NAEP Proficient, and NAEP Advanced). In addition, the panel suggested that further contextual information about students who perform below NAEP Basic be collected from teachers and schools so that policymakers, researchers, and the general public have a more robust set of*

variables from which to gain an understanding of these students' educational performance.

Low-End Achievement Levels in State and International Assessments

In addition to the panels and evaluation presented above focused on NAEP, the inclusion and utility of a low-end achievement scale by state and international assessments may provide insight to COSDAM's discussions. Included are key points from a recent report developed by Karla Egan at the request of the Governing Board, and technical documentation related to achievement levels from the Trends in International Mathematics and Science Study (TIMSS) and Program for International Student Assessment (PISA) international assessments:

- Karla Egan produced a report *Describing the Lowest Achievement Levels* summarizing how state and international assessments define achievement levels at the lowest end of achievement scale. The report is available with the COSDAM August 2021 Board materials. Key points include:
    - In general, states include an achievement level description at low end of the achievement scale; however, they typically include only brief descriptions and/or define what students at this level *cannot* do, not what they *can* do.
    - Board staff reviewed documentation and had conversations with assessment staff in two states identified by the report as having relatively expansive achievement level descriptions at the low-end. These states noted the low-end achievement level descriptions were useful for informing student intervention and classroom instruction. They were developed primarily based on theoretical considerations (e.g., descriptions of what students at lower grade levels generally can do) rather than analysis of empirical performance.
    - Egan highlighted PISA and TIMSS international assessments include descriptions at the low-end of the scale. At the time of the report, PISA split the lowest level into two performance categories. Both assessments describe students at these lowest levels in terms of what they do know or can do.
- Exhibit 1.10 of the 2019 TIMSS report presents an example of a Low benchmark summary description for math at grade 4. The first statement reads: "*Students have some basic mathematical knowledge. They can add, subtract, multiply, and divide one- and two-digit whole numbers. They can solve simple word problems. They have some knowledge of simple fractions and common geometric shapes. Students can read and complete simple bar graphs and tables."*
- Chapter 14: Using Scale Anchoring to Interpret the TIMSS 2015 Achievement Scales of the TIMSS 2015 Methods and Procedures describes an effort to anchor items from the 2015 TIMSS assessments to each of the four achievement levels, including the Low level. For math, 43 of the 238 items (18%) at grade 4 anchored to Low, and only 4 of 209 (2%) anchored to low at grade 8. For science, 12 of 180 (7%) and 6 of 233 (3%) anchored to Low at grades 4 and 8, respectively. This indicates that thought TIMSS includes a Low achievement level description, the assessment includes only few items at this level for science and grade 8 math.

- Chapter 15: Proficiency Scale Construction describes the methodology used to develop the PISA reporting scales. PISA has defined three achievement levels for reading to describe what students at the low-end can do. The lowest of these achievement levels, 1c, is the least defined of the levels, and represents a very basic level of reading comprehension.
- PISA has addressed challenges with achievement differences between countries, specifically low-income countries with a high number of students falling at the low end of the scale, by developing a separate assessment - PISA for Development (PISA-D). The intent of PISA-D is to increase these countries' use of PISA assessments by creating a more meaningful scale for their populations.

COSDAM Discussion

During the March 2022 COSDAM meeting, Chair Suzanne Lane will provide background information and present goals and guiding discussion questions for considering an official Below Basic achievement level. The committee will discuss potential benefits, costs, and next steps for exploring this idea.