NATIONAL CENTER FOR EDUCATION STATISTICS
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Volume I

Supporting Statement

NAEP Survey Assessments Innovations Lab (SAIL)

Dynamic Assessments

Cog Labs and Medium Scale Study

OMB#1850-0803 v.275

October 2020

Table of Contents

Attachments:

Volume II – Dynamic Assessment Protocols
Appendices – Communication Materials

1) Submittal-Related Information

This material is being submitted under the generic National Center for Education Statistics (NCES) clearance agreement (OMB# 1850-0803), which provides for NCES to conduct various procedures (such as pilot tests, cognitive interviews, and usability studies) to test new methodologies, question types, or delivery methods to improve survey and assessment instruments and procedures.

2) Background and Study Rationale

The National Assessment of Educational Progress (NAEP) is a federally authorized survey, by the National Assessment of Educational Progress Authorization Act (20 U.S.C. §9622), of student achievement at grades 4, 8, and 12 in various subject areas, such as mathematics, reading, writing, science, U.S. history, civics, geography, economics, and the arts. NAEP is conducted by NCES, which is part of the Institute of Education Sciences, within the U.S. Department of Education. The primary purpose of NAEP is to assess student achievement in the different subject areas and collect survey questionnaire (i.e., non-cognitive) data to provide context for the reporting and interpretation of assessment results. As part of NAEP's development process, new or improved delivery systems and new innovative assessment items are pretested on smaller numbers of respondents before they are administered to larger samples in pilot or operational administrations. The NAEP Survey Assessments Innovations Lab (SAIL) initiative is a research program set up to explore the potential value to NAEP of developing and conducting design-related research studies to inform item development.

Traditional assessments measure construct mastery. In most models of learning, constructs pass through multiple stages of development:

1. Students are first exposed to concepts
2. Students can use those concepts with scaffolding and support (Zone of Proximal Development[1])
3. Students can use those concepts without scaffolding and support (Zone of Actual Development/mastery)
4. Students can use those concepts atomically (chunking[2])

In particular, there is strong theoretical basis that stage #2 is a better measure of students' level of development than stage #3. Traditionally, this second stage of development was measured through a 1:1 process known as dynamically-based assessment (DBA). However, prior to DBA, this was impossible to replicate in large-scale test settings. With the advent of intelligent tutoring systems, administration of scaffolded items became practical, but had prohibitively high item development costs for use in scalable assessments. In this study, we will begin to evaluate whether it is possible to make use of interactive, computer-delivered items and process data to conduct assessments which generate comparable psychologically-meaningful measurements as classical in-person dynamic assessments. As a stretch goal, we will investigate whether such assessments may help pinpoint specific knowledge gaps.

These assessments will consist of mathematics items, some of which include scaffolding resources, such as:

- Formula sheets and derivations of key concepts, terminology, and formulas students might have forgotten or not previously learned
- On-demand hints to support problem solving
- A means for checking steps throughout the problem-solving process
- Worked examples of similar problems

[1] See *Mind in Society*, (Vygotsky, 1978), for the cognitive theory behind the Zone of Proximal Development (ZPD) and Zone of Actual Development (ZAD).

[2] For an overview of chunking, see *How People Learn* (Bransford, 2000, pp. 32-33).

The primary goal of this study is to understand whether the item scaffolds are effective and what adaptations are necessary for these scaffolded items and supporting technology to measure the constructs required. A central purpose is developing pedagogical content knowledge (PCK) about what scaffolds are required. The research will consist of two main activities: 1) A set of cognitive interviews, and 2) a medium-scale study where students work on items at their own pace.

Cognitive Interviews

We plan to use three categories of items in the cognitive interviews. Items in all three categories have been encoded electronically and will be computer delivered during the interviews. The categories are:

1) Released Grade 8 NAEP mathematics items, with scaffolds;
2) Released Grade 8 NAEP mathematics items for which scaffolds have not been developed;
3) Non-NAEP mathematics items with scaffolds from a previous pencil-and-paper cognitive interview study.

The scaffolds for each item in the first category constitute a weak hypothesis in that they have not been previously tried with students. Consistent with an agile approach to development, we plan to revise these scaffolds in accordance with what we learn during the cognitive interviews so that they are clear and helpful, while maintaining construct relevance and measurement properties. In addition, the cognitive interviews may provide insight as to where additional scaffolds are needed. The second category of items, while not including scaffolds, may help us understand which scaffolds may be useful to develop, based on the aspects of the items with which students might struggle. In addition, these items may serve as a baseline measure of students' performance. The third category of items and scaffolds has already been through one round of testing and may shed light on issues associated with replacing human interventions by machine and with the delivery system.

In addition, the cognitive interviews will help us:

- Confirm and resolve basic UI/UX usability issues
- Confirm whether we are collecting process data correctly
- Collect student process data from a small sample, with the dual goals of
    - Gaining a qualitative understanding of students' cognitive processes when using scaffolds to respond to items
    - Informing both technology development/integration, and study plans for larger-scale collections

It is worth noting that this is the first time students will be interacting with this mode of digital-based assessment, and based on the researchers' prior experience, in most cases, multiple back-and-forth rounds of such testing are necessary to bring new types of educational technologies into a high-quality state, with each round raising successively the level of sophistication of the nuanced issues (e.g. the first round may show basic bugs in technology and the delivery of the items, while later rounds help polish subtle UX interaction issues).

Once item scaffolds are at a sufficient level of quality, we will conduct the second data collection activity associated with this project, which is a medium-scale study involving a larger number of scaffolded items in which we look at students' interactions with these scaffolded items without the support of an interviewer, either in a classroom setting (in person or virtual), or through the recruitment of larger numbers of students, depending on status of the current pandemic. We may alternate between the two modes, with each informing the other. The planned goals of this second part of the project are:

- Collect sufficient data to allow classification of student performance into categories in a prototype report. Theory predicts that student performance can be classified into three main categories: 1) unable to perform task, 2) able to perform task with scaffolding, and 3) able to perform task without scaffolding.
- Understand how students interact with the system without 1:1 support or supervision. Will students request all the hints available, even if they might be able to solve a problem without them? Will students simply not engage? Will students know how to manage their time, and will we be measuring the right construct? Does this tool work in classroom settings, whether virtual or in person?
- Qualitatively understand how students interact with scaffolding and identify common behavior patterns associated with each category listed in the first bullet. For example, it may be that students classified in the "able to perform task without scaffolds" category also tend to click briefly on vocabulary definitions before moving on, while students classified in the "unable to perform task" category make multiple successive attempts at solving each problem. Reviewing process data will be helpful for this sort of interpretation.
- Identify potential issues associated with reliability or validity, recognizing that larger studies will be necessary to properly characterize reliability, validity, and IRT item parameter estimations. In a medium-scale study, such as the one proposed in the second part of this project, we can still compute *preliminary* CTT statistics.
- Begin to understand the nature of transitions between the categories. As mentioned above, theory predicts three main categories: unable to perform task, able to perform task with scaffolding, and able to perform task without scaffolding. However, the data may not support the theory (e.g., a different number of classification categories may be identified). The data might also help determine the nature of the transitions between categories that are identified.
- Set the stage for future work. The data collected in the medium-scale study will be critical to structuring further stages of research needed and understanding the number of students required for more fully characterizing the nature of scaffolded items psychometrically.

3) Recruitment and Data Collection

Recruitment and Sample Characteristics

For the first study, ETS will recruit a maximum of 60 students in grades 6-10. Most likely we will recruit 30-40 students but may adjust based on results of the first study sessions, and number of rounds necessary. Students will be recruited primarily based on existing ETS subject contact lists, and/or through contacts with local student-focused organizations. These are primarily, but not exclusively, based in the New Jersey area, and because data collection is being done remotely, can be located anywhere in the country.

For the second study, recruitment of 300 students will be done primarily through student-focused organizations. Existing contacts are primarily in the New Jersey area, but outreach may include organizations in other parts of the country.

Separate intake forms will be used for the two studies. Recruitment for the Cog Lab study will begin before recruitment for the medium-scale study, though iterative rounds of recruitment for the Cog Lab study may take place concurrently with the medium-scale study. We anticipate using largely separate recruitment pools, and students will be eligible to participate in one or the other, but not both. Newly recruited students will be cross-checked against the link of past participants.

We will not have specific accommodations for students with disabilities included in this study. That is something that will be built out in future iterations of development. English Language Learner (ELL) students at an intermediate or higher level of proficiency will be invited to participate, but we are not specifically targeting them, and are not able to offer specific language accommodations at the time. While we believe this methodology will be beneficial for ELL students, who may understand mathematical concepts without knowing English mathematical terminology, we need to do this initial level of validation before we explore confirming that hypothesis.

Table 1. Sample Sizes

Respondent Group	Grades 6-10	Total
Students – Cognitive Interviews	60	**60**
Students – Medium-Scale Study	300	**300**
Overall Total	**360**	**360**

Data Collection Process

Cognitive Interviews

Students

Student cognitive interviews will be conducted via videoconferencing over Zoom to comply with social distancing mandates. Each interview will include an interviewer, and an observer will be present. Each cognitive interview session will last 90 minutes and will be scheduled by email to parents who have completed an online intake form to enroll their child. The questions students will see are web-based. Depending on technical limitations, the interviewer will either a) run the platform from their computer, and give the student control of the interface through Zoom[3], b) provide the student with a link to access the assessment from their computer and ask the student to share their screen, or c) run the platform from the interviewer's computer, but provide the student with links to click on to view video-based support materials.

Each participant will first be welcomed by staff, introduced to the interviewer and the observer, and told that he or she is there to help researchers understand how students interact with challenging mathematics content with the help of supporting materials. Then, the interviewers will explain the cognitive interview process.

Protocols for cognitive interviews will include probes for use as students work through item sets and probes for use after students finish answering items (see Volume II). Probes will include a combination of pre-planned questions identified before the session and ad hoc questions that the interviewer identifies as important from observations during the interview, such as clarifications or expansions on points raised by the student. For example, if a student paused for a long time over a particular item, appeared to be frustrated at any point, or indicated an "aha" moment, the interviewer might probe these kinds of observations further, to find out what was going on. In particular:

- If students are at an impasse for an extended period-of-time, the interviewer will attempt to scaffold students to be able to complete that item (and note what scaffolds were provided).
- If students are unable to find or navigate to scaffolds in the system, the interviewer will help the student navigate the UX (and note the UX issue).
- If students use the system incorrectly (e.g., view all the hints before trying), interviewers may intervene (and make note of the intervention).
- Generically, if students run into other issues, interviewers will attempt to intervene and to resolve them.

The welcome script, cognitive interview instructions, and probes for the interviewers are provided in Volume II.

The cognitive interviews will be recorded using screen capture and audio. Due to Zoom functionality, video may also be recorded, but will be immediately erased. However, to protect confidentiality, neither students nor interviewers will appear on the screen capture. Interviewers will also record their own notes separately, such as behaviors (e.g., "the participant appeared confused"), questions posed by students, which support materials are used, and observations of how long various items take to complete.

The types of data collected will include:

- process data from students completing the items (e.g. which scaffolds students navigated to, use of checkers /sub-checkers, etc);
- student reactions to and responses to items;
- responses to generic questions;
- responses to targeted questions specific to the item(s);

[3] For technical details on Zoom remote control feature, see https://support.zoom.us/hc/en-us/articles/201362673-Requesting-or-giving-remote-control

- additional volunteered participant comments; and
- answers to debriefing questions.

The principal goal of this study is to validate and improve scaffolds for released Grade 8 NAEP mathematics items. It is worth noting that this will be the first-time students have interacted with the dynamic assessment system and with the scaffolded NAEP released items. Based on what we learn interacting with students, during the course of the study we may transition to:

1. Items with limited or no scaffolds. For these items, the primary goal is to develop pedagogical content knowledge (PCK) about what scaffolds would be required.
2. Non-NAEP items previously tested in pencil-and-paper cognitive interviews. These are complex multi-concept non-NAEP items. In a cognitive interview protocol, students were provided with items, paper-based scaffolds, and in-person scaffolds, and these items now have digital scaffolds informed by that experience.

Prior to data collection, basic demographic information about students will be collected as part of the intake form.

Medium-Scale Study

Students

Students will respond to the scaffolded items on a web-based platform during a single session. Each student will have access to a list of items. Students will participate individually. Parents will be sent a log-in link for students to access, and students will be instructed to spend 45 minutes completing as many items as they have time for. Responses and process data will be recorded by the system.

Prior to data collection, basic demographic information about students will be collected as part of the intake form.

Analysis Plan

Cognitive Interviews

The primary purpose of the cognitive interviews is to support the item and technology development processes. In particular:

- Basic item validation: Interviewers may note places where scaffolds and items provided may be unclear to students, or contain errors, and we plan to use this data to correct those errors.
- Scaffold creation: Both for items with and without scaffolds, we plan to identify common student knowledge gaps and misconceptions pertaining to these items, and to use these findings to either create or improve scaffolds. We want to avoid scaffolds which allow students to solve items without understanding (e.g., by pattern matching), but to include enough scaffolds that students who would be capable of solving such items in a traditional dynamic assessment can also do so here.
- Process data validation: Verify that the process data collected from the system is technically correct and analyzable. We plan to look over the process data, and confirm that we can extract basic surface features, such as which scaffolds were used and for how long.
- UI/UX validation: Interviewers will note where students ran into issues or bugs, and we will use that to resolve issues in the system.
- Preliminary analysis of how cognitive processes (as demonstrated in the interviews) correspond to process data. Are our hypotheses for each item about how students might behave correct?
- Demographic information collected in the intake form will be used to describe the sample.
- Math course and grade information collected on the intake form may be used to help guide interpretation of the data.

Medium-Scale Study

The purpose of the medium-scale study is to understand how to classify student performance when using scaffolds, as well as how students use the scaffolds. To gauge how incorporating scaffolds may affect validity and reliability, we

will compute classical item statistics. Using a qualitative approach, the process data will be mined to address the following questions:

- Can student performance be classified into three categories, as theory suggests, or do the data support a different number of performance categories? What are the features of each of the categories identified?
- Do students use scaffolds in a way that supports learning? In other words, if a student answers an item incorrectly, do they request only a subset of scaffolds, or do they request all of them rapidly? This will be investigated by examining how attempts are interleaved with scaffold requests.
- What do the data suggest about planning for a large-scale study in the future?
- Which scaffolds are used, how often, and for how long with each item? How rapidly do students request scaffolds and hints versus trying without them? How many attempts do students need on checkers and sub checkers?
- Demographic information collected on the intake form will be used identify any significant gaps in the sample, but we do not anticipate having sufficient sample size for any comparison between groups.
- Math course and grade information collected on the intake form may be used to help guide interpretation of the process data.

4) Consultations Outside the Agency

Educational Testing Service (ETS) is the item development, data analysis, and reporting contractor for NAEP and will develop the items and scaffolding resources, analyze results, and draft a report with results. ETS research scientists will recruit participants and administer the mathematics assessment.

5) Justification for Sensitive Questions

Throughout the item and interview protocols development processes, effort has been made to avoid asking for information that might be considered sensitive or offensive.

6) Paying Respondents

Cognitive Interviews

Students will receive a $40 virtual gift card in return for their participation.

Medium-Scale Study

Students will receive a $25 virtual gift card for their participation.

7) Assurance of Confidentiality

The study will not retain any personally identifiable information. Prior to the start of the study, participants will be notified that their participation is voluntary and that all of the information they provide may be used only for statistical purposes and may not be disclosed, or used, in identifiable form for any other purpose except as required by law (20 U.S.C. §9573 and 6 U.S.C. §151).

Cognitive Interviews

Before students can participate in the study, written consent will be obtained from the parents or legal guardians of students less than 18 years of age. Participants will be assigned a unique student identifier (ID) by ETS, which will be created solely for data file management and used to keep all participant materials together. The participant ID will not be linked to the participant name in any way or form. The consent forms, which include the participant name, will be separated from the participant interview files, secured for the duration of the study, and will be destroyed after the final report is released. Cognitive interviews may be audio recorded with screen capture of their interactions with

the assessment. Depending on the platform used, video capture may be necessary in order to record the audio but will be immediately deleted. No names or faces will appear on the screen capture. The only identification included on the files will be the participant ID. The recorded files will be secured for the duration of the study and will be destroyed after the final report is completed.

Medium-Scale Study – Outside of the normal school day

Before students can participate in the study, written consent will be obtained from the parent or legal guardian of students less than 18 years of age. Participants will be assigned a unique student identifier (ID) by ETS, which will be created solely for data file management and used to keep all participant materials together. The participant ID will not be linked to the participant name in any way or form. The consent forms, which include the participant name, will be separated from the participant data files, secured for the duration of the study, and will be destroyed after the final report is released.

8) Estimate of Hourly Burden

Note that due to the iterative, design-based nature of the study, participant burden may change depending on requisite numbers of students to move onto the following stage.

Cognitive Interviews

The estimated burden for recruitment assumes attrition throughout the process.[4] All student cognitive interviews will be scheduled for no more than 90 minutes. Table 3 details the estimated burden for the cognitive interviews.

Medium Scale Study – Out of School Administration

The anticipated total number of student participants in this scenario for the experimental study administration is 300. While students would not participate during class time in this scenario, our recruitment would likely begin with outreach to student-oriented organizations to ask for their assistance sharing information about the study. Initial contact with organization staff will involve the reading of the recruitment email (see Appendix B), which we estimate at 3 minutes or (0.05 hours), and any follow-up communication with ETS as well as time spent sharing the student recruitment letter with their contact lists. We have estimated this at 60 minutes or 1.0 hour. Organizations will be provided with the student-focused recruitment letter to share (Appendix D). We have estimated the reading of this letter to be 5 minutes or 0.08 hours. We anticipate contacting 40 organizations, and that approximately 15 of these will agree to share study information with students. Time for parents to respond if they are interested is estimated at 9 minutes, or 0.15 hours, and time for parents to complete and return the consent form is also estimated at 9 minutes. Students will be able to complete all parts of the study in 45 minutes or 0.75 hours.

[4] Based on our experiences in other similar NAEP studies, the estimated attrition rates for direct student participant recruitment are 33 percent from initial contact to follow-up, 50 percent from follow-up to confirmation, and 40 percent from confirmation to participation for students. We estimate the attrition rates for direct adult participant recruitment for this study are 33 percent from initial contact to follow-up, 20 percent from follow-up to confirmation, and no attrition from confirmation to participation. The estimated attrition rate for the initial youth organization contact for student identification is 25 percent from contact to follow-up.

Table 2. Estimated Hourly Burden for Students and Parents for Dynamic Assessment Activities

Respondent	Number of Respondents	Number of Responses	Hours per Respondent	Total Hours
Cognitive Interviews				
Recruitment				
Student Recruitment via Youth Organizations				
Initial contact	30	30	0.05	2
Follow-up and identify students	20[+]	20	1.0	20
Sub-Total	**30**	**50**		**22**
Parent or Legal Guardian for Student Recruitment				
Initial contact	120	120	0.05	6
Follow-up via e-mail/web form	90[+]	90	0.15	14
Consent and confirmation	60[+]	60	0.15	9
Sub-Total	**120**	**270**		**29**
Participation				
Cognitive Interviews				
Students[+]	60**	60	1.5	90
Parents[+]	60***	60	1.5	90
Sub-Total	**120**	**120**		**180**
Total for Cognitive Interviews	**150**	**440**		**231**
Medium-Scale Study				
Recruitment				
Student Recruitment via Youth Organizations				
Initial contact	40	40	0.05	2
Follow-up and identify students	15[+]	15	1.0	15
Sub-Total	40	**55**		**17**
Parent or Legal Guardian for Student Recruitment				
Initial contact	600	600	0.05	30
Follow-up via e-mail/web form	400*	400	0.15	60
Consent and confirmation	300**	300	0.15	45
Sub-Total	**600**	**1300**		**135**
Participation				
Medium Study				
Students[+]	300	300	.75	225
Sub-Total	**300**	**300**		**225**
Total for Medium-Scale Study	**640**	**1655**		**377**
Total for this submission	**790**	**2095**		**608**

[+] Strictly a subset of the recruitment initial contact group

**Figure represents the maximum of the range of participants.

***Figure represents the maximum number of parents; we anticipate most parents will not observe interviews.

Note: numbers have been rounded and therefore may affect totals.

9) Costs to Federal Government

The total cost of the study is $70,936 as detailed in Table 5.

Table 3. Costs to the Federal Government

Activity	Provider	Estimated Cost
Design, prepare for, and administer cognitive interviews (including recruitment, data collection, data entry)	ETS	$34,115
Support payment of participants for cognitive interviews	ETS	$1,600
Support payment of participants for medium-scale study*	ETS	$7,500
Design and prepare for medium-scale study; recruit and administer medium-scale study (including individual recruitment, data collection, data management)	ETS	$27,721
Total		$70,936

*Amount indicated represents the maximum payment based on the assumption of paying individual students rather than schools

10) Project Schedule

The schedule for this study, including all activities, is provided in Table 5.

Table 4. Project Schedule

Activity *Each activity includes recruitment, data collection, and analyses*	**Dates**
Recruitment Cognitive Interview Activity	October 2020 – May 2021
Cognitive Interview Activity	October 2020 – June 2021
Medium Scale Study Recruitment	November 2020 – May 2021
Medium Scale Study	November 2020 – June 2021