NATIONAL CENTER FOR EDUCATION STATISTICS
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

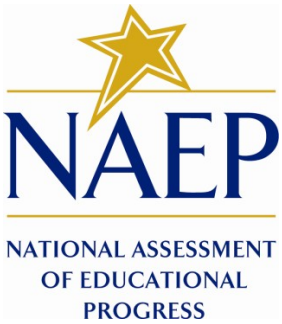
Volume I

Supporting Statement

NAEP Survey Assessments Innovations Lab (SAIL)

Test Assembly Experimental Study

OMB#1850-0803 v.272



September 2020

Table of Contents

Attachments:

Volume II – Math Assessment and Survey Item 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 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, part of the Institute of Education Sciences, in the U.S. Department of Education. NAEP's primary purpose is to assess student achievement in the various subject areas and to also 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, systems of delivery and 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 conducting design-related research studies to inform item development.

Low test taking motivation has serious implications for the validity of test outcomes, particularly for tests with low stakes for individuals (i.e., no personal consequences for the test taker). Methods to minimize the occurrence of low test taking motivation often require additional development and resources (e.g., technology-enhanced items, real time monitoring of response behaviors, payment incentives) that make their implementation challenging at scale. But what if there was a simpler method? We propose that test assembly (i.e., item order) could be a simpler method to increase test taking motivation as it would not require additional development of items or technology. There are a variety of test assembly strategies that could be utilized to address issues of low test-taking motivation (e.g., organization based on item type, content area, or difficulty). However, there is still a need for foundational research that further specifies the existing theory of test taking motivation. In the Demands Capacity Model of test-taking motivation (Wise & Smith, 2011), there are two key sources of variation to predict whether or not motivated responding will occur at the item level. Item demands are related to the characteristics of the items (e.g., difficulty, position, format) that can cause perceived increases in the effort required to respond correctly. Student effort capacity is related to the characteristics of the students that impact their overall motivation to respond effortfully on a test and how that motivation may change as the test proceeds. Although there has been consistent research on test-taking motivation, much of that work has not specified what item and student characteristics are most relevant to test-taking motivation and how those characteristics impact the design of a test to promote motivated responding. Thus, a main goal of the proposed study is to provide evidence to further build the Demands Capacity Model and inform test design decisions to promote motivated responding. Specifically, the proposed study will address the gap in knowledge of:

a. How does motivated responding vary for the same item when it occurs in different item positions?

There is a rich literature on the impact of the same item occurring in different positions from the perspective of item difficulty parameters (e.g., Borghans & Schils, 2012; Debeer & Janssen, 2013; Debeers et al., 2014; Meyers et al., 2008; Zamarro et al., 2019). These studies have found that when the same item occurs earlier in a test it is generally easier than when it occurs later in a test. This finding has been attributed to reductions in motivation and/or fatigue, but there is a lack of research that has directly tested this hypothesis. In addition, while item type and format have been proposed to impact this relationship between item position and motivated responding (Weirich, Hecht, Penk,

Roppelt, & Böhme, 2017), there has only been one study to actually test this hypothesis (Kingston & Dorans, 1984). However, this type of research has not been expanded to new innovative item formats (e.g., technology-enhanced items). Thus, we plan to also explore the interaction of item format and position on motivated responding in the proposed study. This analysis could reveal whether some item formats are more or less resilient to item position effects.

b. How does motivated responding vary for the same block when it occurs in different positions?

Similar to the same item occurring in different positions, there has been research on the impact of block position on item difficulty parameters. However, there has not been direct research on the impact of block position on motivated responding. Thus, the proposed study will address this gap in our current understanding of the impact of test design decisions on test-taking motivation. Of particular interest is the fact that blocks break up a test that would typically be 32-34 items long into two segments of 16-17 items each in NAEP math administrations. This division of the test has the potential to impact test-taking motivation. Prior research has found general trends that motivated responding reduces as the test progresses (e.g., Wise, 2006; Wise, Pastor, & Kong, 2009). However, this research has typically consisted of tests with 30 items or more. Thus, we are also interested to look at the potential changes in motivated responding within a block when it occurs in position 1 or 2. For example, it may be the case that creating a "break" in the test reduces the impact of item position on motivated responding in general or could replenish students' effort capacity at the start of a new block.

c. How does item order impact motivated responding in a free navigation testing environment?

To our knowledge, there has only been one study (Wise & Gao, 2017) that has investigated motivated responding in the context of a free navigation environment, but this study did not explore item position effects. A free navigation environment provides students with the ability to navigate through items (within a block) as they please, which provides choice and autonomy during the testing experience. Choice and autonomy have long been identified as beneficial to motivation and engagement (e.g., Cordova & Lepper, 1996; Deci & Ryan, 1985). It is possible then that the design decision to allow free navigation may have a substantial impact on motivated responding. Thus, we plan to explore how motivated responding occurs within a free navigation environment, whether or not item position effects occur, and how navigation behaviors (e.g., jump around from item-to-item vs. complete items in prescribed order) relate to motivated responding.

Another issue that free navigation and non-required responding present to the study of test-taking motivation is "how should motivated responding be operationally defined?" Typically, motivated responding is operationalized as individual item response times that are above a set time threshold (Wise & Kong, 2005; Wise & Smith, 2016). Conversely, unmotivated responding is operationalized as response times that are below a set time threshold that is meant to represent responses that are too quick for the student to have read the entire item and thoughtfully considered the correct response (i.e., rapid guessing). However, the use of response time thresholds to classify individual responses as motivated or unmotivated was developed and evaluated in testing contexts that typically include only single-selection multiple-choice items, fixed navigation, and required responses. The NAEP administration scenario breaks with all three of these design constraints. This creates a need to understand how motivated responding is operationalized in this type of testing environment. We view this as an opportunity to leverage the rich process data that is afforded by innovative item types, free navigation, and non-required responding to better understand what motivated and unmotivated responding looks like in practice. While this work is exploratory, we are leveraging the knowledge of other research areas to serve as a guide for how to deal with the variety of behaviors that occur (e.g., multiple item visits, skips, omits). Specifically, we will use the framework adopted in eye-tracking research as an initial guide for our exploration (Rayner, 1998).

We plan to address these gaps in knowledge in a study with approximately 390 students in 8th grade. Students will complete a math assessment that is presented in a testing environment that mirrors the functionality of the NAEP interface and the administration of a NAEP assessment. All process data will be captured (e.g., response time, navigation behaviors) and used to investigate the relationship between test assembly and test taking motivation. The outcome of this study will lay the groundwork for future studies in which we investigate the impact of alternative methods for test assembly (e.g., grouping or spacing based on item content or difficulty) on test taking motivation.

3) Recruitment and Data Collection

Recruitment and Sample Characteristics

Students would complete the study outside of normal school hours (e.g., afternoon, evening, weekends) at their homes. The normal school day alterations due to Covid-19 may cause schools to be unwilling or unable to include research studies during the normal school day. Thus, we are preparing for a scenario in which we will need to recruit students individually. In this scenario, ETS will recruit a maximum of 390 8th grade students who are currently enrolled in a mathematics course.

Our goal will be to recruit a diverse sample of students that represents a mixture of demographic targets (e.g., gender, race/ethnicity, free/reduced lunch eligibility). However, recruitment may be more challenging due to the challenges surrounding the COVID-19 outbreak and thus we will not make it a requirement that the recruited sample have specified representation of various demographic subgroups.

Data Collection Process

The study will be conducted online with 8th grade students on their own outside of the normal school day. Due to COVID-19 outbreak concerns we have designed the study in such a way that it does not require an ETS staff member to be present during administration. Thus, the administration of our study would not introduce any new conditions to the student participants that could increase the risk of contracting or spreading COVID-19 to students, their families, or the larger community.

The study will be completed by each student independently on a computer. Thus, the study will not require a proctor to be present either virtually or in-person. The study will require that students have an internet connection but will not require a particular browser or any type of software to be installed on the computer. All materials are designed to be completed within 50-minutes.

The study will involve three parts. The first two parts will be two blocks of math assessment items, with 15 math items in each block. Students will receive a brief video tutorial (approximately 3-minutes) on the assessment interface and available tools (e.g., calculator) prior to beginning the first block of math items. The video tutorial would also instruct students that their participation is voluntary, and they can choose to end their participation at any time during the study. The math assessment items that will be administered are aligned to the NAEP 2019 mathematics objectives for 8th grade students and provide a range of content and difficulty suitable for the target population for this study. Specifically, twenty-eight (93%) of the items align to the 8th grade objectives, with one of the two remaining items aligning to the 4th grade objectives and the other aligning to the 12th grade objectives. The interface and functionality for the two blocks will mirror NAEP administration, in that students can freely navigate between items and are not required to answer an item in order to proceed. Students will have 20-minutes to complete each block. In addition to the math items, students will also be asked one Likert item after the first time they visit each item. The Likert item will address how much effort students put into answering the item on that first visit. The decision to administer the effort Likert item after only the first item visit was done to reduce (a) the interruption to the typical test-taking experience and (b) the time needed to complete the study. However, it is important to note that while there may be missing data in the form of self-reported effort for later item visits, we will still have process data available for each item visit (i.e., visit time and behaviors). Thus, it will still be possible to categorize each item visit as effortful or not.

There will be six conditions in the present study that focus on test assembly (i.e., item organization within a block, block order). The 15 items within each block will be divided into three sets of 5 items each. The three sets of items will only be for study design purposes and will not be visible to the students. In other words, students will be able to freely navigate between all 15 items within a block. The order of sets within a block and order of blocks will be counterbalanced and randomly assigned to participants. Each participant will complete one condition, but all participants will complete the same 30 items. The only difference in participant experience across conditions will be the order in which items and blocks are presented.

After completing the second block, students will proceed to the post-test survey. The post-test survey will ask students to estimate their performance on the two blocks (% of correct responses), overall test difficulty, and overall effort exerted on the test. The post-test survey should take two to three minutes to complete.

The types of data collected will include:

- student responses to math items, Likert items, and post-test survey items;
- process data (e.g., response times, navigation in block, clickstream); and
- demographic information to be collected as part of the Intake Form by parents and legal guardians (see Appendix C).

Analysis Plan

There are multiple parts to our planned analyses:

1. Motivation difference between the test assembly conditions
    a. The motivation analyses will utilize multiple measures of motivation (e.g., effort Likert item, item visit time (in seconds), and effort scale from post-test survey).
    b. The comparisons between test conditions may include multiple parts (e.g., overall motivation level, temporal dynamics of motivated responding across a block, impact of block position on motivated responding, impact of individual item position on motivated responding).
2. Performance differences between the test assembly conditions
    a. The performance analyses will utilize multiple measures of performance (e.g., the quality of responses on individual items and the estimated performance on the post-test survey).
    b. The comparisons between test conditions may include multiple parts (e.g., overall performance, impact of block position on performance, impact of individual item position on performance).
    c. We plan to compare findings from the motivation analyses to those in the performance analyses.
3. Navigation differences between the test assembly conditions
    a. The navigation analyses will utilize multiple measures of navigation (e.g., sequence of item visits, item skips, item revisits).
    b. The comparisons between test conditions may include multiple parts (e.g., item visit sequence overall and within block, individual item visit behaviors).
    c. We plan to compare findings from the motivation analyses to those in the navigation analyses.
    d. We plan to compare findings from the performance analyses to those in the navigation analyses.
4. Sub-group differences between the test assembly conditions
    a. We plan to investigate differences in motivation, performance, and navigation across the six test assembly conditions in relation to student demographics (e.g., gender, race/ethnicity, SES).

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, analyze results, and draft a report with results. ETS research scientists will recruit participants and administer the materials for the proposed study.

5) Justification for Sensitive Questions

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

6) Paying Respondents

To encourage participation in a 50-minute administration session, students (*N*=390) participating after school hours or on the weekend will be offered a $20 virtual gift card as a thank you for his or her or their time and effort.

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).

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. Parents and legal guardians will be emailed ETS prepared consent forms (see Appendix D) and will be asked to scan and email the consent form to ETS project staff. Participants will be assigned a unique student identifier (ID), which will be created by ETS staff and solely for data file management and used to keep all participant materials together. The unique student ID will be structured such that there will be a code for the research project (TA), the study condition (A to F), and the individual student (100-500). For example, the ID TAA101 would represent a student (101) in this study (TA) in condition A. The participant ID will not be linked to the participant name in any way or form. A "crosswalk" roster (ID and student name only) will be created in order to pay students. Student PII is never connected to student data. This is ETS standard operating procedure when collecting student PII for payment purposes. 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

The estimated burden for recruitment assumes attrition throughout the process.[1] Table 1 details the estimated burden for the experimental study activities. Additional explanation for the time estimates is included below.

Table 1. Estimated Hourly Burden for Students and Parents or Legal Guardians for Test Assembly Experimental Study Activities

Respondent	Number of respondents	Number of Responses	Hours per respondent	Total hours
Organization Staff Member				
Flyer review	20	20	0.08	2
Parent or Legal Guardian				
Flyer review	1,000	1,000	0.08	80
Completion of online screening form	500[+]	500	0.15	75
Consent form completion and return	390*	390	0.13	51
Confirmation/acknowledgement to parent via email or letter	390*	390	0.05	20
Sub-Total	**1,000**	**2,280**		**226**
Student				
Pilot	390	390	0.83	324
Total Burden	**1,410**	**2,690**		**552**

[+] Subset of the initial contact group

* Subset with completed screening form

Note: These numbers are rounded to the nearest whole number in the Total Hours column.

[1] Based on our experiences in other similar NAEP studies, we estimate the attrition rates for direct student participant recruitment are 50 percent from flyer review to online screening form, 22 percent from online screening form to consent form, and no attrition from consent form to confirmation.

The anticipated total number of student participants in this scenario for the experimental study administration is 390. All time estimates for contact with parents and legal guardians are based on prior experience conducting NAEP studies. Recruitment for this study will occur across several online platforms (e.g., ETS internal website, social media, and direct email). Flyer review will involve the review of either only the Recruitment Letter/Email (Appendices A and B) in the case of direct or email or the review of both a Recruitment Flyer (Appendix G) and the Recruitment Letter/Email (Appendices A and B). We anticipate distributing 1000 flyers to parents and legal guardians, with an estimated time of 5-minutes or 0.08 hours (see Appendices A, B, and G). Time to fill out the online screening form is estimated at 9-minutes or 0.15 hours (see Appendix C). For those selected to participate and asked to fill out the consent form, the estimated time is 8-minutes or 0.13 hours (see Appendix D). The follow-up email to confirm participation for each student (see Appendix E) is estimated at 3-minutes or 0.05 hours. Students will be able to complete all parts of the study in 50-minutes or 0.83 hours.

In addition to direct recruitment of parents and legal guardians via online platforms, we will also contact organizations that provide services to students outside of normal school hours (e.g., Boys and Girls Club). These organizations will be contacted via email (Appendix F) and will be asked if they would be open to sending information about our study to the parents and legal guardians of their students. The time burden for parents, legal guardians, and students would remain the same as described above.

9) Costs to Federal Government

The total cost of the study is $29,124 as detailed in Table 2.

Table 2. Costs to the Federal Government

Activity	Provider	Estimated Cost
Design and prepare for study administration; administer study (including recruitment, data collection, data entry)	ETS	$21,324
Support payment of participants	ETS	$7,800

10) Project Schedule

The schedule for this study, including all activities, is provided in Table 3.

Table 3. Project Schedule

Activity *Each activity includes recruitment and data collection*	**Dates**
Preparation for experimental study	July – August 2020
Recruitment	September 2020
Communicate with parents and legal guardians to resolve any concerns for student participation (e.g., general process, technology concerns)	October 1 – 15, 2020
Data collection	October 15 - January 31, 2020

References

Borghans, L., & Schils, T. (2012). The leaning tower of pisa: Decomposing achievement test scores into cognitive and noncognitive components. Unpublished manuscript.

Cordova, D., & Lepper, M. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*, 715-730.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164-185.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, *39*(6), 502-523.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8*(2), 147-154.

Meyers, J. L., Miller, G. E., & Way, W. D. (2008). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education*, *22*(1), 38-60.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372-422.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, *41*(2), 115-129.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education*, *19*, 93-112.

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, *30*(4), 343-354.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Understanding correlates of rapid-guessing behavior in low stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*, 185-205.

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendal (Eds.), *High-stakes testing in education: Science and practice in K-12 settings* (pp. 139-153). Washington, DC: American Psychological Association.

Wise, S. L., & Smith, L. (2016). The validity of assessment when students don't give good effort. In G. Brown & L. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 204-220). New York, NY: Routledge.

Zamarro, G., Hitt, C., & Mendez, I. (2019). When students don't care: Reexamining international differences in achievement and non-cognitive skills. *Journal of Human Capital*, *13*(4), 519-552.