Study of the Feasibility of a NAEP Mathematics Accessible Block Alternative

Lizanne DeStefano
University of Illinois at Urbana-Champaign

Jeremiah Johnson
UMass Donahue Institute

August 2013
Commissioned by the NAEP Validity Studies (NVS) Panel

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

This report was prepared for the National Center for Education Statistics under Contract No. ED-04-CO-0025/0012 with the American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

2047_08/13

The NAEP Validity Studies (NVS Panel) was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Peter Behuniak
University of Connecticut

George W. Bohrnstedt
American Institutes for Research

James R. Chromy
Research Triangle Institute

Phil Daro
Strategic Education Research Partnership (SERP) Institute

Lizanne DeStefano
University of Illinois

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Robert Linn
University of Colorado at Boulder

Ina V.S. Mullis
Boston College

Scott Norton
Council of Chief State School Officers

Gary Phillips
American Institutes for Research

Lorrie Shepard
University of Colorado at Boulder

David Thissen
University of North Carolina, Chapel Hill

Karen Wixson
University of North Carolina, Greensboro

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Janis Brown
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
2800 Campus Drive, Suite 200
San Mateo, CA 94403
Phone: 650/843-8100
Fax: 650/843-8200

Executive Summary

This paper describes one of the first efforts by the National Assessment of Educational Progress (NAEP) to improve measurement at the lower end of the distribution, including measurement for students with disabilities (SD) and English language learners (ELLs). Given the need for NAEP to measure the full range of content and skills specified in the frameworks and achievement-level descriptions, the assessments have tended to include many items that students find difficult, and achievement estimates at the lower extreme of the distribution have had relatively large standard errors (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007). Lack of precision at the lower levels represents an important validity issue, however, particularly when NAEP is used as a means of benchmarking and interpreting change in state assessment results over time.

One way to improve measurement at the lower end is to introduce one or more "accessible" blocks into the NAEP Balanced Incomplete Block Design (BIB). Accessible blocks are defined as blocks that are aligned within the NAEP content frameworks, but designed to provide more information about the abilities and skills of students at the lower end of the distribution. The process of creating the NAEP accessible blocks described in this document began in February 2007 with efforts to define what constitutes an accessible block and to design a process to develop mathematics-accessible blocks that are representative of the NAEP content frameworks. Panels of mathematics content experts and item writers were convened to identify major themes and dimensions that affected item difficulty in mathematics and to develop general strategies for reducing difficulty without compromising content and construct validity. This process culminated in the creation of the *Item Modification Guidelines* and *Item Modification Procedures*.

Using the *Item Modification Guidelines* and *Item Modification Procedures*, a sample of Grade 4 and Grade 8 item blocks were modified to create accessible blocks. Two accessible blocks at each grade level, along with the original NAEP blocks from which they were derived (source blocks), were evaluated in a 2010 field test. The purpose of the field test was both to compare the modified blocks with their source blocks and to determine whether the modified items could be successfully placed on the NAEP scale.

This investigation of NAEP accessible blocks served as a proof of concept study in two important ways. First, the creation, application, and expert review of the *Item Modification Guidelines* and *Item Modification Procedures* illustrated that it was possible to develop standard procedures for creating items that were less difficult but still adhered to the content framework. Second, the results from the field test of accessible and operational NAEP blocks indicated that it is indeed feasible to construct accessible blocks that are scalable with the main NAEP assessment and that improve measurement precision at the lower end of the NAEP performance continuum.

The primary study findings were as follows:

- Across all groups and subgroups, there were substantial and similar average gains in the percentage correct by block for the accessible blocks compared with the source blocks.
- There were consistent declines in the number of students omitting items and significant reductions in the percentage of students not reaching items for the accessible blocks compared with the source blocks.

- All accessible items were scalable, and modified items had similar average discrimination and guessing characteristics (a and c parameter estimates) as the source items, while there were significant reductions in item difficulty (b parameter estimates).
- For the lowest performing students, the conditional standard error of measurement was significantly lower for students completing two accessible blocks than for those completing two source blocks.
- Test information functions for books comprised of two accessible blocks were appropriately targeted to the lower end of the performance continuum.

Since this study was conducted, the *Item Modification Guidelines* and *Item Modification Procedures* have been adopted by NAEP and are now routinely used in NAEP item development to improve the quality of all items, not only those intending to be made more accessible. In addition, the accessible block study in mathematics served as the impetus for additional research on improving measurement precision at the lower part of the distribution, including an accessible block study in reading (currently being conducted on behalf of the NAEP Validity Studies Panel [NVS Panel]) and the Knowledge and Skills Assessment (KaSA), an ongoing special study by the NAEP contractor for item development and analysis that has administered accessible blocks to students who would otherwise be excluded from NAEP.

CONTENTS

Background and Overview

This paper describes one of the first efforts by the National Assessment of Educational Progress (NAEP) to improve measurement at the lower end of the distribution, including measurement for students with disabilities (SD) and English language learners (ELLs). Significant numbers of students tend to perform below the *Basic* level on NAEP. For example, on the 2011 assessment, 18 percent of fourth graders performed below *Basic* in mathematics and only 40 percent performed at or above *Proficient*. Very small percentages reached the *Advanced* level. Furthermore, the percentages of students performing in the lower part of the distribution is much greater for many of the demographic groups on which NAEP is required to report by law.

Given the need for NAEP to measure the full range of content and skills specified in the frameworks and achievement-level descriptions, the assessments have tended to include many items that students find difficult, and achievement estimates at the lower extreme of the distribution have had relatively large standard errors (Daro et al., 2007). Lack of precision at the lower levels represents an important validity issue, however, particularly when NAEP is used as a means of benchmarking and interpreting change in state assessment results over time. If state assessment results are showing gains, but NAEP scores remain static for some demographic groups or subject areas, it may be due to NAEP's inability to detect change in the lower performance levels.

Under the current design, NAEP items are organized into blocks, assembled into two-block books, and administered using a Balanced Incomplete Block (BIB) book design. BIB is a complex variant of matrix sampling in which items are administered so that each pair of items is dispensed to a nationally representative sample of respondents in a specific pattern. One way to improve measurement at the lower end is to introduce one or more "accessible" blocks into the NAEP BIB. Accessible blocks are defined as blocks that are aligned within the NAEP content frameworks, but designed to provide more information about the abilities and skills of students at the lower end of the distribution.

Accessible blocks could also be paired with regular blocks and given selectively to students who were previously identified as likely to benefit. (For example, see McLaughlin, Scarloss, Stancavage, and Blankenship, 2005, in which a proposal for using state assessment scores to preassign books is discussed.) The inclusion of an "accessible book," consisting of two accessible blocks, also holds promise as a means for increasing the participation of SD and possibly also ELL students—thereby improving the validity of NAEP as a means of representing the performance of those subgroups. Offering an accessible book option to SDs and ELLs could also reduce the impact of construct irrelevant variance (e.g., readability, language demand, visual distractors) on test results for these subgoups.

The process of creating the NAEP accessible blocks described in this document began in February 2007 with efforts to define what constitutes an accessible block and to design a process to develop mathematics accessible blocks that are representative of the NAEP content frameworks. Panels of mathematics content

experts and item writers were convened to identify major themes and dimensions that affected item difficulty in mathematics and to develop general strategies for reducing difficulty without compromising content and construct validity. This process culminated in the creation of the *Item Modification Guidelines* and *Item Modification Procedures*. (See Appendix A for the final versions of these documents.)

To assess the viability of the guidelines and item modification procedures developed in 2007, seven accessible blocks were created for Grade 4 mathematics by modifying 2007 operational NAEP blocks, and two of these accessible blocks were administered in 2008 in a small pilot test (n=700 per block). The pilot test allowed comparison between the performance of the modified blocks in the pilot test sample and the performance of the parent blocks in the 2007 operational assessment. Results were sufficient to show that the modified items were in fact more accessible to students and motivated plans for a larger 2010 field test in which accessible blocks were again developed from operational NAEP blocks. The purpose of the field test was both to compare the modified blocks with their parent blocks and to determine whether the modified items could be successfully placed on the NAEP scale.

In preparation for the 2010 field test, the study authors conducted the following item development activities:

- Expert item review of 2009 operational items at Grades 4 and 8
- Creation of three more modified blocks of Grade 4 items and eight modified blocks of Grade 8 items (all based on 2009 operational blocks)
- Refinement of the modified blocks via cognitive labs with students (n = 4 per block)
- Expert review of modified items, the Item Modification Guidelines, and the Item Modification Procedures
- Selection of two modified blocks at each grade level for field testing in 2010.

Key Questions

The overarching purpose of the study was to explore the use of modified NAEP blocks as a means of improving measurement of the abilities and skills of students who score at lower end of NAEP performance continuum (including SDs and ELLs). More specifically, we endeavored to address the following three questions:

1. What process can be used to develop mathematics accessible blocks that are representative of the NAEP content frameworks?
2. Are accessible items easier than the unmodified source items?
3. Can accessible items be scaled along with unmodified NAEP items?

Developing Guidelines and Procedures for Preparing Accessible Blocks

The process for developing an operational definition of a mathematics accessible block began in 2007 by convening a panel of content experts**,** including representatives with expertise regarding special education and second language students. The purposes

were to (a) review fourth-grade NAEP items in mathematics and items from other sources such as the Voluntary National Test (VNT)[1] and state assessments, (b) identify construct relevant and irrelevant aspects of the items that contribute to their difficulty, and (c) offer suggestions for how to make the items easier without compromising content/construct validity or alignment with the NAEP framework. (See Appendix B for a list of expert reviewers who participated in 2007.)

The content expert panel assembled for mathematics did not directly address item alignment with the NAEP framework. Rather, it identified factors that increased the difficulty of particular items and proposed strategies for making each item easier without altering the construct being measured. The research team then analyzed the item-specific data generated from this process to identify major themes and dimensions that appeared to account for item difficulty. The next step was to use this information to develop general strategies for reducing difficulty without compromising content and construct validity. That is, the process led to a working model for accessible block construction.

A second expert panel of experienced item writers, special education and second language experts, and content specialists then used the guidelines to modify seven Grade 4 blocks from the 2007 operational assessment.[2] While carrying out its work, the second panel was asked to further develop the scope, clarity, and potential utility of the working model for accessible-block construction, and to examine the extent to which the guidelines provided were consistent with the NAEP framework. These guidelines were the starting point for the accessible-block development for the 2010 field test.

[1] The VNT was never administered operationally. However, a pool of items was developed and piloted tested, and items from this pool were available for analysis in 2007.

[2] Two of these blocks were subsequently evaluated in a 2008 pilot test.

Accessible Block Development

For the 2010 field test, the study group modified a selection of 2009 operational NAEP blocks in order to create paired accessible blocks.

Expert Review of 2009 Operational Items

As a first step in constructing the accessible blocks for the 2010 field test, the research team asked six outside reviewers to evaluate the quality of the mathematical content for items in each of the NAEP 2009 operational blocks proposed as source blocks. Each reviewer was a professor of mathematics, and four panel members had participated in the item review process that occurred during 2007. In addition to the four "veteran" reviewers, two additional reviewers were asked to participate in the panel. These new reviewers provided fresh insight into the item review process and further developed the capacity of the research team to replicate this type of work for future NAEP item review tasks. (See Appendix C for a list of expert reviewers who participated in 2009.)

Each expert reviewer was asked to do three things: (1) rate the mathematical accuracy of every item in each block using the "Item Rating Scale"; (2) comment on the extent to which each item block was congruent with the NAEP framework; and, (3) comment on whether or not each item block was in alignment with the *Item Modification Guidelines*.[3] The Item Rating Scale included values 1, 2, and 3 (with no fractions thereof). A score of 1 meant the item was adequate, a score of 2 meant the item was marginal or somewhat problematic, and a score of 3 meant the item was seriously flawed. (See Appendix D for a fuller description of this scale.)

As in 2007, the review process proved to be a critical step in the process of constructing blocks that were both accessible to the targeted student population and mathematically valid. It should be noted that the "veteran" reviewers were pleased that many of the general recommendations they had made in 2007 for improving the NAEP item pool were reflected in items and blocks under consideration during 2009.

Item Modification

The research team assembled a panel of 10 education professionals, mathematics content specialists, individuals with ELL/SD experience, and assessment specialists to evaluate—and modify as necessary—every item in each of the blocks being considered for inclusion in the 2010 field test. All panel members were required to demonstrate a strong understanding of mathematics and/or mathematics education. (The list of item modification panel members for 2009 appears in Appendix C.)

During most working sessions, the item modification panel was divided into two teams, balanced with respect to mathematical, educational, ELL/SD, and assessment expertise. Each team concentrated its efforts on a subset of the item blocks,

[3] A total of 14 blocks were considered for inclusion in the 2010 field test (eight blocks at Grade 8 and six blocks at Grade 4). Three of the six Grade 4 blocks had already been reviewed and revised during the 2007 analyses and were therefore excluded from the 2009 expert review and item modification.

systematically modifying items in their assigned blocks according to the *Item Modification Guidelines* and *Item Modification Procedures* developed during 2007. A member of the research team facilitated and closely monitored all aspects of the item modification process.

The item modification process largely occurred over a four-week period during March–April 2009 and required approximately 80 hours to complete. During this time, the item modification panel completed several tasks including:

1. Familiarizing its members with the goals of the study, NAEP frameworks, and initial ideas/definitions/strategies for creating accessible blocks.
2. Examining the feasibility and effectiveness of various processes for creating accessible blocks that are aligned with NAEP frameworks while further developing and refining guidelines and recommendations for the creation of accessible blocks.
3. Reviewing and adapting 11 2009 operational blocks by systematically varying items in ways intended to reduce difficulty and increase clarity.
4. Developing new items to replace items that could not be adequately modified.
5. Systematically reviewing, editing, and rating each of the 11 modified blocks to finalize draft accessible blocks suitable for cognitive lab and field-testing activities.
6. Providing recommendations regarding which blocks to include in cognitive lab and field-testing activities.

The item modification panel became more proficient and confident in applying the *Item Modification Guidelines* as its work progressed. In addition, the item modification panel made minor improvements to the *Item Modification Guidelines* and *Item Modification Procedures* to reflect its understanding of "best practice" as the work progressed.

The item modification panel carefully recorded and classified each of the modifications that were recommended for each item. Each modification was classified as being either construct relevant (i.e., directly affecting the level or content of the mathematics being assessed) or construct irrelevant (i.e., dealing with issues of format, context, or clarity). Changes to items were considered "construct relevant" if the modification made to the item was likely to impact the nature or difficulty of the original task. Tables 1 and 2 summarize the specific modifications made to each of the items at each grade level.

Table 1. Summary of Modifications Made to Grade 4 Items

Construct Relevant	75.4%		Construct Irrelevant	93.0%
Cognitive Demand	57.5%		Word Choice	51.1%
Graphics	27.7%		Cues	48.9%
Computational Appropriateness	21.3%		Formatting	25.5%
Context	12.8%		Graphics	21.3%
Alternative Answer Choices	10.6%		Alternative Answer Choices	4.3%
Item Format	4.3%		Computational Appropriateness	4.3%
Grade-Level Appropriateness	2.1%		Extraneous Information	4.3%
Cues	2.1%		Context	2.1%
Word Choice	0.0%			

Note: Percentages are based on total number of Grade 4 items modified during 2009 (*n* = 47). Table does not include data for Grade 4 items modified in 2007.

Table 2. Summary of Modifications Made to Grade 8 Items

Construct Relevant	89.6%		Construct Irrelevant	87.2%
Cognitive Demand	58.4%		Word Choice	42.4%
Graphics	40.8%		Formatting	35.2%
Alternative Answer Choices	24.8%		Cues	26.4%
Context	24.0%		Graphics	17.6%
Computational Appropriateness	19.2%		Alternative Answer Choices	6.4%
Cues	12.8%		Computational Appropriateness	6.4%
Item Format	2.4%		Extraneous Information	3.2%
Word Choice	0.8%		Context	3.2%
Grade-Level Appropriateness	0.0%			

Note: Percentages are based on total number of items in all modified Grade 8 blocks (*n* = 125).

Modifications to graphics, computational appropriateness, context, alternative answer choices, cues, and word choice could be either construct relevant or construct irrelevant; therefore, these categories appear twice in the tables. The following bullet points explain the distinctions between construct-relevant and construct-irrelevant modifications in these categories:

- *Graphics.* A construct-relevant change might include adding, deleting, or substantially altering a graphic provided in the original item stem or alternative answer choices. A construct-irrelevant change might include slight adjustments in graphic placement or content.
- *Computational Appropriateness.* A construct-relevant change might involve the reduction in the number of mathematical steps required to solve a problem. A construct- irrelevant change might involve the elimination of "ugly numbers" from the required calculations.

- *Context.* A construct-relevant change might typically involve removing the context of the problem, while a construct-irrelevant change might typically involve simplifying the description of the context.
- *Alternative Answer Choices.* A construct-relevant change might include, for example, the elimination of an answer choice or a substantial change to one or more of the alternative answer choices that were originally provided. A construct-irrelevant change might include, for example, changing the order in which the answer choices were presented.
- *Cues.* A construct-relevant change might involve the provision of a standard formula (diameter = $2\pi r$), while a construct-irrelevant change might involve bolding or underlining a key word or phrase.
- *Word Choice.* A construct-relevant change might involve, for example, changing one or more key words in an item, while a construct-irrelevant change might involve, for example, changing the tense in which the item is presented (from past tense to present tense).

At Grade 8, the modifications to answer choices included, in a small number of cases, reducing the number of answer choices from five to four. Also, at both grade levels, the format of a few items was changed from short answer to multiple choice. The latter modifications are classified under Item Format.

After the process of item modification was complete, four blocks at each grade level were identified by members of the item modification panel as potential candidates for field testing. These blocks were selected based on several criteria including the following:

1. Items within the block were made easier while retaining the integrity of the original testing objective(s).
2. Items within the block represented an appropriately diverse range of topics/skills in the NAEP framework.
3. Items within the block presented information in multiple ways (e.g., words, pictures, graphs, tables, figures) when appropriate.
4. The block, as a whole, reflected an appropriate and judicious application of the *Item Modification Guidelines*.

Cognitive Labs

The four candidate blocks at each grade level were next subjected to cognitive labs in order to gain insight into how students interpreted and responded to the items and blocks. During the cognitive labs, both an original NAEP block and the parallel accessible block were administered to each student using a counterbalanced design. The observer prompted students to "think aloud" as they completed the item blocks and debriefed students about strategies used once each block was completed. Student work was analyzed to identify strategies and evaluate performance. In total, 13 cognitive labs were conducted with Grade 4 students and 15 cognitive labs were conducted with Grade 8 students. All cognitive lab participants were selected from fourth- and eighth-grade classrooms in Champaign and Urbana, Illinois, in May 2009.

Across blocks and grade levels, students consistently scored higher and required less time to complete the accessible-block version of the assessment. Table 3 summarizes student performance on the Grade 4 blocks and Table 4 provides information regarding average time to completion. Table 5 summarizes student performance on the Grade 8 blocks and Table 6 provides information regarding average time to completion.

Table 3. Average Student Performance in Cognitive Labs—Grade 4

Block	Original Block % Correct	Accessible Block % Correct	Average % Gain
G4-1 (*n* = 3)	90.0%	94.0%	4.0%
G4-2 (*n* = 3)	79.3%	94.7%	15.4%
G4-3 (*n* = 4)	58.3%	84.8%	26.5%
G4-4 (*n* = 3)	33.3%	68.9%	35.6%

Table 4. Average Time to Completion in Cognitive Labs—Grade 4

Block	Original Block Average Time to Completion	Accessible Block Average Time to Completion	Average Time Reduction
G4-1 (*n* = 3)	13.7	12	1.7
G4-2 (*n* = 3)	12.3	6	6.3
G4-3 (*n* = 4)	16.5	9.5	7
G4-4 (*n* = 3)	16.3	11.3	5

Table 5. Average Student Performance in Cognitive Labs—Grade 8

Block	Original Block % Correct	Accessible Block % Correct	Average % Gain
G8-1 (*n* = 4)	50.0%	69.4%	19.4%
G8-2 (*n* = 4)	69.3%	75.0%	5.7%
G8-3 (*n* = 3)	66.6%	86.0%	19.3%
G8-4 (*n* = 4)	68.1%	88.9%	20.8%

Table 6. Average Time to Completion in Cognitive Labs—Grade 8

Block	Original Block Average Time to Completion	Accessible Block Average Time to Completion	Average Time Reduction
G8-1 (*n* = 4)	20.8	15.8	5.0
G8-2 (*n* = 4)	22.5	17.8	4.8
G8-3 (*n* = 3)	23.7	9.3	14.4
G8-4 (*n* = 4)	18.3	15.8	2.5

Analysis of cognitive lab data, including student performance data and time to completion data, support the research team's claim that the cognitive demand of the accessible-block version of each modified block was lower than the cognitive demand of the parallel original block.

Although NAEP does include blocks for which calculators are allowed, none of the calculator blocks were included in the accessible blocks study. It is interesting to note that in some cases students were able to accurately describe a proper strategy for completing an item, but were unable to do so because they did not have access to a calculator.

One block of Grade 8 items requiring the use of a manipulative was modified and included in the cognitive lab activities. Although cognitive lab student scores for this block were not substantially different from others, time to completion and accessibility for some students with disabilities raised concerns.

Student comments on item difficulty generally affirmed that the application of the *Item Modification Guidelines* served the purpose of making items more accessible. Student feedback regarding specific item features was reviewed and incorporated into the final version of the accessible items as appropriate.

Expert Review of Modified Items

Concurrent with the cognitive lab activities, the research team asked the expert review panel to evaluate the quality of the mathematical content of each of the items in each of the modified (accessible) blocks of NAEP items. Each reviewer was given the same instructions as were provided during the initial item review: (1) rate the mathematical accuracy of every item in each block using the "Item Rating Scale"; (2) comment on the extent to which each item block was congruent with the NAEP framework; and, (3) comment on whether or not each item block was in alignment with the *Item Modification Guidelines*.

Again, members of the expert review provided specific, rich information regarding the mathematical quality of modified blocks of items, and their feedback was incorporated into the final versions of the items as appropriate.

Selecting Blocks for Field Testing

Once cognitive lab and expert review activities were complete, the research team carefully reviewed the available evidence and selected two blocks for field testing at each grade level. The research team made every effort to select blocks for field testing that represented a judicious application of the *Item Modification Guidelines*, served as a representative sample of the work of the item modification panel, and provided the targeted student population (i.e., SDs and ELLs) with a reasonable chance of demonstrating their skills and abilities relevant to each of the objectives targeted in each block.

For Grade 4, blocks G4-3 and G4-4 were selected for field testing. The original version of block G4-3 was referred to as "block G4-3A" and the accessible version was referred to as "block G4-3B." The original version of block G4-4 was referred to as "block G4-4A" and the accessible version was referred to as "block G4-4B."

For Grade 8, blocks G8-1 and block G8-3 were selected for field testing. The original version of block G8-1 was referred to as "block G8-1A" and the accessible version was referred to as "block G8-1B." The original version of block G8-3 was referred to as "block G8-3A" and the accessible version was referred to as "block G8-3B."

Field Test

As there was no regularly scheduled administration of mathematics in 2010, the design for field testing and scaling the accessible blocks relied on combining data from the 2010 field test with data from the 2009 operational administration.

At each grade level, the 2010 field test included the two source blocks, S1 and S2; the two accessible blocks, A1 and A2 (where A1 is the modified version of S1 and A2 is the modified version of S2); and two other regular NAEP blocks, S3 and S4. The blocks were arranged in eight books, as follows:

Book	Block 1	Block 2
M181	A1	A2
M182	A2	S1
M183	S1	S2
M184	S2	A1
M185	S4	A1
MI16	S3	A2
M187	A1	S3
M188	A2	S4

Each accessible block thus appeared four times and was paired with every other block except its own source block. Among the regular NAEP blocks, however, the only ones that were paired together were S1 and S2; the rest of the pairings were derived from the 2009 operational data.

Sample

Three thousand cases were planned for the field test at each grade level; a sample size that would provide 375 cases per book, 1,500 cases per each accessible item, and 750 cases per each regular NAEP item. The realized sample was slightly larger than that required by the design: 3,538 cases at Grade 4 (including 372 SDs and 397 ELLs) and 3,608 cases at Grade 8 (including 328 SDs and 250 ELLs).

To facilitate item scaling, the sample obtained for the field test was intended to be representative of the larger sample of students who regularly participate in the NAEP assessment. More precisely, students who are normally excluded from participating in the regular NAEP administration were also excluded from the sample selected for the field test.[4]

[4] The 2008 pilot test included a small sample of Grade 4 students who would have otherwise been excluded from participating in NAEP. Results from the 2008 pilot test indicated that, on average, "otherwise excluded" students were able to correctly answer approximately 50 percent of the items in an accessible block.

Scoring

The validity of short- and extended-response items cannot be assessed without also considering the validity of the scoring guide that is used to assess student performance on those items. Accordingly, the item modification panels prepared draft scoring guides for the modified items, and members of the research team were present during scoring to assist with the finalization of the scoring guides. (Finalization during scoring is necessary because it is nearly impossible for item writers and reviewers to foresee the full range of student responses that may be created.)

Results

After the accessible blocks were developed, field tested, and scored, item, block, and grade-level analyses were completed by ETS using standard NAEP methodology. The primary purpose of these analyses was to evaluate the relative success of the item modification efforts. More specifically, efforts were made to (1) estimate the impact of accessible blocks on student performance (e.g., changes in average percentage correct, percentage omitted, and percentage not reached) by block and item for the full population and several subpopulations of interest (e.g., SDs, ELLs); (2) ensure that each item in each of the accessible blocks could be scaled with regular NAEP items; and (3) investigate reductions in standard error of measurement for various levels of student performance (i.e., theta levels) by grade level. An overview of each of the major analyses completed for the 2010 field test is provided below.

Percentage Correct, Omitted, and Not Reached

Percentage correct, omitted, and not reached were computed for each item in each accessible block and each source block.[5] Average percentage correct, percentage omitted, and percentage not reached also were calculated for each block. If the accessible blocks performed as expected, it was anticipated that the average percentage correct for each accessible block, for the full sample as well as each subpopulation of interest, would be significantly higher than the average percentage correct for the source block. It was also anticipated that there would be no change or some decrease in the rate at which items were omitted, and that students given an accessible block would be as likely, or more likely, to reach the final items in the block than students who were given the original, unmodified block.

Each of these predictions was confirmed by field test data. On average, Grade 4 students scored 32 percent higher on the accessible blocks than on the source blocks, and Grade 8 students scored an average of 26 percent higher on the accessible blocks. For all accessible blocks, a small but significant decrease in the percentage of skipped items was observed. In addition, for both fourth- and eighth-grade blocks, there were significant reductions in the percentage of students not reaching items at the ends of the blocks. Cognitive lab results suggest that at least

[5] In NAEP analyses, missing responses at the end of a block of items are considered not reached items and are treated as if they had not been presented to the respondent. Missing responses to items before the last observed response in a block are considered intentional omissions.

two factors that may contribute to the findings on percentages of items omitted or not reached: (1) On average, it takes students less time to complete an accessible block of items; therefore, students are more likely to attempt each item in the block; and (2) items in an accessible block place a lower cognitive demand on students; therefore, they are less likely to be discouraged by item difficulty.

Table 7 summarizes the percentage correct, omitted, and not reached results by block for Grade 4, and Table 8 summarizes these results for Grade 8.

Table 7. Summary of Percentage Correct, Omitted, and Not Reached Results—Grade 4

	N	% Correct	% Omitted	% Not Reached
G4-3B (Accessible Block)	1,706	77.23	0.97	0.87
G4-3A (Source Block)	927	46.27	1.60	4.24
		+30.96	**-0.63**	**-3.37**
G4-4B (Accessible Block)	1,726	84.96	0.59	1.18
G4-4A (Source Block)	914	48.44	1.58	5.83
		+36.52	**-0.99**	**-4.65**

Table 8. Summary of Percentage Correct, Omitted, and Not Reached Results—Grade 8

	N	% Correct	% Omit	% Not Reached
G8-3B (Accessible Block)	1,787	75.25	0.35	0.61
G8-3A (Source Block)	905	49.99	0.98	3.21
		+25.26	**-0.63**	**-2.60**
G8-1B (Accessible Block)	1,789	72.41	0.79	1.45
G8-1A (Source Block)	913	44.19	1.70	2.55
		+28.22	**-0.91**	**-1.10**

In addition, similar improvements in student performance were observed for SD and ELL populations. More specifically, the average shift in student performance remained relatively consistent regardless of a student's disability or English language proficiency status. Tables 9 and 10 summarize the average improvement in student scores for Grades 4 and 8 students across the disability categorizations reported by NAEP for each block included in the field test. Tables 11 and 12 summarize the average improvement in student scores for Grades 4 and 8 students across various English language proficiency categorizations reported by NAEP for each block in the field test.

Table 9. Summary of Percentage Correct for Students With Disabilities—Grade 4

	IEP Yes	504 Yes	IEP/504 No
G4-3B (Accessible Block)	63.63	73.13	78.48
	+30.06	**+35.52**	**+30.30**
G4-4B (Accessible Block)	74.58	85.06	86.01
	+36.31	**+46.56**	**+35.83**

Note: IEP=individualized education plan; 504=section 504 plan.

Table 10. Summary of Percentage Correct for Students With Disabilities—Grade 8

	IEP Yes	504 Yes	IEP/504 No
G8-3B (Accessible Block)	53.22	76.42	77.12
	+20.65	**+33.61**	**+24.62**
G8-1B (Accessible Block)	51.49	68.06	74.27
G8-1A (Source Block)	26.78	44.12	46.65
	+24.71	**+23.94**	**+27.62**

Note: IEP=individualized education plan; 504=section 504 plan.

Table 11. Summary of Percentage Correct for English Language Learners (ELLs) and Former ELLs—Grade 4

	ELL Yes	ELL No	Former ELL
G4-3B (Accessible Block)	63.72	78.38	83.03
	+29.26	**+30.88**	**+34.65**
G4-4B (Accessible Block)	76.04	85.82	88.45
	+37.77	**+36.39**	**+34.01**

Table 12. Summary of Percentage Correct for English Language Learners (ELLs) and Former ELLs—Grade 8

	ELL Yes	ELL No	Former ELL
G8-3B (Accessible Block)	50.62	76.64	74.85
	+17.44	**+25.19**	**+34.61**
G8-1B (Accessible Block)	52.61	73.95	66.04
G8-1A (Source Block)	24.65	45.83	32.76
	+27.96	**+28.12**	**+33.28**

Item Scaling

Each accessible item was scaled with the full item pool for the NAEP assessment using standard NAEP scaling methodology. That is, a theta value was computed for each item, and students' performance on each item, relative to their estimated proficiency (i.e., theta level), was assessed. For multiple-choice items, NAEP uses a three-parameter model that includes discrimination (a parameter), difficulty (b parameter), and guessing (c parameter). If items in the accessible blocks performed as expected, it was anticipated that one would observe little or no change in the average estimate of item discrimination and guessing parameters. More importantly, the research team expected to observe significant reductions in the average estimate of item difficulty.

Scaling results indicated that all items in the accessible blocks were scalable with the larger pool of unmodified NAEP items. As predicted, accessible items had discrimination and guessing characteristics that were generally similar to their source items, although there were significant reductions in item difficulty. In addition, for the lowest performing students, the standard error of measurement was significantly lower on the accessible blocks than the source blocks.

Figures 1 and 2 summarize the a parameter estimates (discrimination) for accessible and source items (at Grades 4 and 8, respectively). Figures 3 and 4 summarize the c parameter estimates (guessing).

Figure 1. Distribution of Item A Parameters (Discrimination)—Grade 4

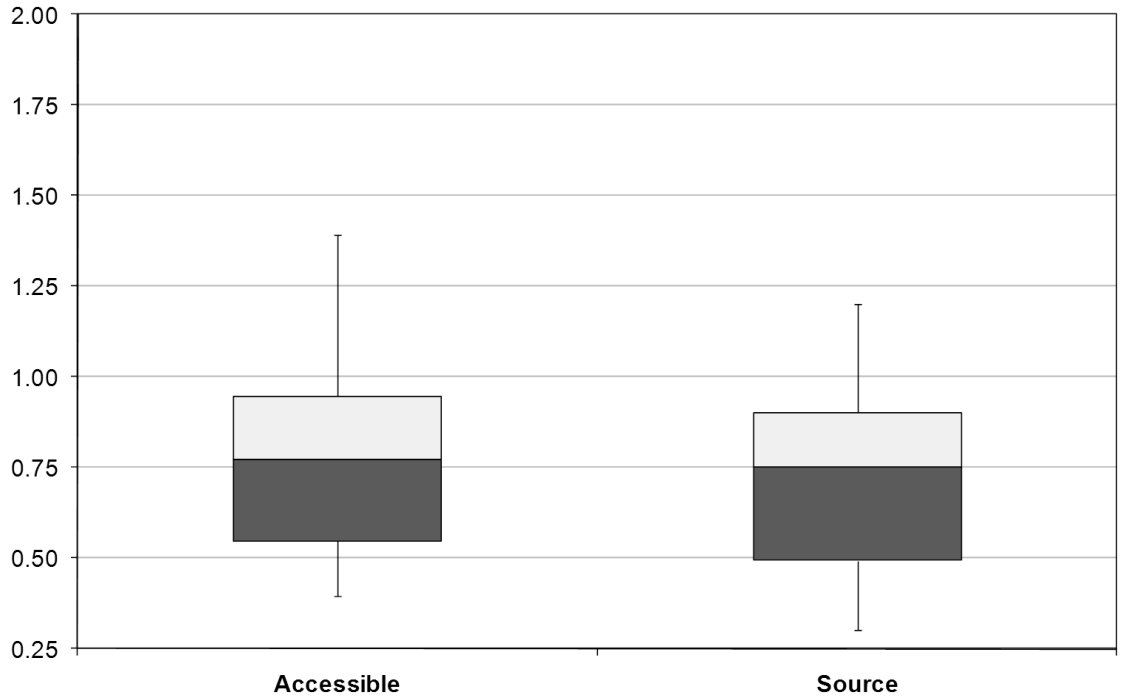


Figure 2. Distribution of Item A Parameters (Discrimination)—Grade 8

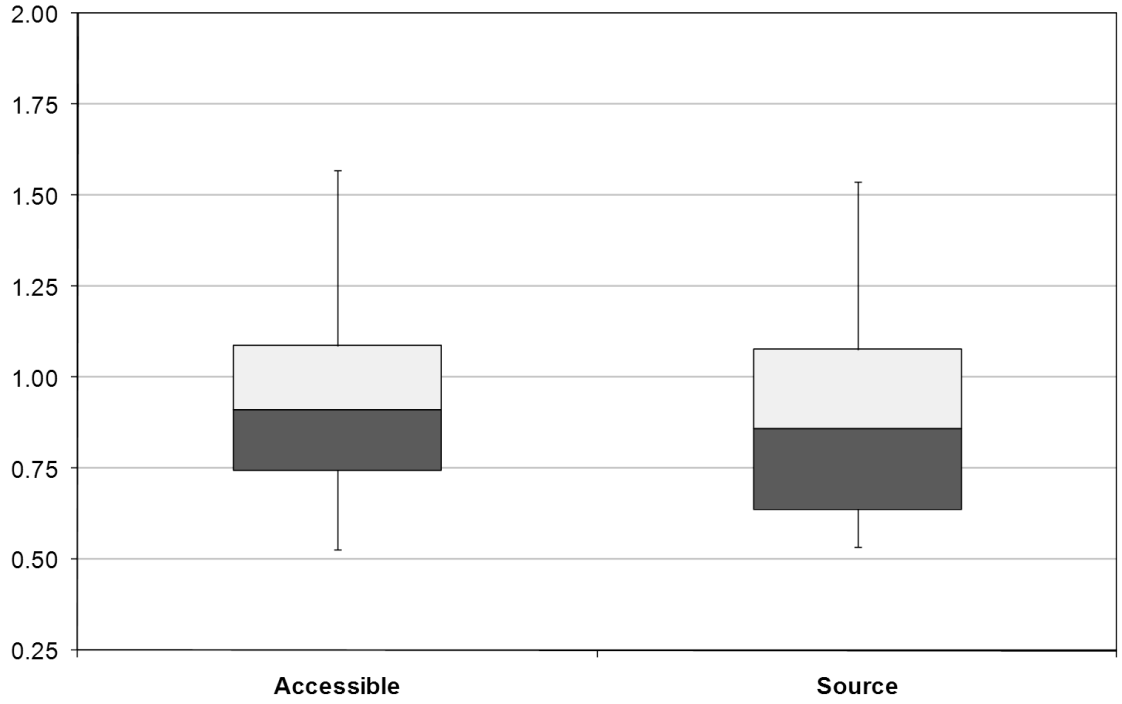


Figure 3. Distribution of Item C Parameters (Guessing)—Grade 4

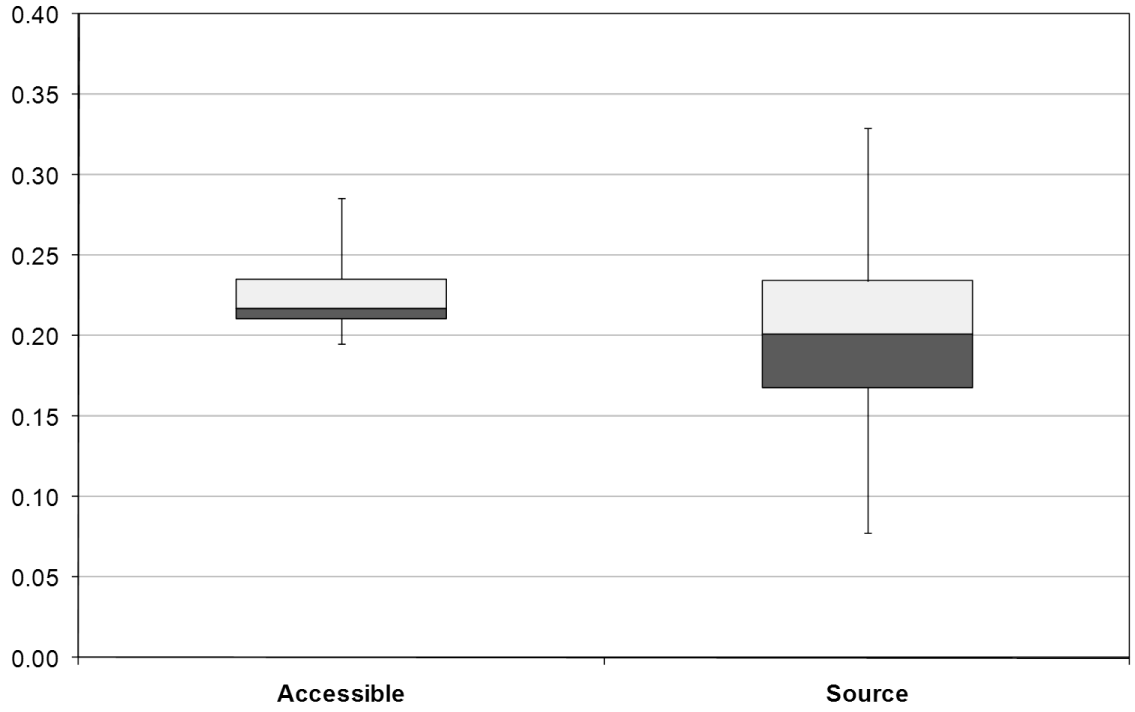
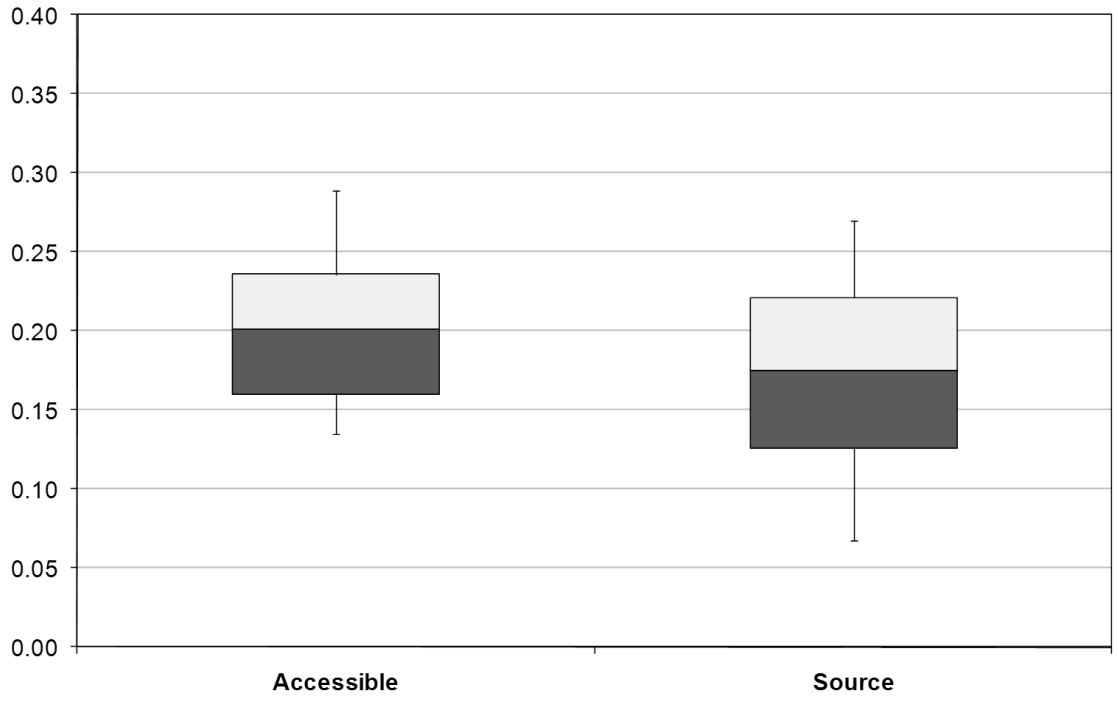


Figure 4. Distribution of Item C Parameters (Guessing)—Grade 8



Figures 5 and 6 summarize the b parameter estimates for the accessible and source blocks, and also show the ability distributions for the sampled Grade 4 and 8 students. These figures illustrate a significant difference in average item difficulty for accessible and source items. On average, accessible items were significantly less difficult than the source items from which they were derived. These figures also illustrate that, overall, items included in the source blocks are relatively well aligned with the estimated ability of the general student population and items included in the accessible blocks are relatively well aligned with the estimated ability of students who perform on the lower levels of the NAEP performance continuum.

Figure 5. Distribution of Item B Parameters (Difficulty) Compared With Student Ability Distribution—Grade 4

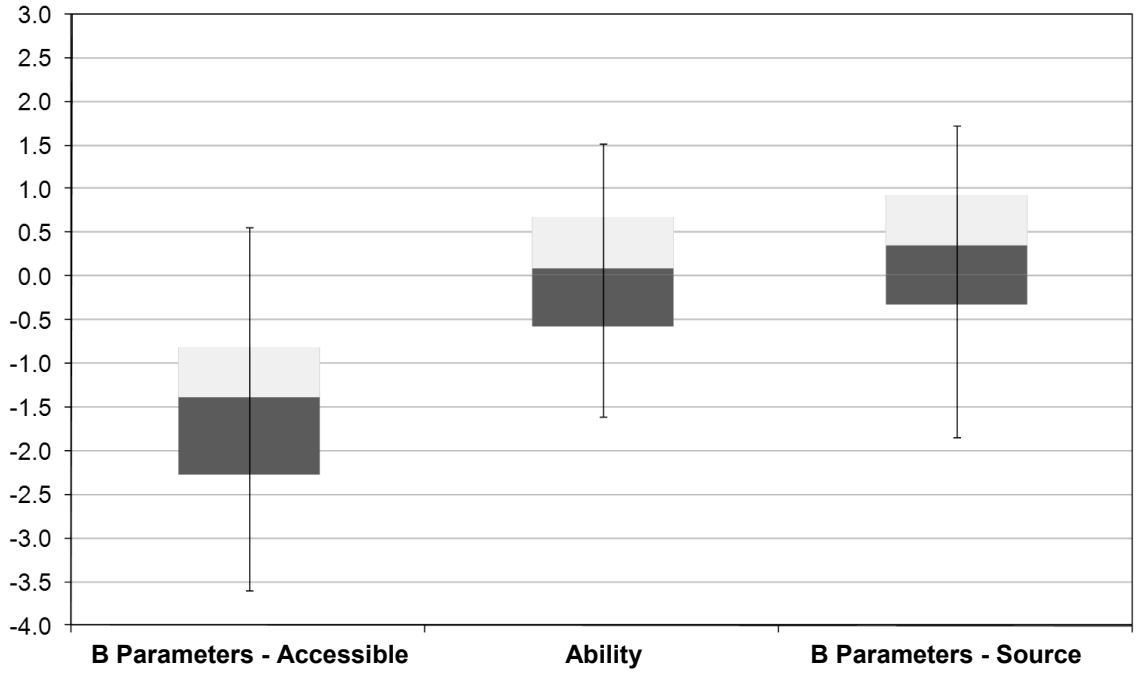
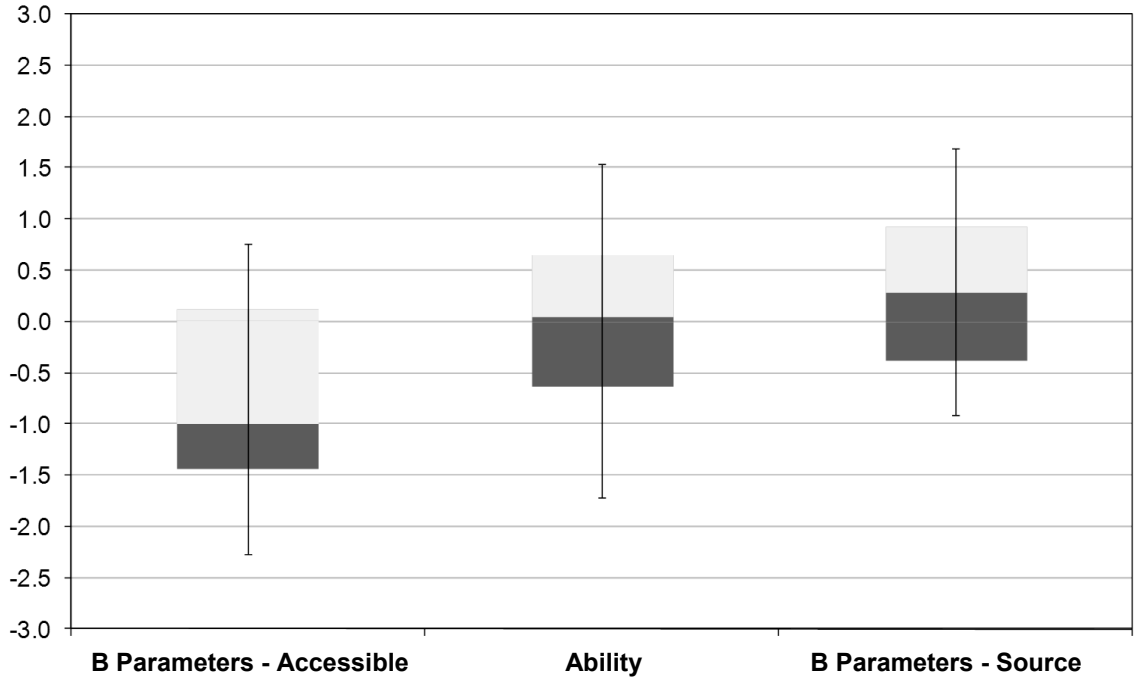


Figure 6. Distribution of Item B Parameters (Difficulty) Compared With Student Ability Distribution—Grade 8



Additional analyses were completed to determine the part of the ability distribution for which the modified and source blocks provided the most information. More specifically, at each grade level, the test information curve for the test book that contained two source blocks (book M183) was compared with the test information

curve for the test book that contained two accessible blocks (book M181). If accessible blocks performed as anticipated, one would expect that the book containing the two accessible blocks would provide more information for students at the lower end of the performance continuum than would the book containing the two original blocks.

Figures 7 and 8 illustrate the observed ability distribution for Grades 4 and 8 students (respectively), and superimpose three information curves on those distributions. An estimated information curve is provided for "regular book 183" (two source blocks) and "accessible book 181" (two accessible blocks). The third test information curve (labeled "Overall") represents the average test information across all eight books in the field test. From these figures, it is clear that the estimated information gathered for students at the lower levels of the NAEP performance continuum is much greater for the two accessible NAEP blocks than for the two source blocks. As expected, the "overall" test information curve falls between the other two.

Figure 7. Ability Distribution and Test Information by Book Type—Grade 4

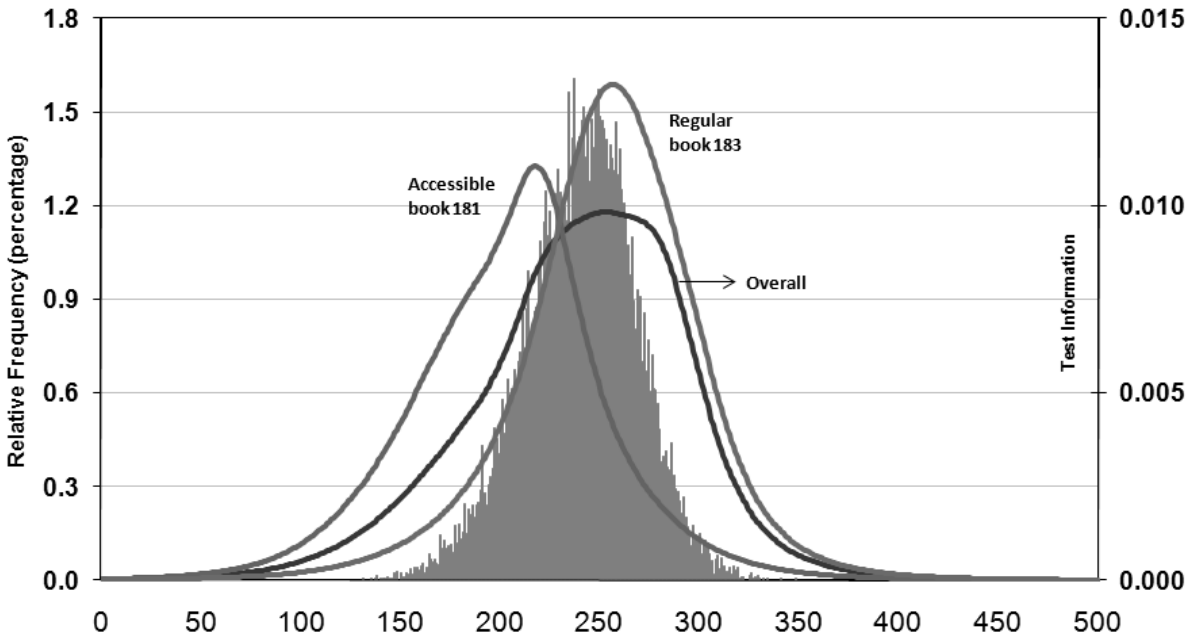
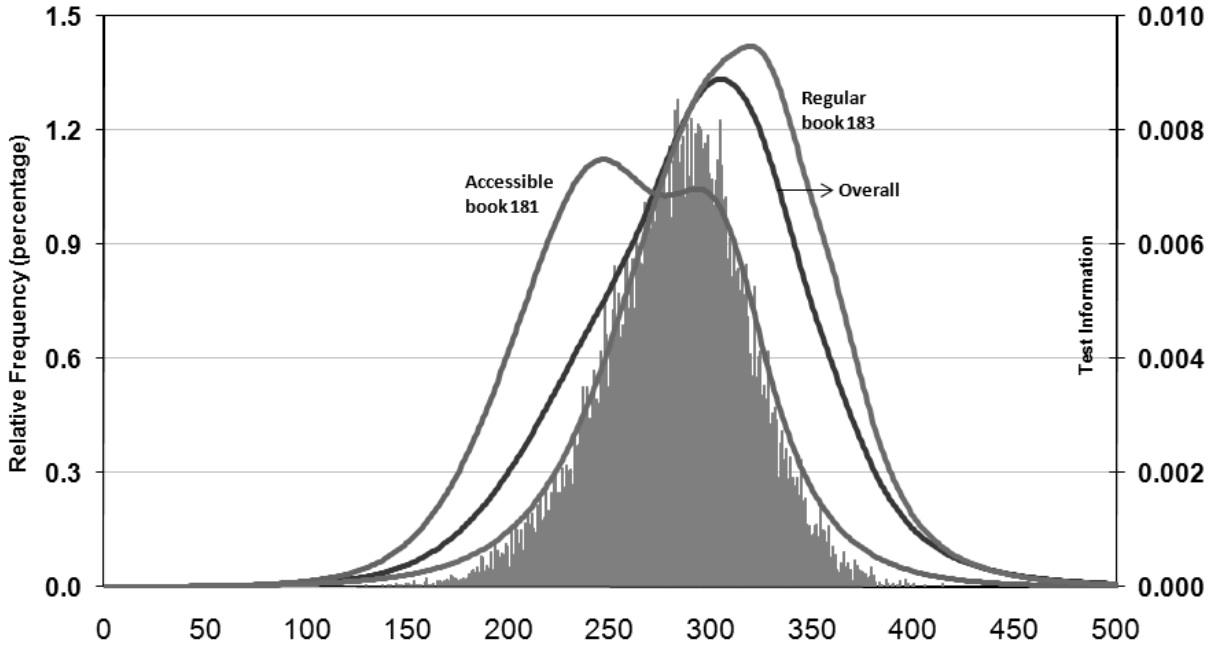


Figure 8. Ability Distribution and Test Information by Book Type—Grade 8



Of course, the amount of information provided by the assessment across the performance continuum is closely related to the estimated standard error of measurement (or more precisely, the conditional standard error of measurement, or CSEM). Because the accessible blocks were designed to provide more information about students at the lower end of the NAEP performance continuum, one would expect to observe an increase in the estimated reliability of students' scores in this range (i.e., a decrease in the observed standard error of measurement for lower performing students). In fact, the research team had anticipated significant reductions in the standard error of measurement on the order of 20–30 percent for these students.

Figures 9 and 10 again illustrate the observed ability distribution for Grades 4 and 8 students (respectively), and superimpose three estimated CSEM curves on those distributions. Estimated CSEM curves are provided for "Regular book 183" and "Accessible book 181," while the "Overall" CSEM curve represents the average CSEM across all eight books in the field test. These figures illustrate that, across students with the lowest estimated ability levels, an accessible book provides a significantly lower measurement error than a book comprised of the two source blocks.

Figure 9. Ability Distributions and Conditional Standard Error of Measurement (CSEM) by Book Type—Grade 4

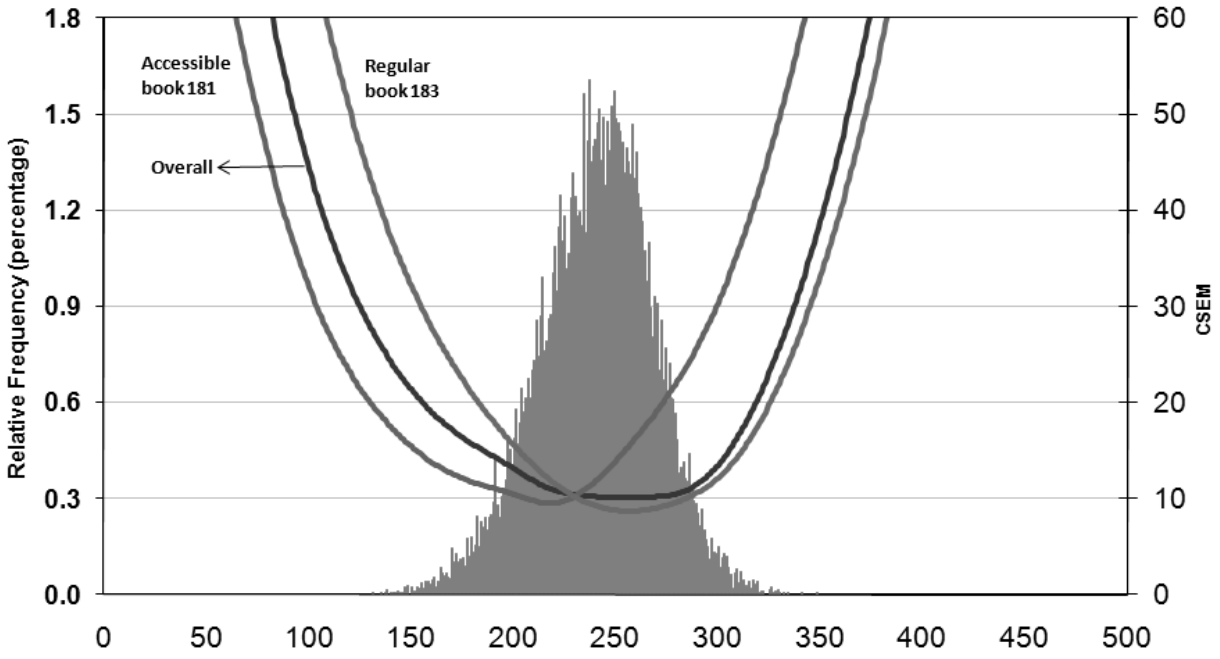
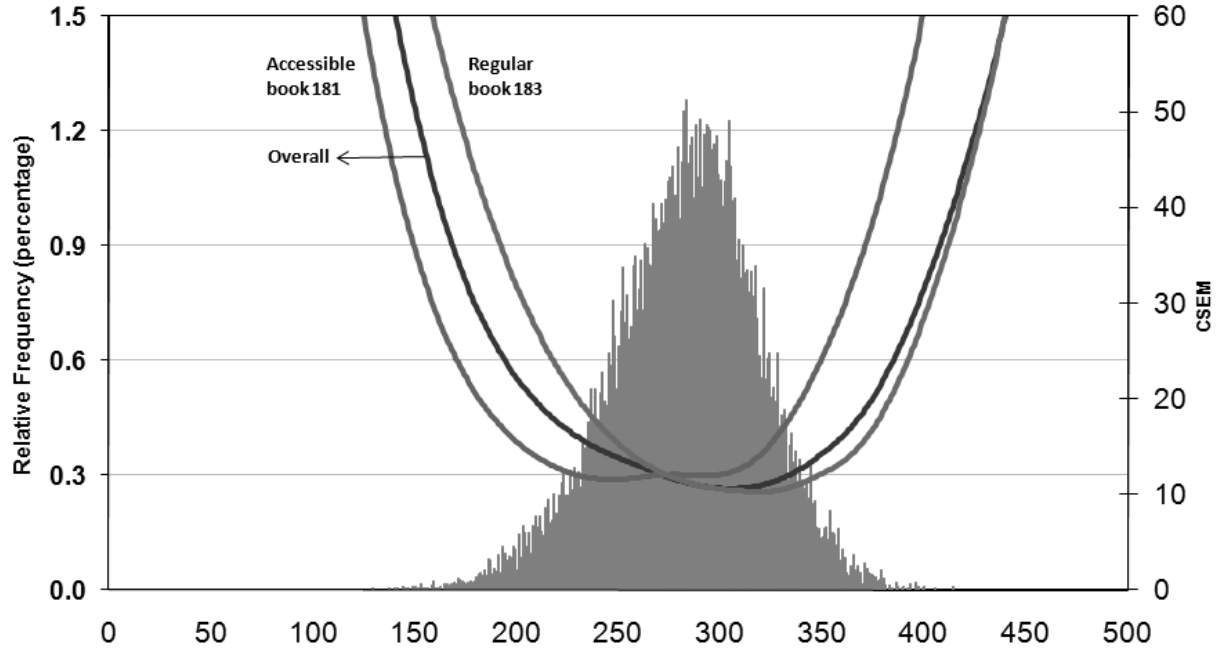


Figure 10. Ability Distributions and Conditional Standard Error of Measurement (CSEM) by Book Type—Grade 8



Tables 13 and 14 provide point estimates of the reduction in the CSEM across the observed ability distribution for Grades 4 and 8 students. These tables show that for students falling below the 25th percentile, accessible books have the potential to

provide a significant reduction in observed measurement error, on the order of 20–40 percent.

Table 13. Conditional Standard Error of Measurement by Percentile by Book Type—Grade 4

Book Type	5th Pctl.	10th Pctl.	25th Pctl.	Median	75th Pctl.	90th Pctl.	95th Pctl.
Overall	14.2	12.6	10.7	10.2	10.1	10.3	10.8
Source	18.3	15.3	11.6	9.3	8.7	9.1	9.7
Accessible	10.8	9.9	9.8	12.6	16.9	21.5	24.7
	-41%	**-35%**	**-18%**				

Table 14. CSEM by Percentile by Book Type—Grade 8

Book Type	5th Pctl.	10th Pctl.	25th Pctl.	Median	75th Pctl.	90th Pctl.	95th Pctl.
Overall	17.7	15.2	13.0	11.1	10.6	11.4	12.8
Source	26.3	20.6	14.7	11.6	10.5	10.3	10.6
Accessible	12.1	11.6	11.9	12.0	12.4	16.4	20.5
	-54%	**-44%**	**-19%**				

Conclusion

This investigation of NAEP accessible blocks served as a proof-of-concept study in two important ways. First, the creation, application, and expert review of the *Item Modification Guidelines* and *Item Modification Procedures* illustrated that it was possible to develop standard procedures for creating items that were less difficult but still adhered to the content framework. Second, the results from the field test of accessible and operational NAEP blocks indicated that it was indeed feasible to construct accessible blocks that are scalable with the main NAEP assessment and improve measurement precision at the lower end of the NAEP performance continuum.

The primary study findings were as follows:

- Across all groups and subgroups, there were substantial and similar average gains in percentage correct by block for the accessible blocks compared with the source blocks.
- There were consistent declines in the number of students omitting items and significant reductions in the percentage of students not reaching items for the accessible blocks compared with the source blocks.
- All accessible items were scalable, and modified items had similar average discrimination and guessing characteristics (a and c parameter estimates) as the source items, while there were significant reductions in item difficulty (b parameter estimates).
- For the lowest performing students, the CSEM was significantly lower for students completing two accessible blocks than for those completing two source blocks.
- Test information functions for books comprised of two accessible blocks were appropriately targeted to the lower end of the performance continuum.

When the NAEP accessible block study in mathematics was undertaken in 2007, the initial goal was to improve measurement of achievement for students at the lower end of the continuum. Results from this study have been used for several purposes over the past few years. First, the *Item Modification Guidelines* and *Item Modification Procedures* are now routinely used in NAEP item development to improve the quality of all items, not only those intended to be made more accessible. Second, the accessible block study in mathematics served as the impetus for additional research on improving measurement precision at the lower part of the distribution, including an accessible block study in reading (commissioned by the NVS Panel) and the Knowledge and Skills Assessment (KaSA), an ongoing special study by the NAEP contractor for item development and analysis that has administered accessible blocks to students who would otherwise be excluded from NAEP. Finally, the concept of accessible items is increasingly relevant as the National Center for Education Statistics moves towards computer-based testing and considers a multistage design for NAEP. The procedures for constructing accessible blocks and lessons learned from this study will be invaluable for ensuring that students at the lower end of the distribution can be adequately measured.

References

Daro, P., Stancavage, F. B., Ortega, A., DeStefano, L., & Linn, R. (2007). *Validity study of the NAEP mathematics assessment: Grades 4 and 8*. A publication of the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

McLaughlin, D. H., Scarloss, B. A., Stancavage, F. B., & Blankenship, C. D. (2005). *Using state assessments to impute achievement of students absent from NAEP: An empirical study in four states*. A publication of the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

APPENDIXES

APPENDIX A
Item Modification Guidelines and Item Modification Procedures

Aligning the Accommodation Block Assessment With the NAEP Framework

Similar to standard blocks of NAEP assessment items, all accommodation blocks should be developed so that they are aligned with the content expectations defined by the 2005 NAEP Mathematics Framework. Unlike standard blocks of NAEP assessment items, there will be less variability in the level of complexity of items in a NAEP accommodation block. Drawing on Webb and others, five interrelated dimensions are considered in structuring the NAEP assessment so that it is aligned with the NAEP framework:

1. The match between the content of the assessment and the content of the framework: The assessment as a whole should reflect the breadth of knowledge and skills covered by the topics and objectives in the framework.
2. The match between the complexity of mathematical knowledge and skills on the assessment and in the framework: The assessment as a whole should represent the balance of levels of mathematical complexity at each grade level as described in the framework. However, an accommodation block is meant to provide important statistical information about students at the lower end of the performance continuum. Therefore, it is appropriate for an accommodation block to contain items that assess students' ability to perform tasks associated with Basic and Proficient levels of achievement.
3. The match between the emphasis of the assessment and the emphasis of topics, objectives, and contextual requirements in the framework: The assessment should represent the balance of content and item formats specified in the framework and give appropriate emphasis to the conditions in which students are expected to demonstrate their mathematics achievement, reflecting the use of calculators, manipulatives, and real-world settings.
4. The match between the assessment and how scores are reported and interpreted: The assessment should be developed so that scores will reflect both the framework and the performance described in the NAEP achievement levels.
5. The match between the assessment design and the characteristics of the targeted assessment population: The assessment should give all students tested a reasonable opportunity to demonstrate their knowledge and skills in the topics and objectives covered by the framework (with a special emphasis on providing students at the lower end of the performance continuum an opportunity to show what they know and are able to do).

Item Modification Guidelines

These guidelines identify the major themes and dimensions of construction that should be addressed when modifying blocks of NAEP items to create an accommodation block. The guidelines are meant to aid item modifiers in assessing relevant and irrelevant aspects of the item's construct that contribute to the overall difficulty and accessibility of the item. The guidelines offer common strategies for reducing difficulty without compromising content, construct validity, or alignment with the NAEP framework.

These guidelines should be applied judiciously. Their application may vary from item to item depending on the measurement intent of the item. Generally, these guidelines should be followed unless the targeted construct of the item precludes doing so.

Word Choice

Careful word choice is an essential component of quality item construction. Word choice refers to language used within the statement of a problem, as well language used in the alternative answer choices. Careful word choice should be a central consideration during the item modification process.

- **Clarity**—Word choice throughout all items should be unambiguous and concise. It is more important for item wording to be clear than for it to be precise. For example, avoid the ambiguous phrase "about how much" when writing problems that require estimation or rounding.
- **Plain Language**—Plain language, as a writing and editing tool, is designed to clearly convey meaning without altering what an item is intended to measure. All items should use plain language. Even when the intent of the item is for the student to define, recognize, or use mathematics vocabulary correctly, the surrounding text should be in plain language. Plain language should increase access and minimize confusion.
- **Terminology Appropriateness**—Terminology used should be current and relevant to a broad population. Use of outdated technology or terminology, can distract from the content of a problem.
- **English as a Second Language Considerations**—Use of commonly accepted and culturally nonspecific words, phrases, and terminology is encouraged whenever possible. Be careful of literal interpretations of items. When using words with multiple meanings, make sure the intended meaning is clear. Avoid ambiguous words, such as "if," "could," "may," or "can." Use high-frequency words as much as possible. Avoid the word "not" whenever possible.
- **Parallel Item Construction**—Item wording should provide parallel syntactic construction. Use of the present tense verb is preferred. Wording within and between the statement of a problem and its possible answer choices (including distracters) should be consistent in tense and vocabulary.
- **Brevity and Simplicity**—Questions should be in brief, 'simple' form. Compound sentences should be written as two short sentences.

- **Grammar**—Present tense and active voice should be used whenever possible. Minimize paraphrasing. Avoid pronouns. Avoid colloquialisms.

Alternative Answer Choices (Distracters)

Alternative answer choices include the solutions presented in a multiple-choice item, as well as the acceptable answers for an open-ended item. Alternative answer choices may be presented in multiple formats (e.g., numbers, text, graphics, charts). Use of these formats can increase item access. However, if used or constructed improperly, they can add confusion to the item and may distract test takers from the original intent of the item.

For multiple-choice items:

- **Provide Plausible Distracters—**Identify alternative answer choices (distracters) that are plausible, and not unreasonable. The easiest multiple-choice questions should provide students with only one reasonably appropriate solution.
- **Provide an Appropriate Number of Distracters—**Make the number of possible answer choices appropriate for the content and context of the problem. The American convention of providing four answer choices is sometimes inappropriate or unreasonable.
- **Provide a Range of Distracters—**Offer students a diverse set of answer choices. This may reduce confusion and testing error. Items requiring rounding or estimation are sometimes clearer when a wide range of answer choices is provided.

For open-ended items:

- **Allow for Multiple Response Types**—Allow students to show their answers through illustrations, diagrams, formulas, or words.

Item and Block Format

Item and block format is the layout, design, and arrangement of information within and between each item in a block. Careful item and block formatting can improve the clarity of an item and the block as a whole.

- **Format Consistency**—Use the same structure for paragraphs throughout the assessment as much as possible (e.g., topic sentence, supporting sentences, and concluding sentence). Be sure that the item format does not add ambiguity to the solution.
- **Separate Information as Appropriate**—Split multiple ideas into separate sentences, statements, or lines to decrease the complexity of an item.
- **Item Spacing**—Provide liberal spacing throughout an item. Double spacing makes word problems easier to read and understand. Double spacing alternative answer choices aids in visual and cognitive processing and discrimination. Separate the main question in an item (How, what…?) from the rest of the information presented in the item.

- **Answer Spacing**—Provide appropriate space for an answer. Too little or too much space for an answer can falsely suggest an answer of a certain length.
- **Clarity**—Use format to clarify text. Use bullets, spacing between pieces of text, and boxing of text to emphasize or separate information.
- **Item Separation**—Provide a clear distinction between each item. Some NAEP items provide information (e.g., a graph, chart) before the statement of the problem. In such cases, the item should always begin on a new page in order to provide a clear distinction between problems.

Graphics

Graphics, such as pictures, charts, and diagrams, are visual images reflecting information. Graphics can be very effective in supporting text, illustrating mathematical concepts, and increasing item access. If used improperly, however, graphics can add substantial confusion and distract test takers from the intent of the item. Graphics should be used judiciously.

- **Clarity**—Visuals should be clear and precise. Adding a visual may clarify the measurement intent of an item.
- **Mathematical Accuracy**—Visuals should utilize standard mathematical notation and formatting.
- **Simplicity**—Visuals should only contain necessary information. Remove unnecessary graphics. Avoid misleading graphics, such as charts with inconsistent scales.
- **Completeness**—Visuals should provide a representation of the important parts of the item. Visuals should mirror and parallel the wording and expectations of the problem.

If a visual within a given item is adding to the unintended difficulty of the item, it should be altered or removed.

Appropriate Use of Context

Contextual information includes problem scenarios, explanations, specific directions, and background text. Using contextual information can place mathematical concepts in more realistic conditions and provide background information that test takers may need. However, the contextual information should not interfere with the mathematics being assessed. It should not be a barrier to a student's ability to demonstrate his or her mathematical knowledge. Contextual information should be included only if the item is intended to assess mathematics in context.

- **Use Plain Language—**Use plain language as much as possible.
- **Increase Clarity—**Use manipulatives and/or graphics to increase item clarity.
- **Use Relevant Contexts—**Use contexts only if they are meaningful to the mathematics being assessed.

- **Provide Appropriate Contexts**—Use contexts that are appropriate for the grade level being assessed.
- **Use Familiar Contexts**—Avoid contexts that may confuse or be unfamiliar to some students taking the assessment.
- **Provide Accurate Contexts**—Avoid contextual information that could interfere with the measurement of the intended skill.

Extraneous Information

Extraneous information includes all portions or aspects of an item that are unessential to the mathematics being assessed. This includes any inconsequential context. Extraneous information should be eliminated from all items in an accommodation block.

- **Eliminate Extraneous Information.**
- **Provide Manipulatives Judiciously**—Only provide manipulatives when absolutely necessary (e.g., it may or may not be appropriate to test students' ability to visualize information using manipulatives).
- **Consider Item Context**—Provide students with units of measure only as necessary or appropriate for the context of an item. Including units of measure can be unnecessarily confusing.
- **Calculator Usage**—Do not ask students if they used a calculator for an item that obviously does not require its use.

Cues

Cues are components of item construction that give key information related to the problem. Cues can also provide information related to incorrect answer choices. Cues can serve to clarify the intent of an item. Item writers should carefully consider how cues are used in each item.

- **Provide Descriptive Titles**—Identify the goal or topic of a problem with a title when appropriate. This is especially helpful for presenting word problems that require multiple pieces of information.
- **Provide Visual Cues—Bold**, *italicize*, underline, or CAPTIALIZE key words and phrases including:
  - Directions (e.g., Solve, COMPUTE, **Explain**)—Directions should always come at the **beginning** of a problem.
  - Operational words and phrases (e.g., **Add**, *Subtract*, Find the product)
- **Clarify Answer Requirements**—Cue students about the number and type of solution(s) they should provide (e.g., written description, graphical representation). This is especially important in open-response items that could be solved using multiple approaches.
- **Avoid Deceptive Cues**—Do not mislead students to perform inappropriate operations.

- **Provide Definitions When Appropriate**—It may be appropriate to provide a brief definition, example, or illustration of a mathematical concept if doing so does not compromise the objective of the assessment item.
- **Use Cues to Clarify Item Intent**—Remember, the objective and intent of all testing items should be as clear as possible.

Computational Appropriateness

Each item on the NAEP mathematics assessment is assigned a mathematical complexity rating (low, moderate, high). The task asked of the student should reflect an appropriate computational level. Generally, it is possible to reduce the computational complexity of an item while preserving its alignment with the NAEP framework.

- **Assess Computational Complexity**—Do not require students to perform calculations that are unnecessarily difficult. Calculations should not distract from the general idea being assessed in any given item.
- **Gauge Time Constraints**—Do not require students to perform calculations that are unnecessarily time consuming. Calculations should not distract from the "flow" of the testing experience. Remember that TIME is a precious resource during the testing experience.
- **Computational Progression**—Do not require students to perform counterintuitive operations.
- **Encourage Mathematical Accuracy**—Do not ask students to estimate or round when an exact calculation is appropriate or easier.
- **Calculator Use**—Items should be constructed with calculator use/availability in mind. Computational complexity should be appropriate to the testing context. Remember, the availability of a calculator should not increase the complexity of a problem.

Grade-Level Appropriateness

Item modifiers should identify the objective(s) being assessed by each NAEP item as well as the grade level at which it is meant to be assessed. Items in an accommodation block should be constructed to assess at or below the grade level under consideration. For example, a fourth-grade accommodation block should not contain items that are constructed to assess a learning objective at an eighth-grade level. In most cases, the NAEP framework provides leveled descriptions of each learning objective. Students being assessed using an accommodation block should not be asked to perform a task at a level higher than is appropriate for their grade.

Cognitive Demand

Cognitive demand is a term used to refer to the overall difficulty of an item. Several components contribute to the cognitive demand of any given item. For the purpose of creating an accommodation block, item writers should carefully consider factors that may unnecessarily increase the cognitive demand of an item.

- **Assessing Multiple Objectives**—Assessing multiple objectives in a single item generally increases the cognitive demand of an item. An accommodation block should limit the number of items that assess multiple objectives.
- **Multiple Steps**—When possible, reduce the number of steps required to correctly answer an item while preserving the integrity of the objective being assessed.
- **Multiple Answers**—Limit the number of items that require multiple answer components.

Item Modification Procedures

1. Begin with a predeveloped block of NAEP assessment items. There are several potential benefits to working with an existing NAEP block.
    - The block meets NAEP standards.
    - There may be information regarding item difficulty (e.g., percentage of students correctly answering the item, percentage of students selecting each alternative answer choice).
    - It may be possible to compare field test results with existing data.
2. Thoroughly review the document titled Item Modification Guidelines.
    - Each member of the item modification panel should have sufficient time to read and discuss the *Item Modification Guidelines*. The team should be presented with sample comparisons of original NAEP items with modified NAEP items and then be allowed to "practice" applying the recommendations on a few released NAEP items. This conversation should allow team members to become more comfortable and familiar with item modification guidelines and procedures.
3. Each member of the item modification panel should be given approximately 30 minutes to perform an initial individual review of each block. That is, panel members should spend a short amount of time reading over each item, familiarizing themselves with the block. During this review, each panel member should note:
    - Item and block clarity.
    - The diversity of NAEP objectives assessed by the block.
    - The difficulty of the items (percentage correct, percentage for each distracter).
    - Issues related to item quality (e.g., Are there errors? Do some items seem awkward or inappropriate for the grade level under consideration?).
    - Issues related to students with disabilities (SDs) and English language learner (ELL) students (e.g., vocabulary and wording), particularly the use of calculators and manipulatives.
    - The balance of multiple-choice and short-response items.
4. Members of the item modification panel should briefly discuss their thoughts from the initial item review. This conversation should be relatively brief (15–20 minutes). The following questions may be used to guide discussion:
    - Are there concerns about block or item clarity?
    - Is the block balanced? Is there a broad range of NAEP objectives assessed by the block, or are some learning objectives over- or under-represented by the block?
    - Are accessibility concerns effectively addressed?
    - Is the use of manipulatives/calculators necessary/appropriate?
    - Are all instructions clear?

5. The item-by-item review should include the following steps:
    - Modify each item as a group. It may be beneficial to have a large display of the item under consideration (i.e., use a projector).
    - Identify issues and concerns regarding the formatting, context, and accessibility of each item.
    - Carefully consider/modify the cognitive demand of each item. Each item must be addressed on a case-by-case basis and considered in the context of the block as a whole. It is generally useful to refer to information regarding item difficulty (e.g., percentage of students correctly answering the item, percentage of students selecting each alternative answer choice) for this task.
    - Carefully review and apply the *Item Modification Guidelines*.
    - Record/comment on recommended modifications to each item for future reference.
    - Record and classify the types of modifications that are recommended for each item. Use the document titled "Item Modification Record" to complete this process for each item.

    Note: It takes an average of 20–30 minutes to review each item. However, some items require less time to review (15 minutes) and others require more time to review (50 minutes).
6. Compile all item modification recommendations. It is helpful to have each member of the panel submit his or her modified version of each item to the panel coordinator. Doing so often reveals misunderstandings or misinterpretations of group decisions regarding item modification. It is also helpful to have the group select one version of each modified item to serve as the representative sample of the panel's work for that item. It is convenient to use this representative sample of modified items as a reference for future editing and review procedures. It may be necessary to create an "editor ready" (i.e., clean copy) of some of the items.
7. Rereview all items in the block.
    - Note the degree of item modification on the Block Summary Sheet. Please refer to the document titled "Item Modification Rating Scale" for this task. This scale describes three levels of item modification, which may be useful for characterizing the overall degree of block modification.

APPENDIX B
2007 Expert Panel Members

Item Review Panel

Patrick Callahan
University of California, Office of the President

Lizanne DeStefano,
University of Illinois at Urbana-Champaign

Arthur (Art) Duval
University of Texas, El Paso

Roger Howe
Yale University

Wilfried Schmid
Harvard University

Item Modification Panel

Lizanne DeStefano *(Director)*
University of Illinois at Urbana-Champaign

Jeremiah Johnson *(Coordinator)*
University of Illinois at Urbana-Champaign

Hsin-Mei Huang
University of Illinois at Urbana-Champaign

Renee Lemons
University of Illinois at Urbana-Champaign

Travis Wilson
University of Illinois at Urbana-Champaign

APPENDIX C
2009 Expert Panel Members

Item Review Panel

Peter Braumfield
University of Illinois at Urbana-Champaign

Patrick Callahan
University of California, Office of the President

Arthur (Art) Duval
University of Texas, El Paso

Roger Howe
Yale University

Randy McCarthy
University of Illinois at Urbana-Champaign

Wilfried Schmid
Harvard University

Item Modification Panel

Lizanne DeStefano *(Director)*
University of Illinois at Urbana-Champaign

Jeremiah Johnson *(Coordinator)*
University of Illinois at Urbana-Champaign

Theresa Bryant
University of Illinois at Urbana-Champaign

Jacqueline Bunn
University of Illinois at Urbana-Champaign

Holly Downs
University of Illinois at Urbana-Champaign

Aaron Hill
University of Illinois at Urbana-Champaign

Renee Lemons
University of Illinois at Urbana-Champaign

Jason Pound
University of Illinois at Urbana-Champaign

Tony Se
University of Illinois at Urbana-Champaign

Kathleen R. Smith
University of Illinois at Urbana-Champaign

Guy Tal
University of Illinois at Urbana-Champaign

APPENDIX D
Item Rating Scale

Item Rating Scale

Each NAEP item should assess mathematical content. In addition, items should assess the student's ability to reason with the content. The assessment should give all students tested a reasonable opportunity to demonstrate their knowledge and skills in the topics and objectives covered by the framework. A special emphasis should be placed on providing students at the lower end of the achievement spectrum an opportunity to show what they know and are able to do.

PLEASE RATE THE MATHEMATICAL ADEQUACY OF EACH ITEM USING THE FOLLOWING SCALE.

1. **Adequate**

   The problem is posed clearly. Any student who learned the mathematics of the task should be able to understand what is being asked. There are no unreasonable hidden assumptions. The context, language, and/or graphics used to pose the problem do not create unnecessary challenges that are unrelated to the mathematics. The problem, along with its response set or scoring rubric, does not contain mathematical errors.

2. **Marginal**

   The item is somewhat problematic. It may work as intended for many students, but defects in the item may unnecessarily lead to error or frustration for some students. In some cases, a simple edit may be sufficient to render the item adequate.

3. **Seriously Flawed**

   The item fails substantially on one or more of the following criteria: (a) It is undermined by hidden assumptions that are unfair to the student; (b) the context is confusing and misleading in ways that are not related to what is being measured; (c) the language and graphs present unnecessary obstacles to understanding what is being posed; or (d) there are mathematical errors in the problem or in its response set or scoring.