Validity Study of the NAEP Mathematics Assessment: Grades 4 and 8

Phil Daro, *University of California, Berkeley*

Frances Stancavage, *American Institutes for Research*

Moreiça Ortega, *American Institutes for Research*

Lizanne DeStefano, *University of Illinois*

Robert Linn, *University of Colorado at Boulder*

Conducted by the NAEP Validity Studies (NVS) Panel
September 2007

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

The NAEP Validity Studies Panel was formed by the American Institutes for Research under contract with the National Center for Education Statistics. Points of view or opinions expressed in this paper do not necessarily represent the official positions of the U.S. Department of Education or the American Institutes for Research.

The 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:

Albert E. Beaton
Boston College

Peter Behuniak
University of Connecticut

George W. Bohrnstedt*
American Institutes for Research

James R. Chromy
Research Triangle Institute

Phil Daro*
University of California, Berkeley

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

Donald M. McLaughlin
Statistics and Strategies

Ina V.S. Mullis
Boston College

Jeffrey Nellhaus*
Massachusetts State Department of Education

P. David Pearson
University of California, Berkeley

Lorrie A. Shepard*
University of Colorado at Boulder

David Thissen
University of North Carolina at Chapel Hill

*Steering Committee member, Validity Study of the NAEP Mathematics Assessment

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
1070 Arastradero Road, Suite 200
Palo Alto, CA 94304-1334
Phone: 650/ 843-8192
Fax: 650/ 858-0958

Technical Work Group for the Validity Study of the NAEP Mathematics Assessment:

Cathy Brown
Consultant

Jan de Lange
University of Utrecht, Netherlands

Wade Ellis
West Valley Community College

Kaye Forgione
Achieve, Inc.

Roger Howe
Yale University

Wilfried Schmid
Harvard University

Norman Webb
University of Wisconsin

Executive Summary

Since its founding in 1963, the National Assessment of Education Progress (NAEP) has made a unique contribution to our understanding of American education. It is the only source of information on the educational attainment of all U.S. students, and it is the only vehicle through which states can compare the progress of their students against a common standard. The current main NAEP mathematics trend line extends back to 1990, although there have been two limited revisions to the framework and corresponding incremental changes in the item pool since that time.[1] The NAEP mathematics framework was last updated in 2001 for the 2005 assessment.

In spring 2006, the NAEP Validity Studies (NVS) Panel was asked by National Center for Education Statistics (NCES) to undertake a validity study of the current NAEP mathematics assessment. In particular, NCES asked the NVS Panel to answer the following questions:

1. Does the NAEP framework offer reasonable content and skill-based coverage compared to the assessments of states and other nations?
2. Does the NAEP item pool and assessment design accurately reflect the NAEP framework?
3. Is NAEP mathematically accurate and not unduly oriented to a particular curriculum, philosophy, or pedagogy?
4. Does NAEP properly consider the spread of abilities in the assessable population?
5. Does NAEP provide information that is representative of all students, including students who are unable to demonstrate their achievements on the standard assessment?

Because the framework for grade 12 mathematics was under revision at the time, the validity study was limited to grades 4 and 8.

Approach

To gather information that could address the research questions, the panel undertook a number of expert reviews. Question 1 was addressed by asking a committee of mathematicians and mathematics educators to compare the NAEP framework to the standards and test blueprints of six states that were selected to exemplify the varied approaches to mathematics education found among the states. To these state documents were added standards from two high-performing countries (i.e., Singapore and Japan), Achieve, and the National Council of Teachers of Mathematics.

[1] NAEP also maintains a long-term trend line in mathematics that goes back to 1972–73. It is the main NAEP mathematics assessment, however, that is the focus of this validity study.

Question 2 was addressed by asking another, larger group of mathematicians and mathematics educators to review the full 2007 NAEP item pool and rate the extent to which the item pool accurately represents the body of grade-appropriate content knowledge described by the framework.

For question 3, mathematicians with varying perspectives on the current curriculum controversies related to school mathematics reviewed all items in the 2005 and 2007 NAEP item pools. The NAEP items were intermingled with a random sample of state test items drawn from the 40+ states that had posted released test items on the Web. All items were rated blind for mathematical quality and classified (based on the mean mathematicians' ratings) as adequate, marginal, or seriously flawed.

Questions 4 and 5 were addressed by members of the NVS Panel with special expertise in psychometrics and special populations, respectively.

Findings

The organizations that make up the NAEP system are now, and have always been, joined in a serious learning community. This study is part of the NAEP system and part of the way it learns about itself and improves. Consequently, this report provides a great deal of detail about what could be improved in the NAEP mathematics assessment. The reader should not construe this proliferation of detail as a summative judgment against the NAEP system. Indeed, the NAEP mathematics assessment has been, and remains, an important and invaluable tool for monitoring what U.S. children know and can do in mathematics.

1. The central finding of the validity study is that ***the NAEP mathematics assessment is sufficiently robust to support the main conclusions that have been drawn about U.S. and state progress in mathematics since 1990.***

   NAEP results show achievement in mathematics rising steadily over the years for all subgroups, although gaps among subgroups persist. Validity issues uncovered by this study tended to be local in nature—affecting a particular set of items on a particular subscale. It is reassuring to observe that the gains across the five NAEP subscales are reasonably parallel. That is, there is no evidence that overestimation or underestimation of gains in some one part of NAEP is driving overall trends at either grade level.

2. ***The NAEP framework is reasonable.*** In general, the choices made by the NAEP framework are reasonable when judged against those of the states and nations chosen for comparison. The choices in each content area are generally similar to those made by members of the comparison group. Exhibit A highlights the ways in which NAEP's choices of content are similar to, or different from, the choices of the comparison standards, by content area.

Exhibit A. Summary of content area emphases in the NAEP framework compared to selected comparison standards

Compared to others, the NAEP framework has:		
	Grade 4	**Grade 8**
Number Properties and Operations	Typical emphasis, less number line	Typical emphasis, less squares and square roots, more decimals and fractions
Measurement	More below grade-level content	More below grade-level content, less connections to other content areas
Geometry	More transformations and symmetry, less parallel and perpendicular lines	More content
Data Analysis and Probability	Typical emphasis	More sampling and experiments
Algebra	More patterns, less quantitative relationships	Typical for pre-algebra, (does not cover algebra I), more broad in specifying functions

3. However, ***the NAEP framework and specifications do not provide as much guidance for test developers as they could.*** The framework and specifications dictate relative weights (in percent of items) at the highest hierarchic level, the five content areas, but they provide no guidance on relative priorities across or within subtopics.

   Furthermore, the NAEP framework and specifications are not as well illustrated with exemplar items as are several of the standards in our comparison group, including some of the state standards, the Achieve expectations, and the standards of the two nations.

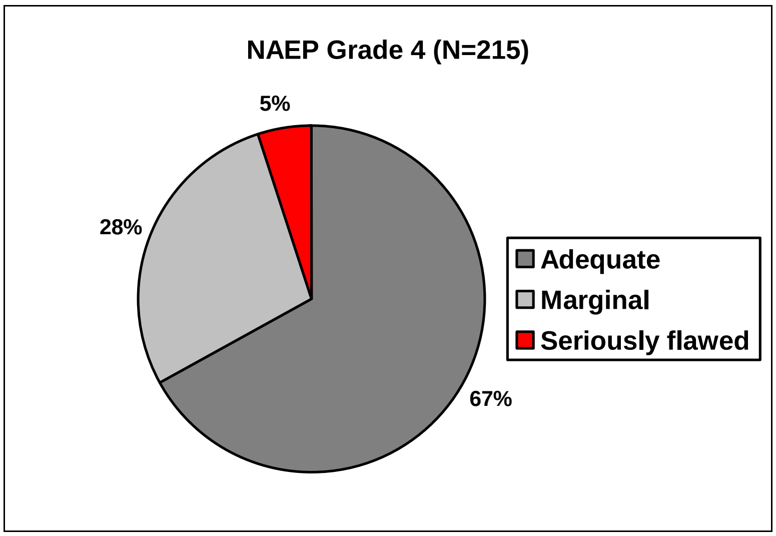
4. ***The NAEP item pool broadly aligns with the framework with some important exceptions.*** All of the items fit somewhere in the framework, and the item counts closely match the prescribed distributions for the five content areas, which is the only level at which the framework stipulates priorities. Nevertheless, there is room for improvement. Virtually every content area at both grade levels had at least one subtopic where the majority of reviewers judged the item set to be lacking on one or more of the three dimensions of alignment used in this report: focus, balance, or reach.[2] The greatest areas of concern were concentrated at grade 8. In particular, at grade 8, there was fairly unanimous criticism of

   - the poor focus and balance of the item set in number properties and operations, and
   - the under-representation of high-complexity items in algebra and measurement.

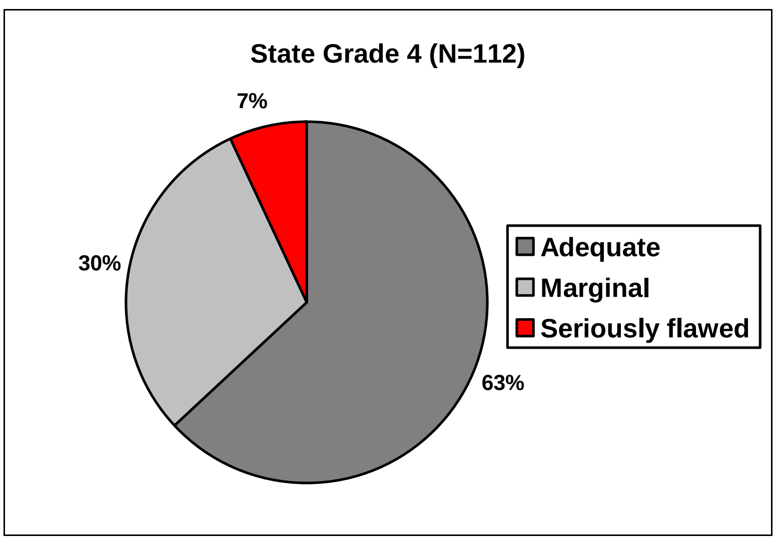
[2] A well-aligned item set is focused on the most important knowledge and know-how in each subtopic, balanced across the range of knowledge and know-how in each content area and subtopic, and reaches to span easier and less advanced, as well as harder and more advanced, aspects of the content in each subtopic.

5. ***Item quality is typical of large-scale assessments but could be better.*** Overall, item quality is typical of large-scale assessments and of sufficient quality to support interpretation of NAEP mathematics scores, but improvements can and should be made. Exhibits B and C display the results of the item-quality analysis for grades 4 and 8. These item classifications are based on the mean ratings of the mathematicians who participated in the study.

Exhibit B. Percentage of adequate, marginal, and seriously flawed NAEP and state items at grade 4

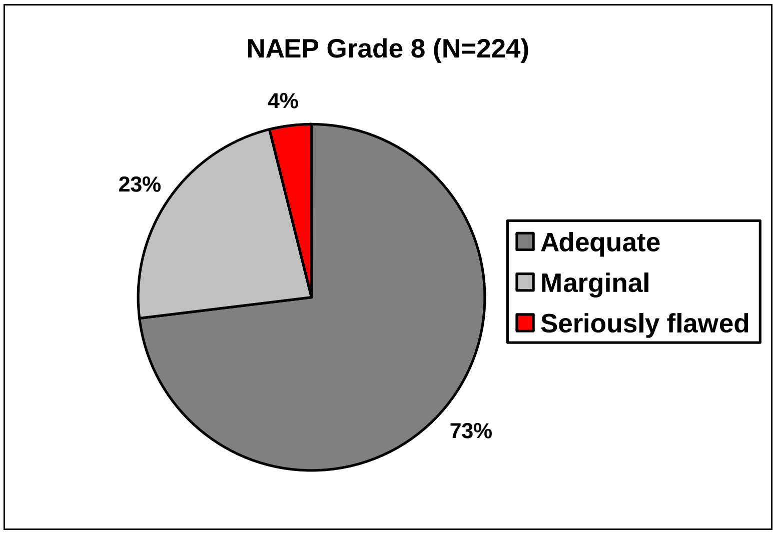


NOTE: NAEP items represent combined 2005 and 2007 item pools.

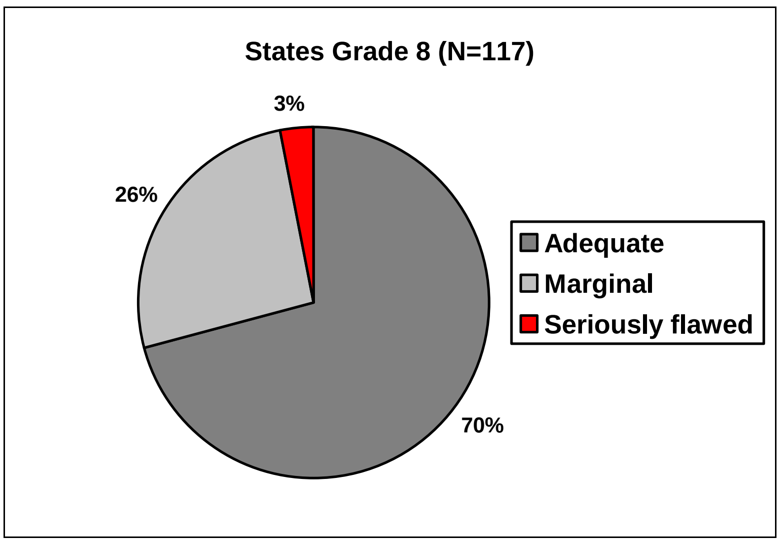


NOTE: State items are a random sample of items from the most recent test forms or item sets released on the Web by 40+ states.

Exhibit C. Percentage of adequate, marginal, and seriously flawed NAEP and state items at grade 8



NOTE: NAEP items represent combined 2005 and 2007 item pools.



NOTE: State items are a random sample of items from the most recent test forms or item sets released on the Web by 40+ states.

As the exhibits show, very similar percentages of items from NAEP and from the comparison sample of states (the latter drawn randomly from the 40+ states with released items on the Web) were classified adequate, marginal, or seriously flawed. The similarity in classifications between NAEP and the state samples indicates that the mathematicians were reacting to common practices in U.S. large-scale assessment, rather than to practices specific to NAEP. At grade 4, nearly all of the seriously flawed NAEP and state items were concentrated among pattern items in the content area of algebra.

The marginal classification encompasses many different kinds of item quality problems, some more serious than others. Nevertheless, the substantial number of items in this classification points to room for improvement.

6. ***Measurement precision is good over a broad range of proficiency but could be better for lower-achieving students.*** For most of the five subscales, and at both grade levels, the standard error of measurement is relatively low for a wide range of achievement. These findings offer positive evidence of NAEP's capacity for accurate reporting of student achievement, especially given that most NAEP reporting is based on the overall mathematics scale (a weighted average of the five subscales). The overall mathematics scale has stronger measurement properties than any one of its constituent subscales.

   Nevertheless, there is room for improvement. Measurement precision is weakest at the bottom of the achievement scale, in a range that includes the performance of large percentages of students from groups of high policy significance.

Recommendations

A number of recommendations flow from this study. Some are consistent with changes already being implemented or are being planned for future testing cycles. Taken together, the recommendations hold strong promise for improving the quality of assessment, not only within the NAEP program, but for U.S. education overall.

1. ***Sharpen the framework***

   The National Assessment Governing Board, which has legislative responsibility for specifying the assessment content, should review and sharpen the current framework.

   A. **Focus: don't worry about leaving things out; worry about targeting the most important things.** When the Governing Board next updates the framework, it should consider reducing the number of objectives. At the same time, it should sharpen the language of the objectives to give test developers a better target rather than using language that tries to include all possibilities.

   B. **Explicitly address high priority issues that cut across content areas.** A revised framework should also provide general guidance on such high priority issues as the extent to which the assessment should include content from earlier grade levels and the approximate proportion of items to be written using the various types of numbers (i.e., whole numbers, fractions, decimals, negative numbers, rates, ratios, and percents).

2. ***Provide detailed implementation plans***

The framework is a public policy document that describes the Governing Board's vision of mathematics assessment to a broad audience. Greater specificity is required for the contractors who develop assessment items under NCES' supervision.

A. **Translate the higher level guidance provided by the framework into detailed implementation plans.** Before beginning item development, NCES should create a formal, written implementation plan for each assessment cycle that translates the higher level guidance provided by the framework. The implementation plan should be developed as quickly as possible after a framework is in place in order to maximize the time available for item writing and review.

B. **Make priorities explicit.** The implementation plan should include, among other things, specification of the relative priorities of the different assessment topics. However, merely allocating percentages of items to content areas is too broad. A reasonable sampling of the mathematics domain will require guidance at each hierarchic level of the framework.

3. ***Define a larger role for exemplar items***

It is time to advance the practice and technology of using exemplar items to communicate expectations. The range and number of items available from released state items, international tests, Achieve, the Dana Center, the Mathematics Diagnostic Testing Project, the Shell Centre, the Freudenthal Institute, national tests (e.g., Japan, Singapore), and other sources is now very large.

A. **Provide ample examples of items.** To clarify their intent, both the Governing Board (in the framework) and NCES (in the implementation plan) should make generous use of example items. Example items are most useful when they are annotated to clearly explain their relationship to the prose descriptions of content.

While individual item exemplars are important as a guide to item quality, sets of exemplars can also be used to clarify the desired attributes of the *item pool*. NCES should compile a coherent body of items to exemplify the intended focus, reach and balance of the assessment. Furthermore, to avoid inbreeding a house style, both the individual and compiled example items should be drawn from multiple sources (e.g., states, nations, and research and development centers), not just from NAEP's past.

B. **Encourage the establishment of a Web-based open bank of released items.** NCES and the Governing Board should encourage the Institute of Education Sciences to support the development and ongoing maintenance of a Web-based open bank of released items. The items should be harvested from as many sources as possible and indexed to a common framework. Such an item bank would both provide exemplars to support NAEP development (as described above) and also serve as an important resource for the states.

4. ***Improve quality assurance for the overall item pool and for individual items***

   Ongoing quality assurance is the particular responsibility of NCES, which has recently undertaken initiatives similar to those described below. NCES should continue and expand upon these current efforts.

   A. **Monitor and manage the focus, balance, and reach of the item pool** across and within the subtopic level of the framework. Once the priorities across assessment topics are clearly specified in the implementation plans, NCES should create routines that monitor the overall item pool each time item blocks are replaced.

   B. **Subject all items to expert review**. The review process should focus on applying individual expertise rather than reaching agreement. Mathematicians, language experts, cognitive scientists, access specialists, and mathematics educators should all be part of the review process, with the expectation that these different types of reviewers will all notice different things. Once the expert critiques have been documented, an independent resolution and revision process should be carried out by NCES.

5. ***Attend particularly to the following aspects of item quality***

   Through the process of research and review, NCES should attend particularly to the following aspects of item quality.

   A. **Sustain attention to the mathematical quality of the items.** Mathematical quality requires that the mathematical content of the items be well expressed. It also requires that any implicit assumptions embedded in the items be fair and not require the student to read the mind of the test developer. Items with hidden assumptions are tests of general cleverness or cultural conditioning, not mathematics.

   B. **Improve the quality of the situated mathematics problems.** Setting mathematics problems in imaginary situations is a basic feature of school mathematics throughout the world and from the earliest grades. Such items can help make the mathematics more accessible, and they can also provide opportunities to assess mathematical modeling skills.

   When items using problem situations are developed and reviewed, the following item quality issues should be attended to:

   - The problem context should, insofar as possible, be familiar to all students.
   - The mathematics in the problem situation should have a purpose that will make sense to the student (authenticity).

C. **Improve the measurement of mathematical complexity.** NCES should turn to nations, centers, and states that are working in different assessment traditions in order to explore divergent approaches to assessing high-complexity reasoning. Simply mounting more intense, well-meant efforts in the same tradition as NAEP has already used is not likely to produce good results. Having sampled ideas from other traditions, alternative approaches to the assessment of complexity could then be examined as part of the recommended program of evidence-based research on item design (see recommendation 6).

D. **Minimize non-construct relevant sources of item difficulty.** Item difficulty is a combination of many factors. In addition to mathematical demands, items may embody demands on auxiliary skills (skills that are necessary for demonstrating competency in the domain, such as reading grade-level text) as well as demands that are merely contaminating (for example, deciphering complex graphical displays). Contaminating skill demands should be avoided entirely, and auxiliary skill demands should be managed so that they do not outweigh the mathematical skill demands of the items.

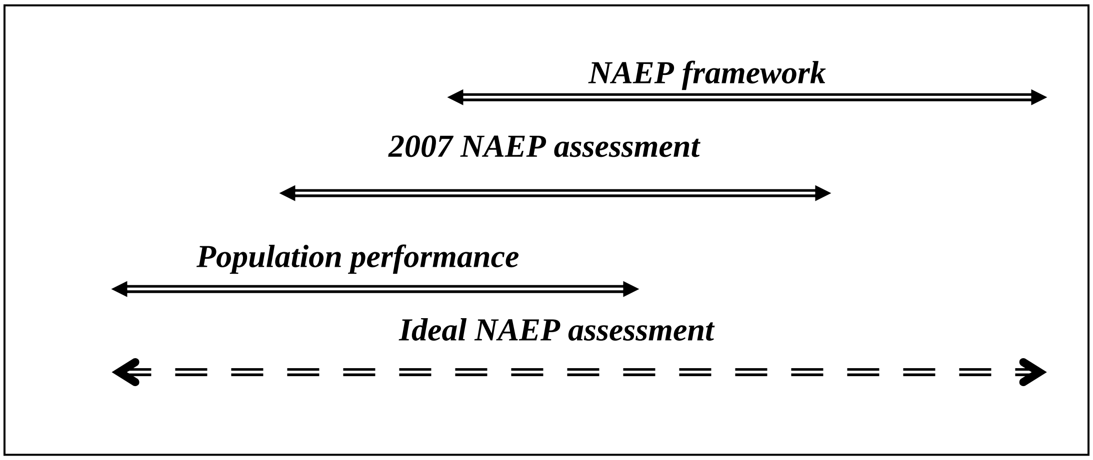
6. ***Undertake a program of evidence-based research on item design***

   Much is known about the psychometric qualities of items as they contribute information to scores constructed through item response theory (IRT) and related methods. Much less is known about item design, student-by-item interactions, and how items relate to the constructs of the domain being assessed (and to the irrelevant domains that contaminate assessment). Resources for research into item performance and construction are seriously underinvested given the importance tests have assumed in the evaluation of the nation’s school systems. It is a recommendation from this study that NCES place research on item quality high on the nation’s education science research agenda.

7. ***Expand the range of item difficulty and curricular reach***

   Comparison of the psychometric properties of NAEP scales to population performance shows that the regions in which the assessment measures with greatest precision are at the leading edge of, if not beyond, where the population is performing. At the same time, comparison of the NAEP item pool to the NAEP framework shows that the mathematics assessment is behind the framework in terms of capturing all of the challenging content implied by the framework. Thus, one can say that the NAEP mathematics assessment is situated “behind” the framework but “ahead” of the population (exhibit D).

Exhibit D. Schematic representation of current and ideal NAEP assessment



Given the mission of NAEP to both lead and reflect, this configuration is probably understandable. However, as an ideal, NAEP should encompass the achievement of the full population—from lowest to highest—and reach from the least to the most advanced content of the framework's domain. To move toward the ideal, the NAEP mathematics assessment needs more easy items, as well as more high-complexity items and more items that reach forward in the curriculum.

8. ***Manage changes in the item pool***

   NAEP must constantly balance the ability to maintain trend lines with the capacity to introduce improvements. A sustained trend line has important policy advantages, particularly given that states are required to track their progress under No Child Left Behind, and these policy considerations have been a major factor in the Governing Board's decisions regarding the extent and timing of framework revisions. The psychometrics of trend measurement also imposes constraints on the rate of change for items in the item pool. Currently NAEP allows no more than 30 percent turnover in items between assessment cycles. Even with assessment cycles scheduled every two years, change—including change aimed at improving the fit to the framework or the quality of the items—is still very slow. NCES should further explore possibilities for accelerating change without compromising trend.

9. ***Move NAEP in the direction of adaptive testing***

   As argued above, the ideal NAEP assessment would provide accurate measurement for the full population of students—from lowest to highest achieving—and also reach from the least to the most advanced content of the domain. However, presenting students with high proportions of items that are either too hard or too easy is both frustrating to the student and a waste of assessment time. Consequently, the Governing Board and NCES should consider the benefits of moving toward some form of adaptive testing, as resources and technology permit.

In sum, NAEP remains a robust measure of mathematics achievement, with a critical role in monitoring educational progress for the nation and the states. The recommendations included in this report are offered in a collegial spirit and with the goal of further improving this important national asset.

Acknowledgements

We would like to thank the many people who contributed to the validity study of the NAEP mathematics assessment. These include the members of the study's steering committee and technical work group: George Bohrnstedt, Cathy Brown, Jan de Lange, Lizanne DeStefano, Wade Ellis, Kaye Forgione, Roger Howe, Gerunda Hughes, Robert Linn, Jeffrey Nellhaus, Wilfried Schmid, Lorrie Shepard, and Norman Webb. Drs. DeStefano and Linn are coauthors of the report, and all the steering committee and technical work group members contributed generously of their time and expertise to help frame the study questions, deliberate the findings, propose recommendations, and review the manuscript for the final report.

In addition, Drs. Jan de Lange (from the technical work group) and Donald McLaughlin (from the NAEP Validity Studies (NVS) panel) contributed essays on aspects of item design which have been included as appendices to this report.

We also thank the mathematicians and mathematics educators who served on our framework comparison committee and our expert panels for alignment analysis and item quality review. All of these individuals are named in the report appendices.

Also thanks to Mark Schneider, the Commissioner of the National Center for Education Statistics (NCES), who requested this independent validity study and provided the support to carry it out. And thanks to the many members of the NCES and Governing Board staff—including Marilyn Binkley, Janis Brown, Peggy Carr, Andrew Kolstad, Alex Sedlacek, and William Tirre at NCES and Mary Crovo at the Governing Board—who provided comments on the report.

Staff from the Educational Testing Service (ETS) were very prompt and helpful in meeting our needs for data and in reviewing the descriptions of NAEP procedures that we included in the report. In particular, Gloria Dione coordinated our data requests and reviewed the manuscript, Jeff Haberstroh reviewed the manuscript, and Mei-Jang Lin provided the standard error of measurement curves used in the report.

Special thanks to Kim Gattis, of the NAEP Education Statistics Services Institute (NESSI) who attended all of the steering committee and expert review meetings to provide us with invaluable information about NAEP processes and procedures.

Finally, thanks to the staff and staff emeritus at American Institutes for Research who produced both this report and the many materials used in the expert reviews. These include Michelle Bullwinkle, Diana Doyal, Phil Esra, and Sandra Smith.

Table of Contents

List of Exhibits

Chapter 1. Introduction

Since its founding in 1963, the National Assessment of Education Progress (NAEP) has made a unique contribution to our understanding of American education. It is the only source of information on the educational attainment of all U.S. students, and it is the only vehicle by which states can compare the progress of their students against a common standard. Assessment results reported by NAEP complement the states' own reports of progress under No Child Left Behind (NCLB) and track the status of achievement gaps for traditionally disadvantaged student groups.

NAEP first assessed mathematics in 1972–73 (the program's fourth year of field operations), and NAEP's long-term trend component has continued an unbroken trend line in mathematics since that time. A second mathematics trend line, now known as main NAEP, was begun in 1990 using an entirely new assessment instrument and offering assessment results for voluntarily participating states as well as for the nation as a whole (Jones & Olkin, 2004). Since that time, the framework that guides the main NAEP mathematics assessment has been updated twice (most recently in 2001 for use in the 2005 assessment), but the changes at grades 4 and 8 were deliberately constrained in order to allow the 1990 trend line for those grade levels to be continued to the present day. This was done out of consideration for the important policy advantages of a sustained trend line for the nation and the states.

The current NAEP schedule includes a mathematics assessment every other year, in which all states and several large urban districts participate.[1]

NAEP is carried out under the guidance of the National Assessment Governing Board and the National Center for Education Statistics (NCES). Over the course of its history, NAEP has frequently sought to improve by studying its own processes, instruments, and procedures. In keeping with this tradition, in spring 2006, NCES asked the NAEP Validity Studies (NVS) Panel, which operates under contract to NCES, to undertake a validity study of the main NAEP mathematics assessment. Since the framework for grade 12 mathematics was under revision at that time, the validity study was limited to grades 4 and 8.

NCES asked the NVS Panel to answer the following questions:

- Does the NAEP framework offer reasonable content and skill-based coverage compared to the assessments of states and other nations?
- Does the NAEP item pool and assessment design accurately reflect the NAEP framework?
- Is NAEP mathematically accurate and not unduly oriented to a particular curriculum, philosophy, or pedagogy?

[1] State participation is required as a condition of Title I funding; districts participate under the Trial Urban District initiative.

- Does NAEP properly consider the spread of abilities in the assessable population?
- Does NAEP provide information that is representative of all students, including students who are unable to demonstrate their achievements on the standard assessment?

A useful way to think about these questions is to map them to an idealized assessment development process as shown in exhibit I-1.

Exhibit I-1. Idealized map of assessment development process

Domain: Mathematics Knowledge and Know-How

Framework: Is coverage reasonable?

Specifications: Is guidance clear?

Item pool: Are the questions of good quality?

Assessment: Is the focus aligned with the framework?

Administration: Is the full range of student abilities assessed?

Scores: Are the measures of student performance accurate and appropriate?

Overview of NAEP

Framework and specifications

Policy for NAEP is set by the National Assessment Governing Board, an independent, bipartisan group whose members include governors, state legislators, local and state school officials, educators, business representatives, and members of the general public. The Governing Board's legislated responsibilities include selecting the subject areas to be assessed and developing assessment objectives and specifications.

To fulfill this mandate, the Governing Board, working through its contractors, produces an assessment framework for each subject area. These frameworks are replaced or updated periodically, balancing the need to stay current with the field against an interest in maintaining trend. As noted, the current NAEP trend line for mathematics goes back to 1990 for grades 4 and 8. The framework, however, was updated prior to the 1996 assessment and again prior to the 2005 assessment.[2]

The framework document is intended to portray the NAEP assessment to a broad audience of educators and the general public as well as to inform the test developer. The

[2] The 2005 framework for grade 12 made more sweeping changes and necessitated a break in the trend line for that grade level. The grade 12 framework is currently undergoing further revisions to align with recent interest in assessing readiness for post-high school activities at grade 12.

framework explicates the structure of the knowledge domain to be assessed, describes the broad outlines of the assessment, defines the achievement levels that will be used to report the assessment, and presents a set of sample questions. A more technical specifications document also is developed by the Governing Board and provided to NCES.

The development of a new or revised framework (and accompanying specifications document) generally requires about 2 years.

Assessment design

Since the mid 1980s, NAEP has employed an assessment design that utilizes student and item sampling to combine broad coverage of the knowledge domain with low respondent burden. The design supports accurate reporting for groups of students, but does not generate reliable scores for individual students.

One element of the design is to develop a large and relatively stable item pool. For example, the 2007 item pool in mathematics includes nearly 170 items per grade level. The size of the item pool allows NAEP to estimate performance in each of five subdomains (content areas) of mathematics, as well as for mathematics overall. The stability of the item pool—only about 30 percent of items are replaced in each assessment cycle—facilitates trend estimation. Balancing stability and change is a constant challenge for the NAEP program.

Items are organized into blocks, each of which typically contains a sampling of the subdomains and cognitive targets to be assessed.[3] Blocks are then assembled into examinee booklets, each containing two blocks of assessment items plus a set of background questions. The assignment of blocks to booklets is done using a balanced incomplete block (BIB) design, which pairs every block with every other block, but does not include all possible orderings of block pairs. In order to enable multiple subject areas to be assessed in the same session, all item blocks are designed to be completed within 25 minutes. In mathematics, this represents approximately 16 to 18 items per block.

Information from all students and all items is combined using Item Response Theory (IRT) methodology to produce achievement estimates for groups of students. Results are reported using either a 300-point or 500-point scale, and basic, proficient, and advanced achievement levels that are set by the Governing Board.[4] By law, NAEP reports results for groups of students at the national and state level defined by race/ethnicity, gender, socio-economic status (as measured by eligibility for free or reduced-price lunch), disability status, and English language learner status.

In order to meet the legislated requirement of 6-month reporting for reading and mathematics, assessments in these subject areas are precalibrated by administering them to smaller samples of students in the year preceding the actual assessment.

[3] Subdomain sampling is less feasible in subject areas such as reading, where an entire block is generally devoted to a single reading passage and associated questions.
[4] Mathematics, like other subject areas that employ a cross-grade scale, uses 500 points.

Assessment development

After the framework and specifications are developed by the Governing Board, these documents are delivered to NCES, which, along with its contractors, is responsible for developing, administering, and scoring the NAEP assessment. As noted above, a large item pool is developed for each assessment and refreshed in accordance with a schedule that is consistent with maintenance of trend. Because the 2005 mathematics framework revision was conceived as an update rather than a break in the trend line, the rate of item replacement remained at the same level prior to, during, and after the period of its introduction.

As new items are developed, they undergo extensive review by multiple parties and over multiple time points. The reviewers include a standing subject-area committee (which has overlapping membership with the planning committee that developed the framework), state representatives, NCES staff, and members of the Governing Board.[5] The reviews for each item block are first conducted prior to pilot testing and are repeated prior to precalibration (or in the case of subjects that are not mandated for 6-month reporting, prior to operational use). However, reviews typically have been focused on the newly developed item blocks for each assessment cycle rather than the item pool as a whole.

Pilot testing and precalibration are carried out during the same annual time window as the operational assessment. Consequently, the entire item development cycle requires approximately 3 years to complete. The low rate of block replacement and the long development cycle combine to create a very long schedule for introducing any significant changes in the composition of the item pool. In the case of mathematics, significant numbers of items in the current item pool have been in operational use since the mid-1990s. This creates some friction between the requirements of the current framework and the composition of the current item pool.

Assessment administration and scoring

The NAEP mathematics assessment is administered to samples of students that are representative of the nation and the states (and of participating large urban districts). Since 2002, the national mathematics sample has been constructed from an aggregate of the state samples.[6] This results in a large overall sample that could support a larger item pool than the current 10-block design. However, there would be significant costs associated with developing, pilot testing, and precalibrating additional item blocks.

Samples are constructed by first drawing representative sets of schools from each state or other participating jurisdiction. Within each sampled school, representative sets of students are sampled from among all the students at the target grade level and then allocated across the subjects that are to be assessed. Student sample lists are reviewed by school representatives to identify any students who have disabilities or English language learner status and to determine whether any of these identified students should be accommodated or excluded from the assessment.

[5] The Governing Board's responsibilities include approving all cognitive and noncognitive NAEP items.
[6] Prior to 2002, two different modes of administration were used for the state and national samples, requiring that separate samples be developed for each purpose.

The assessments are administered by contractor representatives to ensure uniformity of test conditions and security of the item pool. Separate sessions for accommodated students also are provided by the contractor as needed. Students mark their answers directly in the assessment booklets, and completed booklets are shipped to another NAEP contractor for scanning and scoring. Because the mathematics framework requires that approximately half of the assessment time be spent on constructed-response items, the assessment includes substantial numbers of items that must be hand scored after having been scanned as images onto computer files.

Reporting

As noted above, reporting is carried out using the appropriate 300-point or 500-point NAEP scale and the achievement levels set by the Governing Board. Strong efforts are made to release the initial results (for reading and mathematics) within 6 months of the assessment whenever feasible. However, reporting can be delayed when new frameworks or other factors increase the analysis burden. Published reports are relatively brief and focus on national and state trends for the mandated reporting groups. A wide array of additional results is available on the Web and can be accessed using the NAEP data tool. Licenses also are available to researchers who wish to obtain NAEP data files for further analysis.

Organization of this report

The remainder of this report is organized around the five research questions at the heart of the validity study. Chapter 2 discusses the extent to which the NAEP framework offers reasonable content and skill-based coverage when compared to the standards and blueprints used by states and other nations. Chapter 3 reviews the 2007 item pool and considers the extent to which this item pool offers an accurate reflection of the NAEP framework. Chapter 4 considers the quality and mathematical accuracy of the items in the 2005 and 2007 NAEP item pools. Judgments are made absolutely and also in relation to the quality and mathematical accuracy of items randomly sampled from state assessments. Chapter 5 explores the fit of the NAEP assessment to the ability range of the population that takes the assessment. More specifically, the chapter considers the size of the standard error of measurement for each mathematics subscale at different points along the achievement scale. Chapter 6 examines the extent to which NAEP is successful in appropriately including students with disabilities and English language learners when estimating achievement results for the nation and the states. Finally, chapter 7 describes findings that cut across the separate research questions, and it presents a set of recommendations for enhancing the quality of future NAEP mathematics assessments.

Chapter 2. Does the NAEP Framework Offer Reasonable Content and Skill-based Coverage Compared to the Assessments of States and Other Nations?

As explained in chapter 1, the content for each NAEP assessment is described in a framework document that is developed under the supervision of the National Assessment Governing Board. Currently the operative framework for grades 4 and 8 is the *Mathematics Framework for the 2005 National Assessment of Educational Progress* (National Assessment Governing Board, 2004). The *Mathematics Framework,* which is intended to serve an audience of interested educators and policymakers as well as the assessment developers, is approximately 80 pages in length. It organizes the assessment content into five content areas and prescribes the distribution of items across content areas. It also includes brief narrative descriptions (five or six paragraphs) and lists of objectives for each of the five content areas; the lists of objectives are categorized by subtopic within content area and further organized into matrices that allow the reader to trace the evolution of content across grade levels.[7] Finally, the *Mathematics Framework* addresses the distribution of items by format (multiple choice, short constructed response, and extended constructed response) and by a dimension called mathematical complexity. A small number of sample items are included to illustrate the different item formats and the three levels of mathematical complexity.

A second document, the *2005 NAEP Mathematics Assessment and Item Specifications* (National Assessment Governing Board, 2003) provides some additional guidance for item writers. This guidance includes:

- a set of general principles of good item writing and more specific guidance on item writing considerations for English language learners and for students with disabilities;
- brief elaborations of allowable content, which have been added to approximately one third of the individual objectives in the framework;[8] and
- a larger set of 57 sample items (compared to 14 sample items in the framework), which are classified by content objective and level of complexity.[9]

The *Assessment and Item Specifications* does not, however, add any further amplification of the item distribution guidelines provided by the *Mathematics Framework*. As noted

[7] The matrices of subtopics and objectives within each content area are reproduced in appendix A.
[8] For example, for the grade 8 objective on determining the theoretical probability of simple and compound events in familiar or unfamiliar contexts, the specifications add the further guidance: “use familiar contexts such as number cubes, flipping coins, spinners.”
[9] In the content area of measurement, the *Specifications* also includes a page of general guidelines that address the attributes, units, instruments, conversions, and formulas that are appropriate for the assessment.

previously, these guidelines only specify the distribution of items at the level of the five content areas.

Approach

To address the reasonableness of the content and skill-based coverage defined by the NAEP framework and specifications, we compared these documents to the standards and test blueprints of six states that were selected to exemplify the varied approaches to mathematics education found among the states. To these state documents were added standards from two high-performing nations (Singapore and Japan), the Achieve *MAP Mathematics Expectations* (Achieve, n.d.), and the National Council of Teachers of Mathematics (NCTM) *Curriculum Focal Points for Prekindergarten through Grade 8 Mathematics* (NCTM, 2006). *Focal Points* is a recent publication of the NCTM issued in response to criticisms that standards in the United States are too broad and sprawling—a criticism sometimes expressed as “a mile wide and an inch deep.”

The purposes of the various documents differ in important ways. The state and national standards are meant to inform a wide audience about what students should know and be able to do at each grade level. These standards are used to guide the development and adoption of instructional materials; instructional planning from the classroom level to district level; and the design of assessments at all levels, including formative assessments, report cards, and state tests. The NAEP framework, in contrast, has the sole purpose of guiding the construction and interpretation of the NAEP assessment.

States also produce test blueprints, which are derived from their standards. The blueprints stipulate how many and what types of items are needed for each part of the domain of content described by the standards, as well as describing other features of the assessment design. Such stipulations are embedded within the NAEP framework (and supporting specifications document).

Furthermore, because NAEP assesses only at grades 4, 8, and 12—while states assess at every grade from 3 through 8 plus high school—the NAEP framework for a particular grade level might be expected to have more reach into earlier grades. Therefore, while the primary comparisons were carried out within grade level, if a topic was found in NAEP, but not in states, earlier grades from the states were searched to determine if prior coverage explained the absence.

In the first step of the analysis, a framework comparison committee, which included a mathematician, two mathematics educators, and a mathematics standards expert, was formed to assist with the comparisons (see appendix B). The committee members compared the NAEP *Mathematics Framework* to standards documents for California (California Department of Education, 2007), Georgia (Georgia Department of Education, 2006), Indiana (Indiana Department of Education, 2007), Massachusetts (Massachusetts Department of Education, 2000), Texas (Texas Education Agency, 2007), Washington (Office of the Superintendent of Public Instruction, State of Washington, 2006), and Singapore (Ministry of Education Curriculum Planning and Development Division, 2001).

Using the protocol reproduced in appendix C, the committee members answered the following questions for each of the five NAEP content areas:

1. Is NAEP missing something in this content area?
   - Describe what is missing by citing text from the state standard that expresses it best.
   - Indicate where each of the six states and Singapore includes this content in its standards (if at all).
   - Rate how important you think it is that this content be included on NAEP: rate the omission as of minor importance, moderate importance, or major importance.

2. Is NAEP overemphasizing something in this content area?
   - Describe what is overemphasized by citing the NAEP objective(s) in which the over-emphasized content appears.
   - For any topic that you consider overemphasized in NAEP, rate its emphasis in each of the six states and Singapore.

Differences identified by the committee members were then reviewed and interpreted by staff who worked on all aspects of the validity study. These staff also added comparisons to the Japanese standards (Nagasaki et al., 1990), the Achieve and NCTM documents, and the test blueprints for the six states included in the standards comparison.

Findings

Distribution by content area

We begin our discussion of findings with a broad comparison between the numbers of items allocated to each content area by the NAEP framework and by the test blueprints from each of the six comparison states. The comparisons are limited to the six states since the comparison nations, NCTM *Focal Points*, and Achieve did not have tests or test blueprints at these grade levels.

The grade 4 results are presented in exhibit II-1. Here we see that NAEP, like all of the comparison states except Washington, spends most items on number properties and operations in fourth grade.

The only area where NAEP spends a much higher percentage of items than the others is measurement. Measurement is second only to number properties and operations in NAEP, while the comparison states all turn to algebra or geometry after number properties and operations. Furthermore, only NAEP weights measurement more than geometry. All of this suggests the need for a close look at how the NAEP measurement objectives compare to the treatment of measurement elsewhere.

NAEP is comparable to the six comparison states in data and probability.

Exhibit II-1. Grade 4 allocation of test items by content area

<table>
<tr><th></th><th colspan="7">Percent of items</th></tr>
<tr><th>Content Area</th><th>NAEP</th><th>CA</th><th>GA</th><th>IN[1]</th><th>MA</th><th>TX</th><th>WA[2]</th></tr>
<tr><td>Number Properties and Operations</td><td>40</td><td>48</td><td>50</td><td>39</td><td>35</td><td>26</td><td>17</td></tr>
<tr><td>Measurement</td><td>20</td><td rowspan="2">18[3]</td><td>12</td><td>14</td><td>12.5</td><td>14</td><td>17</td></tr>
<tr><td>Geometry</td><td>15</td><td>22</td><td>14</td><td>12.5</td><td>14</td><td>17</td></tr>
<tr><td>Data Analysis and Probability</td><td>10</td><td>6</td><td>10</td><td>0</td><td>20</td><td>10</td><td>17</td></tr>
<tr><td>Algebra</td><td>15</td><td>28</td><td>7</td><td>14</td><td>20</td><td>17</td><td>17</td></tr>
<tr><td>Reasoning and Process Skills</td><td>n/a[4]</td><td>n/a[4]</td><td>n/a[4]</td><td>19</td><td>n/a[4]</td><td>19[5]</td><td>15[5]</td></tr>
</table>

[1]Indiana tests in the fall, so the fourth-grade test is based on third grade standards.
[2]Washington gives ranges of item counts; table entries are within rages.
[3]California combines geometry and measurement.
[4]NAEP, California, Georgia, and Massachusetts embed reasoning and process skills in content areas.
[5]Texas and Washington incorporate multiple content areas in reasoning skills.

SOURCE: National Assessment Governing Board, *Mathematics Framework for the 2005 National Assessment of Educational Progress*, 2004.
California State Board of Education, *California standards test, mathematics blueprint*, 2002. Retrieved May 30, 2007 from http://www.cde.ca.gov/ta/tg/sr/documents/math1105.doc
Georgia Department of Education, Testing Division, *Content Weights for the CRCT GPS-Based CRCT*, 2007. Retrieved May 30, 2007 from http://public.doe.k12.ga.us/ci_testing.aspx?PageReq=CI_TESTING_CRCT
Indiana Department of Education, *Statewide Testing for Educational Progress, ISTEP+ Academic Standards for Grades 3 and 7,* 2004. Retrieved May 30, 2007 from http://www.doe.state.in.us/istep/welcome.html
Massachusetts Department of Education, 2005 MCAS technical report, 2006. Retrieved July 15, 2007 from http://iservices.measuredprogress.org/MCAS2005TechReport.pdf
Texas Education Agency, *Texas Essential Knowledge and Skills (TEKS)*, Chapter 111, Subchapter A, 2007. Retrieved July 2, 2007 from http://www.tea.state.tx.us/teks
Office of the Superintendent of Public Instruction, State of Washington, *Test Specifications for the Washington Assessment of Student Learning Grade 4 Mathematics,* 2005. Retrieved May 30, 2007 from http://www.k12.wa.us/assessment/WASL/MathTestItemSpec.aspx

In eighth grade, NAEP shifts to algebra as its most important area, while number properties and operations ranks second. These two are the typical emphases across the comparison states (exhibit II-2). The unusual emphasis on number properties and operations versus algebra in California may be due to the partitioning of California test takers into students who take algebra (not shown) versus students not enrolled in algebra (shown).

NAEP places more emphasis on the combination of measurement and geometry than most of the comparison states. As at grade 4, NAEP is comparable to the comparison states in data and probability.

Exhibit II-2. Grade 8 allocation of test items by content area

<table>
<tr><th></th><th colspan="7">Percent of items</th></tr>
<tr><th>Content Area</th><th>NAEP</th><th>CA[1]</th><th>GA</th><th>IN[2]</th><th>MA</th><th>TX</th><th>WA[3]</th></tr>
<tr><td>Number Properties and Operations</td><td>20</td><td>37</td><td>22</td><td>25</td><td>26</td><td>20</td><td>15</td></tr>
<tr><td>Measurement</td><td>15</td><td rowspan="2">17[4]</td><td>0</td><td>13</td><td>13</td><td>10</td><td>15</td></tr>
<tr><td>Geometry</td><td>20</td><td>11</td><td>11</td><td>13</td><td>14</td><td>15</td></tr>
<tr><td>Data Analysis and Probability</td><td>15</td><td>14</td><td>17</td><td>17</td><td>20</td><td>16</td><td>15</td></tr>
<tr><td>Algebra</td><td>30</td><td>32</td><td>50</td><td>17</td><td>28</td><td>20</td><td>15</td></tr>
<tr><td>Reasoning and Process Skills</td><td>n/a[5]</td><td>n/a[5]</td><td>n/a[5]</td><td>17</td><td>n/a[5]</td><td>20[6]</td><td>25[6]</td></tr>
</table>

[1]California entries are based on the California general mathematics test for students who did not take algebra in eighth grade.
[2]Indiana tests in the fall, so the fourth-grade test is based on third grade standards.
[3]Washington gives ranges of item counts; table entries are within ranges.
[4]California combines geometry and measurement.
[5]NAEP, California, Georgia, and Massachusetts embed reasoning and process skills in content areas.
[6]Texas and Washington incorporate multiple content areas in reasoning skills.

SOURCE: National Assessment Governing Board, *Mathematics Framework for the 2005 National Assessment of Educational Progress,* 2004.
California State Board of Education, *California standards test, mathematics blueprint,* 2002. Retrieved May 30, 2007 from http://www.cde.ca.gov/ta/tg/sr/documents/math1105.doc
Georgia Department of Education, Testing Division, *Content Weights for the CRCT GPS-Based CRCT,* 2007. Retrieved May 30, 2007 from http://public.doe.k12.ga.us/ci_testing.aspx?PageReq=CI_TESTING_CRCT
Indiana Department of Education, *Statewide Testing for Educational Progress, ISTEP+ Academic Standards for Grades 3 and 7,* 2004. Retrieved May 30, 2007 from http://www.doe.state.in.us/istep/welcome.html
Massachusetts Department of Education, 2005 MCAS technical report, 2006. Retrieved July 15, 2007 from http://iservices.measuredprogress.org/MCAS2005TechReport.pdf
Texas Education Agency, *Texas Essential Knowledge and Skills (TEKS)*, Chapter 111, Subchapter A, 2007. Retrieved July 2, 2007 from http://www.tea.state.tx.us/teks
Office of the Superintendent of Public Instruction, State of Washington, *Test Specifications for the Washington Assessment of Student Learning Grade 4 Mathematics,* 2005. Retrieved May 30, 2007 from http://www.k12.wa.us/assessment/WASL/MathTestItemSpec.aspx

Grain size and explicitness

The NAEP framework, like all standards and frameworks, is an uneven mix of specific and general descriptions of mathematics. Grain size can impact interpretation and influence the number of items spent by an assessment on a particular area of mathematics. For example, one framework might use half a dozen objectives to spell out an area of mathematics, while another uses just one objective for the same area. In the absence of other guidance, the area with more objectives will tend to get a greater proportion of the items.

A comparison of how NAEP, Massachusetts, Washington, Singapore, and Achieve set expectations for adding fractions in fourth grade illustrates variations in explicitness (exhibit II-3).[10] Adding fractions is an important case to examine because fourth grade is

[10] For this comparison, we use the NAEP specifications, which add some additional specificity beyond that offered in the framework.

early in the progression of instruction on this topic. Although many states introduce the addition of fractions by fourth grade, general computational fluency with fractions is not expected until fifth or sixth grade in most states (Reys et al., 2006). This is also true in Singapore and Japan. The NCTM *Focal Points* stress other aspects of learning fractions at fourth grade, including the study of equivalent fractions, which is not explicit in NAEP (but is explicit in most of our comparison states).

In exhibit II-3, NAEP is the least explicit on adding fractions, specifying only that items should focus on “common and decimal fractions.” Massachusetts makes its aim clear by grounding the addition of “common” fractions in concrete objects or visual models as suits the early stage of study about this topic. Washington uses examples that make explicit the types of tasks (including classroom tasks) that students are expected to perform in meeting the expectation for adding fractions.

Singapore explicitly excludes sums with more than two different denominators and limits denominators to 12 or less. The Achieve expectations are the most explicit (and ambitious). Like Singapore, Achieve provides specific notes about the expectations that place them in a context and limit the demand.

Exhibit II-3. Comparison of NAEP, Massachusetts, Washington, Singapore, and Achieve objectives for adding fractions at grade 4

NAEP	Massachusetts
Add and subtract: • whole numbers, or • fractions with like denominators, or • decimals through hundredths **Include items that are not placed in context and require computation with common and decimal fractions, as well as items that use a context.**	Learning Standards *Students engage in problem solving, communicating, reasoning, connecting, and representing as they:* 4.N.18 Use concrete objects and visual models to add and subtract common fractions.

Washington

Grade-Level Expectation	EXAMPLES
Understand the meaning of addition and subtraction of like-denominator fractions.	EX Represent addition and subtraction of fractions with like-denominators using numbers, pictures, and models including everyday objects, fraction circles, number lines, and geoboards. EX Use joining, separating, part-part-whole, and comparison situations to add and subtract like-denominator fractions. EX Translate a given picture or illustration into an equivalent symbolic representation of addition and subtraction of like-denominator fractions. EX Select and/or use an appropriate operation to show understanding of addition and subtraction of like-denominator fractions.

Exhibit II-3. Comparison of NAEP, Massachusetts, Washington, Singapore, and Achieve objectives for adding fractions at grade 4 (cont.)

Singapore

Topics/Outcomes	**Remarks**
a) add and subtract • like fractions • related fractions	• Denominators of given fractions should not exceed 12 • **Exclude** sums involving more than 2 different denominators

Achieve

4.6 Add and subtract simple fractions	Notes:
4.6a Add and subtract fractions by rewriting them as equivalent fractions with a common denominator. • Solve addition and subtraction problems with fractions that are less than 1 and whose denominators are either (a) less than 10 or (b) multiples of 2 and 10, or (c) multiples of each other. • Add and subtract lengths given as simple fractions (e.g., 1/3 + 1/2 inches). • Find the unknowns in equations such as: 1/8 + [ ] = 5/8 or 3/4 - [ ] = 1/2.	Note: The idea of common denominator is a natural extension of common multiples introduced above. Addition and subtraction of fractions with common denominators was introduced in Grade 3. Note: To keep calculations simple, do not use mixed numbers (e.g., 3 1/2) or sums involving more than two different denominators (e.g., 1/3 + 1/2 + 1/5). Also, do not stress reduction to a 'simplest' form (because, among many reasons, such forms may not be the simplest to use in subsequent calculations).

SOURCE: National Assessment Governing Board, *2005 NAEP Mathematics Assessment and Item Specifications*, 2003.
Massachusetts Department of Education, *Mathematics Curriculum Framework*, 2000.
Office of the Superintendent of Public Instruction, State of Washington, *Mathematics Grade Level Expectations*, 2006.
Singapore Ministry of Education Curriculum Planning and Development Division, *Primary Mathematics Syllabus*, 2001.
Achieve, Inc., *MAP Mathematics Expectations*.

Furthermore, when it comes to distributing items across topics, some state blueprints specify the distribution at finer grain sizes than NAEP does. NAEP does not have the equivalent of a test blueprint, and therefore the NAEP test developer makes the de facto decision on how many items to spend on each subtopic and each objective. NCES, which oversees the test developer, could provide more guidance by specifying the allocation of items across and within subtopics (and not just by content area). In the absence of such specification, the test developer has no criteria for deciding focus or balance at these grain sizes.

Complexity and reasoning standards

All standards documents reviewed for this study identify certain cross-cutting objectives like mathematical reasoning or problem solving that are meant to characterize the kinds of thinking a student is expected to do with the mathematical knowledge being assessed. Some standards documents have a distinct set of standards for these process goals, often accompanied by an explanatory essay. As was shown in exhibits II-1 and II-2, some of our comparison states specify percentage distributions of items for reasoning standards in addition to the content domains. And, as the state documents explain, items that demand more complex reasoning often involve mathematical content from more than one content area.

NAEP takes a different approach. NAEP defines three levels of complexity (high, medium, and low) to characterize the demands that different items place on the thinking and performance of the test taker. For example, high-complexity items are said to:

> *…make heavy demands on students, who must engage in more abstract reasoning, planning, analysis, judgment, and creative thought. A satisfactory response to the item requires that the student think in abstract and sophisticated ways. Items at the level of high complexity may ask the student to do any of the following:*
>
> - *Describe how different representations can be used for different purposes.*
> - *Perform a procedure having multiple steps and multiple decision points.*
> - *Analyze similarities and differences between procedures and concepts.*
> - *Generalize a pattern.*
> - *Formulate an original problem, given a situation.*
> - *Solve a novel problem.*
> - *Solve a problem in more than one way.*
> - *Explain and justify a solution to a problem.*
> - *Describe, compare, and contrast solution methods.*
> - *Formulate a mathematical model for a complex situation.*
> - *Analyze the assumptions made in a mathematical model.*
> - *Analyze or produce a deductive argument.*
> - *Provide a mathematical justification.* (National Assessment Governing Board, 2004, p. 43)

NAEP attempts to make complexity an attribute of items rather than response to items, but the descriptions of complexity make frequent reference to what the test taker is doing through the use of language like “abstract reasoning” and “generalize.” In effect, NAEP addresses similar reasoning expectations to those addressed in other standards.

How these expectations are to be realized in the assessment items is somewhat mysterious. More examples of items at different levels of complexity would help, particularly if there is an explanation of why an example is judged to be high versus medium (or medium versus low) in complexity. Singapore, for example, provides a more complete explanation of these types of expectations (see pages 1 through 5 of *Secondary Mathematics Syllabuses, Ministry of Education*, Singapore, 2006). Achieve also gives examples of multi-concept problems with high cognitive demand.

Grade level

All the comparison standards address each grade level. NAEP alone has fourth grade and eighth grade without grade-level neighbors. Because of this, one would expect NAEP to be less exclusive about grade-level content than the comparison standards. This was the case, up to a point. For example, NAEP includes topics at fourth grade that many states do not include at fourth grade because they cover them in third grade. On the other hand, some topics, such as fluent addition of fractions, which states tend to include in standards for grades 5 and 6, receive relatively less emphasis in grade 8 NAEP because they are further off grade level.

For the most part, those off-grade topics that were included in NAEP were judged appropriate because of the importance of the topics and the imbalance that would be apparent from their omission. Nonetheless, including such topics could lead to issues in the management of rigor and accessibility.

Eighth grade presents a special problem for both state assessments and NAEP. Most states now encourage students to take algebra I in eighth grade. In many states, half the eighth-grade students take algebra. That being so, what should be included in the framework for eighth-grade tests? Some states (e.g., California) offer two tests at eighth grade: one for students enrolled in algebra, and one for students not yet enrolled in algebra. (It is the non-algebra California test that is referenced in exhibit II-2.) Given this situation, how should the NAEP framework handle content from algebra I?

Detailed comparisons, grade 4

Number properties and operations

At grade 4, the content that NAEP includes under number properties and operations is typical, although NAEP is less specific about equivalent fractions and places less emphasis on the number line than other standards in our comparison group. The six states, NCTM *Focal Points*, Japan, and Singapore all give more attention to the number line than NAEP. Some address the number line in the number properties and operations standards, while others address it in geometry, in algebra, or across several content areas.

NAEP limits multiplication to 2-digit by 2-digit numbers (except for items in blocks that allow calculators). Many of the comparison standards require 2-digit by 3-digit or 2-digit by multi-digit multiplication. Some strategies that work for 2-digit numbers do not generalize well to multi-digit numbers. Multi-digit numbers require more general methods, so this difference is not trivial.

Measurement

The NAEP measurement content area, on which more grade 4 items are spent than any other content area except number properties and operations, differs in some respects from the treatment of measurement content in the comparison standards. Like NAEP, all the comparison standards (including those of Singapore and Japan) include solving problems with area and perimeter as a central focus at fourth grade. Some explicitly include volume, weight, time, and money. Most (California is the exception) also include a focus on problems with units, simple unit conversion, and understanding concepts of units. This is consistent with the NCTM *Focal Points*, which emphasizes area and units. Texas goes the furthest in building the mathematical foundation for work in science and technology (exhibit II-4).

Exhibit II-4. Texas objectives for measurement at grade 4

Texas Knowledge and Skills	The student is expected to:
(4.11) Measurement. The student applies measurement concepts. The student is expected to estimate and measure to solve problems involving length (including perimeter) and area. The student uses measurement tools to measure capacity/volume and weight/mass.	(A) estimate and use measurement tools to determine length (including perimeter), area, capacity and weight/mass using standard units SI (metric) and customary; (B) perform simple conversions between different units of length, between different units of capacity, and between different units of weight within the customary measurement system; (C) use concrete models of standard cubic units to measure volume; (D) estimate volume in cubic units; and (E) explain the difference between weight and mass.
(4.12) Measurement. The student applies measurement concepts. The student measures time and temperature (in degrees Fahrenheit and Celsius).	(A) use a thermometer to measure temperature and changes in temperature; and (B) use tools such as a clock with gears or a stopwatch to solve problems involving elapsed time.

SOURCE: Texas Education Agency, *Texas Essential Knowledge and Skills (TEKS)*, Chapter 111, Subchapter A, 2007. Retrieved July 2, 2007 from http://www.tea.state.tx.us/teks

Given the large investment of items on measurement in NAEP, there is reason to be concerned with the way in which measurement is treated in the NAEP framework. Along with objectives typical of the comparison states' and nations' standards at grade 4, the NAEP framework includes a number of objectives that are more typical of earlier grades. (See exhibit II-5 for a selection of NAEP measurement objectives that are found in earlier grades elsewhere.) There may be good reasons for this choice, given NAEP's purpose, but items assessing these objectives dilute the overall rigor of NAEP.[11] The mathematicians who reviewed NAEP items (see chapter 4) observed that there were too many items spent on these off-grade-level objectives in measurement.

[11] NAEP is a difficult test for the population assessed. However, easier and off-grade content should be introduced more systematically. In particular, there appears to be no good justification for concentrating off-grade content in measurement.

Exhibit II-5. Selected grade 4 NAEP measurement objectives that other frameworks often place in earlier grades

1) Measuring physical attributes

(a) Identify the attribute that is appropriate to measure in a given situation.

(b) Compare objects with respect to a given attribute, such as length, area, volume, time, or temperature.

(c) Estimate the size of an object with respect to a given measurement attribute (e.g., length, perimeter, or area using a grid).

(g) Select or use appropriate measurement instruments such as ruler, meter stick, clock, thermometer, or other scaled instruments.

2) Systems of measurement

(a) Select or use appropriate type of unit for the attribute being measured such as length, time, or temperature.

SOURCE: National Assessment Governing Board, *Mathematics Framework for the 2005 National Assessment of Educational Progress,* 2004.

In many of the comparison states and in Singapore, calculations drawn from measurement are explicitly addressed within the measurement section. Singapore details what to include, and what to exclude, in the study of units and simple unit conversions; at grade 4 the focus is on getting the concept, and complicated calculations are excluded.

NAEP makes a distinction between number properties and operations items with a measurement context and measurement items with numbers. All items classified in the measurement content area depend on some specific knowledge of measurement. Thus, the framework states: "…an item that asks the difference between a 3-inch and a 1 ¾ -inch line segment is a number item, while an item comparing a 2-foot segment with an 8-inch line segment is a measurement item." (National Assessment Governing Board., 2004, page 19)

The distinction seems aimed at classifying items according to the aspect of mathematics in the items that is most demanding. But the balancing of different mathematical demands is a general issue that cuts across all content areas. Accenting it here may have led to a different treatment for measurement than for other content areas.

Geometry

Geometry is difficult to compare because it has the widest variation across states in what is taught at which grade level. At grade 4, the choice of topics and emphases in geometry is no more unique in NAEP than in any of our comparison states or nations. That said, NAEP emphasizes symmetry more than the comparison states, but no more than Singapore. NAEP also emphasizes transformations, which is recommended in the NCTM *Focal Points,* but not emphasized by the states in our sample.

NAEP has less emphasis on parallel lines and angles than do the comparison states and nations.

The comparison nations, Japan and Singapore, place more emphasis on three-dimensional geometry, including two-dimensional representations of three-dimensional figures, than NAEP or the states.

Data analysis and probability

NAEP's treatment of data analysis and probability at grade 4 is very typical of the treatment afforded in the standards used for comparison.

Algebra

When compared to other standards in our comparison group, NAEP's grade 4 algebra standards place a different set of emphases within the "pattern" subtopic. One way to understand this is to distinguish between two important kinds of mathematical competencies. One set involves analyzing the relationship between two quantities that vary together. For example: If tables have 4 legs, how many legs do 2 tables have, 10 tables have, and *n* tables have? This is one of the foundations of the concept of function. Call these the "quantities vary together" competencies. NAEP emphasizes these less than others in the comparison group at grade 4. The second set of competencies involves analyzing sequences of numbers or objects that grow in some regular way. For example: If the pattern 19, 22, 25, 28, 31, _ continues to increase by the same rule, what will the next number be? At fourth grade, these competencies focus on determining the "next step" or expressing a rule for the next step. NAEP emphasizes these competencies more than most. In the subtopic of patterns, relations and functions, the NAEP framework has five objectives. Four of them are about patterns or sequences and one is about the relationship between quantities.

The treatment of patterns in the Massachusetts, Texas, and Washington standards are shown in exhibit II-6. As can be seen, the attention to patterns is more focused and limited than in NAEP. Furthermore, the connection between patterns and other mathematics (number operations and relations between quantities) is explicit.

Exhibit II-6. Massachusetts, Texas, and Washington objectives for patterns at grade 4

Massachusetts

Learning Standards	
	Students engage in problem solving, communicating, reasoning, connecting, and representing as they:
4.P.1	Create, describe, extend, and explain symbolic (geometric) and numeric patterns, including multiplication patterns like 3, 30, 300, 3000, ….
4.P.2	Use symbol and letter variables (e.g., Δ, x) to represent unknowns or quantities that vary in expressions and in equations or inequalities (mathematical sentences that use =, <, >).
4.P.3	Determine values of variables in simple equations, e.g., 4106 – ∇ = 37, 5 = ○ + 3, and □ – ○ = 3.
4.P.4	Use pictures, models, tables, charts, graphs, words, number sentences, and mathematical notations to interpret mathematical relationships.
4.P.5	Solve problems involving proportional relationships, including unit pricing (e.g., four apples cost 80¢, so one apple costs 20¢) and map interpretation (e.g., one inch represents five miles, so two inches represent ten miles).
4.P.6	Determine how change in one variable relates to a change in a second variable, e.g., input-output tables.

Exhibit II-6. Massachusetts, Texas, and Washington objectives for patterns at grade 4 (cont.)

Texas

Patterns, relationships, and algebraic thinking.
4.6) The student uses patterns in multiplication and division. (A) use patterns and relationships to develop strategies to remember basic multiplication and division facts (such as the patterns in related multiplication and division number sentences (fact families) such as 9 x 9 = 81 and 81 ÷ 9 = 9); and (B) use patterns to multiply by 10 and 100.
(4.7) The student uses organizational structures to analyze and describe patterns and relationships. The student is expected to describe the relationship between two sets of related data such as ordered pairs in a table.

Washington

Grade Level Expectation	**Examples**
Describe a rule for a pattern with a single arithmetic operation.	EX Identify or generate a rule for a pattern with a single arithmetic operation in order to extend or fill in parts of the pattern. EX Show growing patterns using objects or pictures and explain the rule. EX Determine the operation that changes the elements of one set of numbers into the elements of another set of numbers such as using a function machine. EX Explain why a given rule fits a pattern based on a single arithmetic operation in the rule.

SOURCE: Massachusetts Department of Education, *Mathematics Curriculum Framework,* 2000.
Texas Education Agency, *Texas Essential Knowledge and Skills (TEKS),* Chapter 111, Subchapter A, 2007. Retrieved July 2, 2007 from http://www.tea.state.tx.us/teks
Office of the Superintendent of Public Instruction, State of Washington, *Mathematics Grade Level Expectations,* 2006

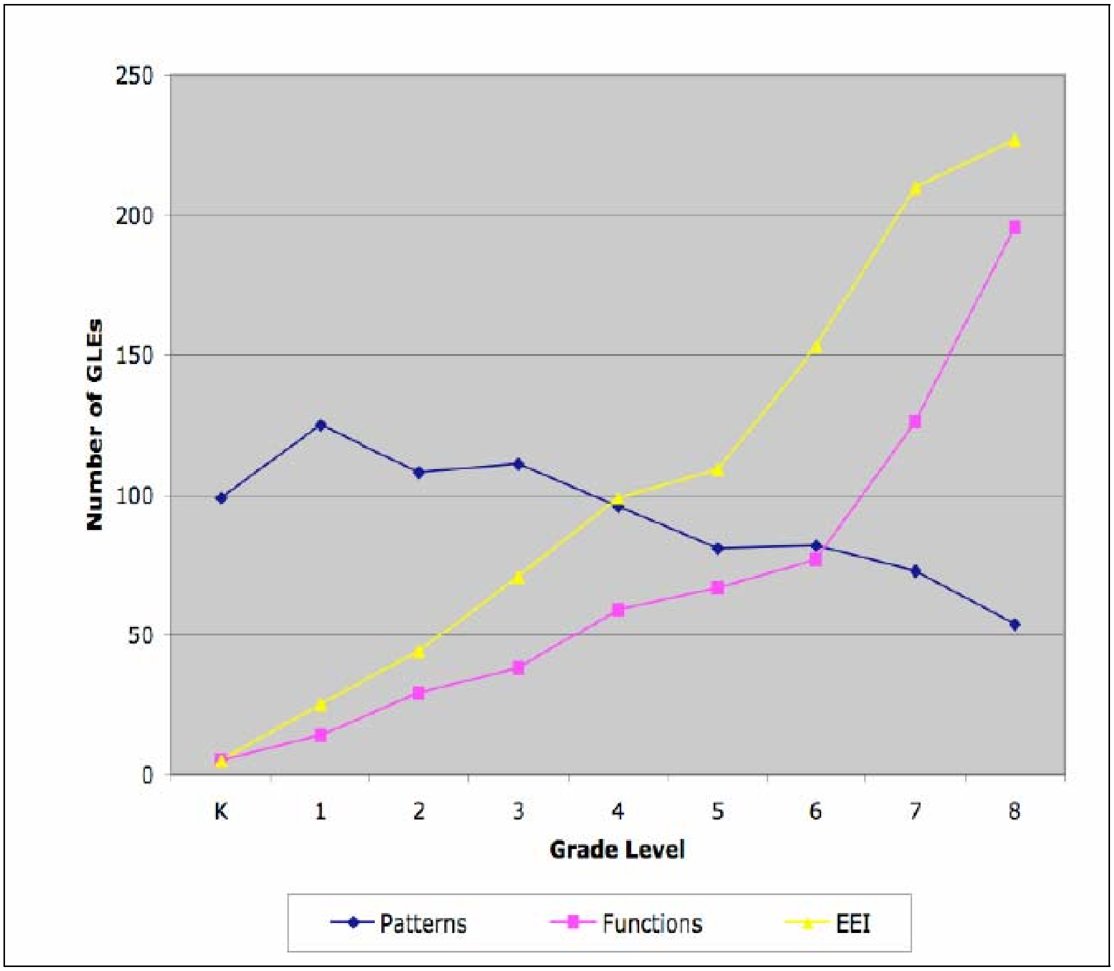
The Japanese standards emphasize learning to "…represent and investigate the relations between two quantities that vary (together)..." The representations identified are tables of ordered pairs and graphs. The Japanese also call for "representing" and "interpreting" mathematical relations in quantitative expressions. Patterns are not mentioned explicitly, although they are implicit in tables of ordered pairs.

Singapore does not treat algebra as a separate part of its standards. It does make the study of applied quantities (e.g., money, measures, and mensuration) explicit. In this context, students are expected to use a mathematical understanding of the relations between quantities to solve problems in context. As a result, units (e.g., centimeters, units of money, grams) play an important role in the problems.

NAEP's treatment of sequences and patterns in fourth grade emphasizes the rule that gets the next term (recursive rule). However, the mathematical foundation for the target concept of functions in later grades is closer to the Japanese emphasis on the relationship between two quantities (where one varies as a function of the other). Furthermore, the mathematicians who addressed mathematical accuracy in our study (see chapter 4) found NAEP items related to recursive-rule patterns flawed much more frequently than items in any other area. For both these reasons, NAEP may want to reconsider its item balance within the subtopic of patterns, relations, and functions.

Looking more broadly across the algebra content area, it is informative to examine the ways in which different standards treat the development of algebra content across grade levels. Reys et al. (2006) compared the number of grade-level expectations in state frameworks that related to three areas of algebra K–8: patterns, symbolic algebra (equations, expressions, and inequalities), and functions. As shown in exhibit II-7, the authors found patterns dominant through grade 3. At grade 4, symbolic algebra—often referred to as "generalization of arithmetic" in this context—catches up to patterns, with functions still trailing. Functions overtake patterns in grade 6, as symbolic algebra continues to ascend. The NAEP framework, although it only addresses grades 4 and 8, would fit this national picture.

Exhibit II-7. Number of grade-level expectations (GLEs) in state standards that address patterns, functions, and equations, expressions and inequalities (EEI), across grade levels



SOURCE: Center for the Study of Mathematics Curriculum, University of Missouri, *The intended mathematics curriculum as represented in state-level curriculum standards: Consensus or confusion?* 2006.

Similar to the Japanese standards, the NCTM *Focal Points* use "four legs for every dog" as an example of a pattern that arises from one quantity varying with another, at third grade. But in fourth grade, the *Focal Points* place the emphasis on recognizing, extending, and developing rules for sequences.

The Japanese also emphasize learning to represent and interpret mathematical relations in expressions with the four operations, parentheses, and the equal sign. Similarly, NAEP, along with the six comparison states, includes objectives that amount to this content.

Detailed comparisons, grade 8

Number properties and operations

While the reviewers of the NAEP item pool in our study's alignment analysis (described in chapter 3) commented explicitly that they liked the way that the NAEP framework handled the number properties and operations content area at grade 8, there are some differences between NAEP and the comparison states that merit consideration when the NAEP framework is revised and when results are interpreted.

The treatment of properties of number and operations (subtopic 5) in NAEP is off the common aim. In most of the comparison states, the emphasis is on extending the basic properties of arithmetic into algebra (e.g., the distributive property, inverse operations, identity property) and into more complicated situations. NAEP, in contrast, emphasizes topics from number theory: odd and even numbers, primes and factorization, and divisibility. This amounts to a different focus and, given the importance of preparing for algebra, one that deserves reconsideration.

Another difference is that NAEP has less emphasis and specificity on roots and exponents. Some of the comparison states are more explicit about operations with roots and the inverse relationship between square roots and squaring. The treatment of rational numbers in NAEP includes more topics from earlier grade levels (e.g., place value) in the subtopic of number sense than the states do. Given NAEP's broad purpose of measuring progress, and its large gap between fourth and eighth grades, this may be appropriate.

Measurement

Here, as in fourth grade, NAEP includes some objectives that states have at earlier grade levels, particularly within subtopic 1, measuring physical attributes. For example, consider objective 1b: compare objects with respect to length, area, volume, angle measurement, weight, or mass.

In chapter 3 of this report, the reviewers judged that NAEP items on measuring physical attributes were well aligned with the framework. However, they also agreed that the complexity of these items was too low. Given that more than 20 items in the 2007 NAEP item pool were spent on this subtopic, this is worthy of examination.

In particular, with the NAEP framework in measurement skewing toward content from lower grades, students who are learning more mature content may not show their progress, and NAEP may be underestimating achievement. Furthermore, if students are taught this content in earlier grade levels in most states, then NAEP is only measuring the knowledge that has been retained. Numerous studies show that retention of learning declines over time for students in the lower half of the achievement distribution. For this reason, NAEP also may be underestimating what those (lower achieving) students previously learned.

Geometry

As in fourth grade, the comparison states and nations vary more in geometry than in other areas as far as which grade level teaches which topics. Nevertheless, it is evident that NAEP has more geometry than the comparison states or nations (or the NCTM *Focal Points*) in eighth grade. But the NAEP objectives are mostly so sweeping that they could include content mostly from earlier grades. This is particularly evident in the subtopic on dimension and shape.

Also, as with fourth grade, NAEP is stronger in its demands for transformational geometry and weaker in its demands for angles and parallel lines than some of the comparison standards, including the NCTM *Focal Points*. NAEP does not include any constructions. Most of the comparison states also assign constructions to high school. The comparison nations, however, include constructions such as perpendicular bisectors and angle bisectors at grade 8.

Because the objectives are so sweeping, and because there is no other documentation to clarify these objectives, guidance to the test developer is particularly weak in the geometry content area. Evidence from chapter 3 of this report indicates the reviewers were not pleased with the item pool in this content area.

Data analysis and probability

NAEP's data analysis and probability content area is more ambitious than that of the comparison states or nations. However, the comparison states and nations, as well as the NCTM *Focal Points*, align with NAEP in stressing the use of mean, median, and mode to understand a data set or distribution. Singapore is very similar to NAEP for data, except it does not include sampling and experiments. The sampling and experiments subtopic is less emphasized in all the comparison states except California, which is similar to NAEP.

The NAEP probability subtopic also is more ambitious than that of the comparison states. Japan and Singapore do not have probability at eighth grade.

Algebra

The NAEP framework does not assume that students have completed algebra I or the equivalent by eighth grade. Although, in many states, nearly half the eighth-grade students are taking algebra I, few have completed the course when eighth-grade NAEP is administered in January and February. All comparisons in this report are to state grade-standards that precede algebra I. Japan and Singapore do not have an algebra I course, per se. They distribute the content of algebra I over grades 6, 7, 8, and 9 (side-by-side with the content of geometry, so that their students go directly to advanced algebra in grade 10).

The scope of NAEP algebra is comparable to that of the comparison states and nations. The subtopic on equations and inequalities is specific and comparable to the most specific standards from the states. In contrast, the treatment of functions is painted with a broad brush. The core concepts of functions are broken up across three subtopics:

1. Patterns, relations, and functions
2. Algebraic representations
3. Variables, expressions, and operations

This may account for the fact that neither the idea nor the word *variable* appears in any NAEP objective related to functions. One might argue that "variable" is entailed in the use of the word function, but this just points to the (odd) grain size at which the NAEP framework portrays and slices this important topic. This leads to a portrayal of functions that is neither concise nor specific.

Here, for example, are the Georgia Standards related to functions and graphs at grade 8 (Georgia Department of Education, 2006):

M8A3. Students will understand relations and linear functions.

a. Recognize a relation as a correspondence between varying quantities.
b. Recognize a function as a correspondence between inputs and outputs where the output for each input must be unique.
c. Distinguish between relations that are functions and those that are not functions.
d. Recognize functions in a variety of representations and a variety of contexts.
e. Use tables to describe sequences recursively and with a formula in closed form.
f. Understand and recognize arithmetic sequences as linear functions with whole number input values.
g. Interpret the constant difference in an arithmetic sequence as the slope of the associated linear function.
h. Identify relations and functions as linear or nonlinear.
i. Translate among verbal, tabular, graphic, and algebraic representations of functions.

M8A4. Students will graph and analyze graphs of linear equations and inequalities.

a. Interpret slope as a rate of change.
b. Determine the meaning of the slope and y-intercept in a given situation.
c. Graph equations of the form y = mx + b.
d. Graph equations of the form ax + by = c.
e. Graph the solution set of a linear inequality, identifying whether the solution set is an open or a closed half-plane.
f. Determine the equation of a line given a graph, numerical information that defines the line or a context involving a linear relationship.
g. Solve problems involving linear relationships.

And here is NAEP (National Assessment Governing Board, 2004, pp. 33–35):

1) Patterns, relations and functions

a) Recognize, describe, or extend numerical and geometric patterns using tables, graphs, words, or symbols.
b) Generalize a pattern appearing in a numerical sequence or table or graph using words

or symbols.

c) Analyze or create patterns, sequences, or linear functions given a rule.
e) Identify functions as linear or nonlinear or contrast distinguishing properties of functions from tables, graphs, or equations.
f) Interpret the meaning of slope or intercepts in linear functions.

2) Algebraic representations

a) Translate between different representations of linear expressions using symbols, graphs, tables, diagrams, or written descriptions.
b) Analyze or interpret linear relationships expressed in symbols, graphs, tables, diagrams, or written descriptions.
c) Graph or interpret points that are represented by ordered pairs of numbers on a rectangular coordinate system.
d) Solve problems involving coordinate pairs on the rectangular coordinate system.
e) Make, validate, and justify conclusions and generalizations about linear relationships.
g) Identify or represent functional relationships in meaningful contexts including proportional, linear, and common nonlinear (e.g., compound interest, bacterial growth) in tables, graphs, words, or symbols.

3) Variables, expressions, and operations

b) Write algebraic expressions, equations, or inequalities to represent a situation.
c) Perform basic operations, using appropriate tools, on linear algebraic expressions (including grouping and order of multiple operations involving basic operations, exponents, roots, simplifying, and expanding).

Is the difference one of style, emphasis, or content? Certainly, Georgia's standards would provide clearer guidance to the test developer. Conversely, NAEP's framework—unless further delineated through supporting documentation—would allow wide latitude. Looking ahead to chapter 3 of this report, the reviewers of the item pool were moderately satisfied with the fit of the algebra items to the framework, but commented that there was too little emphasis on creating patterns, sequences, or linear functions from rules (objective1c), comparing linear and non-linear functions (objective 1e), and interpreting the meaning of slopes and intercepts (objective 1f). The reviewers also felt that the graphing items fell short on complexity and on tapping conceptual understanding. It is hard to say that the way the framework was written caused these problems, but it easy to see that more explicit illustration and explanation could prevent such problems in the future.

Summary

In general, the choices made by the NAEP framework are reasonable, when judged against the states and nations chosen for comparison. The choices in each content area are generally similar to those in the comparison group. Nevertheless, comparisons between the NAEP framework and other standards reveal some differences in content and approach, and these differences raise questions because they define options for future directions in NAEP. Exhibit II-8 highlights, by content area, the ways in which NAEP's choices of content are similar to, or different from, the comparison standards.

Exhibit II-8. Summary of content area emphases in the NAEP framework compared to the comparison standards

Compared to others, NAEP Framework has:		
	Grade 4	**Grade 8**
Number Properties and Operations	Typical emphasis, less number line	Typical emphasis, less squares and square roots, more decimals and fractions
Measurement	More below grade-level content	More below grade-level content, less connections to other content areas
Geometry	More transformations and symmetry, less parallel and perpendicular lines	More content
Data Analysis and Probability	Typical emphasis	More sampling and experiments
Algebra	More patterns, less quantitative relationships	Typical for pre-algebra, (does not cover algebra I), more broad in specifying functions

One issue faced by NAEP is how to handle content from grades earlier than the tested grade level. Since NAEP only assesses fourth and eighth grades, a lot of mathematics is off grade level. It is reasonable, therefore, for NAEP to include fractions content from grades 5, 6, and 7 in the eighth-grade assessment. But is it reasonable to include, as indicated in exhibit II-8, below grade-level content in measurement? This question of what below grade-level content to include is important for two reasons. First, anything included consumes assessment time that could have been used for other content. Second, the NAEP assessment is already difficult for the assessed population, so items are needed to get better information about low performing students. But should this be information about measurement? Perhaps some easy arithmetic is a more important topic. If so, some below grade-level arithmetic objectives might make more sense than below grade-level measurement objectives, assuming the focus on below grade level should be constrained.

In addition, there is more to the construction of an assessment framework than choice of content. The reviewers who judged how well the NAEP item pool assesses the NAEP framework (see chapter 3) commented positively on the quality of the framework. Nevertheless, in some cases, the nature of the framework may have contributed to validity questions raised about other aspects of NAEP (e.g., the alignment of the item pool, the item quality, or the accessibility of the assessment across the range of the population). It appears that the NAEP system can and should provide more specific and clear guidance to the assessment developer. Thus, an important question throughout this chapter was to determine how other framework developers have structured their work in order to provide clear guidance.

The primary purpose of the NAEP framework is to frame guidance for the development of the NAEP assessments. As a foundation document, the framework will serve best if kept simple and focused. The need for greater guidance should be met through supplemental documents and resources that go beyond the current framework and specifications documents. Supplemental resources of this kind have their value, in part, in their details and rich exemplification. Development of such resources needs the detailed attention of experts and staff well versed in the foundations embodied in the NAEP framework, as well as in best practices world wide. Such is not the work of committees.

Findings throughout this study suggest a more focused framework would better serve this purpose. However, the framework serves other important purposes, intended or not. For one, it influences state standards. Reys et al. (2005) found that states reported that the NAEP *Mathematics Framework* had an influence second only to the NCTM *Principles and Standards* on the development of the state's standards. NAEP should consider this secondary use in deciding how to respond to this report. However, given widespread criticism of U.S. curriculum as a "mile wide and an inch deep" compared to other nations, perhaps a better focused NAEP framework would serve both purposes well.

But what does "focus" mean in a framework document? The naive answer is fewer topics. Fewer topics would reduce the dilution of the topics that survived and thus add focus. But the grain size at which the number of topics is reduced is an important consideration. And furthermore, an objective that is written too broadly sweeps many incidental topics into its scope. Therefore, a more considered answer is that each objective and subtopic should be more sharply focused on what is important and what the limits are.

NAEP and states should consider the use of explicit notes that make objectives more focused without introducing excessive verbiage into the objectives themselves. Achieve, Japan, and Singapore make use of such notes about their objectives—often using the notes to limit their focus.

Specifying a domain for assessment should be a specification of targets at which test items are aimed. Objectives, subtopics and content areas should not be viewed as "containers" (or categories) into which test items are sorted. The question is not, "How can an objective/subtopic/content area be written to include every possible item that might be allowable?" The important question is, "How can an objective/subtopic/content area be written to focus on what is most important?" To do this, the rhetoric describing mathematical targets should be amply illustrated with sample items drawn from many sources.

NAEP specifies the percentage of items to be spent on each of its five content areas. It also specifies the percentage contribution to total score for each of three levels of complexity. Given problems identified in chapters 3 and 4, the five content areas may be too broad a level to carry specification of priorities. In addition to the content areas, items could be allocated at finer grain sizes: subtopics and objectives. Each level in the hierarchy refers to some mathematical coherence that is deemed important. Further, mathematics problems often require mathematics from multiple objectives, subtopics, and content area. Thus, aggregating item allocations from the objective level is not equivalent to setting allocations at each hierarchical level. The latter allows for specifying multi-objective, multi-subtopic and multi-content area item allocations and is particularly relevant for mid- and high-complexity items.[12]

Better guidance for the test developer (and for NAEP's clients) cannot be accomplished by improving a single document, the *Mathematics Framework*. Such guidance is best thought of as residing in a combination of documents that serve related purposes at different levels of detail. While some illustration with items is needed within the framework itself, much more is needed in a cross referenced compilation drawn from many sources (not just released NAEP items, but also items from states and other nations).

[12] However, under the present design, every NAEP item must be assigned to a single content area for scaling.

Chapter 3. Does the NAEP Item Pool and Assessment Design Accurately Reflect the NAEP Framework?

As described in earlier chapters, the NAEP mathematics assessment is based on an ambitious content framework, developed specifically to guide the assessment. The framework is organized into five broad subdivisions, or content areas. These content areas are further subdivided (at grade 8) into 20 subtopics and more than 100 objectives,[12] and thus represent a formidable measurement challenge. However, the item pool used to measure this framework is also ambitious, comprising nearly 170 items at each grade level in 2007. Such a large item pool is made possible because of the large samples of students available to NAEP and because of the NAEP assessment design, which distributes items across students and then combines information from all items and all students using IRT scale methodology.

The five content areas of the framework are displayed in exhibit III-1. Exhibit III-2 illustrates the subtopic and objectives levels of the framework with a short excerpt from the content area of number properties and operations. The full set of subtopics and objectives is provided in appendix A. In the example given here, *1) number sense* is a subtopic, while *a)* and *b)* are the first two objectives under number sense. Note that the content area and subtopic are applicable to all three grades, but the objectives are separately worded for each grade, and not all objectives apply to every grade. These first two objectives, for example, do not apply to grade 12. In total, the number sense subtopic has six objectives at grade 4, eight objectives at grade 8, and five objectives at grade 12.

Exhibit III-1. Percentage distribution of items by grade and content area

Content Area	Grade 4 (%)	Grade 8 (%)
Number Properties and Operations	40	20
Measurement	20	15
Geometry	15	20
Data Analysis and Probability	10	15
Algebra	15	30

SOURCE: National Assessment Governing Board, *Mathematics framework for the 2005 National Assessment of Educational Progress*, 2004.

[12] At grade 4, there are 19 subtopics and 65 objectives.

Exhibit III-2. Example of a NAEP subtopic and objectives

A. Number properties and operations

1) Number sense		
GRADE 4	**GRADE 8**	**GRADE 12**
a) Identify the place value and actual value of digits in whole numbers.	a) Use place value to model and describe integers and decimals.	
b) Represent numbers using models such as base 10 representations, number lines, and two-dimensional models.	b) Model or describe rational numbers or numerical relationships using number lines and diagrams.	

NOTE: Only the first two objectives are shown here. The number sense subtopic has six objectives at grade 4, eight objectives at grade 8, and five objectives at grade 12.

SOURCE: National Assessment Governing Board, *Mathematics framework for the 2005 National Assessment of Educational Progress*, 2004.

A sizable literature on alignment already exists (see for example, Bhola et al. (2003), Porter (2002), Rothman et al. (2002), and Webb (1999)). In determining how we would evaluate the alignment between framework and item pool for this study, we acknowledged several considerations. First, the NAEP assessment is substantially different from most other large scale assessments (especially state assessments) in both purpose and design, and it is therefore difficult to identify benchmarks for how thoroughly the NAEP assessment should, in any given year, cover the content of its framework.

Second, the question of how well the item pool aligns to the framework can be evaluated in two distinct directions:

1. *Does each item fit the framework?* and
2. *How well does the item pool assess the framework?*

We address both questions, but place the greater emphasis on the second.

Does each NAEP item fit the framework?

The answer to this question is "yes." Every item has been classified to a single objective by the test developer, and a number of expert committees have reviewed and concurred with these classifications over the course of the regular NAEP item development cycle.

Furthermore, at a gross level, the item pool also matches the *distributions* specified in the framework. The NAEP framework specifies distribution of content only at the level of the five content areas. A review of the item classifications provided by the test developer confirms that the 2007 NAEP item pool adheres closely to the distribution prescribed by the framework. Across both grade levels and all 5 content areas, only 1 of the 10 item counts deviates from the criterion by as much as 3 percentage points (see exhibits III-3 and III-4). In addition, the distribution of item types—multiple choice, short constructed response, and extended constructed response—is well balanced across the content areas.

Exhibit III-3. 2007 mathematics item as classified by test developer, grade 4

Content Areas and Subtopics	Total Items	Short Constructed Response	Extended Constructed Response	Multiple Choice	Proportion of Objectives Covered
Number Properties and Operations	65	19	3	43	19/20
Measurement	35	7	0	28	9/10
Geometry	26	7	2	17	12/15
Data Analysis and Probability	20	7	1	12	9/9
Algebra	20	5	1	14	10/11
Total	**166**	**46**	**6**	**114**	**59/65**

Exhibit III-4. 2007 mathematics item as classified by test developer, grade 8

Content Areas and Subtopics	Total Items	Short Constructed Response	Extended Constructed Response	Multiple Choice	Proportion of Objectives Covered
Number Properties and Operations	37	8	1	28	15/27
Measurement	28	4	0	24	10/13
Geometry	32	10	1	21	16/21
Data Analysis and Probability	26	7	2	17	11/22
Algebra	45	9	2	34	14/18
Total	**168**	**38**	**6**	**124**	**66/101**

How well does the item pool assess the framework?

Approach

The answer to this second alignment question—how well does the item pool assess the framework—is more complicated. One could settle for examining the *distribution of items classified by objective or subtopic*. But this is too low a standard. Five items can be classified as fitting a subtopic, but none assess what is most important about the subtopic. This study therefore chose a tougher standard and asked whether the item pool, taken as a whole, accurately represents the body of grade-appropriate content knowledge described by the framework. “Accurate representation” was operationalized to include:

- *Focus* on the most important knowledge and know-how in each subtopic (as defined by the objectives, but also by prioritizing knowledge on which other knowledge builds), rather than wasting items on marginal knowledge;
- *Balance* across the range of knowledge and know-how in each content area and subtopic; and
- *Reach* to span easier and less advanced, as well as harder and more advanced, aspects of the content in each subtopic. For an item pool to exhibit reach, it should allow all students to show what they have learned.

We also considered the fit of the item pool to a second, cognitive, dimension specified by the framework—mathematical complexity. This dimension, which has three levels in the NAEP framework (low, moderate, and high) is intended to capture "aspects of knowing and doing mathematics, such as reasoning, performing procedures, understanding concepts, or solving problems." An "ideal" distribution on this dimension is defined by the framework as one in which half of the assessment score is based on items of moderate complexity, with the remainder of the score based equally on items of low and high complexity (National Assessment Governing Board, 2004).

The evaluation of the item pool was carried out by a panel of expert reviewers, who were brought together for a two and one-half day meeting in February 2007. Eleven reviewers participated at each grade level; these reviewers included mathematicians, mathematics curriculum specialists, mathematics assessment specialists, and teachers (see appendix D). Several of the reviewers also had particular expertise in the delivery of mathematics instruction for students with disabilities and English language learners.

Reviewers were given a short training on the framework and provided with copies of both the *Mathematics Framework* (National Assessment Governing Board, 2004) and the *Assessment and Item Specifications* (National Assessment Governing Board, 2003). The latter expands, to a modest extent, on the content definitions provided in the framework. In addition, they received copies of all items in the 2007 item pool, along with the item scoring rubrics, where applicable.[13] For convenience, the items were sorted into the content area and subtopic classifications assigned by the developer, although reviewers were not constrained to consider items only for the subtopics or content areas into which they had been classified. Using the directions and rating form reproduced in appendix E, reviewers then rated the focus, balance, and reach of the body of items supporting each *subtopic*. Both the number and quality of the items were taken into consideration.

In addition, for each *content area*, reviewers considered the balance of the items *across* subtopics and the sufficiency of low-, moderate-, and high-complexity items. All ratings were made using a 4-point scale describing how well the body of items met the criterion:

1 = met very well
2 = met well enough

---

[13] Approximately half of NAEP testing time is spent on constructed response items, which are hand scored according to rubrics developed specifically for each item.

3 = not met well enough
4 = not met (met poorly)

In making their judgments, reviewers were instructed to accept the NAEP framework as given, and to use the information in the NAEP framework and specifications (especially the detailed list objectives for each grade level) to formulate an understanding of the content meant to be included. At the same time, they were intended to draw on their own individual and collective professional expertise to judge what constitutes the most core ideas (focus) and the appropriate balance and reach in each area at their grade level.

Evidence indicates that the three dimensions of focus, balance, and reach did function at least somewhat independently. For example, consider the ratings assigned to the subtopic of estimation at grade 4. For this subtopic, there were six different patterns of ratings across the 11 reviewers (exhibit III-5).[14]

Exhibit III-5. Pattern of ratings across dimensions for estimation, grade 4

Pattern of ratings	Focus	2	2	2	2	3	3
	Balance	2	2	3	3	3	2
	Reach	2	3	2	3	2	3
	Number of reviewers assigning each pattern	2	2	1	1	3	2

After making all of the ratings associated with a particular content area, the reviewers were asked to comment on the particular strengths and weaknesses of the item pool in that content area and to support their comments with illustrative items, drawn either from NAEP itself, or from a pool of alternative example items. The alternative example items included released items from six states, three countries, the Shell Centre, and two international assessments.[15] Because we wanted to provide the reviewers with a range of divergent examples to stimulate their thinking, some of these examples were from assessments that were not strictly comparable to NAEP in terms of grade level, assumptions about prior curriculum, or length of task.

Finally, the reviewers were asked to note, for each content area, whether there were defects or ambiguities in the framework that made it a poor tool for judging the item pool in that content area.

Each reviewer recorded his or her individual ratings and comments. However, the review process was designed to incorporate table-level discussions before individual ratings were finalized. For the subsequent analysis, individual ratings were first dichotomized into “met” (ratings 1 and 2) and “not met” (ratings 3 and 4). We then computed the

[14] None of the raters used the extreme categories of 1 (met very well) or 4 (not met) for this subtopic.
[15] The six states were CA, IN, MA, NC, TX, and WA. The countries were Singapore, Japan, and the Netherlands; the Shell Centre test was the Balanced Assessment in Mathematics; and the international tests included TIMSS and PISA.

percentage of reviewers, for each content area/subtopic and dimension, who considered the criterion to be "met."

Overall findings for focus, balance, and reach

In this section, we briefly overview the results by tabulating the numbers of subtopics and content areas rated as having met the criteria for focus, balance, and reach, by different percentages of expert reviewers. In the next two sections, we proceed to a more detailed look at the ratings and comments for each grade level, by content area. After that, we offer a separate discussion of ratings on the complexity dimension.

As shown in exhibit III-6, at grade 4 there was good consensus as to the adequacy of the item bank for about half of the 19 subtopics. That is, at least two thirds of the expert reviewers agreed that the item bank met the "focus" criterion for 9 out of 19 subtopics, met the "balance" criterion for 8 out of 19 subtopics, and met the "reach" criterion for 11 out of 19 subtopics.[16] There was another sizable group of subtopics on which the opinions of the reviewers was mixed, and there were a few subtopics on which the general consensus of the expert reviewers was that the item bank did *not* adequately represent the framework. Specifically, there were two subtopics for which less than one third of the reviewers agreed that the "focus" criterion was met, as well as three subtopics for which less than one third agreed that the "balance" criterion was met, and four subtopics for which less than one third agreed that the "reach" criterion was met.

Exhibit III-6. Number of subtopics (N=19) rated as having met criterion for focus, balance, and reach by different percentages of reviewers: grade 4

Rated as "met" by	Focus	Balance	Reach
≥ 2/3 of reviewers	9	8	11
< 2/3 but ≥ 1/3 of reviewers	8	8	4
< 1/3 of reviewers	2	3	4

NOTE: Ratings based on 2007 mathematics item pool.

For ratings carried out at the content area level (exhibit III-7), at least two thirds of the reviewers agreed that there was good balance across subtopics for three of the content areas. Reviewers were mixed in their evaluation of the adequacy of balance across subtopics for the remaining two content areas.

[16] See preceding section for our operational definitions of these three aspects of alignment.

Exhibit III-7. Number of content areas (N=5) rated as having met criterion for balance across subtopics, by different percentages of reviewers: grade 4

Rated as "met" by	Balance across subtopics
≥ 2/3 of reviewers	3
< 2/3 but ≥ 1/3 of reviewers	2
< 1/3 of reviewers	0

NOTE: Ratings based on 2007 mathematics item pool.

Turning to grade 8, the proportion of subtopics for which there was good consensus as to the adequacy of the item bank was similar to grade 4 (about half of the subtopics), but the proportion of subtopics for which there was general consensus as to the *in*adequacy of the item bank was greater than at grade 4. Thus, at grade 8, there were six subtopics for which less than one third of the reviewers agreed that the "focus" criterion was met, as well as seven subtopics for which less than one third agreed that the "balance" criterion was met, and five subtopics for which less than one third agreed that the "reach" criterion was met (see exhibit III-8).

Exhibit III-8. Number of subtopics (N=20) rated as having met criterion for focus, balance, and reach by different percentages of reviewers: grade 8

Rated as "met" by	Focus	Balance	Reach
≥ 2/3 of reviewers	8	10	10
< 2/3 but ≥ 1/3 of reviewers	6	3	5
< 1/3 of reviewers	6	7	5

NOTE: Ratings based on 2007 mathematics item pool.

Finally, exhibit III-9 shows that for grade 8, as for grade 4, there was good consensus as to the balance across subtopics in three content areas. There was one content area in which reviewers differed in their appraisal of balance across subtopics, as well as one content area where they agreed that there was not sufficient balance across subtopics.

Exhibit III-9. Number of content areas (N=5) rated as having met criterion for balance across subtopics, by different percentages of reviewers: grade 8

Rated as "met" by	Balance across subtopics
≥ 2/3 of reviewers	3
< 2/3 but ≥ 1/3 of reviewers	1
< 1/3 of reviewers	1

NOTE: Ratings based on 2007 mathematics item pool.

Based purely on these tabulated ratings, it would appear that the expert reviewers viewed the focus, balance, and reach of the NAEP item bank with tempered approval. That is, there were many areas in which they agreed that the bank offered good coverage for the content described in the framework. At the same time, there were a number of other areas, particularly at grade 8, where alignment was *not* considered adequate.

Unfortunately, the unique character of NAEP does not lend itself to a comparison against similarly situated tests, so it is difficult to say whether NAEP is doing better or worse than other tests in this regard. Some insights can be gained, however, from the more detailed findings in the following sections—particularly regarding the extent to which reviewers were or were not able to find, among the alternative example items, specific examples of better practice in areas where they judged NAEP to be lacking.

Grade 4 findings on balance, focus, and reach, by content area

In this section we review the grade 4 results by content area. This allows a more detailed examination of the specific content areas and subtopics that were rated high or low by the expert reviewers. In addition, the accompanying comments and example items provide more information about the specific features of the NAEP item pool that drew the reviewers' attention. Note that reviewers were encouraged to make comments about things that NAEP does well, as well as about areas in which they found NAEP to be lacking. However, it would appear that the demand characteristics of the task were such that reviewers were more likely to provide detailed notes about shortcomings than about strengths, even in areas that they rated highly.

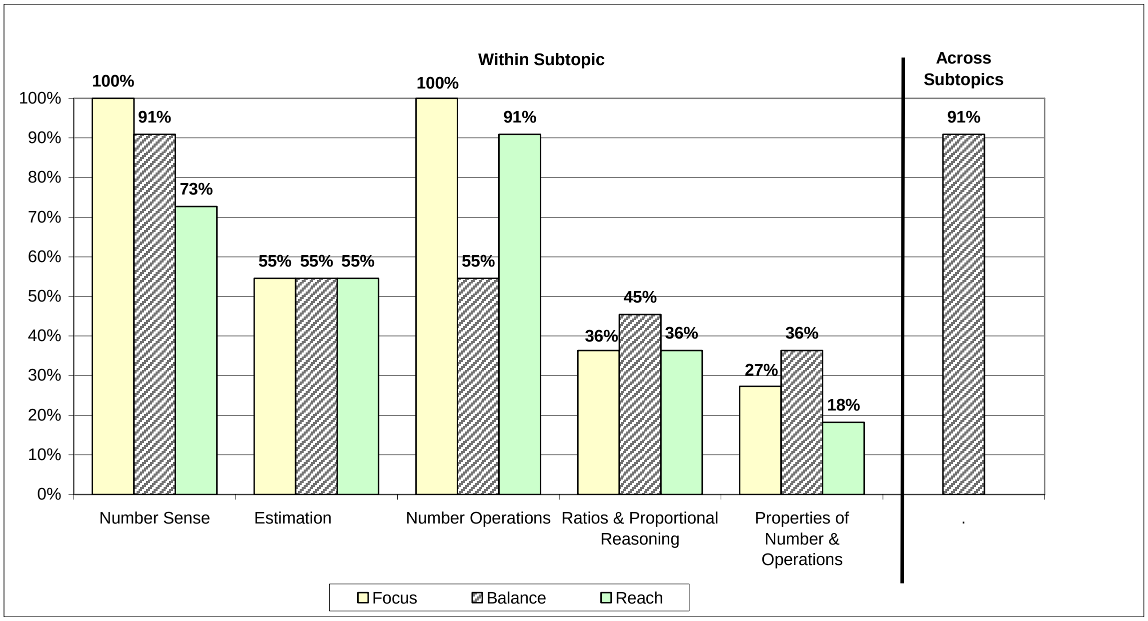
Reviewers selected examples from the NAEP test itself, as well as from the pool of alternative example items. Some of these examples are included in this report to clarify the reviewers' judgments. However, among the NAEP examples, we are only able to reproduce those which happen to come from blocks that were withdrawn from operational use after the 2007 assessment.[17]

Number properties and operations

At grade 4, the NAEP framework assigns the greatest proportion of items (40 percent) to number properties and operations, which is divided into five subtopics: number sense, estimation, number operations, ratios and proportional reasoning, and properties of number and operations. Exhibit III-10 presents the judgments of the expert reviewers concerning the focus, balance, and reach of the 65 items that NAEP uses to assess this content area.

[17] NAEP items are secure while they continue to be used in operational assessments. However, about 30 percent of the item blocks are replaced after each assessment cycle, and many of these items are then released to the public. All released NAEP items can be viewed by using the NAEP questions tool at http://nces.ed.gov/nationsreportcard/itmrls/.

Exhibit III-10. Grade 4 number properties and operations: Percentage of reviewers rating as having met criterion



NOTE: Ratings based on 2007 mathematics item pool.

Although the numbers of items varied substantially across subtopics, ranging from 24 items in the most populous subtopic to 3 items in the least populous, the expert reviewers were satisfied with the way that the items had been distributed across subtopics. Ninety-one percent of the reviewers agreed that balance across subtopics (shown on the right side of the figure) met criterion. Within subtopic, the ratings of the reviewers were more varied.

Number sense. This was one of the more heavily-represented subtopics (in terms of numbers of items assigned) in number properties and operations, and more than two thirds of the expert reviewers agreed that the subtopic met the criteria for all three dimensions.

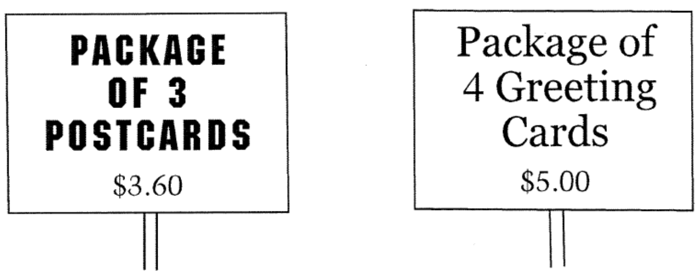
Estimation. Reviewers were more mixed in their evaluation of this subtopic, being almost evenly split between those who considered the criteria met and those who did not. It is worth noting that, although the same percentage of reviewers considered the criterion met on each of the three rating dimensions, it was not the same reviewers who gave a passing score on each dimension.

Reviewers were concerned that too many of the estimation items focused on making estimates appropriate to a given situation (A2b), and too few focused on verifying the reasonableness of results (A2c).[18] Moreover, they noted that it is difficult to write multiple-choice assessment items for A2b that actually tap estimation. Too often, it was felt, the items encouraged students to simply compute an exact answer and then back out to the closest corresponding answer choice. No examples from other tests were identified that overcame this problem.

Number operations. This was another subtopic that was relatively heavily-represented in the item pool. There was good agreement as to the adequacy of focus and reach, but reviewers were more divided on the dimension of balance. Those who were critical of the balance thought that the subtopic should include more work with fractions and decimals and at least some items that simply tap computational skills without being embedded in word problems. (Some reviewers noted that students with poor reading or English skills would have more trouble demonstrating the mathematics they knew if they always had to contend with word problems.)

In addition, some reviewers thought that there were items in the set that were made difficult by “busy” or difficult wording. A NAEP example is given in exhibit III-11.

Exhibit III-11. A NAEP item in which difficulty is increased by “busy” format



Rico bought 10 cards, which cost $12.20 before tax. How many packages of each type did he buy?

______________Packages of postcards

______________Packages of greeting cards

Explain how you know your answer is correct.

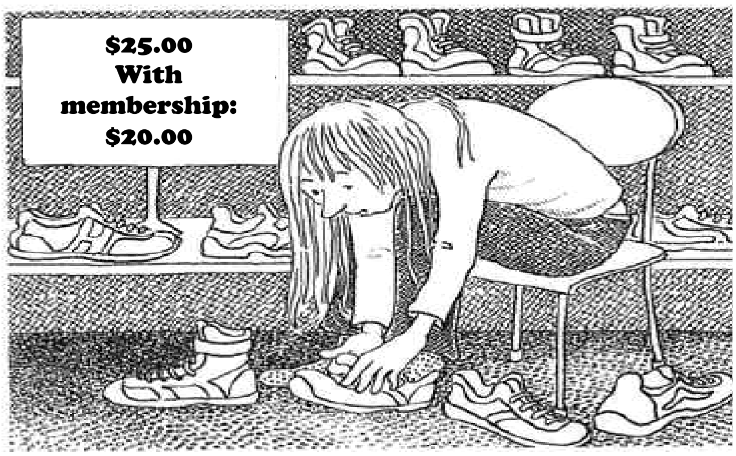
Rico said that one postcard is cheaper than one greeting card. Show that Rico is correct.

SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2007 Mathematics Assessment*, Grade 4, Block B1M7 #16.

[18] See appendix A for the complete text of the objectives. The number A2b refers to objective b in subtopic 2 (Estimation) in content area A (Number properties and operations). Note also that the judgment of the expert reviewers as to where the items were concentrated does not necessarily correspond to the official NAEP classification of these items. We deliberately withheld the official item classification at the objective level to discourage the rating task from devolving into an exercise in classification matching.

It can be argued that the “busy” wording in exhibit III-11 is necessary in order to situate the mathematical task in an authentic context. Such context can be better provided by using pictorial representations. In this way it is possible to add more clues to assist the reader while decreasing the total amount of text to be read. A good example is the Dutch item in exhibit III-12, although the mathematical contents of the item (calculating a percent) is outside the scope of the grade 4 framework.

Exhibit III-12. A Dutch item in which a pictorial representation is used to provide context



What percentage discount do you get when you are a member?

A 5 % C 25%

B 20% D 80%

SOURCE: Central Institute for Test Development (CITO), *Final Primary Education Test*, Math Task 2, # 14.

Ratios and proportional reasoning. There were very few items assigned to this subtopic, which has only a single objective at grade 4. Several reviewers were uncertain as to how narrowly to interpret the objective, and this uncertainty likely contributed to the low estimation of the subtopic by more than half the reviewers. Specifically, since the objective only mentions using simple ratios to describe problem situations (A4a), they weren’t sure how to evaluate items that could be solved by a proportion, such as the following example adapted from a North Carolina item.[19]

[19] Recall that the items under review were all used in the 2007 assessment, and only those NAEP items that have been released since 2007 can be used in this report. In some cases it was possible to illustrate the reviewers’ concerns with similar items taken or adapted from other sources.

Exhibit III-13. A state item that can be solved by a proportion, but not by a simple ratio

Nora needs 2 eggs for every cake she bakes. How many eggs does she need for 12 cakes?

A 2
B 6
C 12
D 24

SOURCE: Adapted from North Carolina Department of Education, End-of-Grade Tests, Grade 4, Math Sample Items, Goal 5, 2006.

Properties of number and operations. A high proportion of reviewers had concerns about this subtopic. In particular, they were dismayed that more than a third of the items were devoted to identifying odd and even numbers (A5a), and they felt that there were too few items on explaining or justifying a mathematical concept or relationship (A5f) and applying basic properties of operations (A5e)—although here they also thought that the framework was unduly vague about what basic properties were meant to be included. No alternative examples were offered for A5f, but a number of examples were offered for A5e, including the California item shown in exhibit III-14, which addresses the relationship between multiplication and division.

Exhibit III-14. A recommended state item for assessing knowledge of basic properties of operations

Justin solved the problem below. Which expression could be used to check his answer?

$$3\overline{)1364}^{\,454r2}$$

A (454 x 3) + 2

B (454 x 2) +3

C (454 +3) x 2

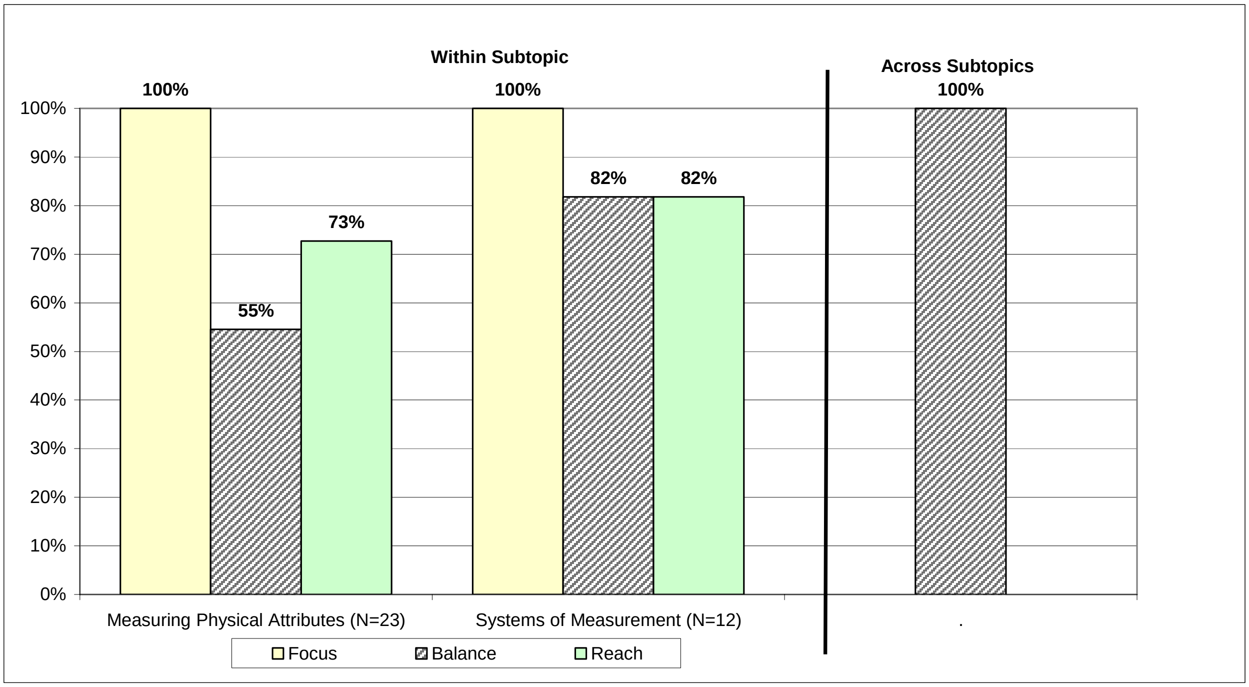
D (454 + 2) x 2

Source: California Department of Education, *California Standards Test, Released Test Questions*, Grade 4, # 17, 2005.

Measurement

This area accounts for 20 percent of the items at grade 4 and includes two subtopics: measuring physical attributes and systems of measurement. Exhibit III-15 shows how the 35 items in this content area were rated by the reviewers. Reviewers were generally positive about this content area, with several commenting that the area was well done; and everyone rated the balance across subtopics as meeting criterion.

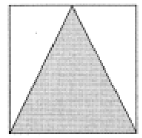
Exhibit III-15. Grade 4 measurement: Percentage of reviewers rating as having met criterion



NOTE: Ratings based on 2007 mathematics item pool.

Measuring physical attributes. All of the reviewers agreed that the items for this subtopic displayed a proper focus, and the majority also agreed that the subtopic met the criteria for balance and reach. Concerns about balance were intertwined with concerns about complexity (discussed later), with some reviewers asking for fewer straightforward measurement problems and more items that addressed the topic conceptually. One reviewer highlighted the positive example of the NAEP item shown in exhibit III-16, noting that this item did a good job of getting at the reasoning underlying the determination of area.

Exhibit III-16. A recommended NAEP item for assessing reasoning about measurement



The area of the shaded triangle is 4 square inches. What is the area of the entire square?

A 2 square inches
B 4 square inches
C 8 square inches
D 16 square inches

SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2007 Mathematics Assessment,* Grade 4, Block Z1M9 #19.

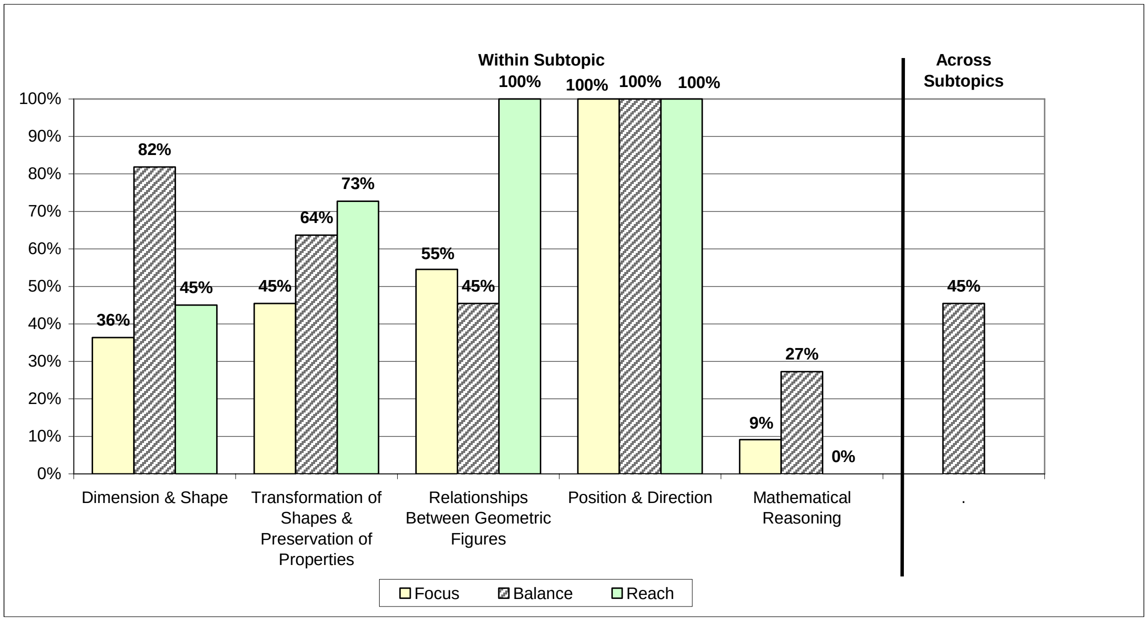
Of course, there are limits to the extent that student reasoning can be measured using multiple-choice items, and one member of our technical work group (TWG) pointed out that this item would be even better in open response format. With the multiple-choice format, a student with generally good reasoning, but who focused (incorrectly) on the white areas of the square, rather than the entire square, would earn the same score as a student who did not know how to approach the problem at all.

Systems of measurement. This subtopic received positive ratings on all three dimensions from virtually all of the reviewers.

Geometry

Fifteen percent of the grade 4 items are assigned to this content area, which includes five separate subtopics: dimension and shape, transformation of shapes and preservation of properties, relationships between geometric figures, position and direction, and mathematical reasoning. As shown in exhibit III-17, only 45 percent of the reviewers were satisfied with the distribution of items across subtopics, and a number of critical comments were made regarding the content area as a whole. These included comments that there were too many items, among the 26 classified as geometry, which related to identifying and describing simple plane figures; that the items did not ask students to demonstrate their knowledge in different ways; and that there was an overemphasis on questions that were just vocabulary (e.g., an item that shows an original and a transformed version of a two-dimensional figure and asks students to name the transformation).

Exhibit III-17. Grade 4 geometry: Percentage of reviewers rating as having met criterion



NOTE: Ratings based on 2007 mathematics item pool.

Dimension and shape. Although most of the reviewers rated this subtopic as balanced, only 36 percent rated it as meeting the criterion for focus, and only 45 percent judged it as meeting the criterion for reach. Criticisms included a lack of items that described real-world objects (C1b) and a lack of items on the attributes of two- and three-dimensional shapes (C1f). With regard to the latter, one reviewer suggested the Trends in International Mathematics and Science Study (TIMSS) item shown in exhibit III-18 as a good example of an item that asks students to recognize the mathematical definition of a shape.

Exhibit III-18. A recommended TIMSS item for assessing students' ability to recognize the mathematical definition of a shape

All of the pupils in a class cut out paper shapes. The teacher picked one out and said, "This shape is a triangle." Which of these statements MUST be correct?

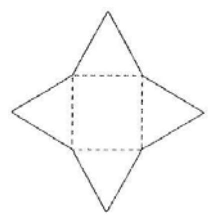
A The shape has 3 sides.
B The shape has a right angle.
C The shape has equal sides.
D The shape has equal angles.

SOURCE: International Association for the Evaluation of Educational Achievement, *Trends in International Mathematics and Science Study (TIMSS)* Assessment, Grade 4, M011022, #10, 2003.

Transformation of shapes and preservation of properties. This subtopic exposed differences among reviewers regarding the proper role of formal mathematics vocabulary in testing. Some reviewers wanted to see more emphasis on vocabulary, while others felt that emphasizing vocabulary tended to produce questions that were *just* about vocabulary. Nearly three quarters of the reviewers felt that the subtopic met the criterion for reach, but several also noted that there were some extremely low level items in the set. (Reach, of course, should extend in both directions and include easier and less advanced, as well as harder and more advanced, items.)

Relationships between geometric figures. Comments were mixed for this subtopic. Some reviewers complained that there were a few poorly written items in this relatively small set, which were even confusing for adults. On the other hand, there was 100 percent consensus that the subtopic met the criterion for reach, and one reviewer commended the subtopic as having good questions that "demonstrate a range of complexity and force students to apply what they learned." Anther reviewer recommended that this subtopic receive greater emphasis since it is directly tied to finding areas and volumes and "generally solving problems by taking apart and analyzing." This same reviewer suggested that there be more items like the NAEP item in exhibit III-19.

Exhibit III-19. A recommended NAEP item for assessing students' ability to recognize two-dimensional faces of three-dimensional objects



What three-dimensional shape could be made by folding the figure above on the dotted lines until the points on the triangles meet?

A Triangle
B Pyramid
C Cube
D Cone

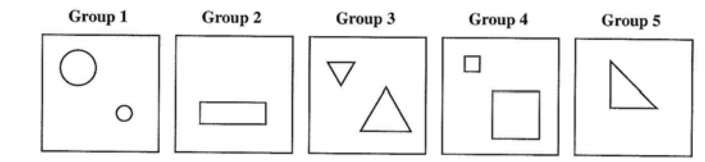
SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2007 Mathematics Assessment*, Grade 4, Block B1M7, #10.

Position and direction. Although there were very few items assigned to this subtopic, the subtopic was rated as meeting the criteria for all dimensions by all reviewers.

Mathematical reasoning. Reviewers were confused by the name of this subtopic, which suggested broad application across the mathematical domain (or at least geometry). In fact, however, there is only a single, specific objective at this grade level, which requires students to distinguish the objects in a collection that satisfy a given geometric definition and explain their choices (C5a). The 2007 item bank only contains a single item that is classified here, and all of the reviewers were very disappointed with the quality of the item. One reviewer suggested a constructed response item from Massachusetts (see exhibit III-20) that offers a much better opportunity for students to show what they know about distinguishing the objects in a collection that satisfy a geometric definition. In addition, the Massachusetts item was commended as being accessible to students with different levels of achievement.

Exhibit III-20. A recommended state item for assessing students' ability to distinguish objects in a collection that satisfy a geometric definition

Natasha sorted eight shapes into five groups as shown below.



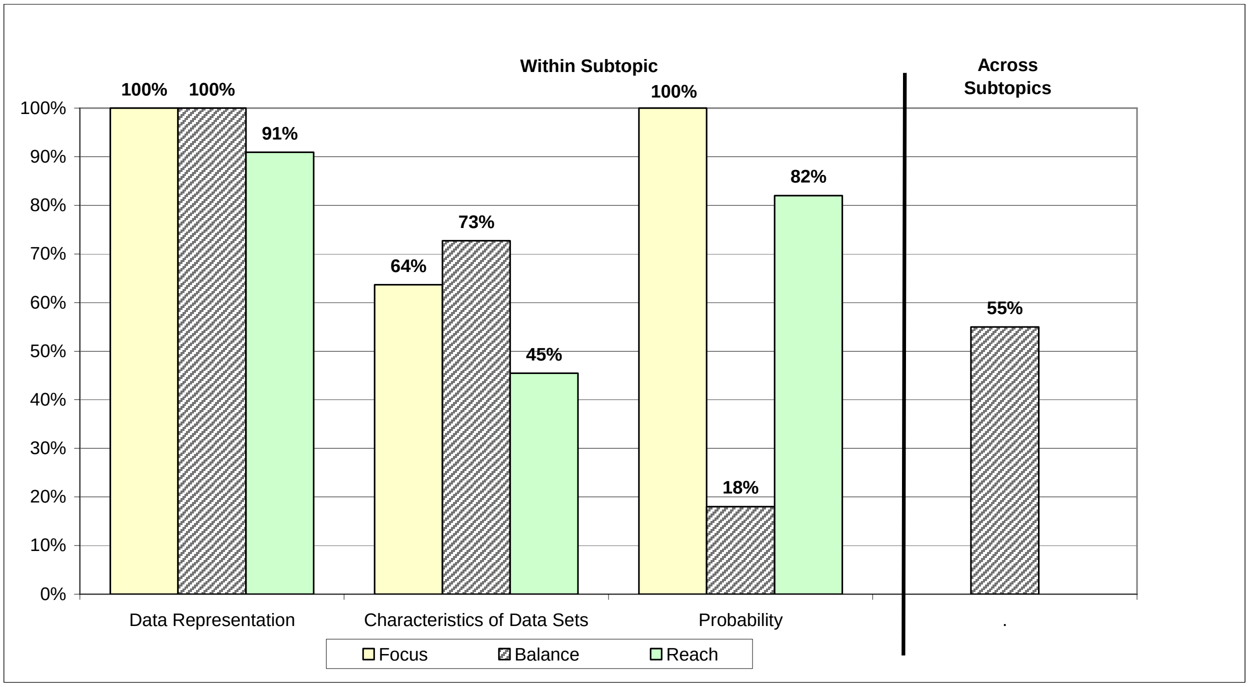
a. Explain how Natasha sorted the shapes into these groups.
b. Why do you think Natasha did not put the shape in Group 5 with the shapes in Group 3?
c. Which two groups could be combined? Explain your answer using geometric facts.

SOURCE: Massachusetts Department of Education, Massachusetts Comprehensive Assessment System, Grade 4, #17, 2004.

Data analysis and probability

This is the smallest content area at grade 4, with the framework assigning only 10 percent of items to the area. The content area includes three subtopics at grade 4—data representation, characteristics of data sets, and probability, and just over half of the reviewers rated the 20 data analysis and probability items in the item pool as balanced across subtopics. Those who did not consider the items well balanced noted that probability, with approximately half of the items, was overemphasized for this grade level. The ratings are shown in exhibit III-21.

Exhibit III-21. Grade 4 data analysis and probability: Percentage of reviewers rating as having met criterion



NOTE: Ratings based on 2007 mathematics item pool.

Data representation. All of the reviewers rated this subtopic as meeting the criteria for focus and balance, and all but one of the reviewers also rated it as meeting the criterion for reach. Some reviewers did note that there was not as much variety in types of data representations as they would have preferred, with more than one third of the items based on pictographs and none using circle graphs.

Characteristics of data sets. Reviewers were divided on the adequacy of items for this subtopic, which had only a few items. Of the three dimensions, the fewest reviewers (45 percent) rated the subtopic as meeting criterion on reach since none of the items were very challenging.

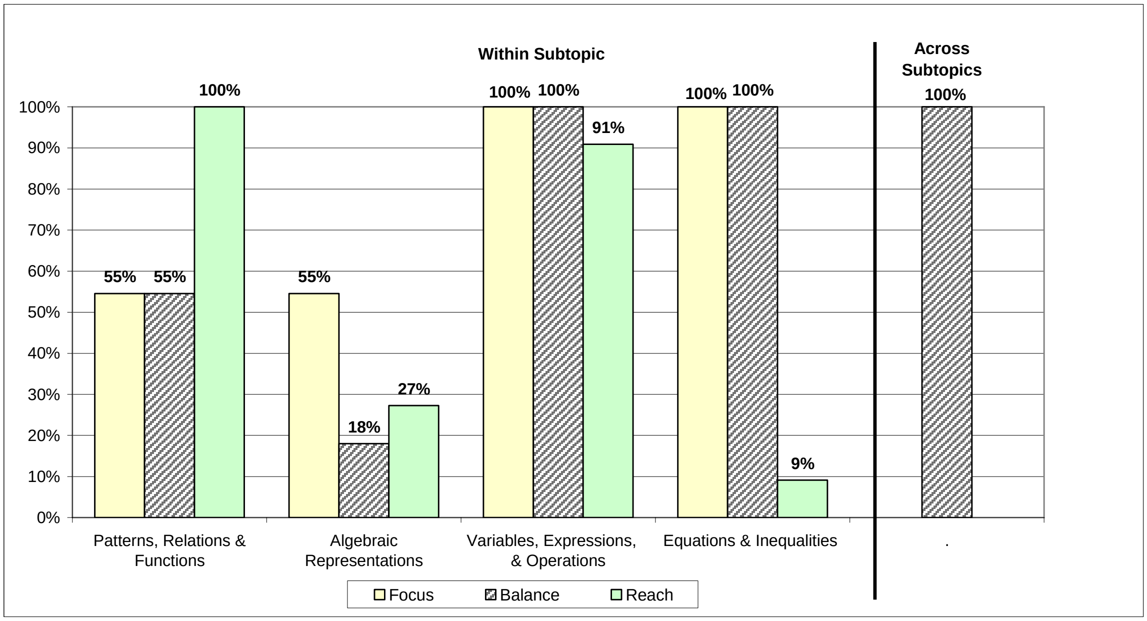
Probability. All the reviewers agreed that the items in this subtopic met the criterion for focus, and most agreed that they met the criterion for reach. On the dimension of balance, however, only 2 of the 11 reviewers felt that the item set met criterion. Specifically, reviewers felt that the items were too heavily weighted toward determining a simple probability (D4b). In addition, there were a number of items that the reviewers felt did not fit the objectives as written. For example, objective D4b calls for determining a simple probability from a context that includes a picture, but some of the items did not have pictures.

Algebra

Algebra is allocated 15 percent of the NAEP items at grade 4 and comprises four subtopics: patterns, relations, and functions; algebraic representations; variables, expressions, and operations; and equations and inequalities. As can be seen in exhibit III-

22, algebra was considered balanced across subtopics by all of the reviewers (although, somewhat inconsistently, several reviewers called for more items on the last two subtopics). Several reviewers commented that the distinctions in the framework between the last two subtopics (variables, expressions, and operations and equations and inequalities) were not clear at grade 4.

Exhibit III-22. Grade 4 algebra: Percentage of reviewers rating as having met criterion



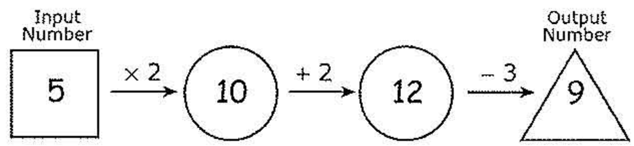
NOTE: Ratings based on 2007 mathematics item pool.

Patterns, relations, and functions. At fourth grade, nearly all of the objectives in this subtopic relate to patterns or sequences. Reviewers were divided in their evaluation of whether or not this subtopic—which was relatively heavily-represented in the item pool for algebra—met the criteria for focus or balance. All, however, agreed that it met the criterion for reach. Those reviewers who were not satisfied with the focus and balance of the subtopic noted that there were too many items devoted to recognizing and extending patterns and not enough on constructing or explaining a rule (E1b). Moreover, this was a subtopic that divided the mathematicians from some of the other reviewers, with the mathematicians complaining about the inclusion of pattern items that "are not really math problems in that you can't justify a single answer mathematically."

Examples of pattern or relation items that were acceptable to all reviewers include the TIMSS item shown in exhibit III-23 and the Japanese item shown in exhibit III-24. (The Japanese item, however, was intended for a higher grade level and would likely have to be modified for grade 4.)

Exhibit III-23. A TIMSS pattern item that was acceptable to all reviewers

A number machine takes a number and operates on it. When the Input Number is 5, the Output Number is 9, as shown below.



When the Input Number is 7, which of these is the Output Number?

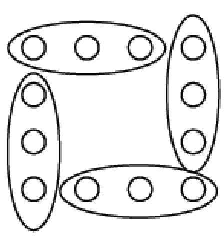
A 11
B 13
C 14
D 25

SOURCE: International Association for the Evaluation of Educational Achievement, *Trends in International Mathematics and Science Study (TIMSS)* Assessment, Grade 4, M031190, 2003.

Exhibit III-24. A Japanese pattern item that was acceptable to all reviewers

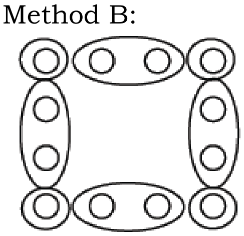
Akira made a square with 4 marbles on each side. She expressed the total number of marbles by applying the two methods described in the figures below:

Method A:



Equation: 4 x 3

Method B:



Equation: 4 x 2 + 4

We would like to count the total number of marbles when the square has seven marbles on each side. Using each of Methods A and B described above, how can we express the total number of marbles in a figure and in an equation?

SOURCE: National Institute for Educational Policy Research, Summary of findings about student achievement on particular types of problems and goals in mathematics and arithmetic, #5, 2006.

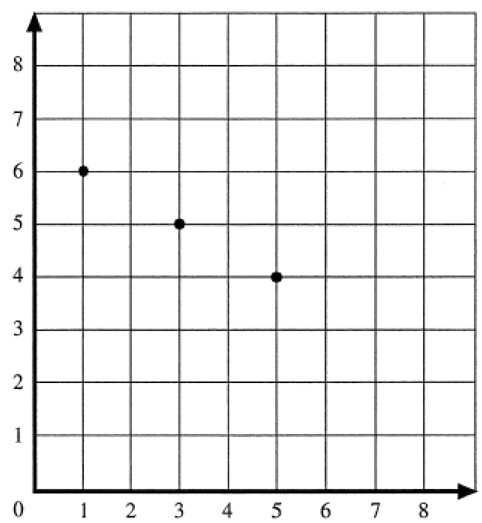
Also in this subtopic, reviewers failed to find items that fit the one objective that is not specifically tied to patterns or sequences—recognize or describe a relationship in which quantities change proportionally (E1e). However, some reviewers pointed out that two of the items classified under ratios and proportional reasoning in the number properties and operations content area would satisfy this objective.

Algebraic representations. Reviewers were divided in how they rated this subtopic for focus, but more than two thirds agreed that the subtopic was deficient in balance and reach. Of greatest concern was the absence of any items using conventional coordinate graphs, but reviewers also complained that some of the items—on translating between different forms of representation—could as easily have been classified in the next subtopic since they involved translation into algebraic expressions.

An example of a conventional coordinate grid problem that is very simple and unadorned, but which several reviewers liked for that very reason, is the California item shown in exhibit III-25.

Exhibit III-25. A recommended state item for assessing understanding of a coordinate grid

Chu plotted 3 points on a grid. The 3 points were all on the same straight line.



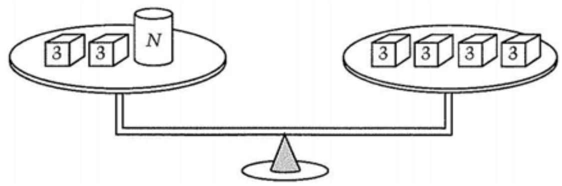
Chu wants to plot another point on the line. What could be its coordinates?

A (2, 5)
B (4, 4)
C (6, 3)
D (7, 3)

SOURCE: Adapted from California Department of Education, *California Standards Test, Released Test Questions*, Grade 4, #38, 2005.

Variables, expressions, and operations. Although it had few items, this subtopic got positive ratings on all three dimensions from nearly all the reviewers, and the only recommendation offered by the reviewers was that more items be devoted to the subtopic. An example of a NAEP item that several reviewers liked was the balance scale problem shown in exhibit III-26. To these reviewers, the item was exemplary because it offered good scaffolding. However, a member of the TWG argued that the item was actually seriously flawed because the natural way to answer the question is by visual inspection and does not require the construction of a number sentence. Therefore, the number sentences are inauthentic and imposed as a convention of the testing situation.

Exhibit III-26. A NAEP item on which there was disagreement as to whether the graphic provided scaffolding or undermined the intended solution strategy



The weights on the scale above are balanced. Each cube weighs 3 pounds. The cylinder weighs N pounds. Which number sentence best describes this situation?

A $6 + N = 12$
B $6 + N = 4$
C $2 + N = 12$
D $2 + N = 4$

SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2007 Mathematics Assessment*, Grade 4, Block B1M7, #4, 2007.

Equations and inequalities. This was another subtopic that had few items in the item pool and for which several reviewers would have liked to see more items. All reviewers agreed that the subtopic met the criteria for focus and balance, but that it failed to meet the criterion for reach since the only items assigned to it were very simple.

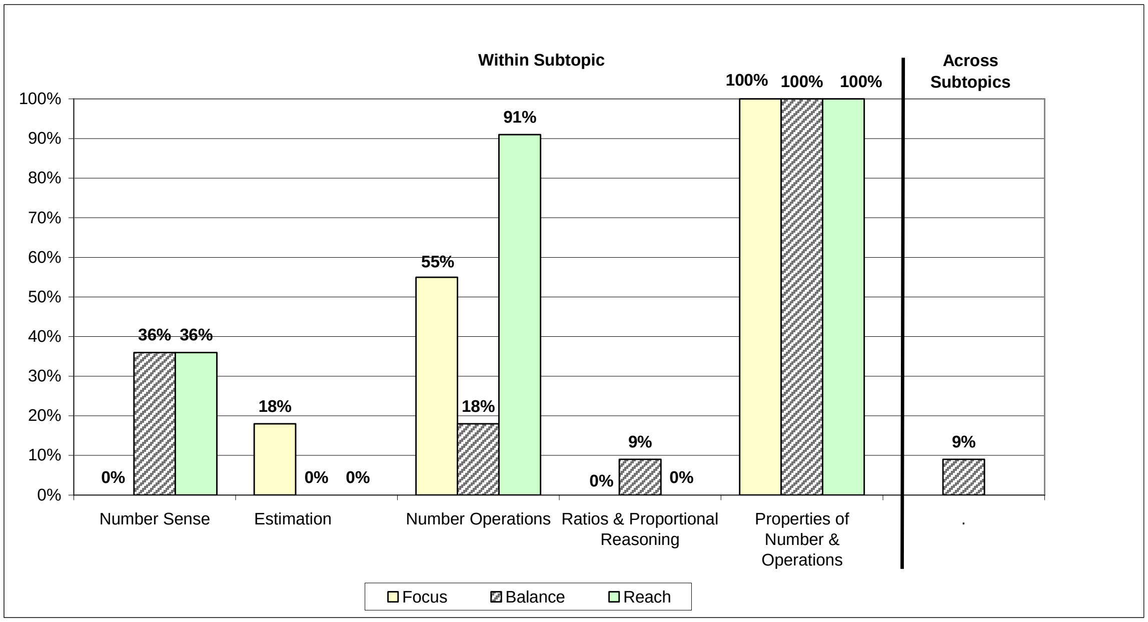
Grade 8 findings on balance, focus, and reach, by content area

We now turn to the content area results for grade 8. As was discussed in the overview of results section, grade 8 reviewers agreed upon positive ratings for about the same number of subtopics as grade 4 raters. However, the grade 8 raters also achieved consensus on negative ratings for a number of subtopics—more than at grade 4. There were also some notable differences between grades in the specific content areas and subtopics that earned positive or negative ratings.

Number properties and operations

For grade 8, the relative emphasis on number properties and operations is substantially decreased compared to grade 4, with the framework specifying that 20 percent of grade 8 items be allocated to this content area. As at grade 4, there are five subtopic areas within number properties and operations; the grade 8 reviewers were consistent in rating the balance across these subtopics as failing to meet criterion (exhibit III-27). Several reviewers commented that the NAEP framework for this content area was very well written, but that there were disappointing gaps in the coverage afforded by the 37 items classified here.

Exhibit III-27. Grade 8 number properties and operations: Percentage of reviewers rating as having met criterion



NOTE: Ratings based on 2007 mathematics item pool.

Number sense. All of the reviewers agreed that this subtopic failed to meet the criterion for focus, and nearly two thirds agreed that the subtopic also fell short with regard to balance and reach. Reviewers complained that there were too many items devoted to low level ideas about place value or very simple area models of fractions. As one reviewer explained, “What’s important in number sense is traveling between and applying multiple representations of rational numbers,” but this is lacking in the item set at grade 8. Another complaint was that the “meaningful contexts” in which the framework specifies that certain objectives be situated (especially A1e, recognize, translate between, or apply multiple representations of rational numbers…in meaningful contexts) was also lacking. Several reviewers suggested the Singapore item shown in exhibit III-28 as properly addressing both the need for mixing types of rational numbers and for placing items in meaningful contexts.

Exhibit III-28. A recommended Singapore item for assessing students' ability to translate between different types of rational numbers

> A retailer purchased 4 cartons of glasses. In each carton, there were 5 boxes of glasses. There were 40 glasses in each box. He found that 10% of the glasses were broken in 2 of the cartons and 1/5 of them were broken in the third carton. How many unbroken glasses had he left?
>
> Answer:____________
>
> SOURCE: SNP Panpac, *PSLE Mathematics*, Singapore, 2002.

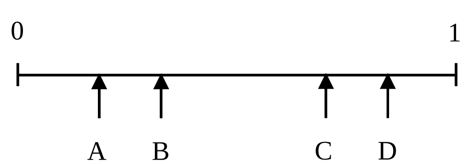
One reviewer commented that the large number of objectives in this subtopic (eight objectives), and the lack of explicit ranking among objectives (as noted elsewhere, the framework does not have a vehicle for expressing relative ranking within content area), may have allowed relatively minor areas to get more emphasis than they deserve simply because they are easier to test.

Estimation. Very few items were assigned to this subtopic, and all of the reviewers agreed that the subtopic failed to meet the criteria for balance and reach. Nine of the 11 reviewers also agreed that it failed to meet the criterion for focus. The reviewers wanted to see more items for this subtopic, and they particularly wanted to see items that addressed the establishment and use of benchmarks (A2a). Like the reviewers at grade 4, they also wanted to avoid estimation items in which students could just answer the question by working out the exact answer.

A suggestion for benchmarking was to include an item like the Dutch item shown in exhibit III-29, but with harder numbers such as 1/10, 3/25, 7/15, 9/16, or 27/30.

Exhibit III-29. A recommended Dutch item for assessing estimation through benchmarking

> Where is ¾ on this line?



> SOURCE: Central Institute for Test Development (CITO), Final Primary Education Test, Math Task 1, # 9.

Number operations. In this subtopic, reviewers agreed that the items had satisfactory reach, but not much balance. They were divided in their rating of focus. The concerns (which impacted both focus and balance) were that nearly all of the fairly large group of items devoted to this topic were on solving application problems (A3g). There was little attention to the more conceptual objectives such as providing a mathematical argument to explain operations with fractions (A3e) or interpreting rational number operations and the relationships between them (A3f). The Singapore item shown in exhibit III-30 was suggested as one way to address A3f.

Exhibit III-30. A recommended Singapore item for assessing relationships between rational number operations

Yiyuan was supposed to divide a number by 5. Instead, he multiplied it by 5 and obtained an answer 10.5. If he had done what he was supposed to do, what should the correct answer be?

Answer:____________

SOURCE: SNP Panpac, *PSLE Mathematics*, Paper 4, #36, 2003.

Ratios and proportional reasoning. Ratios and proportions was the third subtopic in number properties and operations where the majority of reviewers agreed that none of the dimensions met the criteria. The criticisms stated that there were too few items in this subtopic, and the ones that were there were too routine. One reviewer noted that "the focus in eighth grade is on rates, proportional reasoning, and percents, but this is not evident in the choice of items." (However, subsequent comments did acknowledge that proportional reasoning appeared in the objectives for several content areas and that its overall representation in the assessment was therefore greater than could be judged from number properties and operations alone.)

A large number of example items was offered, including many that addressed percents as well as ratios or rates. An Indiana item was singled out as a relatively simple item that dealt with percent *decrease* (exhibit III-31).

Exhibit III-31. A recommended state item for assessing students' ability to compute a percent decrease

Last year, the freighter *Mariposa* carried 20 million tons of cargo. This year, the *Mariposa* carried 16 millions tons of cargo. What is the percent decrease in the amount of cargo carried by the Mariposa from last year to this year?

A 20%
B 25%
C 36%
D 40%

SOURCE: Indiana Department of Education, *Indiana statewide testing for educational progress (ISTEP+)* grade 8 sampler, #4, 2006.

Properties of number and operations (7 items). This was the one subtopic in number properties and operations and that received positive ratings from all of the reviewers. Reviewers remarked that there were "great items" on this subtopic. Two examples are shown in exhibit III-32.

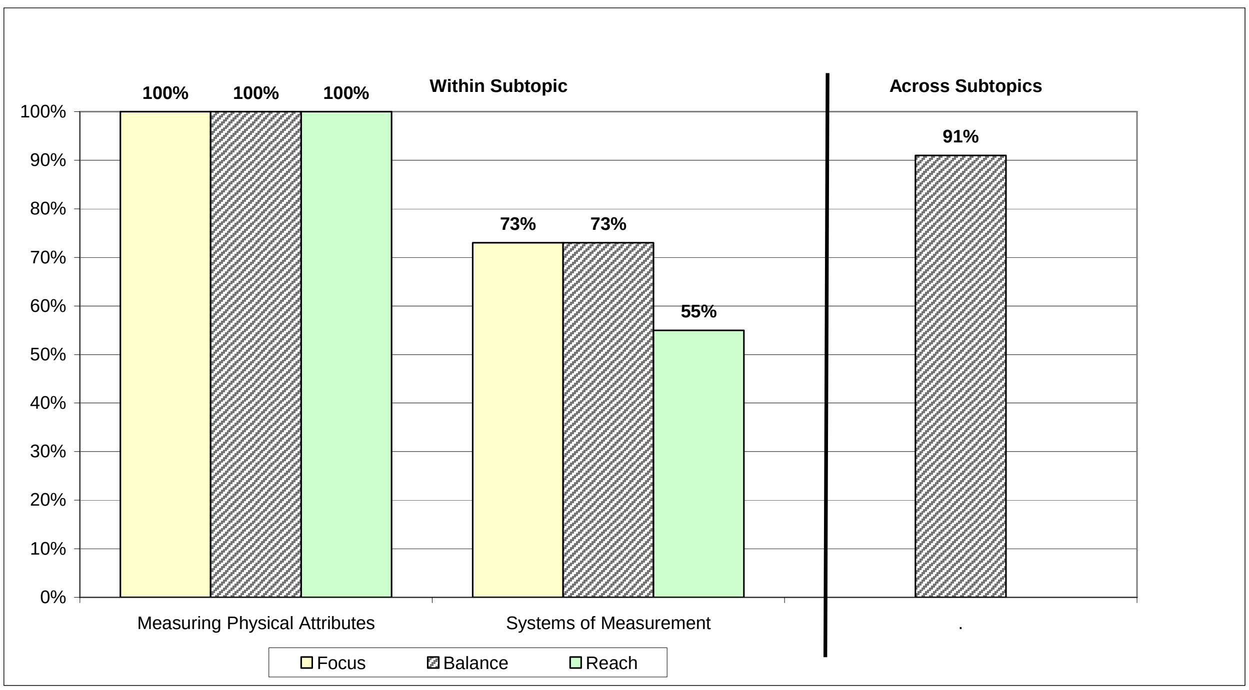
Exhibit III-32. Two recommended NAEP items for assessing properties of number and operations

Which of the following must be true about the sum of any two prime numbers greater than 2? A The sum will be even. B The sum will be odd. C The sum will be a prime number. D The sum will be a multiple of 3. E The sum will be a multiple of 5. SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2007 Mathematics Assessment*, Grade 8, Block Z23M8B #5, 2007.	The sum of three numbers is 173. If the smallest number is 23, could the largest number be 62? A Yes B No Explain your answer in the space below: SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2007 Mathematics Assessment*, Grade 8, Block Z23M8B #9, 2007.

Measurement

With 15 percent of items devoted to measurement, this is the second content area that has less emphasis at grade 8 than at grade 4. Twenty-eight measurement items are included in the grade 8 item pool. The grade 8 reviewers were satisfied with most of the content and with the balance across subtopics (exhibit III-33).

Exhibit III-33. Grade 8 measurement: Percentage of reviewers rating as having met criterion



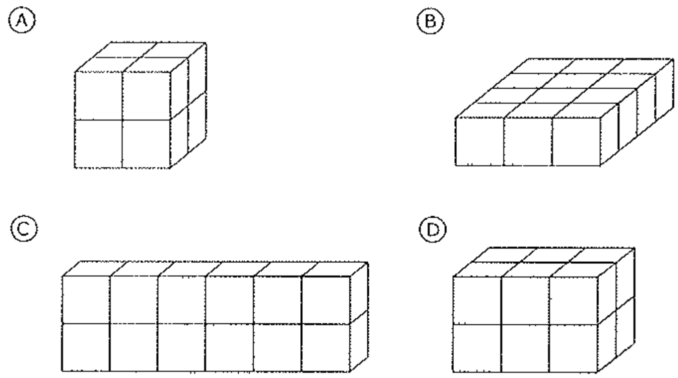
NOTE: Ratings based on 2007 mathematics item pool.

Measuring physical attributes. All of the reviewers judged this subtopic—which included the bulk of the measurement items—to have met criterion on focus, balance, and reach.

Many also provided examples from other tests that would do a better job of tapping a higher level of complexity. Three examples are given in exhibits III-34–III-36. The first item, taken from TIMSS, addresses the comparison of objects with respect to volume (B1b). The second, taken from a book of Singapore public school leaving exams, addresses indirect measurement (B1k); and the third, taken from the Washington state assessment, addresses the surface area of a cylinder (B1j).

Exhibit III-34. A recommended TIMSS item for assessing the students' ability to compare objects with respect to volume

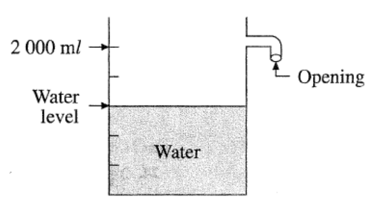
All the small blocks are the same size. Which stack of blocks has a different volume from the others?



SOURCE: International Association for the Evaluation of Educational Achievement, *Trends in International Mathematics and Science Study (TIMSS)* Assessment, Grade 8, M012013, 2003.

Exhibit III-35. A recommended Singapore item for assessing indirect measurement

A rectangular metal block measuring 16 cm by 14 cm by 6 cm is put into the container shown below.

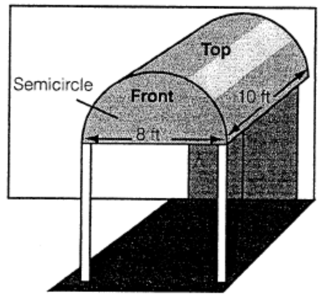


(a) What is the volume of the metal block?
(b) How much water will flow out of the container when the metal block is put in?

SOURCE: SNP Panpac, *PSLE Mathematics,* Paper 4, #50, 2003.

Exhibit III-36. A recommended state item for assessing the students' ability to compute the surface area of a cylinder

Bella Restaurant is building a curved awning for the entrance to their restaurant. They need materials for only the top and the front of the awning.



Area of a circle = πr^2
Circumference of a circle = πd

Find the surface area of the awning to determine the total amount of canvas necessary to make the awning.

Show your work using words, numbers, and/or pictures.

Be sure to label your answer.

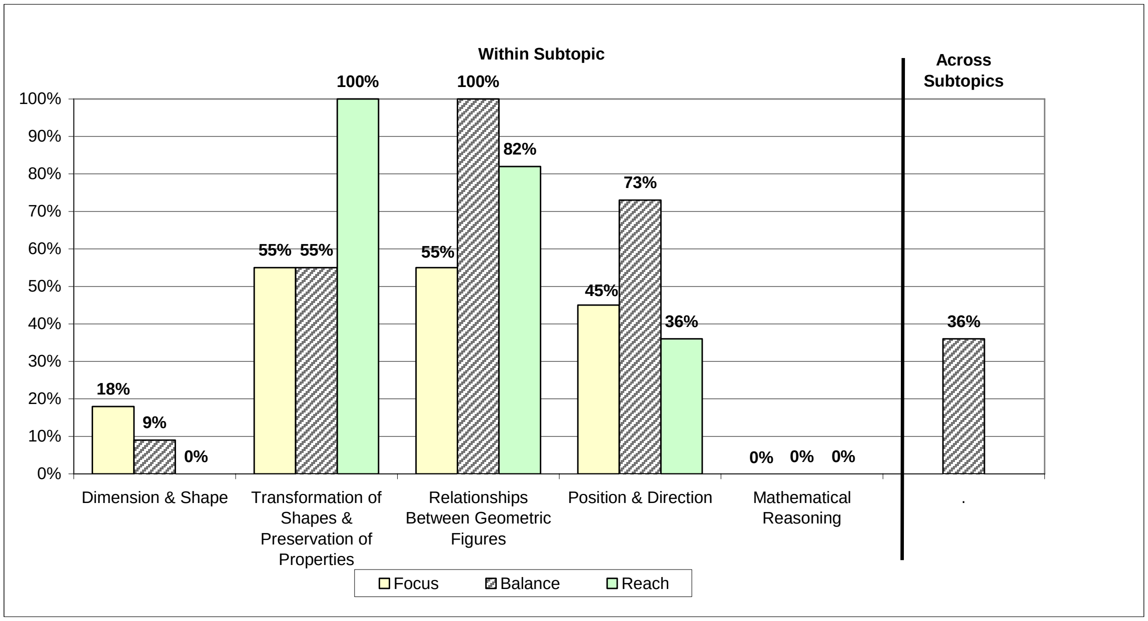
SOURCE: Washington State Department of Education, *Washington Assessment of Student Learning (WASL)*, Mathematics Grade 8, Sample Test Booklet, #13, 2006.

Systems of measurement. Reviewers were somewhat more divided in their reactions to this subtopic than they were to the first subtopic in measurement. Nevertheless, nearly three quarters of the reviewers judged the subtopic to have met criterion on focus and balance, while more than half judged it to have met criterion on reach.

Geometry

At eighth grade, this content area is increased to 20 percent of the item total, and the 2007 item pool for geometry contains 32 items. As shown in exhibit III-37, the grade 8 reviewers were not very well satisfied with the distribution across subtopics in geometry: nearly two thirds felt that the content area had not met criterion on this dimension.

Exhibit III-37. Grade 8 geometry: Percentage of reviewers rating as having met criterion



NOTE: Ratings based on 2007 mathematics item pool.

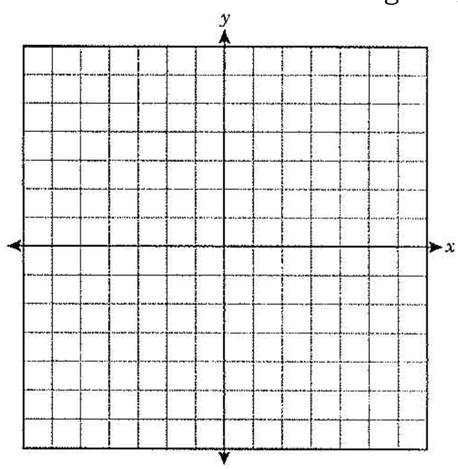
Dimension and shape. Nearly all of the reviewers rated this subtopic as not having met criterion on any of the dimensions. The reviewers complained that there were too many items in this subtopic that simply asked students to identify shapes or to name or count faces, edges, or vertices. Correspondingly, they felt that the item set was lacking in items that asked students to draw or sketch polygons and other figures from a written description (C1d), as well as items that required students to represent or describe a three-dimensional situation in a two-dimensional drawing from different views (C1e). A Washington state item was proposed as an example of the former (exhibit III-38), while a Texas item was offered as an example of the latter (exhibit III-39).

Exhibit III-38. A recommended state item for assessing students' ability to draw polygons from a written description

Burke is using a coordinate grid to draw a rhombus. He selected three points: *A*(1, 3), *B* (1, -1), and *C* (-1, 1).

- Plot the ordered pairs listed above and label them *A, B,* and *C.*
- Plot the missing vertex of the rhombus and label it *D.*
- Connect the four points to make it a rhombus.

You must use a ruler or straightedge.

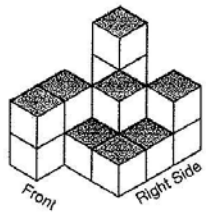


Write the coordinates of point *D:* ____________

SOURCE: Washington State Department of Education, *Washington Assessment of Student Learning (WASL)*, Mathematics Grade 8, Sample Test Booklet, Grade 8, Sample Test Booklet #5, 2006.

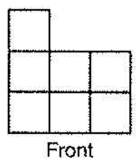
Exhibit III-39. A recommended state item for assessing students' ability to represent a three-dimensional situation in a two-dimensional drawing from different views

Melody made a solid figure by stacking cubes. The solid figure is shown below.

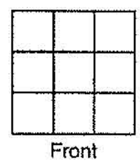


What drawing best represents a front view of this solid figure?

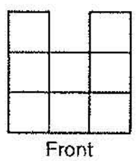
A



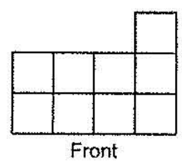
C



B



D



SOURCE: Texas Education Agency, *Texas Assessment of Knowledge and Skills*, Grade 8 Mathematics Online Test, #50, 2006.

Transformation of shapes and preservation of properties. Reviewers were more divided in their evaluation of this subtopic. Just over half of the reviewers rated the subtopic as having met the criteria for focus and balance, while all of the reviewers agreed that it had met the criterion for reach. Reviewers were split on the issue of whether the subtopic included enough items on similarity and proportional reasoning (objectives C2e and C2f). Some argued that the level of treatment was sufficient for students not taking a formal geometry course, while others thought that these objectives should receive more coverage (although not, at this grade level, with an emphasis on formal, procedural solutions).

Relationships between geometric figures. Reviewers were divided on the question of whether this subtopic met the criterion for focus, but they were in better agreement that the criteria for balance and reach were met successfully. Some of the reviewers advocated for a greater emphasis on the use of the Pythagorean theorem to solve problems (C3d), characterizing this as one of the "big ideas" in the grade 8 curriculum.

Position and direction. For this subtopic, there was, again, no clear consensus on how the item set should be judged. Forty-five percent of reviewers thought that the subtopic met the criterion for focus, 73 percent thought it met the criterion for balance, and 36 percent thought it met the criterion for reach. Reviewers particularly wanted to see more items on intersections of figures in the plane (C4b) and cross sections of solids (C4c).

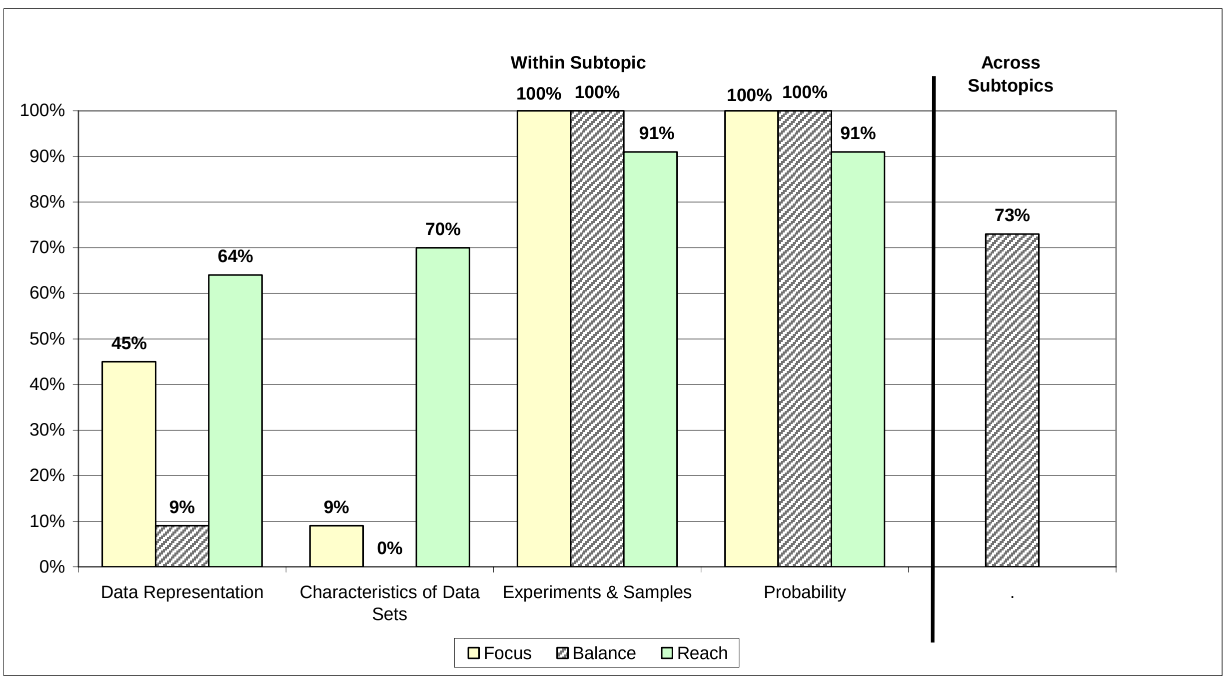
Mathematical reasoning. As is also true at grade 4, this subtopic contains only a single objective—in this case, making and testing a geometric conjecture about regular polygons. The 2007 item pool does not contain any items classified to this objective. Clearly, the subtopic of mathematical reasoning therefore failed to pass the criteria for focus, balance, or reach. However, reviewers did not find the stated objective to be particularly compelling or central (although they did question an organizational structure that would leave an entire subtopic empty in one or more assessment cycles), and they did not offer any suggestions of items from the alternate example set that could fulfill the objective.

Data analysis and probability

This content area is allocated 15 percent of the total assessment, and it is represented by 26 items in the 2007 assessment. This is the only content area that has an additional subtopic at grade 8, which is not represented at grade 4. The new subtopic is experiments and samples.

Nearly three quarters of the reviewers rated the balance across data analysis and probability subtopics as adequate (exhibit III-40).

Exhibit III-40. Grade 8 data analysis and probability: Percentage of reviewers rating as having met criterion



NOTE: Ratings based on 2007 mathematics item pool.

Data representation. The reviewers were divided in their evaluation of the focus and reach for this subtopic, but they were agreed that the subtopic lacked balance. Specific concerns included too many items devoted to reading or interpreting data (D1a) and too few that actually require students to complete a graph and then solve a problem using data in the graph (D1b). Also missing were problems that required working across data sets (in D1c) and items that compare and contrast data representations (D1d and D1e). A Washington state item (exhibit III-41) was nominated as an example that made meaningful use of multiple data sets to solve a problem.

Exhibit III-41. A recommended state item for assessing students' ability to use multiple data sets to solve a problem

The Associated Student Body (ASB) at Baker Middle School conducted a survey to determine which assemblies the school should schedule for next year. The tables show the options and costs of options for March and April.

Assembly Options for March and April

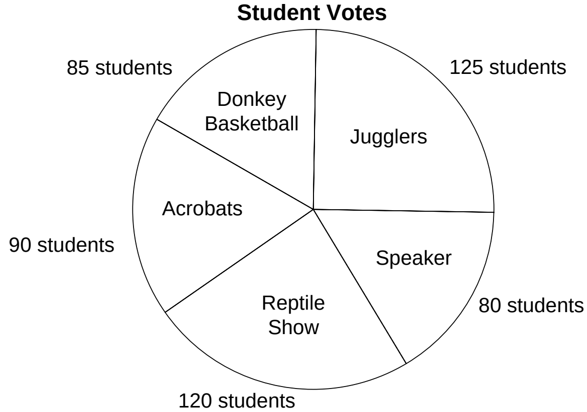
March Assembly	April Assembly
Jugglers	Speaker
Reptile Show	Acrobatics
Donkey Basketball	

Cost for Each Assembly Option

Guest Assembly	Cost
Jugglers	$1,000.00
Reptile Show	$800.00
Acrobats	$700.00
Donkey Basketball	$500.00
Speaker	$200.00

Each of the 500 students voted for one of the five choices. The circle graph shows the results of their votes.

Student Votes



The ASB must select **one** assembly for each month. They want to spend as **much** of their $1,500 budget as possible without going over $1,500.

Organize all of the information give in order to determine which assemblies the school should schedule for March and April. Make a proposal to the ASB and include the following:

- All possible combinations of assemblies for March and April
- The cost of each combination
- A recommendation for the March and April assemblies
- A reason why your recommendation is appropriate using information from each table or chart.

Show your work using words, numbers, and/or pictures.

SOURCE: Washington State Department of Education, *Washington Assessment of Student Learning (WASL)*, Mathematics Grade 8, Sample Test Booklet, #8, 2006.

Data representation was also a subtopic in which the reviewers found some of the language in the framework confusing. They were not sure how (or if) to draw a distinction between “data” and “graph.” Some objectives refer to both, but others refer only to “data,” and—based on the item set—there would appear to be an intention to include “graph” as a form of data. A second point of confusion was the overlap between D4d and D4e. Both refer to judging the effectiveness of a data representation, and the two objectives seem to cover overlapping ground.

Characteristics of data sets. Nearly all the reviewers evaluated this subtopic as not having met the criteria for focus or balance. Reach received a more favorable review, with 70 percent of reviewers judging the items for the characteristics of data sets subtopic to have met this criterion. Reviewers pointed to the fact that, among the small number of items assigned to this subtopic, too many called for primarily procedural or recall skills. Measures of central tendency also received too much attention at the expense of other characteristics of data sets. Furthermore, while three of the five objectives appeared to offer a good basis for challenging student work—identifying the impact of outliers (D2c), comparing data sets describing the same characteristic in two different populations (D2d), and fitting a line to scatter plot (D2e)—all of the items were concentrated on the other two objectives.

While, as noted, the reviewers felt that objectives D2c, D2d, and D2e held a great deal of promise for challenging items, they were not able to find examples of such items from other tests to recommend to NAEP.

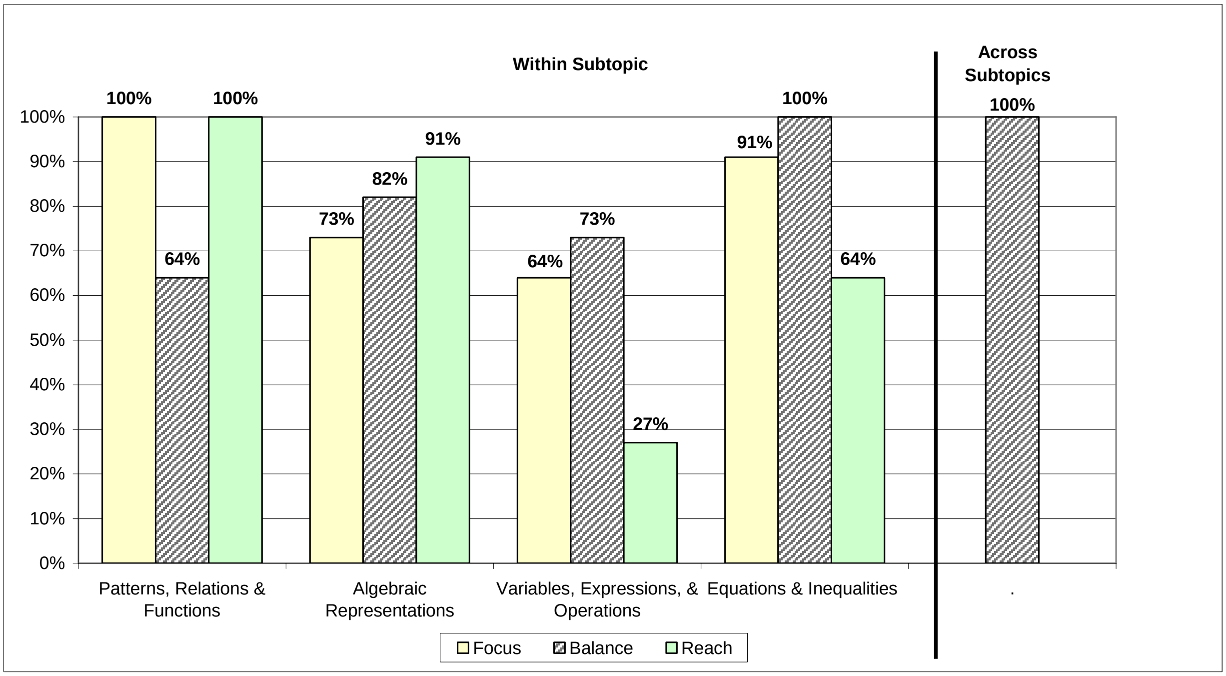
Experiments and samples. Although there were very few items in this subtopic, all of the reviewers rated the subtopic as having met the criteria for focus and balance, and all but one reviewer rated it as having met the criterion for reach.

Probability. This subtopic was also evaluated positively, with all of the reviewers rating it as having met the criteria for focus and balance, and all but one rating it as having met the criterion for reach.

Algebra

Algebra is the most heavily weighted content area in the NAEP framework at grade 8, with 30 percent of items. The 2007 item pool has 45 items in this area. As shown in exhibit III-42, all of the reviewers rated the balance across subtopics as adequate.

Exhibit III-42. Grade 8 algebra: Percentage of reviewers rating as having met criterion



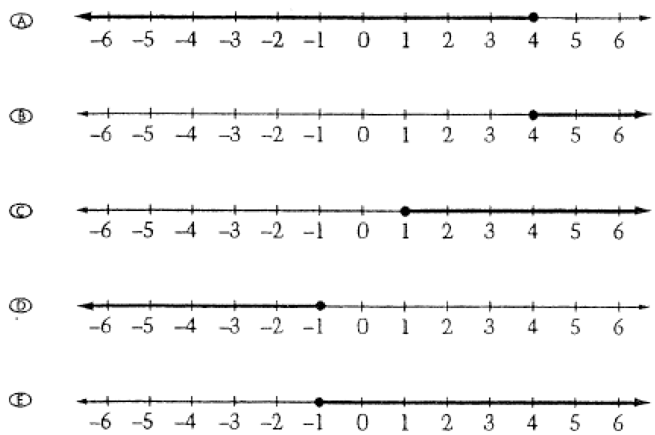
NOTE: Ratings based on 2007 mathematics item pool.

Patterns, relations, and functions. All of the reviewers considered that the item set for this subtopic met criterion on focus and reach. Their assessment of balance was more mixed, with slightly less than two thirds rating the subtopic as having met criterion on this dimension. The dissenting reviewers felt that there was too much emphasis on recognizing, describing, or extending patterns (E1a) and generalizing patterns (E1b), and too little emphasis on creating patterns, sequences, or linear functions from rules (E1c), comparing linear and nonlinear functions (E1e), and interpreting the meaning of slopes and intercepts (E1f).

Algebraic representations. This was one of the more heavily-represented subtopics in algebra at grade 8. Most of the reviewers felt that this subtopic met criterion on the three dimensions of focus, balance, and reach, but they also thought that the items fell short on complexity and on tapping conceptual understanding. Reviewers further noted that several of the NAEP items in this subtopic could be modified to better tap conceptual understanding by simply making them constructed response rather than multiple choice. One such item is the NAEP item shown in exhibit III-43.

Exhibit III-43. A NAEP item that would do a good job of assessing conceptual understanding if converted to a constructed response format

Which of the following is a graph of $2x - 5 \geq 3$?



SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2007 Mathematics Assessment*, Grade 8, Block Z23M8B, #10, 2007.

Variables, expressions, and operations. This subtopic has only two objectives at grade 8. Although all of the reviewers acknowledged that there were items addressing each of these objectives, some of the reviewers felt that the item set was weak on core aspects of the second objective—performing basic operations on linear algebraic functions (E3b). The reviewers wanted to see more emphasis on order of operations (described as important at this grade level, both for algebra and arithmetic) and exponents. The greatest concern with this subtopic, as with all the subtopics in algebra, was the lack of challenging items, as well as the fact that too many items could be answered by working backwards from the answer options and therefore did not really measure the intended skill.

A number of examples of alternative items were offered, including two problem situations from the Singapore examinations, shown in exhibit III-44 that could be used as the basis for items in which students write algebraic equations and then use the equations to solve the problems.

Exhibit III-44. Two examples of problem situations from the Singapore examinations that could be used as the basis for items requiring students to write and solve algebraic equations

A pencil costs \$$q$ and a pen costs 80 cents more. How much does 3 pencils cost and 2 pens cost?

Yuhui and Peirong have the same amount of money. After Yuhui spent \$72 and Peirong spend \$115, Yuhui has twice as much money as Peirong. How much money did each of the girls have at first?

SOURCE: SNP Panpac, *PSLE Mathematics*, Paper 4, #20 and #46, 2003.

Exhibit III-45 shows an example of a more challenging order of operations item, taken from the Texas assessment, while exhibit III-46 shows an example of a California item that taps conceptual understanding of exponents.

Exhibit III-45. A recommended state item for assessing students' understanding of order of operations

A set of parentheses is missing from the expression below.

15 – 5 + 7 × 2 + 4

Which of the following expressions has the parentheses in the correct place for the expression to equal 52?

A 15 – (5 + 7 × 2) + 4
B (15 – 5 + 7) × 2 + 4
C 15 – (5 + 7 × 2 + 4)
D 15 – 5 + 7 × (2 + 4)

SOURCE: Texas Education Agency, *Texas Assessment of Knowledge and Skills (TEKS)*, Grade 8 Mathematics Online Test, #8, 2006.

Exhibit III-46. A recommended state item for assessing conceptual understanding of exponents

Which expression below has the same value as x^3?

A $3x$
B $x \div 3$
C $x * x * x$
D $3x * 3x * 3x$

SOURCE: California Department of Education, *California Standards Test*, *Released Test Questions*. Grade 8, #33, 2005.

Equations and inequalities. Almost all the reviewers rated this subtopic as meeting criterion on focus and balance, but opinion was more divided on reach. As was true for the previous subtopic, reviewers were concerned that that so many of the items were straightforward “plug and chug” exercises.

Findings for complexity

Besides rating focus, balance, and reach at the subtopic level, and balance across subtopics at the content area level, reviewers were also asked to rate the extent to which each content area contained an adequate supply of low-, moderate-, and high-complexity items. The definitions of the three levels of complexity were taken from the NAEP framework and are presented here in exhibit III-47.

Exhibit III-47. NAEP definitions of complexity

Low Complexity
This category relies heavily on the recall and recognition of previously learned concepts and principles. Items typically specify what the student is to do, which is often to carry out some procedure that can be performed mechanically. It is not left to the student to come up with an original method or solution. The following are some, but not all, of the demands that items in the low-complexity category might make: o Recall or recognize a fact, term, or property. o Recognize an example of a concept. o Compute a sum, difference, product, or quotient. o Recognize an equivalent representation. o Perform a specified procedure. o Evaluate an expression in an equation or formula for a given variable. o Solve a one-step word problem. o Draw or measure simple geometric figures. o Retrieve information from a graph, table, or figure.
Moderate Complexity
Items in the moderate-complexity category involve more flexibility of thinking and choice among alternatives than do those in the low-complexity category. They require a response that goes beyond the habitual, is not specified, and ordinarily has more than a single step. The student is expected to decide what to do, using informal methods of reasoning and problem-solving strategies, and to bring together skill and knowledge from various domains. The following illustrate some of the demands that items of moderate complexity might make: o Represent a situation mathematically in more than one way. o Select and use different representations, depending on situation and purpose. o Solve a word problem requiring multiple steps. o Compare figures or statements. o Provide a justification for steps in a solution process. o Interpret a visual representation. o Extend a pattern. o Retrieve information from a graph, table, or figure and use it to solve a problem requiring multiple steps. o Formulate a routine problem, given data and conditions. o Interpret a simple argument.

Exhibit III-47. NAEP definitions of complexity (cont.)

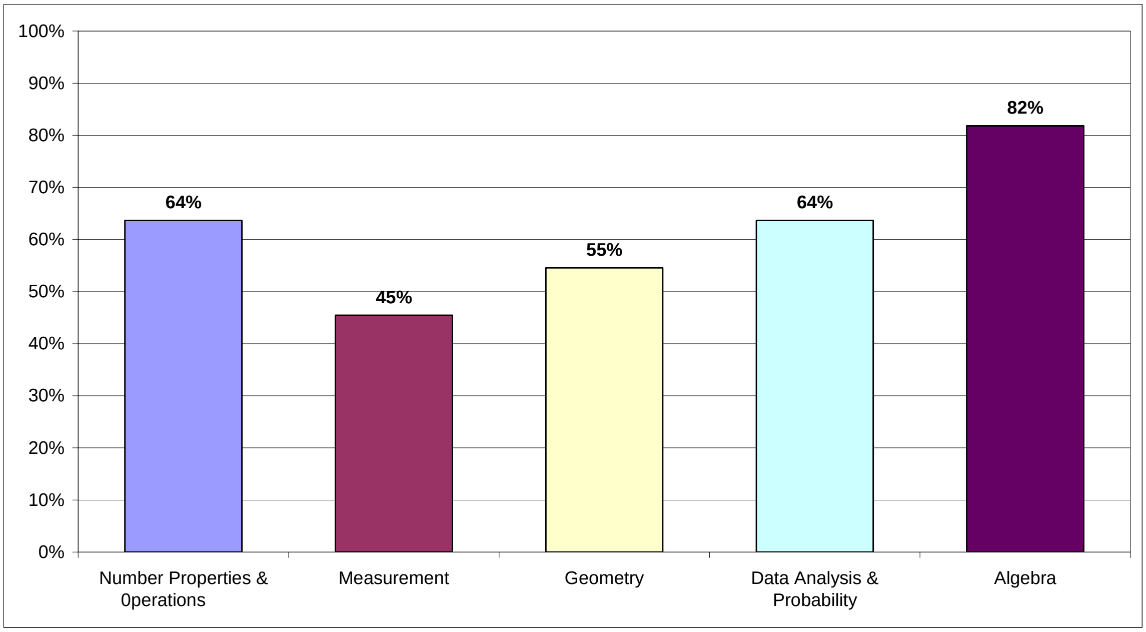
High complexity
High-complexity items make heavy demands on students, who must engage in more abstract reasoning, planning, analysis, judgment, and creative thought. A satisfactory response to the item requires that the student think in abstract and sophisticated ways. Items at the level of high complexity may ask the student to do any of the following: o Describe how different representations can be used for different purposes. o Perform a procedure having multiple steps and multiple decision points. o Analyze similarities and differences between procedures and concepts. o Generalize a pattern. o Formulate an original problem, given a situation. o Solve a novel problem. o Solve a problem in more than one way. o Explain and justify a solution to a problem. o Describe, compare, and contrast solution methods. o Formulate a mathematical model for a complex situation. o Analyze the assumptions made in a mathematical model. o Analyze or produce a deductive argument. o Provide a mathematical justification.

SOURCE: National Assessment Governing Board, 2004.

Although the framework suggests that a quarter of the total assessment score should be based on high complexity items, the test developer’s own classifications only place 5 of the 166 fourth-grade items and 4 of the 168 eighth-grade items in this category. The reviewers that participated in our alignment exercise were actually more forgiving in their estimation, although they still expressed concerns about the complexity level of the assessment in many of the content areas, particularly at grade 8.

Exhibit III-48 displays the percentages of grade 4 reviewers who rated each content area as offering sufficient representation of high complexity. As can be seen, the reviewers were divided in their reactions on this dimension, and there was considerable variation in their level of consensus across content areas. The percentage of reviewers who rated a given content area as having adequate representation of high complexity items varied between a low of 45 percent for measurement and a high of 82 percent for algebra.

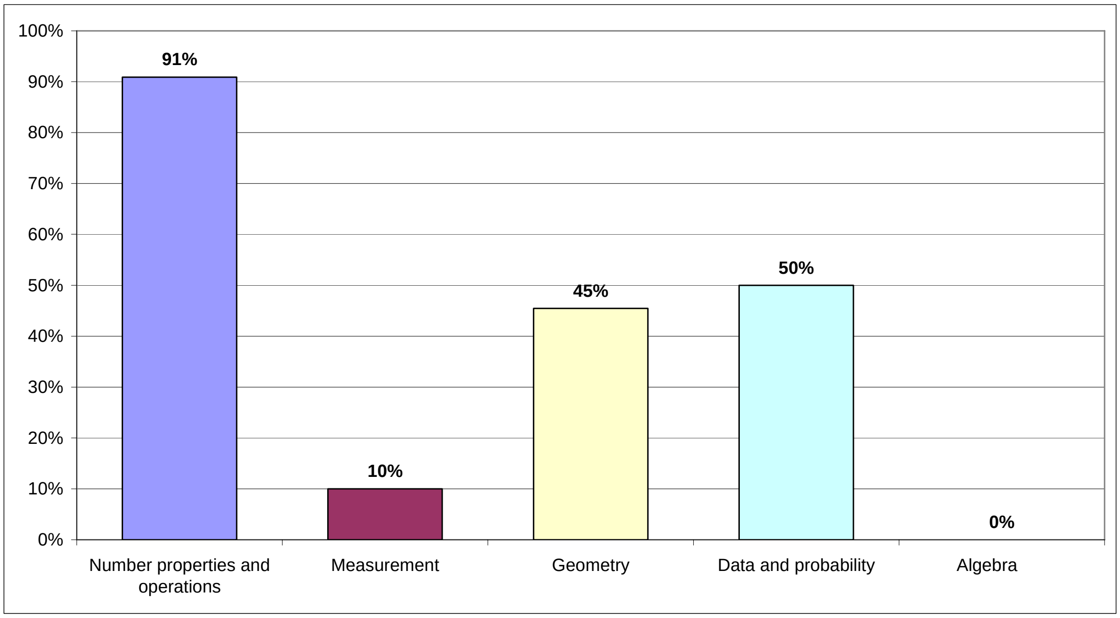
Exhibit III-48. Percentage of reviewers judging high complexity to be adequately represented in each content area, grade 4



NOTE: Ratings based on 2007 mathematics item pool.

While, as noted, the grade 4 reviewers were mostly divided in their judgments, the grade 8 reviewers had good consensus regarding their evaluations of high complexity in three of the five content areas. As can be seen in exhibit III-49, they were nearly unanimous in their agreement that number properties and operations met the criterion for high complexity, while measurement and algebra did not. Their opinions were more divided regarding the adequacy of high-complexity items in geometry and in data and probability.

Exhibit III-49. Percentage of reviewers judging high complexity to be adequately represented in each content area, grade 8



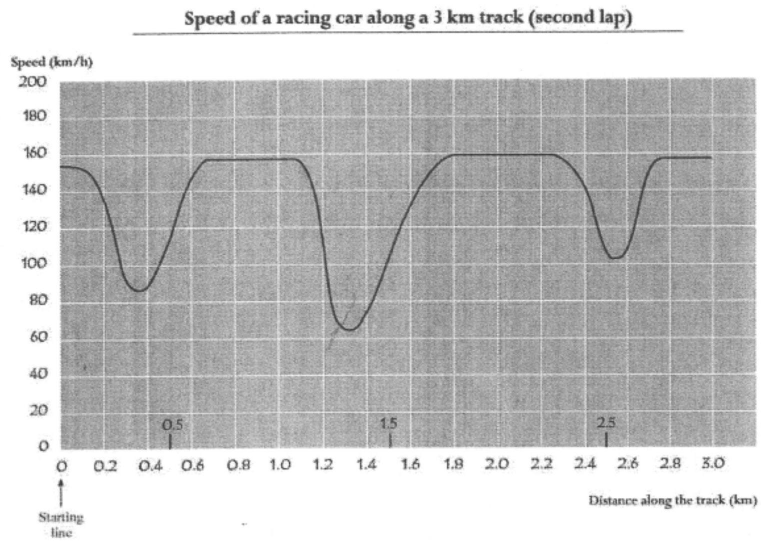
NOTE: Ratings based on 2007 mathematics item pool.

Interestingly, the ratings for high complexity did not track very well with the ratings on focus, balance, and reach. This is less evident at grade 4, but it is still the case that grade 4 measurement got some of the most consistently favorable ratings for focus, balance, and reach, but had the fewest reviewers judging it adequate for high complexity. At grade 8, the divergent patterns are much more pronounced. For example, grade 8 number properties and operations, which received positive ratings for high complexity, was not judged well on most of the subtopic ratings or on balance across subtopics. On the other hand, grade 8 algebra, which received uniformly negative ratings for high complexity, had consistently positive ratings on most other dimensions and most subtopics. As TWG member de Lange points out in an essay included in appendix F, complexity is not synonymous with difficulty, and assessments should strive to have high complexity items that distribute across the achievement scale.

The design constraints of the standard NAEP mathematics block, which typically includes 16 to 18 items and is timed at 25 minutes, create a serious challenge for the construction of high complexity items. High complexity items are not necessarily high difficulty items, but they frequently demand responses that take more time to complete. It may also be easier to access high complexity when several items are written to one integrated problem situation. In this way, the problem set can include some straightforward items that provide scaffolding for the more challenging items. Reviewers identified two such examples of multi-part tasks from other assessments that systematically build from lower complexity to higher complexity within a single problem context. The PISA task (exhibit III-50) is all multiple choice, while the Balanced Assessment in Mathematics task (exhibit III-51) is all constructed response. Although these seem quite different in format from the typical NAEP item, NAEP does include some multi-part problem sets, as well as a substantial percentage of constructed response items.

Exhibit III-50. A multiple-choice item set from PISA that builds from low to high complexity

This graph shows how the speed of a racing car varies along a flat 3 kilometer track during its second lap.

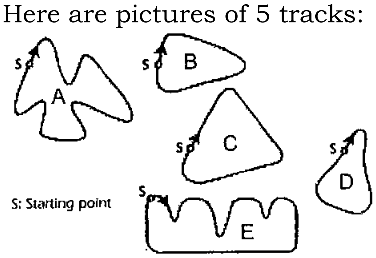


1. What is the approximate distance from the starting line to the beginning of the longest straight section of the track?
   A 0.5 km
   B 1.5 km
   C 2.3 km
   D 2.6 km

2. Where was the lowest speed recorded during the second lap?
   A At the starting line
   B At about 0.8 km
   C At about 1.3 km
   D Halfway around the track

3. What can you say about the speed of the car between the 2.6 km and 2.8 km marks?
   A The speed of the car remains constant.
   B The speed of the car is increasing.
   C The speed of the car is decreasing.
   D The speed of the car cannot be determined by the graph.

4. Here are pictures of 5 tracks:



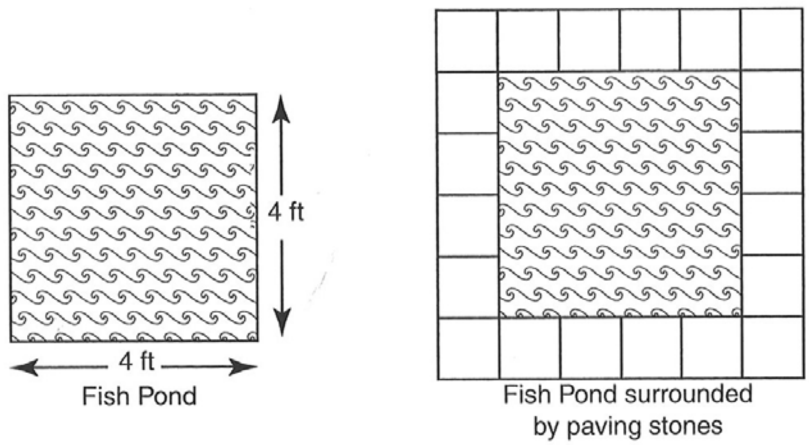
   Along which one of these tracks was the car driven to produce the speed graph shown earlier?

SOURCE: Organization for Economic Cooperation and Development, Program for International Student Assessment (PISA), *Mathematical Literacy Assessment*, Section 5.2, Unit 3.

Exhibit III-51. A constructed response item set from the Balanced Assessment in Mathematics that builds from low to high complexity

Fish Ponds

This problem gives you the chance to:
- find a number pattern in real spatial context and express the rule
- extend the rule to two variables



Chris works at a garden center that sells square fish ponds and paving stones.

The paving stones are squares with sides one foot long.

1. Use the diagram above to figure out how many paving stones are needed to surround a fish pond that is 4 feet by 4 feet. ____________

2. Chris begins to make a table to show how many paving stones are needed to surround square ponds of different sizes. Fill in the empty boxes in the table.

Side of pond in feet	**1**	**2**	**3**	**4**	**5**
Number of paving stones	**8**				

Continued on next page.

Exhibit III-51. A constructed response item set from the Balanced Assessment in Mathematics that builds from low to high complexity (cont.)

3. How many paving stones are needed to surround a fish pond that is 20 feet by 20 feet? Explain how you figured it out.

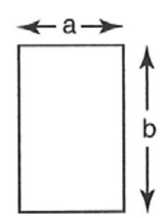
4. Chris has 48 paving stones. Find the size of the largest square pond the paving stones can surround. Explain how you figured it out.

5. The garden center sells many different sizes of square fish ponds.

Write down a rule that will help Chris figure out how many paving stones are needed to surround square ponds of different sizes.

6. The garden center decides to sell rectangular ponds.

Find a rule that will help Chris figure out how many paving stones are needed to surround rectangular ponds of different sizes.



SOURCE: CTB/McGraw Hill, *Balanced Assessment in Mathematics*: Practice booklet 8B, 2001.

Distribution of items by number type

The NAEP framework does not give much guidance regarding the appropriate balance of items across types of numbers since most of the objectives, if they mention number type at all, are inclusive rather than restrictive. For example, at grade 4, objective A1j calls for ordering or comparing whole numbers, decimals, or fractions. In considering the adequacy of the item pool, several members of the study steering committee and the TWG asked for a review of the distribution of items by number type. The findings, which are summarized in exhibit III-52 for grade 4, show that, at this grade level, 19 (11 percent) of the 166 items contain fractions, while 19 contain some other type of non-integer rational number. The fraction and decimal items are primarily concentrated in two content areas: number properties and operations, and measurement.

Exhibit III-52. Grade 4 distribution of items by number type

	No numbers	Whole numbers	Fractions	Decimals	Percents, rates, ratios	≥2 types of rational numbers[1]
Number Properties and Operations	1	42	12	8	2	0
Measurement	8	20	5	1	1	0
Geometry	17	9	0	0	0	0
Data Analysis and Probability	1	11	1	1	6	0
Algebra	5	14	1	0	0	0
Total	**32**	**96**	**19**	**10**	**9**	**0**

[1]Non-integer

NOTE: 2007 mathematics item pool.

Results for grade 8 are very similar, as can be seen in exhibit III-53. Eighteen (11 percent) of the 168 items contain fractions, and 26 contain some other form of non-integer rational numbers.[20] As at grade 4, the largest concentration of fraction and decimal items is in number properties and operations, but the remaining items are spread fairly evenly across the other content areas.

[20] This statement double counts the one item that contains both a fraction and a second form of non-integer rational numbers.

Exhibit III-53. Grade 8 distribution of items by number type

	No numbers	Whole numbers	Fractions	Decimals	Percents, rates, ratios	≥2 types of rational numbers[1]
Number Properties and Operations	2	14	8	9	3	1
Measurement	6	17	4	2	0	0
Geometry	16	14	2	0	0	0
Data Analysis and Probability	6	12	2	1	5	0
Algebra	2	35	2	4	2	0
Total	**32**	**92**	**18**	**15**	**10**	**1**

[1]Non-integer

NOTE: 2007 mathematics item pool.

Summary

In summary, the expert reviewers judged the NAEP item pool to be broadly aligned with the framework. However there were some important areas of concern, particularly at grade 8, where there was fairly unanimous criticism of

- the poor focus and balance of the item set in number properties and operations, and
- the under-representation of high-complexity items in algebra and in measurement.

In addition, virtually every content area at both grade levels had at least one subtopic where the majority of reviewers judged the item set to be lacking in focus, balance, or reach. It is likely that these problems arise, at least in part, out of features of the framework that were discussed in chapter 2. That is, the framework includes 65 objectives at grade 4 and more than 100 objectives at grade 8. Yet no guidance is provided—either in the framework or elsewhere—that specifies how to set priorities among the objectives. In the absence of such guidance, items can drift toward objectives that are easier to measure, and item selections can be made to satisfy psychometric properties of the test without regard to the impact on content distribution. For example, reviewers noted that, at grade 4, more than one third of the items in the important subtopic of properties of number and operations were very easy items about distinguishing odd from even numbers. Including these items probably helped to balance the difficulty of the test, but at the expense of an odd distortion of content coverage.

More guidance also is required with regard to the appropriate distribution of number types. The framework is not prescriptive with regard to number type, and about half of the reviewers felt that the NAEP item pool was deficient with regard to fractions and other non-integer rational numbers. (As noted, 11 percent of the items at each grade level involve fractions. It is not clear whether this distribution is intentional.)

Finally, more has to be done to incorporate high-complexity items into the item pool. The framework calls for about one quarter of the assessment score to be based on high complexity items, and a substantial number of objectives describe competencies which seem to demand high-complexity items for adequate measurement. Yet the item classifications provided by the test developer only designate five grade 4 items and four grade 8 items as being high complexity, and many (but not all) of the expert reviewers found high complexity to be lacking in virtually every content area. Appendix F presents a brief essay by Jan de Lange, a member of the study's TWG, which describes a conceptual framework for increasing complexity without necessarily increasing item difficulty.

Chapter 4. Is the Assessment Mathematically Accurate and Does It Strike an Appropriate Balance Between Competing Curricula, Philosophies, and Pedagogies?

A high-quality mathematics item demands, from the student, knowledge of mathematics and the know-how to reason with mathematics. It does not demand a general ability to decipher complicated presentations or guess what the test maker is looking for. The presentation of the item should be consistent with correct mathematical language available to the student at the grade level being assessed.

Reasoning with mathematics can include analyzing a situation to identify relevant quantities and expressing their relationship mathematically, but items should not present unnecessary challenges to test takers that are unrelated to mathematical performance. Such inappropriate challenges can include inaccurate or poorly specified mathematics, unreasonable hidden assumptions; misleading language, graphics, or contexts; irrelevant complexities; or other cognitive challenges not related to the NAEP framework.

One must keep in mind that K-8 assessment items are written for, and read by, children. The demands of mathematical quality must accommodate the demands of communicating with children in the target age range. On the one hand, attention to mathematical quality can produce items that are easy to understand because language is precise and extraneous challenges have been eliminated. On the other hand, such efforts can end up making items unnecessarily difficult by requiring the student to read and comprehend too much explicit specification. It is not always straightforward to decide the quality of items written for fourth- and eighth-grade students. Judgments must be made.

A related challenge arises from the fact that the mathematics problem is a peculiar genre of text with its own conventions and assumptions. These genre conventions must be learned. For example, it is not until more advanced courses in high school that acceleration of motion can be modeled mathematically. In earlier grades, speeds are constant; "… a train leaves the station traveling an average speed of 50 mph…" is interpreted by convention as meaning a constant speed of 50 mph, ignoring speeding up and slowing down. This is a conventional word problem assumption. Is it reasonable to expect test takers to understand this assumption?

Approach

Five mathematicians (listed in appendix G) with experience in school mathematics and assessment were assembled to examine NAEP item used in the 2005 and 2007 assessment cycles. The mathematicians were deliberately selected to represent a spectrum of perspectives on current controversies related to school mathematics.

As a frame of reference for interpreting the results, a random sample of items from state tests was shuffled into the deck of NAEP items. The identity of the source was concealed.

The state items were sampled from the most recent released tests or item sets of all of the 40+ states that post items on the Web. This sample of items can be thought of as representing current practice in large-scale assessment.

The mathematicians reviewed items organized into packets of items with similar content. Each reviewer rated each item in their packets as 1 = “adequate,” 2 = “marginal,” or 3 = “seriously flawed,” where the rating categories were defined as follows:

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**
   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 graphics 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 rubric.

At least two mathematicians rated each item. Two packets—one containing many of the number properties and operations items from grade 4, and the other containing many of the algebra items from grade 8—were selected for review by all five of the mathematicians. After each reviewer rated the items in a packet independently, the reviewers compared ratings, discussed differences, and had the opportunity to change ratings. In addition to assigning ratings, reviewers wrote comments to document the reason for each marginal or seriously flawed rating. (See appendix H for a more comprehensive description of the rating procedure.)

At the end of the rating process for each grade level, whole group discussions were held addressing the study question of mathematical quality. The discussions were recorded, noted, and analyzed, and they contributed to the interpretation of the findings.

Our overall approach was designed to preserve legitimate differences of perspective rather than to train to a standard that would produce consistent ratings. Indeed, in this study, differences were considered informative, as were agreements.

After the ratings were compiled, the mean rating for each item was calculated.[21] Then, using the mean ratings, items were designated as “adequate,” “marginal,” or “seriously flawed,” using the following rule:

Mean Rating	Summary Designation
1.0 – 1.4	Adequate
1.5 – 2.4	Marginal
2.5 – 3.0	Seriously flawed

Items that at least one reviewer rated “3” and at least one reviewer rated “1” were also tagged as strong disagreements for additional analysis.

Our procedure weighs against classifying items as adequate since a mean of 1.5 typically arose from an equal number of adequate (1.0) and marginal (2.0) ratings.[22] We chose to designate such items as “marginal” in order to keep them in the pile for further analysis. Our goal was to maximize opportunities for identifying potential improvements, and the reader should take this bias into consideration. (Further breakdowns of the mean-rating distributions *within* categories are also included in the discussion of findings, below.)

Classification distributions were calculated and compared for the total set of NAEP items and the total set of state items. NAEP and state classification distributions were also calculated within each of the five NAEP content areas, and content areas with the highest frequencies of marginal and seriously flawed items were studied further in an effort to identify general issues.

Findings

As noted above, the rating procedures required the mathematicians to support their marginal or seriously flawed ratings with comments, but no systematic comments were collected for the adequate items. Consequently, this chapter has more information about items that were seriously flawed or marginal and less information about items that were adequate. This should not mislead the reader into an unwarranted negative judgment about the overall assessment.

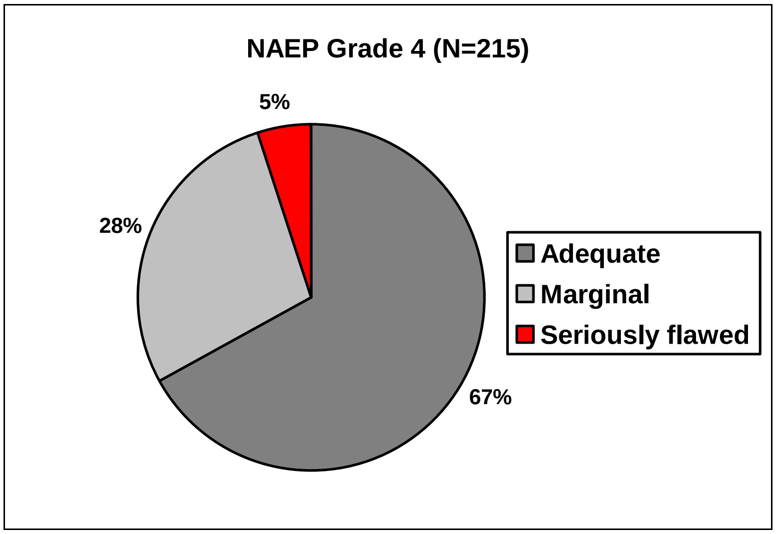
Exhibits IV-1 and IV-2 show the overall findings. Five percent of NAEP items were designated as seriously flawed mathematically at grade 4, and 4 percent were designated seriously flawed at grade 8. The state items were classified as 7 percent seriously flawed in fourth grade and 3 percent seriously flawed in eighth grade. For marginal items, NAEP had 28 percent at grade 4 and 23 percent at grade 8, while the state sample had 30 percent

[21] NAEP includes a certain number of cross-grade blocks in which some, but not all, of the items appear at both fourth and eighth grade. In our procedure, these items were rated twice, once with the grade 4 items and once with the grade 8 items. The items did not always earn the same average score at both grade levels. This could be partly the result of different expectations for different grade levels. It could also reflect the differing perspectives of the raters who happened to be assigned the items at each grade level.

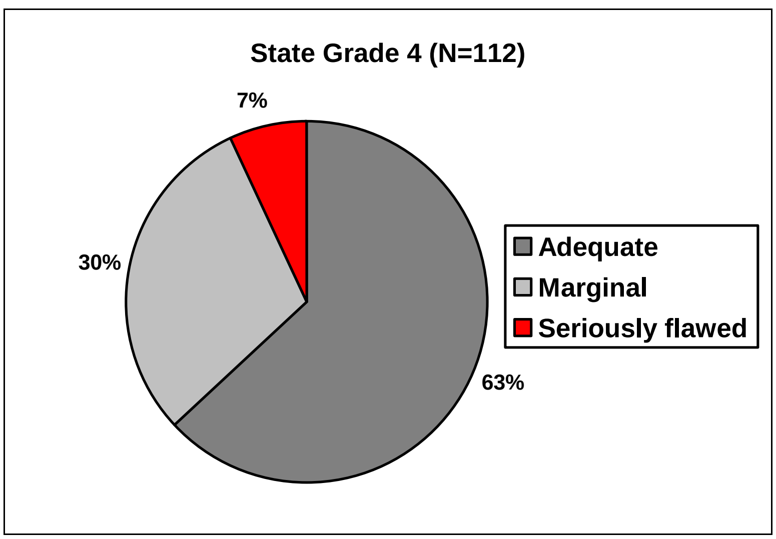
[22] Three ratings of “adequate” and one rating of “seriously flawed” would also produce a mean of 1.5.

at grade 4 and 26 percent at grade 8. By this estimation, NAEP is less flawed than some critics have suggested, but it is also less than perfect mathematically. The substantial number of marginal items in NAEP and the states is cause for concern. Marginal items may well be leading to underestimates of achievement, although this study did not produce empirical evidence on this possibility.

Exhibit IV-1. Percentage of adequate, marginal, and seriously flawed NAEP and state items at grade 4

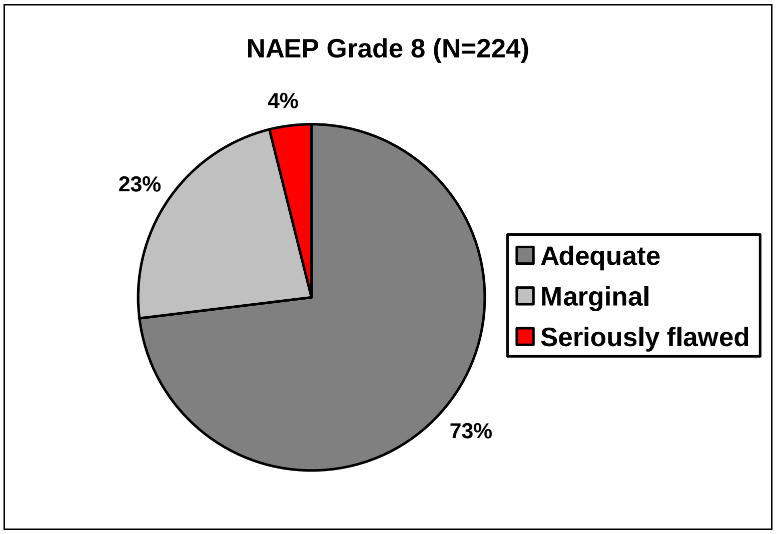


NOTE: NAEP items represent combined 2005 and 2007 item pools.

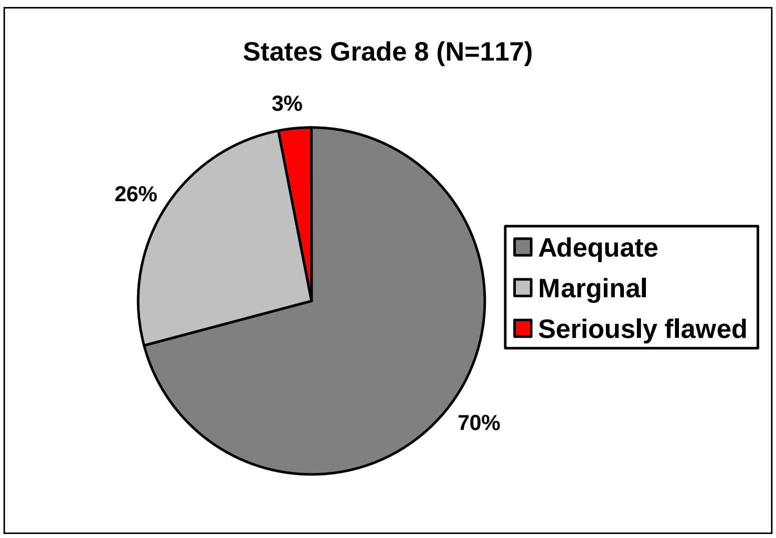


NOTE: State items are a random sample of items from the most recent test forms or item sets released on the Web by 40+ states.

Exhibit IV-2. Percentage of adequate, marginal, and seriously flawed NAEP and state items at grade 8



NOTE: NAEP items represent combined 2005 and 2007 item pools.



NOTE: State items are a random sample of items from the most recent test forms or item sets released on the Web by 40+ states.

Exhibit IV-3 allows one to examine the mean mathematician ratings at a finer grain size. About three quarters of the items designated as adequate had a mean rating of 1.0, meaning that they had been judged adequate by all the mathematicians who reviewed them. The majority of the items in the marginal category had mean ratings less than 2.0, meaning that these items had been rated as adequate by at least one of the mathematicians who reviewed them.[23]

[23] Recall that there were between two and five mathematician reviewers for each item.

Exhibit IV-3. Percentage of NAEP and state items by mean mathematicians' rating

Classification	Adequate		Marginal			Seriously Flawed	
Mean Rating	1.0	1.1-1.4	1.5-1.9	2.0	2.1-2.4	2.5-2.9	3.0
Grade 4							
% NAEP Items	49	18	17	10	<1	5	1
% State Items	46	16	21	9	1	5	2
Grade 8							
% NAEP Items	54	18	13	8	2	2	2
% State Items	58	12	16	9	1	3	1

NOTE: NAEP items represent combined 2005 and 2007 item pools. State items are a random sample of items from the most recent test forms or item sets released on the Web by 40+ states.

These overall similarities in classifications between NAEP and the state samples indicates that the mathematicians were reacting to common practices in U.S. large-scale assessment, rather than to something specific to NAEP. Furthermore, as shown in exhibits IV-4 and IV-5 (below), NAEP and the state samples also demonstrate parallel profiles *across content areas* in the distribution of item classifications. This parallelism further supports the interpretation that certain widespread assessment practices, affecting about 5 percent of items, are seriously flawed in the view of mathematicians.

Although comparison with a random sample of items from 40+ states indicates NAEP is typical, some states may have higher quality items than NAEP, and some states may have lower quality items. Our analysis did not compare states with each other because it was not possible to compare so many item sets within the frame of the study.

Is it possible or likely that the presence of seriously flawed or marginal items could have altered overall NAEP results? Some of the flaws categorized as "serious" are the mathematical equivalent of grammatical errors: students can still understand the problem situation and answer the questions, so the results are not affected. Still, there is something unacceptable about having such errors on a test. Other types of serious flaws, however, could alter results by creating real obstacles for test takers. The mathematicians also were clear that many of the items they rated as marginal exhibited construct-irrelevant difficulties that could affect performance for some test takers.

Classifications by content area

Exhibit IV-4 shows classifications for items within content area for fourth grade. Nine of the 11 seriously flawed items in grade 4 NAEP are in the algebra content area. The state items parallel this pattern: six of the eight seriously flawed items in the state sample are in algebra. This suggests that there is a widespread source of flaw that is not specific to NAEP, but typical in item development for large scale assessments in the United States.

When grade 4 marginal NAEP items are examined, measurement and geometry have more issues than the other content areas. For states, the marginal items are most evident in measurement. While NAEP was somewhat cleaner in number properties and operations, the states were cleaner in algebra.

Exhibit IV-4. Number of grade 4 NAEP and state test items classified as adequate, marginal, or seriously flawed, by content area

NAEP Items	Adequate	Marginal	Flawed	Total
Number Properties and Operations	67	18	2	87
Measurement	27	15	0	42
Geometry	19	15	0	34
Data Analysis and Probability	18	6	0	24
Algebra	12	7	9	28
Total	**143**	**61**	**11**	**215**
STATE Items	**Adequate**	**Marginal**	**Flawed**	**Total**
Number Properties and Operations	35	13	0	48
Measurement	7	10	2	19
Geometry	13	4	0	17
Data Analysis and Probability	7	5	0	12
Algebra	8	2	6	16
Total	**70**	**34**	**8**	**112**

NOTE: NAEP items represent combined 2005 and 2007 item pools. State items are a random sample of items from the most recent test forms or item sets released on the Web by 40+ states. One NAEP geometry item was inadvertently left out of the rating process.

The results for grade 8 are found in exhibit IV-5. At this grade level, data analysis and probability has a high proportion of marginal or seriously flawed items, 15 out of 32 for NAEP, and 7 out of 15 for states.

Exhibit IV-5. Number of grade 8 NAEP and state test items classified as adequate, marginal, or seriously flawed, by content area

NAEP Items	Adequate	Marginal	Flawed	Total
Number Properties and Operations	42	9	0	51
Measurement	27	6	4	37
Geometry	34	8	3	45
Data Analysis and Probability	17	14	1	32
Algebra	43	15	1	59
Total	**163**	**52**	**9**	**224**
STATE Items	**Adequate**	**Marginal**	**Flawed**	**Total**
Number Properties and Operations	27	8	0	35
Measurement	12	4	2	18
Geometry	15	10	0	25
Data Analysis and Probability	8	6	1	15
Algebra	20	3	1	24
Total	**82**	**31**	**4**	**117**

NOTE: NAEP items represent combined 2005 and 2007 item pools. State items are a random sample of items from the most recent test forms or item sets released on the Web by 40+ states.

What are the flaws?

Pattern problems in algebra

The seriously flawed algebra items were examined, along with reviewer comments. All the seriously flawed algebra items (nine in fourth grade and one in eighth grade) related to *patterns*. In addition to the mathematical quality of these items, several mathematicians made the point that there were too many of them, regardless of whether they were mathematically adequate. Thus, not only the density of flaws, but the enthusiasm for pattern items (as reflected in the number on the test) was criticized. In fourth grade especially, it was agreed by the mathematicians that the foundations of algebra needed to be represented with other types of items, such as number sentences of the type: 23 + ? = 30 + 8.

Note that a separate group of teachers, mathematics educators, and mathematicians, who were asked how well the item pool assessed the NAEP framework for algebra at fourth grade, were moderately approving of the way that the algebra items reflected the framework (see chapter 3). In fact, it is the NAEP framework, and not just the item pool, that emphasizes patterns more than the mathematicians would like. Indeed, the analysis of the NAEP framework compared to a sample of state and other nation's standards (chapter 2) shows NAEP placing more emphasis on patterns than the comparison standards do. Nevertheless, it is important to make clear that mathematicians were not opposed to pattern problems per se (although they thought they were overemphasized compared to other algebra topics), but they were very critical of badly posed pattern problems.

In the judgment of the mathematicians, unreasonable hidden assumptions flaw many of the pattern items. In the absence of rules for pattern generation, there are a multitude of possible patterns and possible correct answers. (Thus it is incorrect to say "the" pattern when there are many possible.) Yet many of the pattern items do not explicitly tell the students how the patterns were generated. Rather, the students are expected to share in assuming the same (unspoken) rules as the item writer. One reviewer remarked that the pattern items were like IQ items, which measured the test takers' shared assumptions with the test makers.

Two NAEP pattern items, which are indicative of the flaws found, are reproduced in exhibits IV-6 and IV-7.[24] In the fourth-grade item in exhibit IV-6, the sequence 19, 22, 25, 28, 31, … is given and referred to as "the pattern." The fourth-grade student is expected to assume that the pattern will continue by increasing the number by the same amount at each step. The item does not state this explicitly. Therefore, one reviewer said: "From a mathematician's perspective, this is ill posed." Another said, it "…could be saved by stating the pattern..."

[24] As noted in chapter 3, NAEP items are secure while they continue to be used in operational assessments. However, about 30 percent of the item blocks are replaced after each assessment cycle. Only items that were replaced after the 2005 or 2007 assessments are reproduced here.

One could address the reviewers' concern by editing the item to ask: "If the pattern shown continues to increase by the same amount at each step…" Another acceptable revision would be to pose the question "What rule could make the pattern shown above?" and to offer, as answer choices, a selection of rules like "add three each time." Finally, if this were a constructed response item, a student could be asked to state a rule that explains the pattern shown, and then further asked to apply the rule to find a number at some later step.

Exhibit IV-6. A pattern item that is not adequately specified

19, 22, 25, 28, 31, . . .

If the pattern shown continues, which of the following numbers would be in the pattern?

(A) 38

(B) 39

(C) 40

(D) 41

SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment*, Grade 4, Block Z12M3A #10, 2005.

The eighth-grade pattern item shown in exhibit IV-7 further illuminates the issue bothering the mathematicians. The mathematicians suggested revisions that would make the item acceptable. One suggested: "If you continue this pattern by adding the…" Another mathematician proposed: "Use the pattern…" instead of "According to the pattern suggested…" The point is to get away from guessing the pattern (more appropriate to an IQ test) and focus on what determines the pattern and how is it modeled mathematically. Furthermore, although the examples provided in the item "suggest" that all the sums in the pattern start with the number 1, the question actually asks a more general question: "…how many consecutive odd integers are required to give a sum of 144?" Two consecutive odd integers are all that is required: 71 + 73 =144. Of course, "2" is not offered as a response. To be worded correctly the item should say, "…consecutive odd integers, beginning with 1…"

Would making the language more precise in this way alter student performance on the item? Perhaps not, but the mathematicians still believe the more precise language should be used.

Exhibit IV-7. A pattern item that is not adequately specified

$$1 + 3 = 4$$
$$1 + 3 + 5 = 9$$
$$1 + 3 + 5 + 7 = 16$$
$$1 + 3 + 5 + 7 + 9 = 25$$

According to the pattern suggested by the four examples above, how many consecutive odd integers are required to give a sum of 144 ?

Ⓐ 9

Ⓑ 12

Ⓒ 15

Ⓓ 36

Ⓔ 72

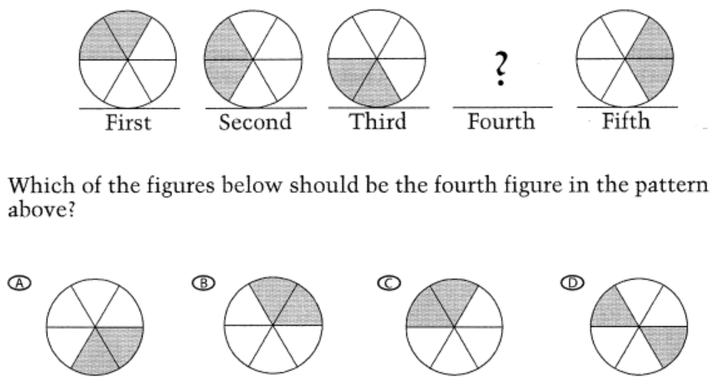
SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment*, NAEP Grade 8, Block Z12M3B #12, 2005.

Although the mathematician's objections can be dealt with by revising the items, the revisions do not necessarily assess the same content as the unrevised items. In the unrevised form, the student has to decipher the pattern and figure out how it extends or applies. In the revised form, the pattern is explicitly described so the student only has to apply the rule.

Other examples set in geometric contexts allow for specifying the process that generates the pattern without specifying the numeric rule. This would satisfy the mathematicians' perspective while still assessing the students' skill at formulating the rule.

For example, the cross-grade NAEP item shown in exhibit IV-8 could be revised to make it acceptable to the mathematicians. The revision is illuminating; one merely needs to insert: "The figure rotates by the same amount each step." This insertion serves to adequately determine the pattern, although it may present reading difficulties to fourth-grade students.

Exhibit IV-8. A pattern item that could be edited to be acceptable while still assessing students' ability to formulate a rule



SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment*, Grades 4 and 8, Block Z12M4A/B #3, 2005.

To illustrate acceptable pattern problems, a search of released items from state tests was made.[25] The following examples meet the requirement of not having unreasonable hidden assumptions.

The fourth-grade Pennsylvania state item in exhibit IV-9 was found adequate by the mathematicians because the rule that generates the pattern is explicitly stated. The item also involves the relationship between an input and output variable, which relates directly to the future study of functions.

[25] There were a number of acceptable pattern items in the 2005-2007 NAEP item pool, but they were still in operational use and therefore could not be displayed in this report.

Exhibit IV-9. A pattern item judged adequate because the rule for generating the pattern is given

The input/output table shows the rule:
Multiply the input number by 3 and then
Add 2.

Input	Output
3	11
5	17
6	20
7	?
9	?

What 2 output numbers are missing in the table?

A 14, 27

B 21, 27

C 23, 26

D 23, 29

SOURCE: Pennsylvania Department of Education, Bureau of Assessment and Accountability, 2006-2007 Mathematics Item and Scoring Samples, Grade 4; item D.1.2.1.

Two other state pattern items that were judged adequate because the rule for generating the pattern was given are shown in exhibits IV-10 and IV-11. These fourth-grade items are from California and Ohio, respectively.

Exhibit IV-10. A pattern item judged adequate because the rule for generating the pattern is given

The numbers in this pattern decrease by the same amount each time.
What are the next three numbers in this pattern?

10, 8, 6, 4, 2, , , .

A 0, -2, -4

B 0, -1, -2

D 0, 2, 4

D 0, 1, 2

SOURCE: California Department of Education, *California Standards Test, 2007 Released Test Questions*, Grade 4, #8, 2007.

Exhibit IV-11. A pattern item judged adequate because the rule for generating the pattern is given

Courtney starts with 12 birdhouses. She makes three new birdhouses each week.

Which pattern shows the number of birdhouses Courtney has at the end of each week?

A. 3,6,9,12
B. 3,15,27,39
C. 12,15,18,21
D. 12,24,36,48

SOURCE: Ohio Department of Education, *Ohio Achievement Tests*, 2005 Mathematics student test booklet, Grade 4, #3, 2005.

Finally, in exhibit IV-12, we see another model for an acceptable pattern item. In this fourth-grade Massachusetts item, the student is asked to supply a possible rule for an input-output table. The item is acceptable because it asks for "a possible rule," rather than "the rule."

Exhibit IV-12. A pattern item judged adequate because it asks for "a possible rule"

An input-output table is shown below.

Input (A)	Output (B)
7	14
12	19
20	27

Which of the following could be the rule for the input-output table?

A. $A \times 2 = B$
B. $A + 7 = B$
C. $A \times 5 = B$
D. $A + 8 = B$

SOURCE: Massachusetts Department of Education, *Massachusetts Comprehensive Assessment System*, Grade 4, 39, 2006.

Unduly complicated presentation

A common reason for judging an item marginal was undue complications in the presentation of the problem. Often, the language was unnecessarily complicated. Sometimes the situation presented had complications disproportionate to the mathematics being assessed. Elaborate contexts for simple questions are inappropriate in a test with limited time. The content has to justify the context.

Items often include a presentation of a problem situation in words, diagrams, and/or symbols. The student faces three challenges:

1. make sense of the situation,

2. understand the question being asked about the situation, and

3. answer the question.

Each of these three challenges can combine, in some mixture, legitimate aspects of the mathematics defined in the framework and construct-irrelevant difficulty. For the mathematician, it is the mathematical *relevance* of the challenges involved in making sense of the situation and understanding the question that determines the contribution of these factors to item quality. The preferences of the mathematicians in this regard differed somewhat from the preferences of the mathematics educators who participated in the validity study. The mathematicians preferred items in which situations were used to test understanding of mathematical concepts. The mathematics educators also liked these types of items, but, in addition, they wanted more items that tested students' skills at using mathematics to make sense of situations that have features typical of real world applications. The NAEP assessments reviewed for this study had few items of the latter type (see chapter 3), as reflected in the lack of high complexity.

The items critiqued in the following paragraphs were defective (that is, unduly complicated) with regard to one or more of the sources of challenge described above.

Some of the items involved geometrical situations. In one case, students were asked to assemble a three-dimensional figure from a paper punch out. This assembly job was, in itself, time consuming with demands of its own. Different students might react differently to the assembly demands. Indeed, in an attempt to show how easy it was, one of our participants proceeded to assemble it incorrectly. Most problematic, however, was the fact that once the figure was assembled, the items that were asked about it were trivial vocabulary questions that did not even take advantage of its three-dimensional nature.

Other items had directions that were too complicated for the amount of mathematical content assessed. In the cross-grade NAEP item shown in exhibit IV-13, the mathematicians felt the directions were much more difficult than the mathematics, at least for fourth-grade students. They rejected the idea that, in problems like this, understanding the directions is part of the mathematics.[26] The reading comprehension of instructions like this is not what a mathematics test should be assessing.

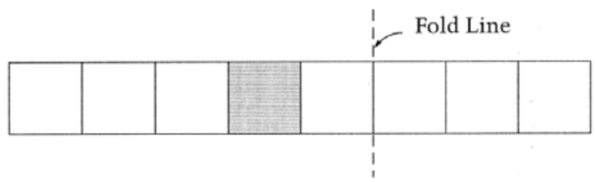
[26] The NAEP framework does not specify following directions as a target of assessment.

Exhibit IV-13. An item in which the directions are more difficult than the mathematics

You may use the paper strip from your packet.

Place an X in one of the squares below so that if the paper strip were folded along the dotted fold line shown, the square with the X could cover the shaded square.

Show your answer on the strip below.



SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment*, NAEP Grades 4 and 8, Block Z12M4A #13/Z12M4B #12, 2005.

To reiterate, the criticism about complicated presentations is not a criticism of items that ask students to formulate mathematical expressions in order to model imaginary situations. It is a criticism of awkward or inconsiderate presentations of the situations.

In the fourth-grade NAEP item in exhibit IV-14, a good problem is made unnecessarily complicated by decorating the problem with an episode about Jan entering numbers in a calculator and forgetting two of them. This introduces reading comprehension hurdles unrelated to the mathematics as well as contamination related to variation in students' background knowledge (e.g., prior experience with calculators will differ; different calculators have different orders of operations).

Exhibit IV-14. An item with unnecessary reading and prior knowledge demands

Jan entered four numbers less than 10 on his calculator. He forgot what his second and fourth numbers were. This is what he remembered doing.

$$8 + \boxed{} - 7 + \boxed{} = 10$$

List a pair of numbers that could have been the second and fourth numbers. (You may use the number tiles to help you.)

________ , ________

List a different pair that could have been the second and fourth numbers.

________ , ________

SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment*, NAEP Grade 4, Block Z12M4A #12, 2005.

There is a difference between decorating a problem with a context (a practice criticized by mathematicians across the spectrum) and presenting a problem situation out of which the mathematics comes (a practice accepted by mathematicians across the spectrum). A simple example of the latter is the fourth grade NAEP item shown in exhibit IV-15.

Exhibit IV-15. An item in which the mathematics arises appropriately out of the problem situation

N stands for the number of hours of sleep Ken gets each night. Which of the following represents the number of hours of sleep Ken gets in 1 week?

Ⓐ $N + 7$

Ⓑ $N - 7$

Ⓒ $N \times 7$

Ⓓ $N \div 7$

SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment*, Grade 4, Block Z1M12 #12, 2005.

Language that is unclear, inconsiderate, or misleading

While the language of the majority of NAEP items was good enough, some items used language that presented difficulties to the student out of proportion to the mathematics being assessed. Such items could have a false negative impact on scores. Unclear language is always imprecise; but imprecise language can be clear, and precise language can be confusing for students at a given grade level. The mathematicians were agreed that the important issue was being clear to the student. Unclear language can lead to false negatives (e.g., the student knew the mathematics being assessed, but misunderstood the question due to poor item construction). Confusing language can also waste time, casting a time shadow over performance on items later in the test.

Sometimes the language in the items was more puzzling than the mathematics. The syntax of the question in the fourth-grade NAEP item shown in exhibit IV-16 would be difficult to process for many students at this grade level.

Exhibit IV-16. An item with unnecessarily difficult syntax

> There are 8 children on a hike. One-fourth of them are wearing hats. How many more would need to put on hats to have all of them wearing hats?
>
> Answer: _______________
>
> SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment*, Grade 4, Block Z12M3A #11, 2005.

Multiple-choice questions are a genre unto themselves. There are inherent difficulties in the genre that can interfere with reading comprehension and contaminate the measurement of mathematics. A simple example is the fourth grade item shown in exhibit IV-17. The item stem begins by asking "which of these…," where the pronoun "these" refers to something not yet stated. Indeed, it refers to a nominal category for which the fourth-grade student may have no vocabulary. Which of these what? However, the mathematics content of this item is so easy (knowing that a meter stick measures length, not temperature, weight, or number of people) that the syntactic puzzle may not make much difference.

Exhibit IV-17. An item with unnecessarily difficult syntax

Which of these could be measured using a meter stick?

Ⓐ The length of a swimming pool

Ⓑ The temperature of the water in a swimming pool

Ⓒ The weight of the water in a swimming pool

Ⓓ The number of people in a swimming pool

SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment*, Grade 4, Block Z1M12 #5, 2005.

An example of imprecise language shows up in a grade 8 Louisiana state item (exhibit IV-18). The reader is expected to assume that *s* represents *the number* of small tables, not, as the item states, "…small tables (s)...," and likewise, that l is the *number* of large tables. Later in the problem the number of people is correctly stated. All of the mathematicians found such incorrect usage irritating and unacceptable regardless of whether students were bothered by it. An essential skill in using mathematics to solve problems is identifying relevant quantities and defining variables. The wording in this problem exemplifies bad language habits. The revision could state, " …a number of small tables, *s*, and a number of large tables, *l*…"

Exhibit IV-18. An item with imprecise language

A restaurant has small tables (s) and large tables (l). Small tables seat four people each, and large tables seat eight people each. Which inequality shows the maximum number of people (p) that can be seated at the restaurant?

O A. $p \geq 8l + 4s$
O B. $p \leq 8l + 4s$
O C. $p > 8l + 4s$
O D. $p < 8l + 4s$

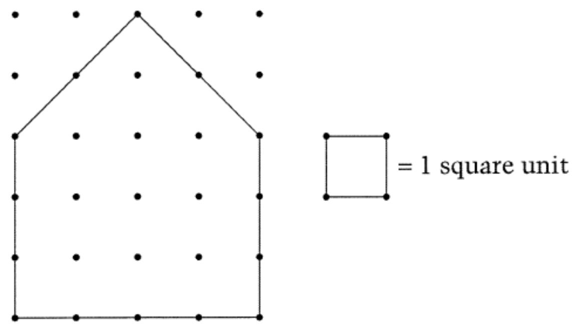
SOURCE: Louisiana State Board of Elementary Education, *Louisiana Educational Assessment Program (LEAP)*, Grade 8 Practice Test, #5, 2006.

A different type of language challenge appears in a grade 8 state item that asks: "…which inequality shows the maximum number…" The answer choices then offer the following expressions of inequalities: "more than or equal to," "less than or equal to," "more than," and "less than." A student has to reconcile "maximum," which is similar to "most," with its actual meaning, which is closer to "can be no more than," or literally, "less than or equal to." This flip in semantic polarity is difficult from a reading perspective, but it is the mathematical point of the problem. No mathematician objected to this language challenge because it is part of the challenge of mathematics.

The following are some additional examples of vague or imprecise language that tended to recur in items and that the mathematicians judged was unfair to students and contaminating to the measurement goals of the test.

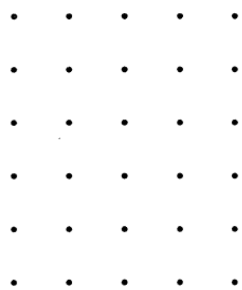
- A common device that the mathematicians disliked was asking students to "…show the steps…" or "…show all the steps…" as though there were one and only one sequence of steps. It is often unclear what steps the test writer has in mind. This can be another case of "guess what's in the item writer's mind."
- The word "about," or the phrases "about how many…," or "about how much…," can be used incorrectly. In a number of items, several of the answer choices could be considered "about" the exact number. But only one choice is "closest" to the exact number. Thus the question "which is closest?" is better. Item writers apparently assume that students will understand that "about," when used on a multiple-choice test, does not mean "about," but means "closest." This is inconsiderate writing and can produce false negatives.
- A number of items ask students which is the "best" without stating for what purpose. This happened with a number of measurement items. Which is best depends on for what purpose.
- In the eighth-grade NAEP item that is shown in exhibit IV-19, students are asked to draw "…a different pentagon…," where the crucial word "different" is not defined. Matters are made worse by the negative, "…be sure that the pentagon you draw does not…" This type of advice makes sense only if you have imagined something the item writer does not want you to imagine.

Exhibit IV-19. An item with imprecise and confusing language



a. What is the area, in square units, enclosed by the pentagon above?

b. On the figure below, draw a different pentagon that has the same area as the one shown. (Be sure the pentagon that you draw does not look like the one shown when it is turned in a different direction).



SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment*, Grade 8, Block Z12M4B #17, 2005.

- Gratuitous words can be unfair to students. Why say “average monthly pay” when there is no average situation? Just say “monthly pay.” “Average” will mislead students into trying to find an average or remember what they learned about averages.

- The phrase “at random” is also abused. Usually “equally likely” is what is meant. “Equally likely” expresses the relevant mathematics and does not require an unnecessary interpretation from the test taker.

- Parallelism in language is considerate. Items should not, for example, refer to the same character as “a friend” in one sentence and “a boy” in another.

Time consuming items

The fourth-grade NAEP item in exhibit IV-20 exhibits another type of flaw. A diligent student, who does not notice the efficient way to think about the problem, will spend too much time on this one item. The student might write all the numbers to 100, and get the correct answer, but spend too much time. This will diminish potential performance on the remainder of the test. Such items may correlate with general cleverness as much as with mathematical knowledge and know-how.

Exhibit IV-20. An item which may be unnecessarily time consuming for some students

A photo album has 100 pages. Carol is numbering the pages 1, 2, 3, and so on.

How many times does Carol have to write the digit 1 ? ____________

Show how you arrived at the solution.

SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment*, Grade 4, Block Z12M3A #18, 2005.

Measurement

Measurement at fourth grade had many items that merely asked what unit or what measuring device fits a situation. The fourth-grade NAEP item shown earlier in exhibit IV-17, for example, essentially assesses whether a student knows what “meter stick” means. This is a tiny amount of mathematical content on which to spend an item, and it is the sort of measurement item expenditure that ties back to concerns about the heavy emphasis on measurement in the NAEP framework. (See chapter 2 for a discussion of how the NAEP framework compares to other standards.)

Agreement among mathematicians

The mathematicians were not trained to agree, as might have been the case if this were a scoring procedure. They were trained on the meaning of the criterion and the issues to be evaluated. Thus, the level of agreement or disagreement is, in itself, evidence of harmony or discord among mathematicians with regard to item quality.

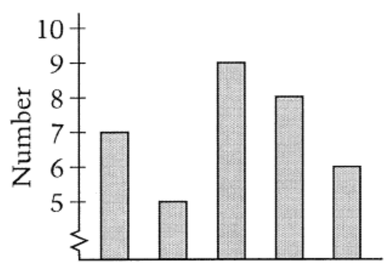
To examine level of agreement, an operational definition of “strong disagreement” was established as being any ratings that are two categories apart for the same item (i.e., one rating of “1” and another of “3”). Although the number of raters per item varied from two to five, almost all of the items on which there were two-category disagreements ended up being designated as “marginal” by virtue of the arithmetic of the ratings.

Given that we deliberately selected mathematicians with a range of perspectives on school mathematics controversies for the rating task, there was little evidence of strong disagreement among raters. We found 8–11 percent strong disagreement across grades and samples.

The agreement extended beyond the ratings themselves to the comments and discussions. For example, on the issue of pattern items, all agreed that some pattern items were legitimate; that NAEP spent too many items on patterns at the expense of other foundational concepts for algebra at fourth grade; that patterns with hidden assumptions were bad practice in item writing; and that asking for the rule that determines the pattern is a good idea. This level of agreement suggests that our findings were not merely a reflection of “math wars” agendas, but reflected judgments of basic item quality independent of the mathematicians’ perspectives on the controversies.

An examination of the items with strong (two-category) disagreements showed that, in many cases, there was agreement about the character of the item, but disagreement about how important a negative feature was. For example, the comments on the NAEP graph-reading item that is shown in exhibit IV-21 ranged from “interpretation” (by a mathematician who rated it adequate); to “bit strange, more ‘hunting/eliminating’ and ‘common sense’ than math” (by a mathematician who rated it marginal); to “this is just weird” (by a mathematician who rated it seriously flawed).

Exhibit IV-21. An item in which the mathematicians disagreed about whether the item was assessing mathematics



Jin made the graph above. Which of these could be the title for the graph?

Ⓐ Number of students who walked to school on Monday through Friday

Ⓑ Number of dogs in five states

Ⓒ Number of bottles collected by three students

Ⓓ Number of students in each of ten clubs

SOURCE: U.S. Department of Education, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment*, Grade 4, Block Z1M12 #11, 2005.

A different example, from the eighth-grade state sample, is shown in exhibit IV-22. This item illustrates the difficulty in judging what assumptions are appropriate at a grade level. The problem assumes a particular orientation in space for the prism. One mathematician worried about the applicability of the item to “twisted shapes,” while another asked: “What is the distinction between ‘side’ and ‘base’?” The third mathematician who rated the item thought the problem was “good.” “Side” and “base” might well be widespread conventional terms in school mathematics, but perhaps they should not be, if they lack mathematical definition. More precisely, the question is about the shapes and numbers of the "faces" of the prism, yet "faces" is not mentioned.

Exhibit IV-22. An item in which the mathematicians disagreed about what assumptions are appropriate at grade level

Which of these describes a triangular prism?

A 4 triangular sides and 1 square base
B 3 triangular sides and 1 triangular base
C 3 rectangular sides and 2 triangular bases
D 4 rectangular sides and 2 square bases

SOURCE: Mississippi Department of Education, *Mississippi Grade Level Testing Program*, Grade 8 Sample Items, #31, 2001.

Summary

NAEP item quality is typical of large-scale assessments overall and in specific content areas. The results of this study were virtually the same for NAEP and a random sample of released state test items.

Ratings provided by mathematicians who were chosen from across the spectrum of possible positions regarding current mathematics curriculum controversies, indicated that only 5 percent of the grade 4 NAEP items and 4 percent of the grade 8 NAEP items were seriously mathematically flawed. Further analysis showed that the flaws in these items were mostly the mathematical equivalent of misspellings on a vocabulary test, or grammatical errors on a reading test. These flaws seem unlikely to affect achievement estimates overall. Nevertheless, there is little excuse for having *any* flaws of this kind on the assessment.

The flaws were concentrated in certain areas: patterns in algebra at fourth grade, and measurement and geometry at eighth grade.

A greater cause for concern is the substantial number of NAP items (and state items) that were classified as marginal based on the mathematicians' ratings—nearly 30 percent at grade 4 and slightly fewer at grade 8. There were a variety of reasons why items ended up classified as marginal. These reasons are described and illustrated in this chapter. In sum, many marginal items were inconsiderate of the test takers and presented construct-irrelevant challenges that often exceeded the modest mathematical challenge of the item. These irrelevant challenges took the form of poorly written text, complicated instructions, misleading presentations and excessive contexts not related to defining or solving the problem. Many were mathematically off base, if not incorrect.

It is beyond the reach of this study to determine empirically the effect of marginal items on performance or on trends. However, most of the issues related to adding irrelevant difficulty to the item. Thus, it is fair to assume that any impact on performance would be negative. Moreover, although this study looked only at 2005–2007 items, there is no reason to think the items we reviewed were worse or better than items on earlier NAEP assessments. Therefore, there is also no reason to think that marginal items have had an impact on NAEP trends. Nevertheless, it would be appropriate to undertake empirical studies on this topic.

The observed problems with item quality are not NAEP specific issues. It may be that these sorts of obstacles are familiar to test takers in this era of high-stakes testing. If so, learned test-taking strategies and skills may compensate, to some degree, for the construct-irrelevant difficulties of these items. But this raises the question: Do we want our students spending their time learning how to guess what a poorly written item means?

Furthermore, it may be that construct-irrelevant skills for handling these items have been acquired by the most test savvy (e.g., educationally advantaged) subpopulation, but not by other subpopulations. It is easy to worry especially about subpopulations with low reading levels. NAEP may be underestimating the mathematics achievement of these populations and overestimating the mathematics achievement of those who are more test savvy.

And finally, NAEP leads by example, as do state tests. Assessment items should exemplify the best in mathematics, not the marginal

Chapter 5. Does NAEP Properly Consider the Spread of Abilities in the Assessable Population?

Other chapters of this report have addressed questions regarding the NAEP framework, the adequacy of the item pool, the mathematical accuracy of the NAEP items, and the appropriateness of the items for different curricula and different instructional philosophies and pedagogies. In this section of the report we address a psychometric question: Does NAEP properly consider the spread of achievement across the assessable population?

The precision with which an assessment measures the achievement of students depends on a number of characteristics of the assessment. It depends on the number of items and on the degree to which items discriminate among students with different levels of achievement. It also depends on the match of the difficulty of the items to the achievement levels of the students being assessed. Other things being equal, the precision of measurement increases as the number of items administered to each student increases. Precision is enhanced when the difficulty of the items is appropriate for the achievement levels of the students being assessed and when the items have good discriminating power.

It would be relatively easy to design an assessment that would have a high level of precision if the target population for NAEP was narrowly defined (e.g., eighth-grade students who were enrolled in an algebra course with a well-defined curriculum). The challenge for NAEP, however, is that the assessment is intended to measure student achievement over a broad range (e.g., the mathematics achievement of all eighth-grade students in the United States regardless of the type and level of mathematics instruction they have experienced).

NAEP scale scores are based on an application of item response theory (Yamamoto and Mazzeo, 1992), which also provides a basis for addressing the question of the relative precision of NAEP for different segments of the population of assessable students. Item response theory allows the estimation of the standard error of measurement for various points along the achievement continuum. For this study, standard error of measurement curves for the 2005 NAEP mathematics assessment were plotted separately for both grade levels and for each of the five content area subscales that comprise the NAEP mathematics assessment (number properties and operations, measurement, geometry, data analysis and probability, and algebra). The resulting standard error of measurement curves were compared to the 2005 distributions of achievement for each of the subpopulations of students that comprise the mandated reporting groups in NAEP. Comparisons of the distributions of student achievement with the standard error of measurement curves provide a means of identifying the achievement levels where the assessment had the greatest precision (smallest standard error of measurement) as well as the levels where the precision was less than might be desired.[27]

The NAEP reporting groups are based on gender, race/ethnicity, eligibility for free or reduced-price lunch, disability status, and English language learner status. Except for gender, each reporting group contrast includes a focal group whose performance distribution is significantly lower than the performance distribution for the population as a whole. Therefore, measurement precision for these subgroups is differentially affected by the standard error of measurement in the lowest part of the achievement continuum.

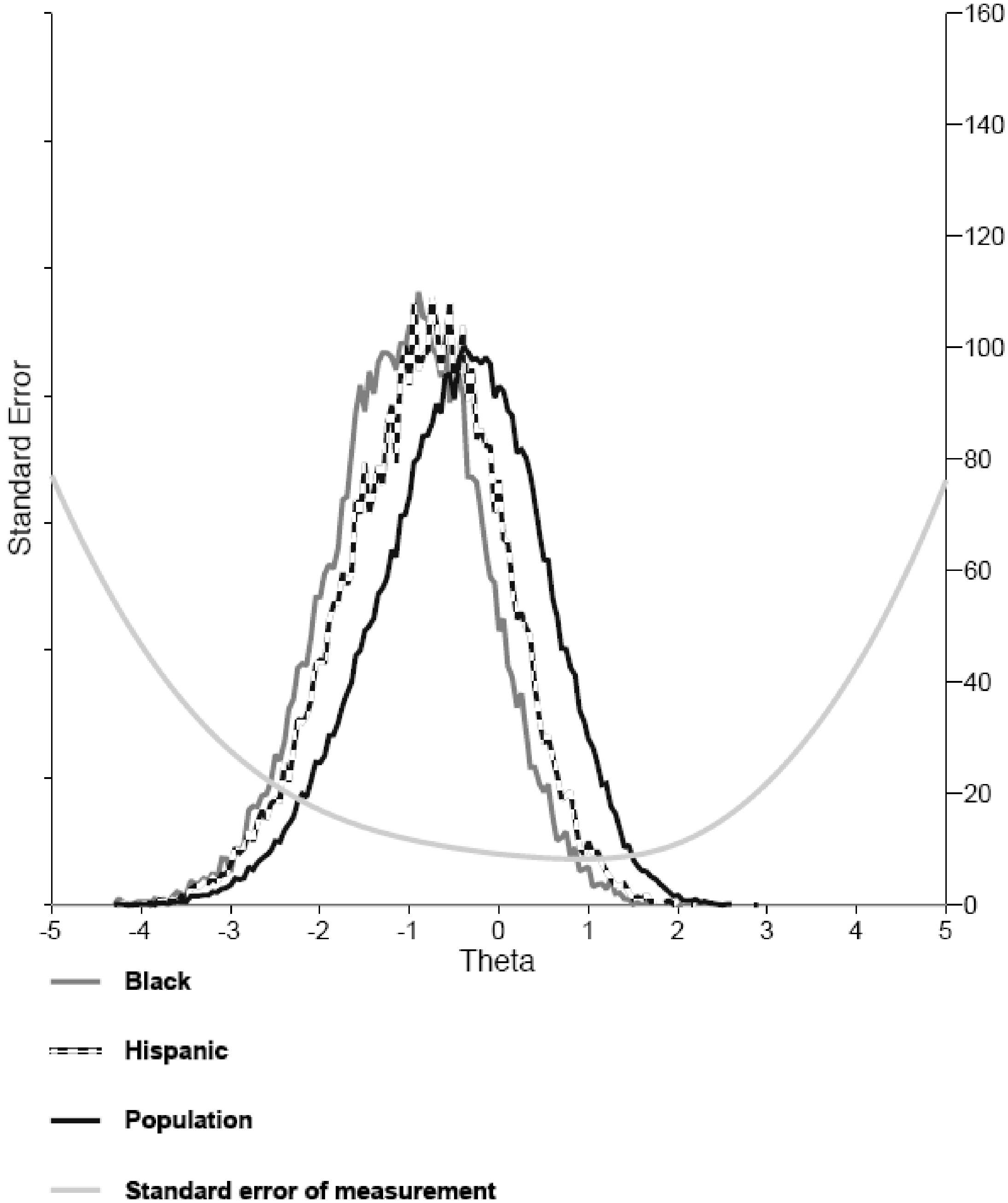
The full set of plots is in appendix I. In this chapter we present some examples drawn from the subscales with the most items at each grade level—number properties and operations at grade 4 and algebra at grade 8. Exhibit V-1, for example, displays the standard error of measurement curve for the grade 4 number properties and operations subscale together with frequency distributions for the population as a whole (gray curve) and separately for black and Hispanic students. As can be seen, the standard error of measurement is relatively small for students with middle and high achievement, but rises fairly sharply in the lowest levels of achievement. Thus, the standard error of measurement is roughly twice as large for the lowest achieving 5 percent or so of the black students as it is for high achieving students.

Exhibit V-2 displays comparable information when the subpopulations are defined by eligibility for free or reduced-price lunch.[28] Similar to exhibit V-1, it can be seen that the standard error of measurement is roughly twice as large for low achieving students who are eligible for free or reduced-price lunch as it is for students with more typical or high achievement.

Standard error of measurement curves and frequency distributions of student achievement for the grade 8 algebra subscale are displayed in exhibits V-3 and V-4. The plots are comparable to those in exhibits V-1 and V-2 in that subpopulations are defined either by racial/ethnic group or by eligibility for free or reduced-price lunch. The most notable difference between the plots for grade 8 algebra and those shown earlier for grade 4 number properties and operations is that, in algebra, the standard error of measurement is at or near its lowest value for a somewhat narrower range of achievement. As was true for grade 4 number properties and operations, the precision of measurement for the grade 8 algebra subscale drops off substantially for the lowest achieving black or Hispanic students and for the lowest achieving students who are eligible for free or reduced-price lunch.

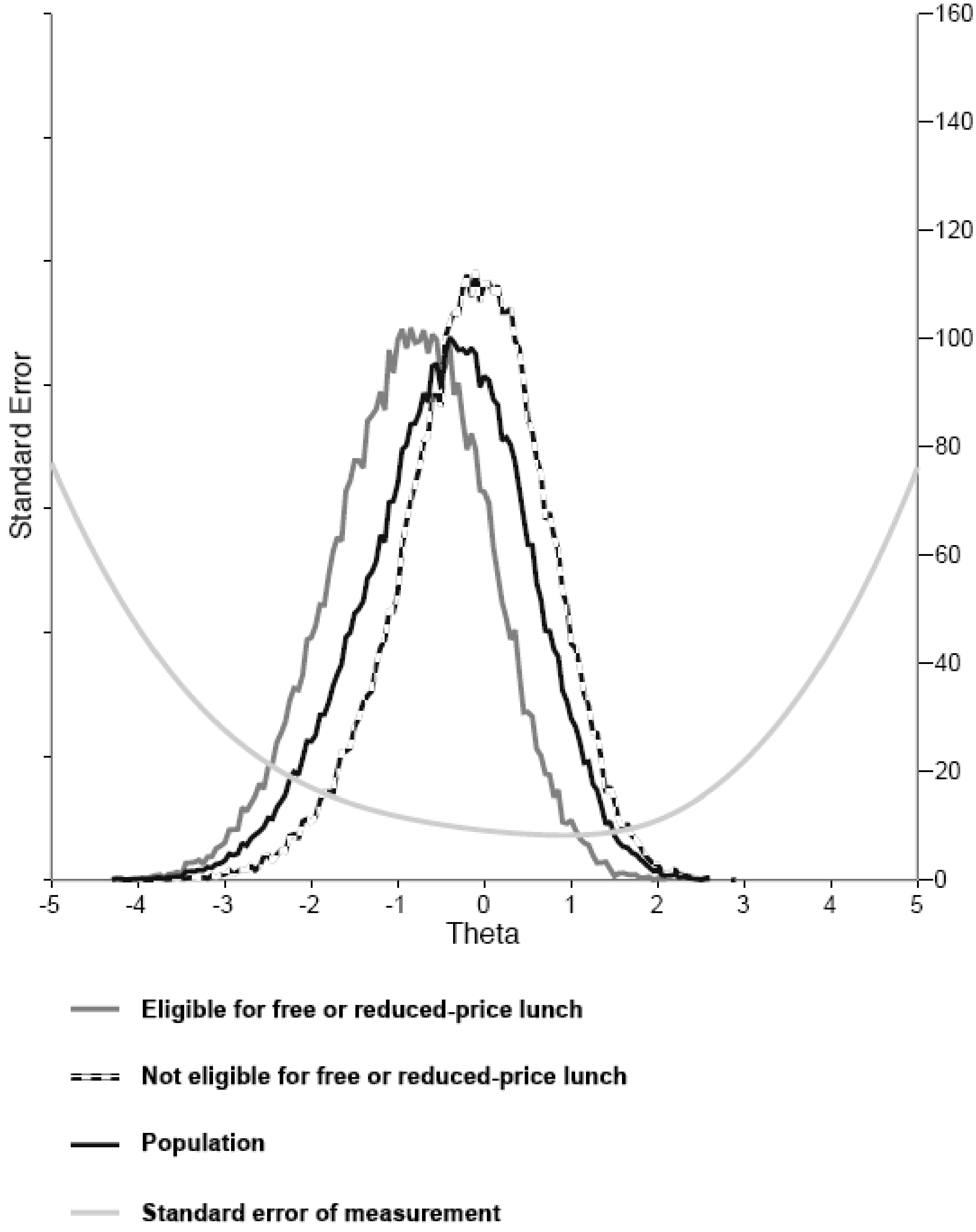
[27] The curves presented here are based on the theta metric, which is used during the subscale analysis. The subscales are then combined into a weighted composite scale that preserves the relative emphasis for each content area as specified in the NAEP framework, and the composite scale values are converted to the 500-point metric used for reporting. Achievement levels are set on the composite scale.
[28] Note that the standard error of measurement curve is the same in exhibits V-1 and V-2, as is the frequency distribution for the total population. Within grade and subscale, only the frequency distributions for the specific subpopulations vary across plots.

Exhibit V-1. Grade 4 number properties and operations subscale, 2005: Standard error of measurement and achievement distributions by race/ethnicity



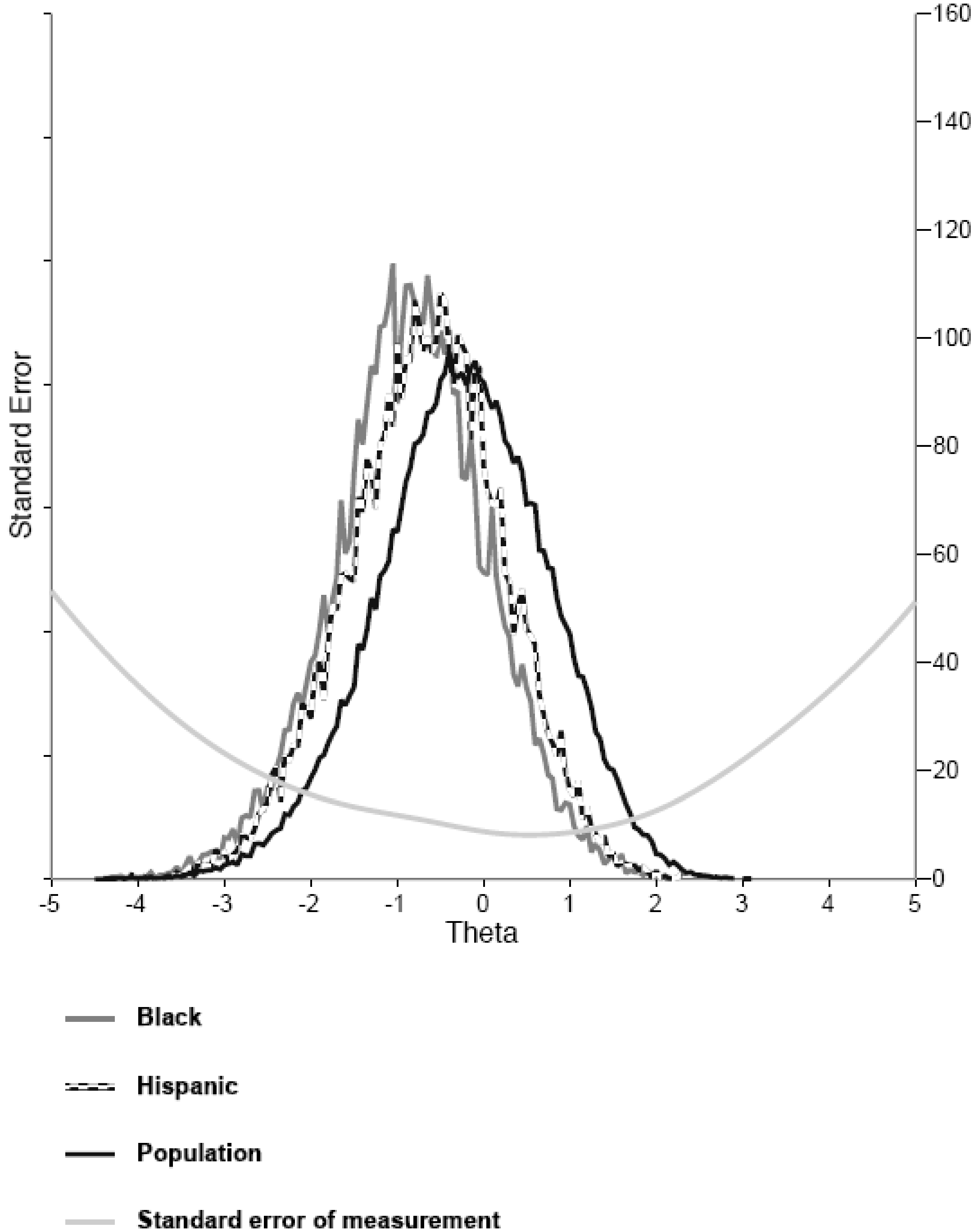
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit V-2. Grade 4 number properties and operations subscale, 2005: Standard error of measurement and achievement distributions by eligibility for free or reduced-price lunch



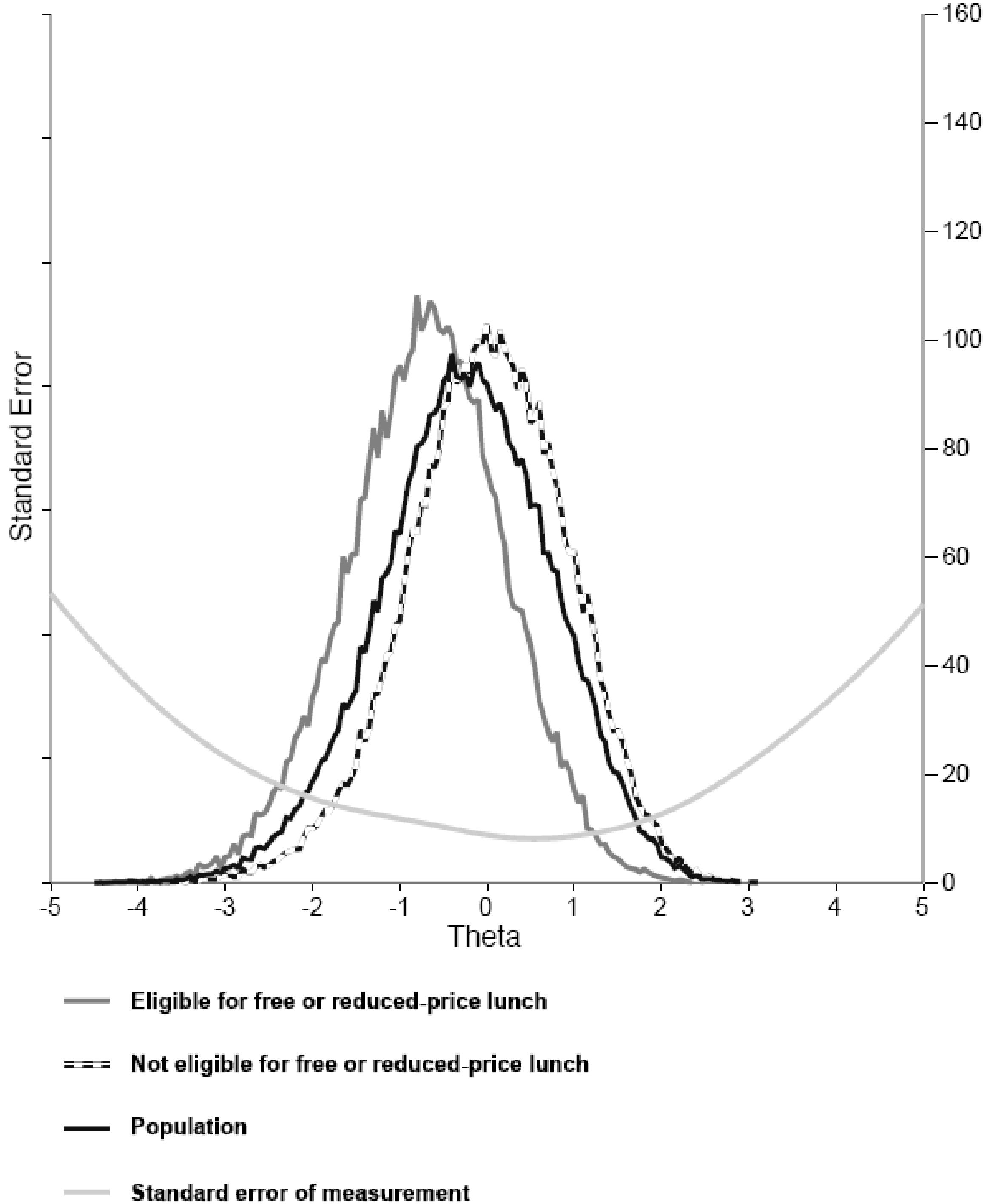
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit V-3. Grade 8 algebra subscale, 2005: Standard error of measurement and achievement distributions by race/ethnicity



Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit V-4. Grade 8 algebra subscale, 2005: Standard error of measurement and achievement distributions by eligibility for free or reduced-price lunch



Note: 2005 NAEP Mathematics Assessment, national sample

A review of the full set of plots in appendix I shows much the same story for each of the other subscales at each grade level. That is, the plots generally show that all of the subscales in the assessment have good precision over a broad range of proficiency. However, there is some variability across subscales in the minimum value of the standard error of measurement and in the width of the achievement range for which the standard error curve stays close to its minimum value.

Summary

In summary, comparisons of the standard error of measurement curves to the distributions of student achievement levels shows that the NAEP mathematics assessment is well targeted to the bulk of the distribution of student achievement at both grade levels. For most of the five subscales, and at both grade levels, the standard error of measurement is relatively low for a wide range of achievement. These findings offer positive evidence of NAEP's capacity for accurate reporting of student achievement, especially given that most NAEP reporting is based on the overall mathematics scale (a weighted average of the five subscales). The overall mathematics scale has stronger measurement properties than any one of its constituent subscales.

Nevertheless, there is room for improvement. Specifically, the measurement precision could be better for low-achieving students. This could be accomplished by adding more low-difficulty items either in the form of an easy block of items, or by sprinkling more low-difficulty items in various blocks. Even greater gains could be made in the precision of measurement by targeting easy blocks to particular groups of students who are expected to have low achievement or by the use of adaptive testing procedures.

Chapter 6. Does NAEP Provide Information That is Representative of All Students, Including Students Who are Unable to Demonstrate Their Achievements on the Standard Assessment?

As the country's most visible and long-standing national assessment, it is critical that NAEP provide information that is representative of all students. The NAEP samples and frameworks are carefully designed with that intention, and patterns of performance of students on the 2005 assessment (NCES, 2005) suggest that NAEP indeed constitutes a rigorous assessment of mathematics achievement that captures performance over much of the student spectrum.

Exhibit VI-1. Percentage of students performing at each of the achievement levels in NAEP mathematics, 2005

Grade	Below Basic	Basic	Proficient	Advanced
4	21%	44%	30%	5%
8	32%	39%	23%	6%

SOURCE: National Center for Education Statistics. *NAEP Data Explorer (main).* Retrieved July 2, 2007 from http://nces.ed.gov/nationsreportcard/nde. Data from public-school-only national sample.

On the other hand, concerns have been expressed regarding the relatively small number of items at lower difficulty levels on NAEP and about the lack of out-of-level or alternate assessment procedures. Both of these factors call into question the extent to which NAEP provides adequate information for students at the lower end of the performance continuum. Furthermore, exclusion rates for NAEP, though in decline since 1996, remain slightly higher than those seen on state assessments. As part of the larger NAEP mathematics validity study, we consider these access issues, assess the extent to which they compromised NAEP's ability to represent all students, and offer suggestions for how they might be addressed.

In our investigation, we obtained data from a variety of sources including:

1. Analysis of NAEP policies and practices designed to promote inclusion of the full range of students in the assessment and resulting participation and inclusion rates;
2. Expert review of the current NAEP item pool to assess coverage of the NAEP mathematics framework and to identify strategies for reducing construct-irrelevant variance and improving the accessibility of the assessment for all students, but particularly for those at the lower end of the performance continuum; and
3. Analysis of the standard error of measurement curves for each NAEP mathematics subscale in comparison to the achievement distributions for mandated reporting groups and for the general population.

Participation and accommodation policies and practices

As a national indicator of student performance, NAEP is committed to including all sampled students in the assessment including students with disabilities (SD) and English language learners (ELL). Consequently, NAEP has implemented a number of procedures designed to promote inclusion and valid representation of these subgroups. Data from the most recent NAEP assessments suggest that, on average, 23 percent of the sampled NAEP population is identified as SD, ELL, or both at grade 4. At grade 8, the corresponding percentage is 19 percent, and percentages for both grade levels have been increasing steadily since 1992 (NCES, 2005, 2007).

Since 1996, NAEP has developed a set of robust procedures to promote inclusion and appropriate accommodation of SD and ELL students in the assessment, and this is reflected in the fact that exclusions have not been rising with identification rates. NAEP preassessment procedures require that schools complete SD and/or ELL questionnaires for each identified SD or ELL student. These questionnaires request information about the student's participation and accommodation on the state test as well as demographic information about the student. During a preassessment visit, the NAEP assessment coordinator reviews these questionnaires and develops a plan for inclusion and accommodation of each student. In the most recent assessments, NAEP has further streamlined this process by incorporating decision trees into the SD and ELL questionnaires to guide decisions about NAEP participation and accommodation based on participation in the regular state assessment. As with other subject areas, participation of SD and ELL students in the NAEP mathematics assessment has increased steadily since 1996. In 2005, 10 percent of the assessed mathematics sample at grade 4 was SD and 8 percent was ELL, compared to 7 percent and 5 percent, respectively, in 1996. The corresponding figures for grade 8 were 10 percent and 5 percent in 2005, compared to 6 percent and 2 percent in 1996.

According to the decision trees for SD and ELL students, if a student takes the regular assessment without accommodations, that student should take NAEP without accommodations. In 2005, 43 percent of all fourth-grade students and 35 percent of eighth-grade students identified as SD and/or ELL students took NAEP without accommodations.

If a student takes the regular state assessment with accommodations and those accommodations are also allowable for NAEP, then that student participates in NAEP with accommodations. During regular administration of the mathematics assessment, NAEP allows for the following accommodations:

- Bilingual dictionary (supplied by school)
- Bilingual booklet (English/Spanish)
- Large print
- Magnification equipment
- Directions signed
- Read aloud occasional word or phrase
- Computer or typewriter to respond

- Special writing tool or template
- Extended time

NAEP also allows a set of mathematics assessment accommodations that require a separate session:

- Test items signed
- Braille version
- Read aloud most or all of test (English or Spanish)
- Respond orally to scribe
- Respond in sign language
- Small group administration
- One-on-one administration
- Breaks during testing
- Administration by school staff[29]

The array of accommodations offered in NAEP is consistent with those offered and used most frequently in state assessment. In the 2005 NAEP mathematics assessment, approximately 43 percent of SD and/or ELL students at fourth grade and 47 percent at eighth grade took NAEP with at least one accommodation.

Since NAEP does not provide an out-of-level or alternate assessment, students who take an alternate or modified regular state assessment are not currently assessed in NAEP. In the 2005 mathematics assessment, 3 percent of the entire sampled population at fourth and eighth grade was excluded. This figure approaches the 2 percent alternate assessment target for state assessment advocated under NCLB. The percent of students excluded from the 2005 NAEP mathematics assessment varied considerably across states from a high of 11 percent (DE—grade 8) to a low of 1 percent.[30] Within the SD and ELL subgroups, approximately 15 percent of all sampled SD students were excluded at fourth grade and 25 percent at eighth grade. Exclusion rates for ELL students tended to be slightly lower, at 10 percent for fourth grade and 17 percent for eighth grade.

Summary and recommendations on policies and practices

Since 1996, NAEP has developed and put in place a robust system to promote participation and valid accommodation of SD and ELL students in the assessment. No alternate assessment option exists within NAEP, and approximately 3 percent of students are excluded. Development of an alternate assessment option for NAEP could permit the inclusion of the entire sampled population in the assessment, thereby increasing representation. NCES should consider incorporating alternate assessment procedures into NAEP.

[29] The array of accommodations in other subject areas is similar except that reading aloud (other than test directions) is not allowed in reading and the Spanish/English bilingual booklet is only available in mathematics.
[30] The state figures are for public schools only. The national figures include both public and nonpublic schools.

Accessibility of NAEP mathematics items

Significant numbers of students tend to perform at the below basic level on NAEP. For example, on the 2005 mathematics assessment, 20 percent of all fourth graders and 31 percent of all eighth graders performed below the basic level. Only 36 percent of all fourth graders and 30 percent of all eighth graders performed at or above the proficient level, and very small percentages reached the advanced level at either grade. Furthermore, the percentages of students performing in the lower part of the distribution is much greater for many of the demographic groups that NAEP is required to report by law. For example, in 2005, 44 percent of public school students with disabilities scored below basic, 56 percent at or above basic, and only 16 percent at or above proficient. Performance was lower at eighth grade, with 69 percent of eighth-grade students with disabilities in public schools performing below basic, and only 7 percent at or above proficient. Patterns were similar for ELL students.

Yet, given the need for NAEP assessments to measure the full range of content and skills specified in the frameworks and achievement level descriptions with relatively few items, NAEP has 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. The tension between coverage of the content specified by the framework and accurate measurement of performance across the full continuum obviously presents a huge challenge to the assessment and its developers, particularly when respondent burden is also considered.

As discussed in chapter 4, a group of mathematicians from around the country met on February 17 and 18, 2007 and systematically reviewed all current fourth- and eighth-grade NAEP mathematics items, focusing on the mathematical accuracy of the items. Reviewers were also asked to respond to the following questions: What factors contribute to the difficulty of these items? How might these items be modified so as to maintain their mathematical accuracy, but reduce construct-irrelevant variance and increase accessibility? Reviewers' comments were synthesized into the suggestions, shown below, for creating NAEP items that are more accessible to the full range of students taking NAEP.

On February 21–24, 2007, a second review panel of mathematics curriculum experts, teachers, and mathematicians (including those with expertise with ELL and SD students) were convened to rate the extent to which the NAEP framework is accurately reflected in the NAEP item pool (see chapter 3). Complexity, as defined by the NAEP framework, was one of the dimensions evaluated. That group concluded that items of low and moderate complexity were generally well represented on NAEP across all content areas, while high-complexity items were frequently lacking.[31] This review panel shared some of the same observations as the previous panel regarding strategies for making items more accessible to all students.

[31] It should be noted that complexity and difficulty are separate constructs. It is entirely possible (although challenging) to write complex items that are not of high difficulty.

The suggestions for accessibility gleaned from both panels include the following:

Increase consistency of wording

- Item wording should provide parallel syntactic construction; e.g., wording within and between the statement of the problem and its possible answers (including distracters) should be consistent.
- Avoid wording that invites multiple interpretations.

Consider clarity and cultural appropriateness in word choice

- Clarity – Word choice throughout all items should be unambiguous and concise. For example, avoid the phrase "about how much" when writing problems that require estimation or rounding. It is generally more important for item wording to be clear than precise.
- Cultural Appropriateness – Use terminology that is current and relevant to a broad population. For example, questions that include outdated or culturally specific technology or terminology can distract from the content of a problem.
- ESL Considerations – Use common words, phrases, and terminology whenever possible. Be conscious of literal interpretations of items.

Reconsider alternative answer choices (distracters)

- Identify alternative answer choices (distracters) that are plausible, but not very reasonable. The easiest multiple-choice questions provide students with only one obvious solution.
- Provide an appropriate *number* of alternative answer choices (distracters). The number of possible answer choices provided for a given item should be determined by the content and context of the problem rather then testing convention. (Four answer choices may not be appropriate in all cases.)
- Items requiring rounding or estimation are sometimes clearer when a wide range of values is provided in the answer choices.

Simplify question format

- Use short, simple sentences whenever possible. Combining multiple ideas into one sentence or statement increases the complexity of a problem.
- Provide appropriate (often 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.
- Provide the appropriate amount of space for each constructed response answer. (The amount of space provided for an answer can falsely suggest that the item requires an answer of a certain length.)
- Visuals should be as clear and precise as possible and should be relevant to the item.

Add cues

- Provide descriptive titles or introductions for an item or set of items. This is especially helpful for presenting word problems that require multiple pieces of information.
- Provide visual cues – **Bold**, *italicize*, <u>underline</u>, or CAPITALIZE key words and phrases including:
  - Directions (e.g., <u>Solve</u>, COMPUTE, **Explain**) – Directions should always come at the beginning of a problem and be clearly denoted.
  - Operational words and phrases (e.g., **Add**, *Subtract*, <u>Find the product</u>).
- Cue students about the number and type of solution(s) they should provide (e.g., written description, graphical representation, et cetera). This is especially important in open response items that could be solved using multiple approaches.
- The objective and intent of all items should be as clear as possible to the student. Avoid deceptive cues. Do not mislead students to perform inappropriate operations.

Consider computational appropriateness

- 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.
- Do not require students to perform calculations that are unnecessarily time consuming. Calculations should not distract from the "flow" of the students' testing experience. Remember that *time* is a precious resource during the testing experience.
- Do not require students to perform counterintuitive operations.
- Do not ask students to estimate or round when exact calculation is necessary or easier.
- Calculator use – Items should be constructed with calculator use/availability in mind. More computational complexity is acceptable when calculators are allowed, but the availability of a calculator can also sometimes increase the overall complexity of a problem.

Reduce extraneous information

- Provide manipulatives only when absolutely necessary. (For example, it may or may not be appropriate to test students' ability to visualize information using manipulatives.)
- Provide students with units of measure when it is necessary or appropriate for the context of item.
- Do not ask students if they used a calculator for an item that obviously does not require its use.

Embedding these considerations into the item development process and including a level of review that seeks to reduce construct-irrelevant variance and increase accessibility of items will improve the validity of assessment results for the entire sample, but particularly for students who may be differentially affected, such as SD and ELLs.

Summary and recommendations on item pool

Expert review indicates that there are a sufficient number of items of low and moderate complexity on NAEP, but issues in wording choice and consistency, construction of distracters, item format, and other features may limit the accessibility of the items, particularly for SD and ELLs. We recommend that item development guidelines and review procedures be developed and implemented to improve item quality and reduce construct-irrelevant factors that influence performance.

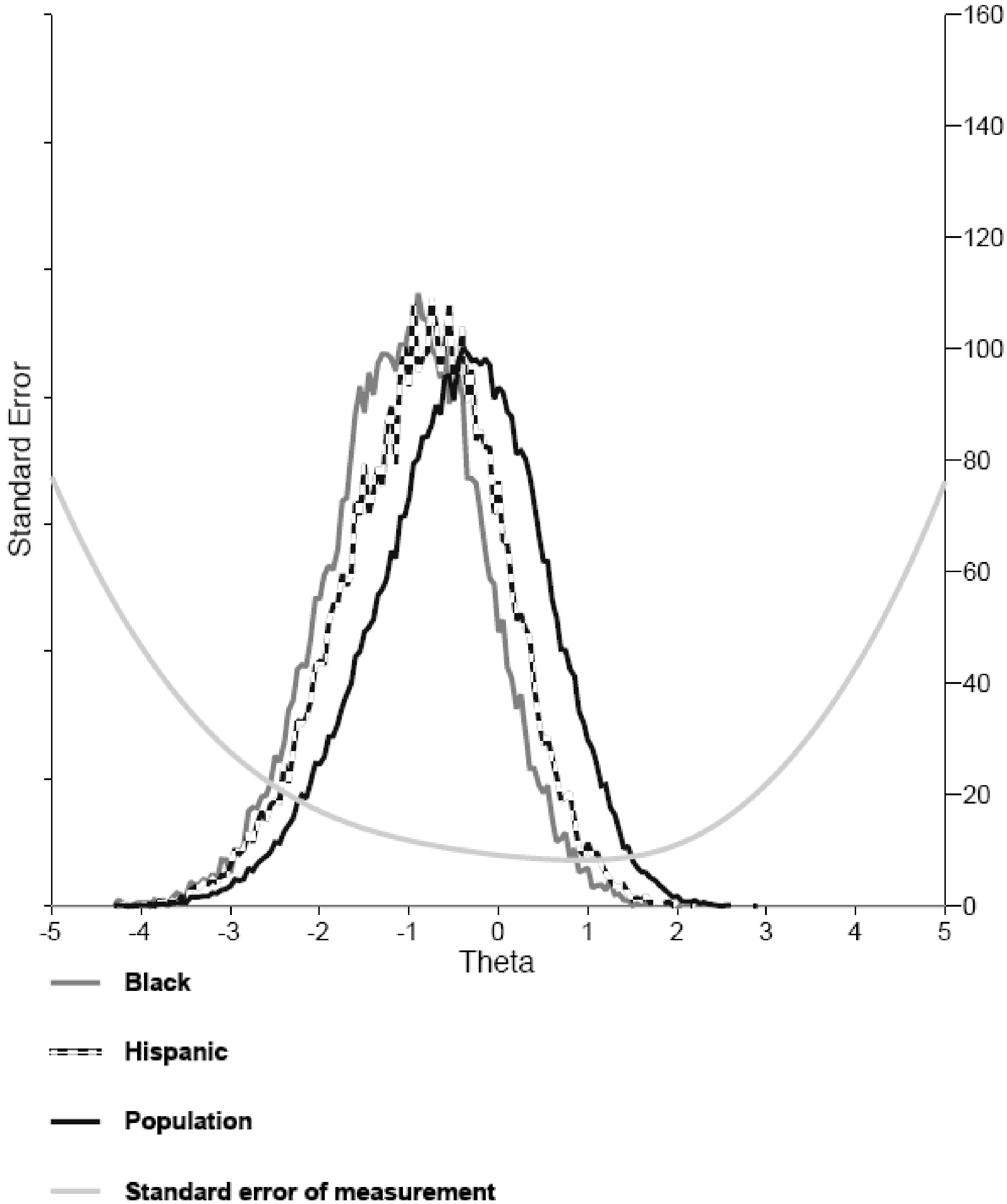
Precision of measurement across the achievement distribution

Given the need for NAEP assessments to measure the full range of content and skills specified in the frameworks and achievement-level descriptions with relatively few items, 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. This creates an important validity issue for NAEP, which is required by law to report on subgroups defined by gender, race/ethnicity, socioeconomic status, disability status, and ELL status. Progress among the lowest performing of these subgroups is often of greatest concern to policymakers, who are striving to attain high achievement for all students, yet NAEP reporting for these subgroups is hampered by decreased precision in the relevant scale range.

The figures in chapter 5 and appendix I illustrate—for each of the five NAEP mathematics subscales—the standard error of test information compared to the distribution of performance for the general population and for the reporting groups specified under NCLB (see example in exhibit VI-2, below). For SD and ELL students across each of the subscales and both fourth and eighth grade, performance falls substantially lower on the theta scale than does performance for the general population. Consequently, a significantly greater proportion of students in these subgroups performs at a theta level where the standard error is larger than for students in the middle, or at the high end of the distribution. This same pattern is seen for students eligible for free or reduced-price lunch, and black and Hispanic students. No gender discrepancies are evident.

At present, no standard exists on which to judge the significance of the discrepancy in size of standard errors, but it seems reasonable to be concerned about such a persistent and dramatic pattern that affects those groups of children around which many intervention efforts are focused. The need expressed by participants in the second expert review for more highly complex items—especially at the eighth-grade level—would likely exacerbate the problem unless NAEP were to rebalance the item pool by increasing the number of items used in the assessment.

Exhibit VI-2. Grade 4 number properties and operations subscale, 2005: Standard error of measurement and achievement distributions by race/ethnicity



Note: 2005 NAEP Mathematics Assessment, national sample

Summary and recommendations on improving precision at lower performance levels

For several years, the NVS panel and other groups have been interested in the use of an “easy block” as a means of improving measurement at the lower levels of the scale. “Easy blocks” could be incorporated into the normal spiral or they could be paired with regular blocks and given selectively to students who were previously identified as likely to benefit. (See, for example, McLaughlin et al. 2005, in which a proposal for using state assessment scores to preassign booklets is discussed.) The inclusion of an “easy booklet,” consisting of two “easy blocks,” also holds promise as a means of increasing the participation of SD and ELLs, and thereby improving the validity of NAEP as a measure of performance for those subgroups.[32] Offering an “easy booklet” option to SD and ELLs could be viewed as an accommodation aimed at improving the validity of assessment results by increasing the amount of assessment information generated, and by reducing the impact of construct-irrelevant variance (readability, language demand, visual distracters, etc.), on assessment results for the SD and ELL subgroups.

[32] In reading, for example, the level and length of the reading passages has been cited as a major barrier to the accessibility of the NAEP assessment for SD and ELL students.

Chapter 7. Conclusions and Recommendations

NAEP is unique in its purposes and circumstances. There are no direct parallels. The findings that follow rest on a foundation of comparisons to the assessment systems of states and other nations that are related but not the same as NAEP in purpose or situation. NAEP, uniquely among the comparison sets of states and other nations used in different parts of this report (see chapters 2 and 4), is low stakes at the student and school levels. It is also the only one of these assessments that uses a matrix sampling design. Because of this design, NAEP has many more items, about 180, compared to states, which typically have about 40. More items allows for better sampling of the domain of knowledge.

This report provides a great deal of detail about what could be improved in the NAEP mathematics assessment. The reader should not construe this proliferation of detail as a summative judgment against the NAEP system. The NAEP mathematics assessment has been, and remains, an important and useful tool for monitoring what U.S. children know and can do in mathematics. Importantly, the organizations that make up the NAEP system are now, and have always been, joined in a serious learning community. This study is part of the NAEP system and part of the way it learns about itself and improves.

Findings specific to each study question can be found in the relevant chapter. In this chapter are findings and recommendations that cut across study questions.

Overall findings

1. **The NAEP mathematics assessment is sufficiently robust to support the main conclusions that have been drawn about U.S. and state progress in mathematics since 1990.**

NAEP results show achievement in mathematics rising steadily over the years for all subgroups, although gaps among subgroups persist. Validity issues uncovered by this study tended to be local in nature—affecting a particular set of items on a particular subscale. It is reassuring to observe that the gains across the five NAEP subscales are reasonably parallel. That is, there is no evidence that overestimation or underestimation of gains in some one part of NAEP is driving overall trends at either grade level.

Comparing the range of NAEP items to the response distribution of the test taking population and the curricular reach of the framework

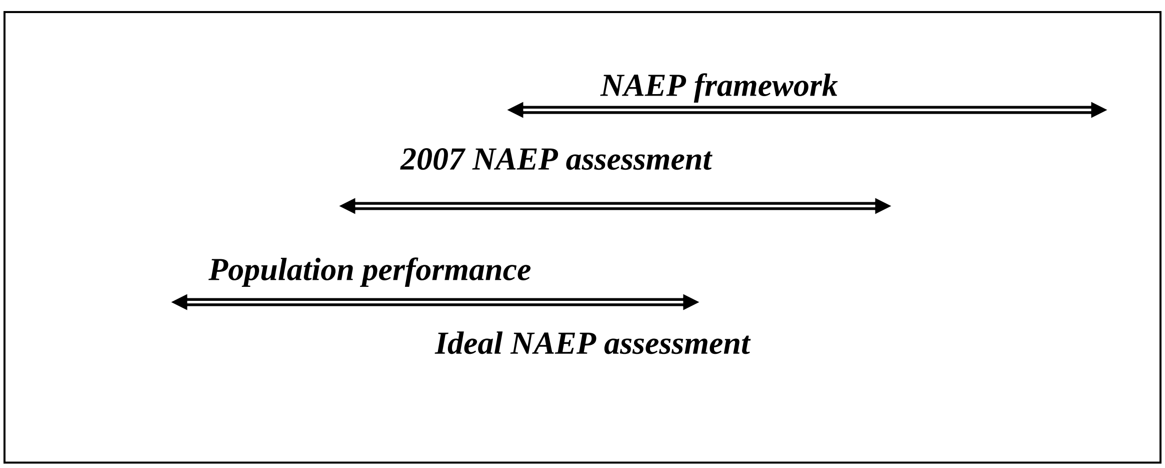
Comparison of the psychometric properties of NAEP scales to population performance shows that the regions in which the assessment measures with greatest precision are at the leading edge, if not ahead of where the population is performing. This is good from the perspective that it increases sensitivity to gains beyond the current level of achievement in the population. It is bad from the perspective that it creates a relative insensitivity to gains in the lower quartile of the population. At the same time, comparison of the NAEP item pool to the NAEP framework shows that the mathematics assessment is behind the framework in terms of capturing all of the challenging content implied by the framework.

Thus, one can say that the NAEP mathematics assessment is situated "behind" the framework but "ahead" of the population.

Given the mission of NAEP to both lead and reflect, this configuration is probably understandable. However, as an ideal, NAEP should encompass the achievement of the full population—from lowest to highest—and should reach from the least to the most advanced content of the framework's domain. Exhibit VII-1 provides a graphic representation of the current—and the ideal—relationship among the framework, the assessment instrument, and the achievement distribution of the population.

- The "NAEP framework" arrow in exhibit VII-1 refers to the range of content expectations in the framework from the least to most advanced content in the domain. This content dimension also encompasses levels of complexity as specified in the framework.
- The "Population performance" arrow refers to a scale score/item difficulty dimension that is based on item responses from the tested population.
- The "2007 NAEP assessment" arrow refers to both and represents the relationship between them as represented by the pool of items on the NAEP assessment instrument. That is, items have content referenced to the domain and difficulty referenced to the population. How closely content and difficulty correlate depends on many things, including opportunity to learn the content and the many validity issues discussed in this report.
- The "Ideal NAEP assessment" arrow shows an assessment instrument that gives good estimates of what the lowest performing students can do, even though doing so may require item content that is less advanced than the framework. It also gives good achievement estimates for the highest performing students and includes the most advanced content in the framework.

Exhibit VII-1. Schematic representation of current and ideal NAEP assessment



The offset between the framework and the 2007 assessment is partly a function of the manner in which NAEP is designed. In order to maintain the trend line, the proportion of new items in any NAEP year is limited. This "blending" method means each new assessment includes many items from old assessments. The framework has evolved incrementally as well. One result of the combination of histories (item pool and framework) is that old items from framework topics that have changed or been replaced will migrate to new or revised topics. Appropriately managing the rate of change is one of the great challenges of the NAEP program.

Even if the trend problem can be solved, the attainment of an "ideal NAEP assessment," with its very wide range of achievement levels and content, will probably require some form of adaptive testing. Adaptive testing adjusts the items administered to a student (or a school) based on available information on performance levels and/or opportunity to learn. In its simplest form, for example, eighth-grade students would take either an algebra I test or an eighth-grade mathematics test (but not both) based on whether they are enrolled in algebra I. This adapts to opportunity to learn. Another example of an adaptive NAEP would employ test booklets that combine one item block randomly representative of the whole assessment with one item block that is customized to student performance on the state test or on a special screener test. This adapts to expected achievement.

Interpreting gains since 1990

NAEP results show achievement in mathematics rising steadily over the years for all subgroups, although gaps among subgroups persist. While all evidence is that these gains are real, there are nevertheless a number of issues with the NAEP assessment that raise questions about the exact size and interpretation of the gains.

- NAEP does not have enough easy items for the bottom segment of the population. NCLB reporting mandates have focused interest on subgroups—defined by socioeconomic status, race/ethnicity, and disability or language learner status—whose performance is centered substantially lower than that of the general population. Larger standard errors in the lower part of NAEP's proficiency distribution may mean that the assessment differentially obscures gains or losses for these students.
- It is valid to infer that students who score in the proficient range on fourth-grade NAEP can handle arithmetic calculations in a range of situations: without and with calculators, and situated in word problems and applications of various kinds. However, because NAEP does not have simple arithmetic calculations on the assessment (as one would find on assessments for second- and third-grade students), it is not valid to infer that low-achieving students who score in the below basic range *cannot* perform simple arithmetic calculations. NAEP provides no direct information on this question. If some students can perform arithmetic, but cannot manage word problems very well, NAEP may be underestimating their achievement. Equally important, users of NAEP results cannot draw valid conclusions about what it means to be "Basic" or "Below Basic" with respect to basic arithmetic. NAEP calculations are pitched at a higher level, as well as being embedded in word problems.
- Unadorned arithmetic computations (performed without calculators) are not necessarily easier than the same computations situated in problem situations. The concrete elements of the problem situation can help some students think through

the solution and make sense of it. The item quality review described in chapter 4, however, found that many of the word problems on NAEP and state tests presented additional hurdles and pitfalls for the student rather than scaffolds. This effect may interact differently with students with different reading proficiencies.

- The NAEP item pool does not reach to advanced topics in the framework in some areas.[33] Because of these deficiencies, gains or losses associated with achievement on advanced topics may be underestimated.
- Similarly, the lack of high-complexity items means little is known about gains or losses in reasoning skills needed for high-complexity items.
- Overspending items in some areas of the domain may exaggerate the effects of small areas of the domain on total scale score. This could exaggerate overall gains, if there were especially large knowledge gains in those small areas, or it could exaggerate overall losses for a similar reason.
- Item writing issues may make items difficult for some students for nonmathematical reasons. These difficulties could produce false negatives and depress scores, in a given year or over several years. This study did not compare item quality across years, so variations in item quality over time are unknown. However, if the same level of item quality has persisted over time, then trends might remain the same, whether or not scores have been depressed overall.

In sum, a number of validity issues could have affected scores, some upward and some downward. Further research would be needed to estimate the size of any of these effects.

Some light can be shed on the possible effects of these issues on trends, however, by looking at the content area subscale scores over time. Exhibit VII-2 shows fourth-grade scores from 1990 to 2005. Scores on all five subscales have moved upward each year. Because all the issues cited above (except lack of high-complexity items) are concentrated in certain content areas, the strong parallels in growth across content areas provide some comfort in trusting the basic story told by NAEP scores across the years.

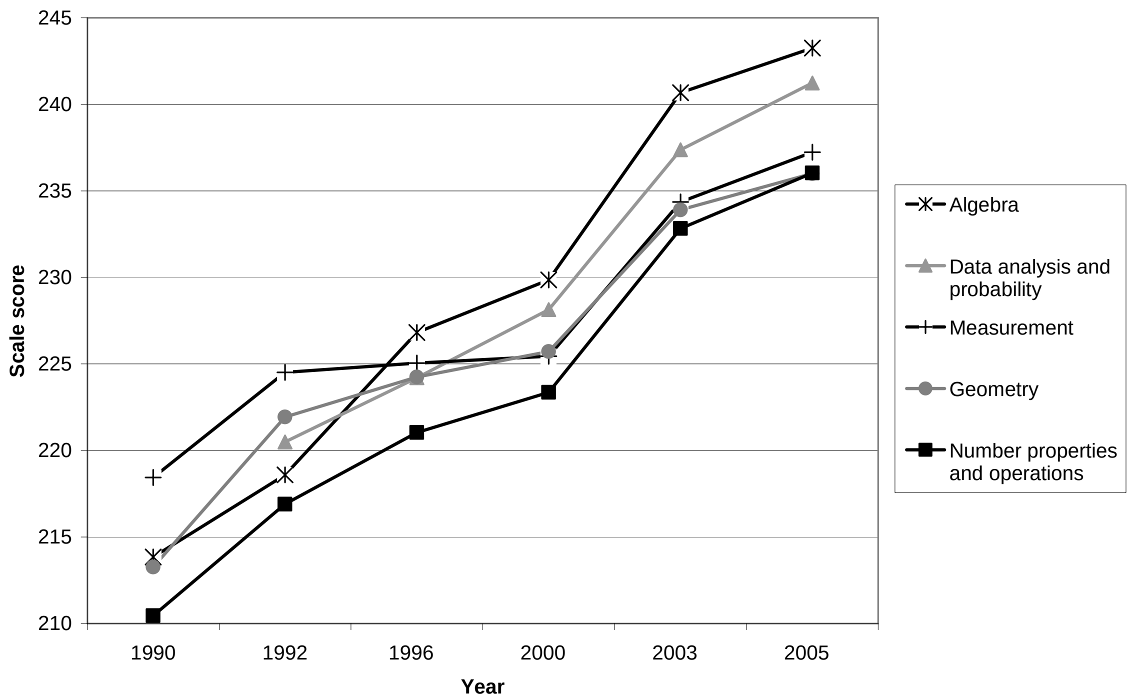
Number properties and operations, the content area with the greatest weight for fourth grade, has a solid trend line upward. Measurement makes the second-largest contribution to the fourth-grade composite mathematics score with 20 percent of the total. Although performance in this content area has also grown each year, growth has been slower. There may be a relationship between the observed low complexity of the measurement items and the flatter rate of gain. Or this pattern of growth may reflect less improvement in teaching and learning measurement over the time period.

The algebra subscale shows strong improvement. Due to the topical distribution of grade 4 algebra items, however, this also can be interpreted as gains in patterns, and does not provide a lot of information about progress on other aspects of algebra. Geometry has not gained as much as algebra. About half the expert reviewers whose work was reported in chapter 3 judged the geometry item pool out of balance, lacking in high complexity, and not well focused (e.g., too many items devoted to identifying or describing shapes and

[33] Note that advanced does not mean difficult; easy items for advanced topics are well known. For example, an easy multiplication problem is more advanced than, but easier than, a hard subtraction problem.

too many that test “just vocabulary”). As with measurement, it is not possible to tell whether the pattern of gains in geometry is related to characteristics of the geometry item pool or to the nature of geometry instruction.

Exhibit VII-2. Mathematics content area scores by year, grade 4



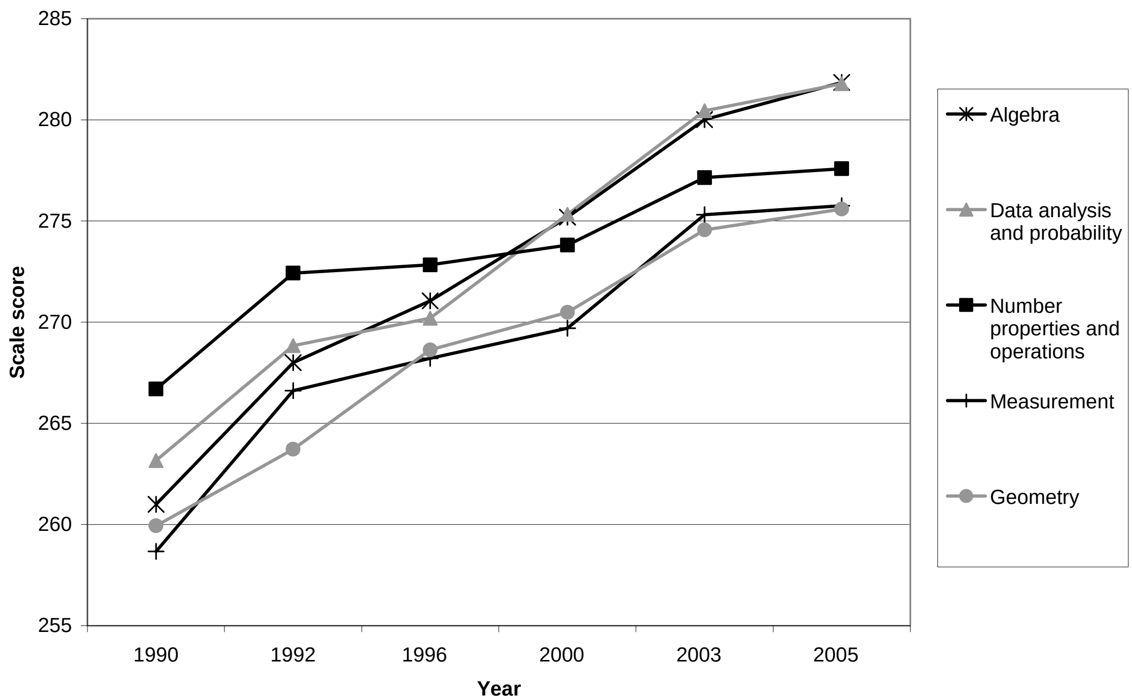
SOURCE: National Center for Education Statistics. *NAEP Data Explorer (main).* Retrieved July 2, 2007 from http://nces.ed.gov/nationsreportcard/nde. Data from national sample of public and private schools.

Exhibit VII-3 shows a similar pattern of subscore gains for grade 8. Overall, the strong parallels across content areas give support to trusting the basic story told by NAEP eighth-grade scores across years.

As at grade 4, however, it is interesting to speculate on the reasons behind the variations in observed growth. Algebra, with the greatest weight in the total score at grade 8, shows the greatest gains. This may reflect more eighth-grade students taking algebra I, as well as the influence of state standards and policies that have made the foundations of algebra a priority in upper elementary and middle grades. In grade 8, it is number properties and operations that runs flatter across years than other content areas. The number properties and operations item set stood out in the alignment review as undersampling grade-level content from the framework. It is possible that students are making gains in this content area that are not being detected by NAEP.

Data analysis and probability is the newest part of the curriculum. One might expect large gains in this content area. Indeed, the gains are almost as great as in algebra. Algebra is weighted to count as 30 percent of the total NAEP score at eighth grade, while data analysis and probability is weighted to count as 15 percent.

Exhibit VII-3. Mathematics content area scores by year, grade 8



SOURCE: National Center for Education Statistics. *NAEP Data Explorer (main).* Retrieved July 2, 2007 from http://nces.ed.gov/nationsreportcard/nde. Data from national sample of public and private schools.

2. The NAEP framework is reasonable.

The NAEP *Mathematics Framework* and accompanying specifications provide a reasonable representation of the domain of fourth- and eighth-grade mathematics when compared to states and other nations. But the framework and specifications are not uniquely reasonable. The study found differences between the NAEP framework and the standards and blueprints used by our sample of six representative states, two high-achieving nations, and two prominent policy bodies. Some of these differences are informative and should be considered in future NAEP mathematics framework updates or revisions.

For example, NAEP places considerable emphasis on measurement, and NAEP distinguishes measurement more from geometry and number properties and operations than did any of the comparison standards. At the same time, NAEP measurement includes more attention to less advanced content. Measurement should get a hard look during the next revision cycle. Other variations between NAEP and the comparison standards are described in chapter 2.

3. Guidance to test developers requires more than the NAEP framework and specifications provide.

The 2005 mathematics framework and specifications specify relative weights (in number of items) at the highest hierarchic level, the five content areas, but provide no guidance on relative priorities across or within subtopics. Moreover, the NAEP framework and specifications are not as well illustrated with exemplar items as are several of the standards in our comparison group, including some of the state standards, the Achieve expectations, and the standards of the other nations. Many uncertainties about what is meant by an objective could be cleared up by exemplar items.

Some NAEP objectives are too sweeping. The framework and specifications are both silent on some decisions that must be made by the test developer. These decisions may therefore end up being driven by the psychometric properties of items without regard for important but unarticulated content issues. For example, in number properties and operations, fourth-grade students are expected to add and subtract:

- whole numbers, or
- fractions with like denominators, or
- decimals through hundredths.

And eighth-grade students are expected to perform computations with rational numbers.

Neither here nor elsewhere does the framework (or specifications) indicate how much attention fractions, for example, should get in the mix of items. In fact, 11 percent of the items at each grade level involve fractions. Is this too few? Most of the mathematicians involved in our review process thought that this was far too few, especially at eighth grade. Other reviewers also observed that fractions were infrequent. The framework leaves one guessing as to what was intended.

4. The NAEP item pool broadly aligns with the framework with some important exceptions.

All of the items fit somewhere in the framework, and the item counts closely match the prescribed distributions for the five content areas, which—as noted—is the only level at which the framework stipulates priorities. Nevertheless, there is room for improvement. Too many items are spent on too little mathematics, while some important areas of mathematics are poorly represented in the item pool.

The details of the item pool strengths and deficiencies are presented in chapter 3. Some important examples of problem areas are listed here.

- There are too few high-complexity items at both grade levels, but especially in grade 8.
- Fourth-grade measurement overspends on low-complexity items and underspends on conceptual content and connections to other content areas.

Too many items ask what measuring instrument to use, what attribute is being measured, and other questions where the vocabulary is the biggest challenge. The large investment in measurement items is not well leveraged to include fractions or decimals used in realistic situations.

- Fourth-grade algebra overspends on recognizing and extending patterns and underspends on constructing or explaining rules, relationships between quantities, and algebraic representations—particularly conventional coordinate graphs.
- While the eighth-grade framework for number properties and operations was reviewed favorably, the item pool was judged disappointing. There were problems with focus and balance in four out of five subtopics, reach in three out of five, and balance across subtopics. Whatever aspects of number properties and operations are measured by this item pool undoubtedly correlate with what the framework intends, but collectively the item pool is missing the targets set by the framework.
- Eighth-grade algebra overspends on routine items that fail to tap conceptual understanding. The entire item pool for this content area was judged to be seriously lacking in high-complexity, or otherwise challenging, items.

5. Item quality is typical of large-scale assessments but could be better.

Overall item quality is typical of large-scale assessment and good enough to support interpretation of the overall NAEP mathematics scores, but improvements can and should be made. NAEP has item quality assurance practices that are more extensive than most other systems. However, these practices rely heavily on either external reviews by committees of stakeholders such as state representatives or Governing Board members, or internal reviews by the test developer, a very large-scale testing company. It should not surprise anyone that the result is an item pool typical of state assessments and other large-scale assessments. While reviews like those currently in place may be necessary, they may not be sufficient to raise the NAEP item pool above the bar for "typical."

Issues in item quality can diminish the validity of the test by allowing student traits other than mathematical knowledge and know-how to influence scores. For example, scores can be opened to influence by general cleverness and test taking savvy (false positives), and by language, cognitive style, stereotype threat, disposition and motivation (false negatives). Most of these influences will lead to underestimates of mathematics achievement for some or all students, although research would be needed to determine the extent of diminished performance. In NAEP and the random sample of state test items that we analyzed, as many as a third of the items were affected by item quality issues such as the following:

- Complicated presentations that require comprehension of representations not related to the framework;
- Word problems in which the stipulated situation merely decorates and confuses the mathematics rather than
  - providing scaffolding (to make the problem more accessible), or

  - demanding that the student use mathematics to make sense of the situation (formulate a mathematical representation of related quantities in the situation);
- Hidden assumptions, especially in pattern items;
- Mathematically incorrect language;
- Language or context that demands nonmathematical prior knowledge likely to be absent for too many test takers; and
- Disproportionate complications relative to the amount of framework content assessed.

6. Measurement precision is good over a broad range of proficiency but could be better for lower achieving students.

For most of the five subscales, and at both grade levels, the standard error of measurement is relatively low for a wide range of achievement. These findings offer positive evidence of NAEP's capacity for accurate reporting of student achievement, especially given that most NAEP reporting is based on the overall mathematics scale (a weighted average of the five subscales). The overall mathematics scale has stronger measurement properties than any one of its constituent subscales.

Nevertheless, there is room for improvement. Measurement precision is weakest at the bottom of the achievement scale, in a range that includes the performance of large percentages of students from groups of high policy significance. This includes students with disabilities, English language learners, students eligible for free or reduced-price lunch, as well as black and Hispanic students. The item quality review found that, although some items (including several from the measurement content area) addressed low level mathematical content, they were nevertheless embedded in complicated language and situations, which raised their overall difficulty.

Recommendations

The reader should be aware that NAEP framework and item development proceeds very methodically on schedules set years in advance. Not only is the rate of change constrained by trend considerations, but the process of change involves many officials and official groups giving advice and approval. NAEP is not just the product of artisans and expert test developers. It is also a consensual process involving many divergent perspectives. It is not a recommendation of this study to make the process more complicated.

Some of the recommendations herein are already planned for future testing cycles. No systematic attempt is made here to identify those. The Governing Board and NCES, the stewards of the NAEP program, can better speak for themselves.

As a result of the findings of this study, the following are recommended:

1. Sharpen the framework

The National Assessment Governing Board, which has legislative responsibility for specifying the assessment content, should review and sharpen the current framework. A more focused framework would form the foundation for better guidance to test developers, and it would set an example for focus that could benefit states that emulate NAEP.

A. *Focus: don't worry about leaving things out; worry about targeting the most important things.*

Reduce the number of objectives. If additional explication is absolutely necessary, subsume explication under an existing objective rather than making it another objective. For example, instead of five pattern objectives, there could be one objective with the five parts subsumed. (However, see below regarding making objectives too broad.)

At the same time, sharpen the language of the objectives to give test developers a better target rather than using language that tries to include all possibilities. Objectives are targets, not containers. Don't worry about what they include, worry about what they say about where the test should be aimed. Containers get vaguer and vaguer as they mean more and more. Targets get sharper and sharper as they define the most important aspect of a topic.

B. *Explicitly address high priority issues that cut across content areas:*

1. Specify the approximate proportion of items to be written using the various types of numbers: whole numbers, fractions, decimals, negative numbers, rates, ratios, and percents.
2. Specify the manner and extent to which straightforward content from earlier grade levels should be included in the assessment.
3. Specify high-priority connections across subtopics or content areas and use them to develop high-complexity items.
4. Specify translations across representations that deserve priority.

2. Provide detailed implementation plans

The framework is a public policy document that describes the Governing Board's vision of mathematics assessment to a broad audience. Greater specification is required for the contractors who develop assessment items under NCES' supervision.

A. *Translate the higher level guidance provided by the framework into detailed implementation plans.*

Before beginning item development, NCES should create a formal, written implementation plan for each assessment cycle that translates the higher level guidance provided by the framework. The implementation plan should be developed as quickly as possible after the framework is in place in order to maximize the time available for item writing and review.

B. Make priorities explicit.

The implementation plan should include specification of the relative priorities of the different assessment topics. However, merely allocating percentages of items to content areas is too broad. A reasonable sampling of the mathematics domain will require guidance at each hierarchic level of the framework.

Why is it necessary to set targets at each level in the framework hierarchy? At first one might think that the most specific level (objectives) can be the targets and that the higher orders: subtopics and content areas, can be taken care of through aggregating up from objectives. This view is naïve and violates the basic structure of mathematical knowledge. It is typical for a single mathematics problem to draw upon knowledge from multiple content areas, subtopics, and objectives. Even a naked long division problem requires subtraction and multiplication, not to mention rational numbers, place value, et cetera. Many important kinds of problems have multiple connections across framework categories and levels. If the assessment items were limited to problems that adhered to single objectives, the sample of problems would misrepresent the domain.

Therefore, we recommend that guidance be written allocating half the items (about 90) at the objective level. Among the remainder, half (about 45 items) should be allocated at the subtopic level so that they can span multiple objectives. Of the remaining items (about 45), more than half (about 25 items) should be allocated at the content area level so that they can span multiple subtopics, and less than half (about 20 items) should be allocated at the framework level so that they can span multiple content areas.

The foregoing recommendation might require changes in the way NAEP is scaled or the way scales are interpreted. NAEP scales are developed for each content area. Every item is assigned to one and only one content area. If some items (less than one eighth) are explicitly referenced as multi-content area, a decision has to be made as to how these items are scaled. In practice, each such item could be identified more with one content area than the others and assigned to that scale.

3. Define a larger role for exemplar items

It is time to advance the practice and technology of using exemplar items to communicate expectations. The range and number of items available from released state items, international tests, Achieve, the Dana Center, the Mathematics Diagnostic Testing Project, the Shell Centre, the Freudenthal Institute, national tests (Japan, Singapore), and other sources is now very large.

A. *Provide ample examples of items*

Both the Governing Board (in the framework) and NCES (in the implementation plan) should make generous use of example items to clarify their intent and help avoid in-breeding a house style.

NCES should, as a matter of principle, compile an eclectic file of example items from all the various sources of released items. These can be used to illustrate the expectations for individual item quality and also to clarify the desired attributes for the total *item pool.* That is, NCES should compile a coherent body of items to exemplify the intended focus, reach, and balance of the assessment.

For greatest utility, NCES should annotate the compilation:

- Select items that align with the framework and explain what is being illustrated from the framework.
- Provide some examples of common types of problems to avoid with reasons why.

B. *Encourage the establishment of a Web-based open bank of released items.*

NCES and the Governing Board should encourage the Institute of Education Sciences to support the development and ongoing maintenance of a Web-based open bank of released items. It should be operated by a third party with technical capability and it should include items from as many sources as possible, indexed to a common framework. A selection of items should be reviewed by and commented upon by mathematicians, educators, and language specialists. Comments could include suggested edits to enhance the items.

Such an item bank would both provide exemplars to support NAEP development (as described above) and also serve as an important resource for the states.

4. Improve quality assurance for the overall item pool and for individual items

Ongoing quality assurance is the particular responsibility of NCES, which has recently undertaken initiatives similar to those described below. NCES should continue and expand upon these current efforts.

A. Monitor and manage the focus, balance, and reach of the item pool across and within the subtopic level of the framework.

Once the priorities across assessment topics are clearly specified in the implementation plans, NCES should create routines that monitor the overall item pool each time item blocks are replaced. The routines should include attention to the focus, balance, and reach of the item pool across and within the subtopic level of the framework. The point of this recommendation is that the pool as a whole has to be evaluated against the framework.

B. Subject all items to expert review.

While better guidance (especially in the form of an annotated compilation of exemplar items as recommended above) will lead to better quality first drafts, review will always be essential. The compilation of exemplar items can also be useful as a tool during the review process, as can guidelines for item accessibility such as are laid out in chapter 6.

What kind of expertise is needed for expert review? Mathematicians, language experts, cognitive scientists, access specialists, and mathematics educators are all needed. They will notice different things and sometimes pull in different directions. An expert review should focus on applying individual expertise rather than reaching agreement. Once the expert critiques are documented, an independent resolution and revision process should be carried out by NCES.

Furthermore, whatever the past process has been for quality assurance, however elaborate and intense, it has been biased toward the typical. The recommendation from this study is for more expertise and design as part of the process. Consensus must underlie the initial guidance, and consensus must also be confirmed once items are on the table. But consensus is no substitute for a high level of technical or “craft” expertise. Cycles of draft, review, and revision also take time. Haste drives out quality. Although this study did not review the procedures used by NCES and the Governing Board, short timelines must be a prime suspect whenever questions of quality arise.

5. Attend particularly to the following aspects of item quality

Through the process of research and review, NCES should attend particularly to the following aspects of item quality.

A. Sustain attention to the mathematical quality of the items.

Mathematical quality requires that the mathematical content of the items be well expressed. Symbolic expressions, tables, graphs, diagrams and ordinary language should be used correctly and considerately for the age of the test takers. Mathematical quality also requires that any implicit assumptions embedded in the items are fair and do not require the student to read the mind of the test developer. Items with hidden assumptions are tests of general cleverness or cultural conditioning, not mathematics.

B. Improve the quality of the situated mathematics problems.

Setting mathematics problems in imaginary situations is a basic feature of school mathematics throughout the world and from the earliest grades. Such items can help make the mathematics more accessible, and they can also provide opportunities to assess mathematical modeling skills.

When items using problem situations are developed and reviewed, the following item quality issues should be attended to:

- The problem context should, insofar as possible be familiar to all students.
- The mathematics in the problem situation should have a purpose that will make sense to the student (authenticity).

C. Improve the measurement of mathematical complexity.

NCES should turn to nations, centers, and states that are working in different assessment traditions in order to explore divergent approaches to assessing high-complexity reasoning. Simply mounting more intense, well-meant efforts in the same tradition as NAEP has already used is not likely to produce good results. Having sampled ideas from other traditions, alternative approaches to the assessment of complexity could then be examined as part of the recommended program of evidence-based research on item design (see recommendation 6).

The NAEP definition of complexity, which is described in chapter 3, was introduced in the 2005 framework as a method for specifying items that demand different kinds and levels of reasoning with mathematics. Prior to using the "complexity" approach, NAEP relied on a more typical matrix of content by process. The process dimension had three classifications: procedural knowledge, conceptual understanding, and problem solving. Other American approaches to capturing cognitive processes in mathematics have included the five proficiencies defined by the National Research Council (2001) in *Adding It Up,* and the NCTM (2000) process standards. Many states have followed NCTM.

No one, however, has yet come upon an approach that resolves all the issues in a dependable way. Clearly, the NAEP system isn't working. High complexity is severely lacking in the NAEP item pool. This is surely due, in part, to the constraints and habits of large-scale assessment. Yet in this study, reviewers found items from the Netherlands, Singapore, Japan, the Shell Centre in England, PISA, and some states that had more satisfactory treatments and traditions of high-complexity items.

D. Minimize non-construct relevant sources of item difficulty.

Item difficulty is a combination of many factors. In addition to mathematical demands, items may embody demands on auxiliary skills (skills necessary for demonstrating competency in the domain, such as reading grade-level text) and demands that are merely contaminating (for example, deciphering complex graphical displays). Contaminating skill demands should be avoided entirely, and auxiliary skill demands should be managed so that they do not out weigh the mathematical skill demands of the items.

6. Undertake a program of evidence-based research on item design

Item development is an art and a science, but not as much a science as it could and should be. Resources for research into item performance and construction is seriously underinvested given the importance tests have assumed in the nation’s school systems. NCES should lead advances in evidence-based item design, not go along with the status quo.

Much is known about the psychometric qualities of items as they contribute information to scores constructed through IRT and related methods. Much less is known about item design, student-by-item interactions, and how items relate to the constructs of the domain being assessed (and to the irrelevant domains that contaminate assessment). The following questions could well be the focus of empirical research, and it is a recommendation from this study that NCES place research on item quality high on the nation’s educational science research agenda:

- What makes an item difficult or easy for students?
- What are the dimensions of construct irrelevant difficulty?
- What are the dimensions of nonconstruct enabling skills (e.g., reading) that are a necessary medium of learning and assessing the constructs?
- What item characteristics exacerbate student-by-item interactions with construct irrelevant challenges?
- How can the reading barrier (e.g., syntax, vocabulary, parallelisms between English expressions and mathematical expressions) be raised or lowered through item design in different parts of the domain?
- What test taker assets can substitute for learning in the domain (and thus create false positive effects on the overall score) as they relate to particular item features?
- How can high-complexity mathematics reasoning be measured on large-scale assessments such as NAEP and (also) state tests?
- How can items be designed to be easier while still focusing on the grade-level domain?

7. Expand the range of item difficulty and curricular reach

The NAEP mathematics assessment needs more easy items, more high-complexity items, and more items that reach forward in the curriculum. These recommendations seem to pull in different directions. In part, this is so; NAEP needs to provide more information about low-performing students (can they add, subtract, multiply, and divide whole numbers?) while at the same time it needs to provide more information about aspects of the framework for which there are too few items (high complexity, fractions, number properties, conceptual understanding of measurement and geometry).

It is worth clarifying the distinction between how advanced an item is in the curriculum and how difficult it is. For example, the following three problems are at the same level of advancement in the curriculum—they are taught around the same time:

a. 9 + 14 = ?

b. 9 + ? = 23

c. ? + 14 = 23

Yet *a* is much easier than *b* or *c*. The underlying concept is the same: addition of one- and two-digit whole numbers, but the challenges are different. In *a*, a student merely executes his or her procedural knowledge of addition. The calculation is set up for the student. In *b* and *c*, the student has to do some reasoning that involves the equal sign and perhaps some manipulation of the number sentence. Therefore, in *b* and *c*, a deeper, but not more advanced, knowledge of addition is assessed.

NAEP needs items of the first type to answer the question: can students perform basic calculations? NAEP also needs items of the second and third type to answer the question: do students understand the properties of number and operations?
A corresponding example from eighth grade will be familiar. Consider the relationship: Cost = price + (tax rate $*$ price). This can be assessed using any of the following problems:

a. How much did Molly pay for a $44 jacket after sales tax? The sales tax was 5%.
b. Molly paid $46.20 for a jacket after the sales tax was added. The sales tax was 5%. How much was the original price?
c. Molly paid $46.20 for a jacket after the sales tax was added. The original price was $44. What was the tax rate?

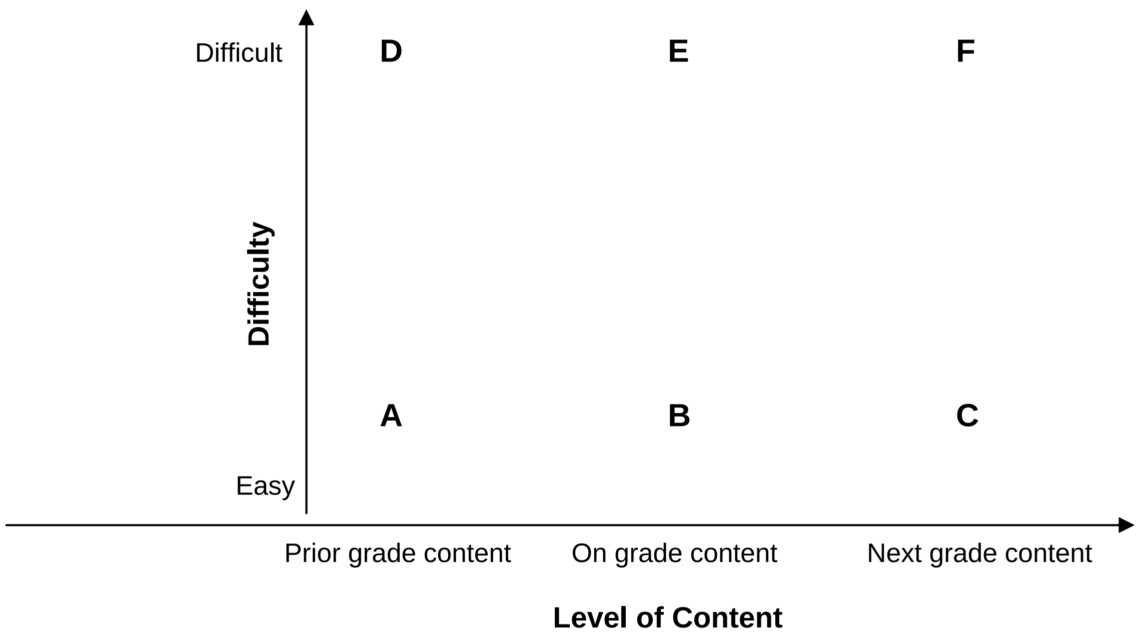
These three problems are from the same lesson or adjacent lessons in most programs. They are equally advanced. Yet *b* and *c* are much higher in difficulty than *a*. *A* can be solved by substitution; *b* and *c* require an understanding of the mathematical structure of the situation (the invariant relationship among the quantities) and the skill to manipulate the quantities (symbolically or otherwise) to find an expression for the unknown. This is a construct-relevant step up in difficulty.

NAEP should use items of type *a* to measure easy or basic understanding of the domain. But it should also use items of types *b* and *c*, which assess a more robust and flexible understanding of the domain.

In Exhibit VII- 4 below, we define an item space using the two dimensions of difficulty level and content level. Items can be high on difficulty and low on curricular demands (D), they can be high on curricular demands and low on difficulty (C), or they can be at any point in between. Ideally, NAEP would have ample items near C to provide a sensitive measure of the most advanced and recently learned content. It would also have ample items near D (assuming construct-relevant difficulty) to measure robustness and flexibility of knowledge. Of course, items near A and B are needed to capture the foci of the domain, and items near E are needed where specified in the framework. Items near F should be avoided because the cumulative cognitive load of recent learning and high difficulty can lead to erratic responses from students.

The findings of this report suggest that the NAEP mathematics assessment lacks sufficient A and B items to accurately measure achievement for many subgroups in the population. It also lacks sufficient C, D, and E items to fully reflect the framework.

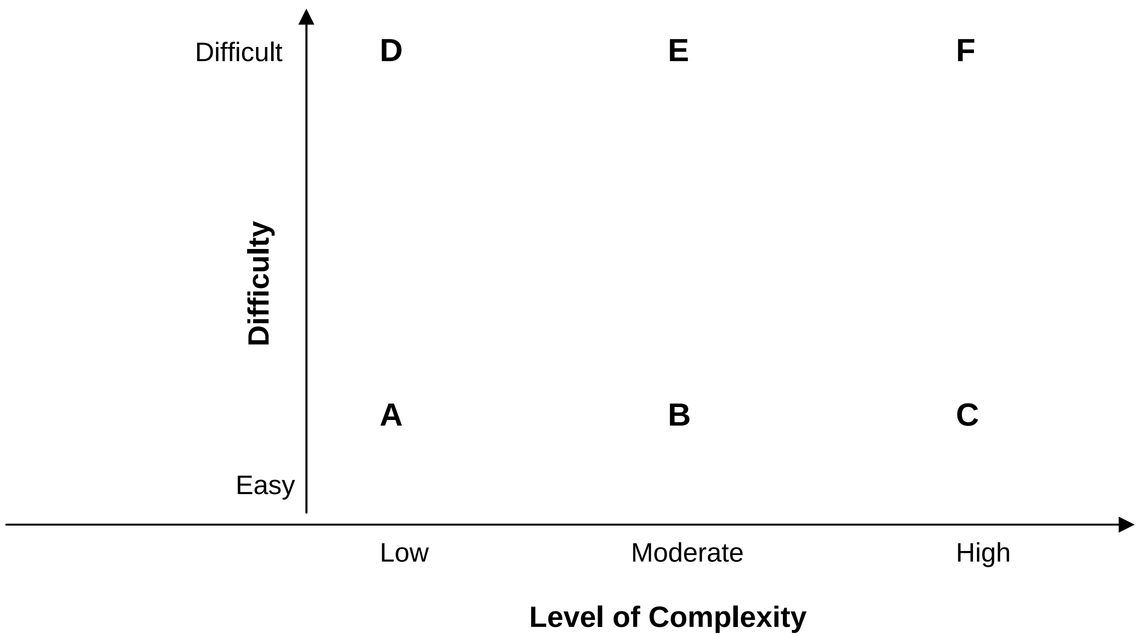
Exhibit VII-4. Difficulty by content level: theoretical distribution



A. Difficulty and complexity

Similar to the item space graphed in exhibit VII-5, an item space can be constructed to display the relationship between difficulty and complexity. In the NAEP framework, complexity is a defined construct in the domain of mathematical competency. It is *not* a synonym for difficulty. There can be items that are both easy and highly complex (C) and items that are both difficult and of low complexity (D).

Exhibit VII-5. Difficulty by complexity level: theoretical distribution



An item that requires the student to comprehend an elaborate problem situation presented as text can fit the definition of mathematical high complexity *if* the student has to formulate a mathematical model for a complex situation. But if the situation is not complex, and the student merely has to solve a word problem requiring multiple steps, albeit after a lot of reading work, then the item is of moderate complexity with high (and irrelevant) reading difficulty.

It is possible to have items near any of the letters in the item space, but more natural to find items near A, B, E, and F. One way of summarizing the findings in this report related to the distribution of complexity on NAEP is to say that NAEP has too many items near D and E, and not enough items near A, B, C and F.

B. Nonconstruct relevant sources of item difficulty

Item difficulty is a combination of many factors: some relevant and some irrelevant to the domain. Examples of appropriate nondomain (auxiliary) challenges include reading considerate text,[34] interpreting clear and grade-appropriate diagrams, comprehending grade-appropriate mathematical language, and writing general academic language. On the other hand, reading inconsiderate text or interpreting poorly constructed diagrams or tables introduces contaminating difficulties.

[34] Armbruster (1984) defines considerate text as text that is well-written, well-organized, and signals the organization of its thought to the reader.

It is also necessary to consider the proper balance between mathematical skills and auxiliary skills so that the auxiliary skill demands do not outweigh the mathematical skill demands. Consider the typical word problem in which the student must comprehend a problem situation from text before specifying, and then solving, the problem mathematically. It is not easy to write items that keep the reading load lower than the mathematics load required to specify and solve the problem.

The annotated compilation of items that is constructed to illustrate the framework should also include analyses of some of the items with respect to the auxiliary and contaminating difficulties that must be overcome for the test taker to respond effectively. Particular priority should be given to explicating auxiliary skills that, like reading, are known to vary substantially in the tested population. In the longer term, empirical studies are needed to determine the contribution of auxiliary skills to the measured mathematics achievement of the population overall and of different subpopulations. Including some items in the assessment that are designed specifically to measure auxiliary skills could generate data that would illuminate these issues. (See the memorandum by McLaughlin in appendix J.)

8. Manage changes in the item pool

NAEP must constantly balance the ability to maintain trend lines with the capacity to introduce improvements. A sustained trend line has important policy advantages, particularly given that states are required to track their progress under NCLB, and these policy considerations have been a major factor in the Governing Board’s decisions regarding the extent and timing of framework revisions. The psychometrics of trend measurement also imposes constraints on the rate of change for items in the item pool. Currently NAEP allows no more than 30 percent turnover in items between assessment cycles. Even with assessment cycles scheduled every two years, change—including change aimed at improving the fit to the framework or the quality of the items—is still very slow. NCES should further explore possibilities for accelerating change without compromising trend.

9. Move NAEP in the direction of adaptive testing

As was stated near the beginning of this chapter, the ideal NAEP assessment would encompass the full population—from lowest to highest achieving—and reach from the least to the most advanced content of the domain. However, presenting students with high proportions of items that are either too hard or too easy is both frustrating to the student and a waste of assessment time. Consequently, the Governing Board and NCES should consider the benefits of moving towards some form of adaptive testing. This could be as limited as providing an easier booklet that could be used as an accommodation but still scaled with the rest of the assessment (if it was composed of two easy blocks that were also included in the regular test spiral).

A more ambitious effort would be to adopt some form of two-stage adaptive testing in which students would be prescreened (perhaps using their state test scores) and then assigned to an appropriate test book in which at least one of the two item blocks was chosen to be easy, moderate, or challenging.

In closing, we repeat the admonition with which we began this chapter: The NAEP mathematics assessment has been, and remains, an important and useful tool for monitoring what U.S. children know and can do in mathematics. Moreover, the organizations that make up the NAEP system are joined in a serious learning community that constantly seeks to improve. We hope that the findings and recommendations in this report will contribute positively to that process.

References

Achieve, Inc. (n.d.) *MAP mathematics expectations*. Retrieved July 2, 2007 from http://www.achieve.org/files/K-8Number.2.05.pdf

Armbruster, B.B. (1984). The problem of "inconsiderate texts." In G.G. Duffy, L.R. Roehler, & J. Mason (Eds.), *Theoretical Issues in Reading Comprehension* White Plains, NY: Longman. (pp. 202-217).

Bhola, D. S., Impara, J. C., & Buckendahl, W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice. 22,* No. 3, 21-29.

California Department of Education. (2007). *Content standards.* Retrieved July 13, 2007 from http://www.cde.ca.gov/be/st/ss/index.asp

California Department of Education (2005). *California standards test, released test questions: grade 4.* Retrieved November 29, 2006 from http://www.cde.ca.gov/ta/tg/sr/documents/rtqgr4math.pdf

California Department of Education (2007). *California standards test, released test questions: grade 7.* Retrieved November 29, 2006 from http://www.cde.ca.gov/ta/tg/sr/documents/rtqgr7math.pdf

California State Board of Education (2002). *California standards test, mathematics blueprint.* Retrieved May 30, 2007 from http://www.cde.ca.gov/ta/tg/sr/documents/math1105.doc

Central Institutes for Test Development (CITO). (2006). Final primary education test. Netherlands: Author.

Chong, R., & Song, A. (2002). *PSLE Mathematics*. Singapore: SNP Panpac.

CTB/McGraw Hill (2001). *Balanced assessment in mathematics: Practice booklet 8B.* Monterey, CA: Author.

De Lange, J. (2007). Aspects of the art of assessment design. In Alan H. Schoenfeld (Ed.). *Assessing Mathematical Proficiency.* Cambridge: Cambridge University Press.

Georgia Department of Education. (2006). *6-8 mathematics Georgia performance standards.* Retrieved July 2, 2007 from http://public.doe.k12.ga.us

Georgia Department of Education, Testing Division (2007). *Content Weights for the CRCT GPS-Based CRCT.* Retrieved May 30, 2007 from http://public.doe.k12.ga.us/ci_testing.aspx?PageReq=CI_TESTING_CRCT

Indiana Department of Education. (2007). *Indiana's academic standards*. Retrieved July 13, 2007 from http://www.doe.state.in.us/standards/welcome2.html

Indiana Department of Education (2006). *Indiana statewide testing for educational progress (ISTEP+) grade 8 sampler*. Retrieved November 29, 2006 from http://www.doe.state.in.us/istep/pdf/ItemSamplers/41359-WEB_08_Sampler_02IN.pdf

Indiana Department of Education (2004). Statewide Testing for Educational Progress, ISTEP+ Academic Standards for Grades 3 and 7. Retrieved May 30, 2007 from http://www.doe.state.in.us/istep/welcome.html

International Association for the Evaluation of Educational Achievement. *TIMMS 2003 mathematics items: Released sets, 4th and 8th grade.* Retrieved July 16, 2007 from http://timss.bc.edu/timss2003i/released.html

Jones, L. V. & Olkin, I. (Eds.). (2004). *The Nation's report card: Evolution and perspectives*. Bloomington, Indiana: Phi Delta Kappa International.

Krishnan, Gopi (date) *Ace it! Math Test Papers 4*. Singapore: SNP Panpac.

Louisiana State Board of Elementary and Secondary Education (2006). *Grade 4 Practice Test.* Retrieved November 29, 2006 from http://www.doe.state.la.us/lde/saa/2032.html

Louisiana State Board of Elementary and Secondary Education (2006). *Grade 8 Practice Test.* Retrieved November 29, 2006 from http://www.doe.state.la.us/lde/uploads/8261.pdf

Massachusetts Department of Education. (2000). *Mathematics curriculum framework.* Massachusetts: Author.

Massachusetts Department of Education. (2005). *VI. Mathematics, Grade 4.* Retrieved November 29, 2006 from http://www.doe.mass.edu/mcas/2005/release/g4math.pdf

Massachusetts Department of Education (2006). *2005 MCAS technical report*. Retrieved July 15, 2007 from http://iservices.measuredprogress.org/MCAS2005TechReport.pdf

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

Ministry of Education Curriculum Planning and Development Division. (2001). *Primary mathematics syllabus*. Singapore: Author.

Ministry of Education Curriculum Planning and Development Division. (2006). *Secondary mathematics syllabuses*. Singapore: Author.

Mississippi Department of Education (2001). *Mathematics, Grade 8 Test.* Retrieved November 29, 2006 from http://www.mde.k12.ms.us/ACAD/osa/D_MCT_G8_Math.pdf

Nagasaki, E., Sawada, T., & Senuma, H. (1990). *Mathematics Program in Japan: Kindergarten to Upper Secondary School*. Japanese Society of Mathematical Education, Tokyo, Japan.

National Assessment Governing Board (2003). *2005 NAEP Mathematics Assessment and Item Specifications.* Washington, DC: Author.

National Assessment Governing Board. (2004). *Mathematics Framework for the 2005 National Assessment of Educational Progress*. Washington, DC: Author.

National Center for Education Statistics. (2007). *The Nations Report Card: U.S. History 2006*. Washington, DC: Author

National Center for Education Statistics. (2005). *The Nations Report Card: Mathematics 2005*. Washington, DC: Author.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics.* Reston, VA: Author.

National Institute for Educational Policy Research. (2006). *Summary of findings about student achievement on particular types of problems and goals in mathematics and arithmetic.* Japan: Author. Retrieved July 2, 2007 from http://www.nier.go.jp/kaihatsu/tokutei/04002030200007001.pdf

North Carolina State Board of Education, Department of Public Instruction. (2007). *North Carolina End-of-Grade Tests*. Raleigh, NC: Author. Retrieved July 2, 2007 from http://www.ncpublicschools.org/accountability/testing/eog/

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. J.Kilpatrick, J. Swafford, and B.Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Office of the Superintendent of Public Instruction, State of Washington. (2006). *Mathematics Grade Level Expectations*. Retrieved July 2, 2007 from http://www.k12.wa.us

Office of the Superintendent of Public Instruction, State of Washington (2005). *Test Specifications for the Washington Assessment of Student Learning Grade 4 Mathematics.* Retrieved May 30, 2007 from http://www.k12.wa.us/assessment/WASL/MathTestItemSpec.aspx

Ohio Department of Education (2005). *Ohio Achievement Tests, Grade 4 Mathematics Student Test Booklet.* Retrieved November 29, 2006 from http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=240&Content=16595

Organization for Economic Cooperation and Development Program for International Student Assessment (PISA). (2007). *Mathematics items 5.2.* Retrieved July 16, 2007 from http://www.peterpappas.com/blogs/pisa-blog/PISA_Math_questions.pdf

Pennsylvania Department of Education Bureau of Assessment and Accountability(2006-2007). 2006-2007 *Mathematics Item and Scoring Sampler, Grade 4.* Retrieved November 29, 2006 from http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2006-2007Gr4MathItemSampler.pdf

Porter, A.C. (2002). Measuring the content of instruction: uses in research and practice. *Educational Researcher, 31,* No. 7, 3-14.

Reys, B.J., Dingman, S., Sutter, A., & Teuscher, D. (2005). *Development of state-level mathematics curriculum documents: Report of a survey*. Columbia, MO: Center for the Study of Mathematics Curriculum, University of Missouri.

Reys, B. J., Dingman, S., Olson, T.A., Sutter, A., Teuscher, D., Chval, K.,et al. (2006). *The intended mathematics curriculum as represented in state-level curriculum standards: Consensus or confusion?* Columbia, MO: Center for the Study of Mathematics Curriculum, University of Missouri.

Rothman, R., Slattery, J.B., Vranek, J.L., & Resnick, L.B. (2002). *Benchmarking and alignment of standards and testing*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. CSE Technical Report 566.

Texas Education Agency. (2007). *Texas Essential Knowledge and Skills (TEKS), Chapter 111, Subchapter A.* Retrieved July 2, 2007 from http://www.tea.state.tx.us/teks

Texas Education Agency. (April 2006). *Texas Assessment of Knowledge and Skills Grade 8 Mathematics Online Test.* Retrieved November 29, 2006 from http://www.tea.state.tx.us/student.assessment/resources/online/2006/grade4/math/4math.htm

Texas Education Agency (2004). Texas *Assessment of Knowledge and Skills (TAKS) Information Booklet Mathematics Grade 4.* Retrieved May 30, 2007 from http://www.tea.state.tx.us/student.assessment/taks/booklets/index.html

U.S. Department of Education, National Center for Education Statistics. *NAEP Data Explorer (main).* Retrieved July 2, 2007 from http://nces.ed.gov/nationsreportcard/nde.

U.S. Department of Education, National Center for Education Statistics. (2007). National Assessment of Educational Program (NAEP) 2007 Mathematics Assessment. Washington, DC: Author.

Washington State Department of Education. (2006). Washington Assessment of Student Learning (WASL), *Mathematics Grade 8, Sample Test Booklet.* Retrieved November 29, 2006 from http://www.k12.wa.us/assessment/WASL/Mathematics/PracticeTests/WAgr8Msmpltst.pdf

Webb, N. L. (1999).*Research Monograph No. 18: Alignment of science and mathematics standards and assessments in four states*. Madison, WI: National Institute for Science Education.

Yamamoto, K. and Mazzeo, J. (1992) Item response theory scale linking in NAEP. *Journal of Educational Statistics, V* 17, 2, Special Issue: National Assessment of Educational Progress, pp. 155-173.

Appendix A:

Objectives for NAEP Mathematics Framework

A. Number Properties and Operations

1) Number sense

GRADE 4	GRADE 8	GRADE 12
a) Identify the place value and actual value of digits in whole numbers.	a) Use place value to model and describe integers and decimals.	
b) Represent numbers using models such as base 10 representations, number lines, and two-dimensional models.	b) Model or describe rational numbers or numerical relationships using number lines and diagrams.	
c) Compose or decompose whole quantities by place value (e.g., write whole numbers in expanded notation using place value: 342 = 300 + 40 +		
d) Write or rename whole numbers (e.g., 10: 5 + 5, 12 – 2, 2 x 5).	d) Write or rename rational numbers.	d) Write, rename, represent, or compare real numbers (e.g., π, $\sqrt{2}$, numerical relationships using number lines, models, or diagrams).
e) Connect model, number word, or number using various models and representations for whole numbers, fractions, and decimals.	e) Recognize, translate between, or apply multiple representations of rational numbers (fractions, decimals, and percents) in meaningful contexts.	
	f) Express or interpret numbers using scientific notation from real-life contexts.	f) Represent very large or very small numbers using scientific notation in meaningful contexts.
	g) Find or model absolute value or apply to problem situations.	g) Find or model absolute value or apply to problem situations.
		h) Interpret calculator or computer displays of numbers given in scientific notation.
	i) Order or compare rational numbers (fractions, decimals, percents, or integers) using various models and representations (e.g., number line).	
j) Order or compare whole numbers, decimals, or fractions.	j) Order or compare rational numbers including very large and small integers, and decimals and fractions close to zero.	j) Order or compare real numbers, including very large or small real numbers.

A. Number Properties and Operations (continued)

2) Estimation

GRADE 4	GRADE 8	GRADE 12
a) Use benchmarks (well-known numbers used as meaningful points for comparison) for whole numbers, decimals, or fractions in contexts (e.g., ½ and .5 may be used as benchmarks for fractions and decimals between 0 and 1.00).	a) Establish or apply benchmarks for rational numbers and common irrational numbers (e.g., π) in contexts.	a) Establish or apply benchmarks for real numbers in contexts.
b) Make estimates appropriate to a given situation with whole numbers, fractions, or decimals by: • knowing when to estimate, • selecting the appropriate type of estimate, including overestimate, underestimate, and range of estimate, or • selecting the appropriate method of estimation (e.g., rounding).	b) Make estimates appropriate to a given situation by: • identifying when estimation is appropriate, • determining the level of accuracy needed, • selecting the appropriate method of estimation, or • analyzing the effect of an estimation method on the accuracy of results.	b) Make estimates of very large or very small numbers appropriate to a given situation by: • identifying when estimation is appropriate or not, • determining the level of accuracy needed, • selecting the appropriate method of estimation, or • analyzing the effect of an estimation method on the accuracy of results.
c) Verify solutions or determine the reasonableness of results in meaningful contexts.	c) Verify solutions or determine the reasonableness of results in a variety of situations including calculator and computer results.	c) Verify solutions or determine the reasonableness of results in a variety of situations including scientific notation, calculator, and computer results.
	d) Estimate square or cube roots of numbers less than 1,000 between two whole numbers.	d) Estimate square or cube roots of numbers less than 1,000 between two whole numbers.

A. Number Properties and Operations (continued)

3) Number operations

GRADE 4	GRADE 8	GRADE 12
a) Add and subtract: • whole numbers, or • fractions with like denominators, or • decimals through hundredths.	a) Perform computations with rational numbers.	a) Perform computations with real numbers including common irrational numbers or the absolute value of numbers.
b) Multiply whole numbers: • no larger than two-digit by two-digit with paper and pencil computation, or • larger numbers with use of calculator.		
c) Divide whole numbers: • up to three-digits by one-digit with paper and pencil computation, or • up to five-digits by two-digits with use of calculator.		
d) Describe the effect of operations on size (whole numbers).	d) Describe the effect of multiplying and dividing by numbers including the effect of multiplying or dividing a rational number by: • zero, or • a number less than zero, or • a number between zero and one, • one, or • a number greater than one.	d) Describe the effect of multiplying and dividing by numbers including the effect of multiplying or dividing a real number by: • zero, or • a number less than zero, or • a number between zero and one, or • one, or • a number greater than one.
	e) Provide a mathematical argument to explain operations with two or more fractions.	
f) Interpret whole number operations and the relationships between them.	f) Interpret rational number operations and the relationships between them.	
g) Solve application problems involving numbers and operations.	g) Solve application problems involving rational numbers and operations using exact answers or estimates as appropriate.	g) Solve application problems involving numbers, including rational and common irrationals, using exact answers or estimates as appropriate.

A. Number Properties and Operations (continued)

4) Ratios and proportional reasoning

GRADE 4	GRADE 8	GRADE 12
a) Use simple ratios to describe problem situations.	a) Use ratios to describe problem situations.	
	b) Use fractions to represent and express ratios and proportions.	b) Use proportions to model problems.
	c) Use proportional reasoning to model and solve problems (including rates and scaling).	c) Use proportional reasoning to solve problems (including rates).
	d) Solve problems involving percentages (including percent increase and decrease, interest rates, tax, discount, tips, or part/whole relationships).	d) Solve problems involving percentages (including percent increase and decrease, interest rates, tax, discount, tips, or part/whole relationships).

5) Properties of number and operations

GRADE 4	GRADE 8	GRADE 12
a) Identify odd and even numbers.	a) Describe odd and even integers and how they behave under different operations.	
b) Identify factors of whole numbers.	b) Recognize, find, or use factors, multiples, or prime factorization.	b) Solve problems involving factors, multiples, or prime factorization.
	c) Recognize or use prime and composite numbers to solve problems.	c) Use prime or composite numbers to solve problems.
	d) Use divisibility or remainders in problem settings.	d) Use divisibility or remainders in problem settings.
e) Apply basic properties of operations.	e) Apply basic properties of operations.	e) Apply basic properties of operations.
f) Explain or justify a mathematical concept or relationship (e.g., ex-plain why 15 is an odd number or why 7–3 is not the same as 3–7).	f) Explain or justify a mathematical concept or relationship (e.g., explain why 17 is prime).	f) Provide a mathematical argument about a numerical property or relationship.

B. Measurement

1) Measuring physical attributes

GRADE 4	GRADE 8	GRADE 12
a) Identify the attribute that is appropriate to measure in a given situation.		
b) Compare objects with respect to a given attribute, such as length, area, volume, time, or temperature.	b) Compare objects with respect to length, area, volume, angle measurement, weight, or mass.	
c) Estimate the size of an object with respect to a given measurement attribute (e.g., length, perimeter, or area using a grid).	c) Estimate the size of an object with respect to a given measurement attribute (e.g., area).	c) Estimate or compare perimeters or areas of two-dimensional geometric figures.
		d) Estimate or compare volume or surface area of three dimensional figures.
		e) Solve problems involving the coordinate plane such as the distance between two points, the midpoint of a segment, or slopes of perpendicular or parallel lines.
		f) Solve problems of angle measure, including those involving triangles or other polygons or parallel lines cut by a transversal.
g) Select or use appropriate measurement instruments such as ruler, meter stick, clock, thermometer, or other scaled instruments.	g) Select or use appropriate Measurement instrument to determine or create a given length, area, volume, angle, weight, or mass.	
h) Solve problems involving perimeter of plane figures.	h) Solve mathematical or real-world problems involving perimeter or area of plane figures such as triangles, rectangles, circles, or composite figures.	h) Solve mathematical or real world problems involving perimeter or area of plane figures such as polygons, circles, or composite figures.
i) Solve problems involving area of squares and rectangles.		
	j) Solve problems involving volume or surface area of rectangular solids, cylinders, prisms, or composite shapes.	j) Solve problems involving volume or surface area of rectangular solids, cylinders, cones, pyramids, prisms, spheres, or composite shapes.
	k) Solve problems involving indirect measurement such as finding the height of a building by comparing its shadow with the height and shadow of a known object.	k) Solve problems involving indirect measurement such as finding the height of a building by finding the distance to the base of the building and the angle of elevation to the top.

B. Measurement (continued)

1) Measuring physical attributes (continued)

GRADE 4	GRADE 8	GRADE 12
	l) Solve problems involving rates such as speed or population density.	l) Solve problems involving rates such as speed, density, population density, or flow rates.
		m) Use trigonometric relations in right triangles to solve problems.

2) Systems of measurement

GRADE 4	GRADE 8	GRADE 12
a) Select or use appropriate type of unit for the attribute being measured such as length, time, or temperature.	a) Select or use appropriate type of unit for the attribute being measured such as length, area, angle, time, or volume.	a) Select or use appropriate type of unit for the attribute being measured such as volume or surface area.
b) Solve problems involving conversions within the same measurement system such as conversions involving inches and feet or hours and minutes.	b) Solve problems involving conversions within the same measurement system such as conversions involving square inches and square feet.	b) Solve problems involving conversions within or between measurement systems, given the relationship between the units.
	c) Estimate the measure of an object in one system given the measure of that object in another system and the approximate conversion factor. For example: • Distance conversion: 1 kilometer is approximately 5/8 of a mile. • Money conversion: US dollar is approximately 1.5 Canadian dollars. • Temperature conversion: Fahrenheit to Celsius	
d) Determine appropriate size of unit of measurement in problem situation involving such attributes as length, time, capacity, or weight.	d) Determine appropriate size of unit of measurement in problem situation involving such attributes as length, area, or volume.	
e) Determine situations in which a highly accurate measurement is important.	e) Determine appropriate accuracy of measurement in problem situations (e.g., the accuracy of each of several lengths needed to obtain a specified accuracy of a total length) and find the measure to that degree of accuracy.	e) Determine appropriate accuracy of measurement in problem situations (e.g., the accuracy of measurement of the dimensions to obtain a specified accuracy of area) and find the measure to that degree of accuracy.
	f) Construct or solve problems (e.g., floor area of a room) involving scale drawings.	f) Construct or solve problems (e.g., number of rolls needed for insulating a house) involving scale drawings.
		g) Compare lengths, areas, or volumes of similar figures using proportions.

C. Geometry

1) Dimension and shape

GRADE 4	GRADE 8	GRADE 12
a) Explore properties of paths between points.	a) Draw or describe a path of shortest length between points to solve problems in context.	
b) Identify or describe (informally) real-world objects using simple plane figures (e.g., triangles, rectangles, squares, and circles) and simple solid figures (e.g., cubes, spheres, and cylinders).	b) Identify a geometric object given a written description of its properties.	b) Use two-dimensional representations of three-dimensional objects to visualize and solve problems involving surface area and volume.
c) Identify or draw angles and other geometric figures in the plane.	c) Identify, define, or describe geometric shapes in the plane and in three-dimensional space given a visual representation.	c) Give precise mathematical descriptions or definitions of geometric shapes in the plane and in three-dimensional space.
	d) Draw or sketch from a written description polygons, circles, or semicircles.	d) Draw or sketch from a written description plane figures (e.g., isosceles triangles, regular polygons, curved figures) and planar images of three-dimensional figures (e.g., polyhedra, spheres, and hemispheres).
	e) Represent or describe a three-dimensional situation in a two-dimensional drawing from different views.	e) Describe or analyze properties of spheres and hemispheres.
f) Describe attributes of two- and three-dimensional shapes.	f) Demonstrate an understanding about the two- and three dimensional shapes in our world through identifying, drawing, modeling, building, or taking apart.	

C. Geometry (continued)

2) Transformation of shapes and preservation of properties		
GRADE 4	**GRADE 8**	**GRADE 12**
a) Identify whether a figure is symmetrical, or draw lines of symmetry.	a) Identify lines of symmetry in plane figures or recognize and classify types of symmetries of plane figures.	a) Recognize or identify types of symmetries (e.g., point, line, rotational, self-congruences) of two and three-dimensional figures.
		b) Give or recognize the precise mathematical relationship (e.g., congruence, similarity, orientation) between a figure and its image under a transformation.
c) Identify the images resulting from flips (reflections), slides (translations), or turns (rotations).	c) Recognize or informally describe the effect of a transformation on two-dimensional geometric shapes (reflections across lines of symmetry, rotations, translations, magnifications, and contractions).	c) Perform or describe the effect of a single transformation on two- and three-dimensional geometric shapes (reflections across lines of symmetry, rotations, translations, and dilations).
d) Recognize which attributes (such as shape and area) change or don't change when plane figures are cut up or rearranged.	d) Predict results of combining, sub-dividing, and changing shapes of plane figures and solids (e.g., paper folding, tiling, and cutting up and rearranging pieces).	d) Describe the final outcome of successive transformations.
e) Match or draw congruent figures in a given collection.	d) Justify relationships of congruence and similarity, and apply these relationships using scaling and proportional reasoning.	e) Justify relationships of congruence and similarity, and apply these relationships using scaling and proportional reasoning.
	f) For similar figures, identify and use the relationships of conservation of angle and of proportionality of side length and perimeter.	

C. Geometry (continued)

3) Relationships between geometric figures		
GRADE 4	**GRADE 8**	**GRADE 12**
a) Analyze or describe patterns of geometric figures by increasing number of sides, changing size or orientation (e.g., polygons with more and more sides).		
b) Assemble simple plane shapes to construct a given shape.	b) Apply geometric properties and relationships in solving simple problems in two and three dimensions.	b) Apply geometric properties and relationships in solving multi-step problems in two and three dimensions (including rigid and non-rigid figures).
c) Recognize two-dimensional faces of three-dimensional shapes.	c) Represent problem situations with simple geometric models to solve mathematical or real-world problems.	c) Represent problem situations with geometric models to solve mathematical or real-world problems.
	d) Use the Pythagorean theorem to solve problems.	d) Use the Pythagorean theorem to solve problems in two- or three-dimensional situations.
		e) Describe and analyze properties of circles (e.g., perpendicularity of tangent and radius, angle inscribed in a semicircle).
f) Describe and compare properties of simple and compound figures composed of triangles, squares, and rectangles.	f) Describe or analyze simple properties of, or relationships between, triangles, quadrilaterals, and other polygonal plane figures.	f) Analyze properties or relationships of triangles, quadrilaterals, and other polygonal plane figures.
	g) Describe or analyze properties and relationships of parallel or intersecting lines.	g) Describe or analyze properties and relationships of parallel, perpendicular, or intersecting lines, including the angle relationships that arise in these cases.

C. Geometry (continued)

4) Position and direction

GRADE 4	GRADE 8	GRADE 12
a) Describe relative positions of points and lines using the geometric ideas of parallelism or perpendicularity.	a) Describe relative positions of points and lines using the geometric ideas of midpoint, points on common line through a common point, parallelism, or perpendicularity.	
	b) Describe the intersection of two or more geometric figures in the plane (e.g., intersection of a circle and a line).	b) Describe the intersections of lines in the plane and in space, intersections of a line and a plane, or of two planes in space.
	c) Visualize or describe the cross section of a solid.	c) Describe or identify conic sections and other cross sections of solids.
d) Construct geometric figures with vertices at points on a coordinate grid.	d) Represent geometric figures using rectangular coordinates on a plane.	d) Represent two-dimensional figures algebraically using coordinates and/or equations.
		e) Use vectors to represent velocity and direction.

5) Mathematical reasoning

GRADE 4	GRADE 8	GRADE 12
a) Distinguish which objects in a collection satisfy a given geometric definition and explain choices.	a) Make and test a geometric conjecture about regular polygons.	a) Make, test, and validate geometric conjectures using a variety of methods including deductive reasoning and counterexamples.

D. Data Analysis and Probability

1) Data representation

GRADE 4	GRADE 8	GRADE 12
The following representations of data are indicated for each grade level. Objectives in which only a subset of these representations is applicable are indicated in the parenthesis associated with the objective.		
Pictographs, bar graphs, circle graphs, line graphs, line plots, tables, and tallies.	**Histograms, line graphs, scatter-plots, box plots, circle graphs, stem and leaf plots, frequency distributions, tables, and bar graphs.**	**Histograms, line graphs, scatter-plots, box plots, circle graphs, stem and leaf plots, frequency distributions, and tables.**
a) Read or interpret a single set of data.	a) Read or interpret data, including interpolating or extrapolating from data.	a) Read or interpret data, including interpolating or extrapolating from data.
b) For a given set of data, complete a graph (limits of time make it difficult to construct graphs completely).	b) For a given set of data, complete a graph and then solve a problem using the data in the graph (histograms, line graphs, scatterplots, circle graphs, and bar graphs).	b) For a given set of data, complete a graph and then solve a problem using the data in the graph (histograms, scatterplots, line graphs).
c) Solve problems by estimating and computing within a single set of data.	c) Solve problems by estimating and computing with data from a single set or across sets of data.	c) Solve problems by estimating and computing with univariate or bivariate data (including scatterplots and two-way tables).
	d) Given a graph or a set of data, determine whether information is represented effectively and appropriately (histograms, line graphs, scatterplots, circle graphs, and bar graphs).	d) Given a graph or a set of data, determine whether information is represented effectively and appropriately (bar graphs, box plots, histograms, scatterplots, line graphs).
	e) Compare and contrast the effectiveness of different representations of the same data.	e) Compare and contrast the effectiveness of different representations of the same data.

D. Data Analysis and Probability (continued)

2) Characteristics of data sets

GRADE 4	GRADE 8	GRADE 12
	a) Calculate, use, or interpret mean, median, mode, or range.	a) Calculate, interpret, or use mean, median, mode, range, interquartile range, or standard deviation.
b) Given a set of data or a graph, describe the distribution of the data using median, range, or mode.	b) Describe how mean, median, mode, range, or interquartile ranges relate to the shape of the distribution.	b) Recognize how linear transformations of one-variable data affect mean, median, mode, and range (e.g., effect on the mean by adding a constant to each data point).
	c) Identify outliers and determine their effect on mean, median, mode, or range.	c) Determine the effect of outliers on mean, median, mode, range, interquartile range, or standard deviation.
d) Compare two sets of related data.	d) Using appropriate statistical measures, compare two or more data sets describing the same characteristic for two different populations or subsets of the same population.	d) Compare two or more data sets using mean, median, mode, range, interquartile range, or standard deviation describing the same characteristic for two different populations or subsets of the same population.
	e) Visually choose the line that best fits given a scatterplot and informally explain the meaning of the line. Use the line to make predictions.	e) Given a set of data or a scatter plot, visually choose the line of best fit and explain the meaning of the line. Use the line to make predictions.
		f) Use or interpret a normal distribution as a mathematical model appropriate for summarizing certain sets of data.
		g) Given a scatterplot, make decisions or predictions involving a line or curve of best fit.
		h) Given a scatterplot, estimate the correlation coefficient (e.g., Given a scatterplot, is the correlation closer to 0, .5, or 1.0? Is it a positive or negative correlation?).

3) Experiments and samples

GRADE 4	GRADE 8	GRADE 12
	a) Given a sample, identify possible sources of bias in sampling.	a) Identify possible sources of bias in data collection methods and describe how such bias can be controlled and reduced.
	b) Distinguish between a random and nonrandom sample.	b) Recognize and describe a method to select a simple random sample.
		c) Make inferences from sample results.
	d) Evaluate the design of an experiment.	d) Identify or evaluate the characteristics of a good survey or of a well-designed experiment.

D. Data Analysis and Probability (continued)

4) Probability		
GRADE 4	**GRADE 8**	**GRADE 12**
a) Use informal probabilistic thinking to describe chance events (i.e., likely and unlikely, certain and impossible).	a) Analyze a situation that involves probability of an independent event.	a) Analyze a situation that involves probability of independent or dependent events.
b) Determine a simple probability from a context that includes a picture.	b) Determine the theoretical probability of simple and compound events in familiar contexts.	b) Determine the theoretical probability of simple and compound events in familiar or unfamiliar
	c) Estimate the probability of simple and compound events through experimentation or simulation.	c) Given the results of an experiment or simulation, estimate the probability of simple or compound events in familiar or unfamiliar contexts.
	d) Use theoretical probability to evaluate or predict experimental outcomes.	d) Use theoretical probability to evaluate or predict experimental outcomes.
e) List all possible outcomes of a given situation or event.	e) Determine the sample space for a given situation.	e) Determine the number of ways an event can occur using tree diagrams, formulas for combinations and permutations, or other counting techniques.
	f) Use a sample space to determine the probability of the possible outcomes of an event.	f) Determine the probability of the possible outcomes of an event.
g) Represent the probability of a given outcome using a picture or other graphic.	g) Represent probability of a given outcome using fractions, decimals, and percents.	
	h) Determine the probability of in-dependent and dependent events. (Dependent events should be limited to linear functions with a small sample size.)	h) Determine the probability of in dependent and dependent events.
		i) Determine conditional probability using two-way tables.
	j) Interpret probabilities within a given context.	j) Interpret probabilities within a given context.

E. Algebra

1) Patterns, relations, and functions

GRADE 4	GRADE 8	GRADE 12
a) Recognize, describe, or extend numerical patterns.	a) Recognize, describe, or extend numerical and geometric patterns using tables, graphs, words, or symbols.	a) Recognize, describe, or extend arithmetic, geometric progressions, or patterns using words or symbols.
b) Given a pattern or sequence, construct or explain a rule that can generate the terms of the pattern or sequence.	b) Generalize a pattern appearing in a numerical sequence or table or graph using words or symbols.	b) Express the function in general terms (either recursively or explicitly), given a table, verbal description, or some terms of a sequence.
c) Given a description, extend or find a missing term in a pattern or sequence.	c) Analyze or create patterns, sequences, or linear functions given a rule.	
d) Create a different representation of a pattern or sequence given a verbal description.		
e) Recognize or describe a relation-ship in which quantities change proportionally.	e) Identify functions as linear or nonlinear or contrast distinguishing properties of functions from tables, graphs, or equations.	e) Identify or analyze distinguishing properties of linear, quadratic, inverse ($y = k/x$) or exponential functions from tables, graphs, or equations.
	f) Interpret the meaning of slope or intercepts in linear functions.	
		g) Determine the domain and range of functions given various contexts.
		h) Recognize and analyze the general forms of linear, quadratic, inverse, or exponential functions (e.g., in $y = ax + b$, recognize the roles of a and b).
		i) Express linear and exponential functions in recursive and explicit form given a table or verbal description.

E. Algebra (continued)

2) Algebraic representations

GRADE 4	GRADE 8	GRADE 12
a) Translate between the different forms of representations (symbolic, numerical, verbal, or pictorial) of whole number relationships (such as from a written description to an equation or from a function table to a written description).	a) Translate between different representations of linear expressions using symbols, graphs, tables, diagrams, or written descriptions.	a) Translate between different representations of algebraic expressions using symbols, graphs, tables, diagrams, or written descriptions.
	b) Analyze or interpret linear relationships expressed in symbols, graphs, tables, diagrams, or written descriptions.	b) Analyze or interpret relationships expressed in symbols, graphs, tables, diagrams, or written descriptions.
b) Graph or interpret points with whole number or letter coordinates on grids or in the first quadrant of the coordinate plane.	c) Graph or interpret points that are represented by ordered pairs of numbers on a rectangular coordinate system.	c) Graph or interpret points that are represented by one or more ordered pairs of numbers on a rectangular coordinate system.
	d) Solve problems involving coordinate pairs on the rectangular coordinate system.	d) Perform or interpret transformations on the graphs of linear and quadratic functions.
c) Verify a conclusion using algebraic properties.	e) Make, validate, and justify conclusions and generalizations about linear relationships.	e) Use algebraic properties to develop a valid mathematical argument.
		f) Use an algebraic model of a situation to make inferences or predictions.
	g) Identify or represent functional relationships in meaningful contexts including proportional, linear, and common nonlinear (e.g., compound interest, bacterial growth) in tables, graphs, words, or symbols.	g) Given a real-world situation, determine if a linear, quadratic, inverse, or exponential function fits the situation (e.g., half-life bacterial growth).
		h) Solve problems involving exponential growth and decay.

E. Algebra (continued)

3) Variables, expressions, and operations		
GRADE 4	**GRADE 8**	**GRADE 12**
d) Use letters and symbols to represent an unknown quantity in a simple mathematical expression.		
b) Express simple mathematical relationships using number sentences.	b) Write algebraic expressions, equations, or inequalities to represent a situation.	b) Write algebraic expressions, equations, or inequalities to represent a situation.
	c) Perform basic operations, using appropriate tools, on linear algebraic expressions (including grouping and order of multiple operations involving basic operations, exponents, roots, simplifying, and expanding).	c) Perform basic operations, using appropriate tools, on algebraic expressions (including grouping and order of multiple operations involving basic operations, exponents, roots, simplifying, and expanding).
		d) Write equivalent forms of algebraic expressions, equations, or inequalities to represent and explain mathematical relationships.

E. Algebra (continued)

4) Equations and inequalities

GRADE 4	GRADE 8	GRADE 12
a) Find the value of the unknown in a whole number sentence.	a) Solve linear equations or inequalities (e.g., $aX + b = c$ or $aX + b = cX + d$ or $aX + b > c$).	a) Solve linear, rational, or quadratic equations or inequalities.
	b) Interpret "=" as an equivalence between two expressions and use this interpretation to solve problems.	
	c) Analyze situations or solve problems using linear equations and inequalities with rational coefficients symbolically or graphically (e.g., $aX + b = c$ or $aX + b = cX + d$).	c) Analyze situations or solve problems using linear or quadratic equations or inequalities symbolically or graphically.
	d) Interpret relationships between symbolic linear expressions and graphs of lines by identifying and computing slope and intercepts (e.g., know in $Y = ax + b$, that a is the rate of change and b is the vertical intercept of the graph).	d) Recognize the relationship between the solution of a system of linear equations and its graph.
	e) Use and evaluate common formulas [e.g., relationship between a circle's circumference and diameter ($C = \pi d$), distance and time under constant speed].	e) Solve problems involving more advanced formulas [e.g., the volumes and surface areas of three dimensional solids; or such formulas as: $A = P(1 + r)^t$, $A = Pe^{rt}$].
		f) Given a familiar formula, solve for one of the variables.
		g) Solve or interpret systems of equations or inequalities.

Appendix B:

Framework Comparison Committee

Susan Hull
Director of Mathematics
The Charles Dana Center

Lucy H. Michal
Director of Mathematics and Science
El Paso Collaborative for Academic Excellence

Ann Shannon
Independent consultant
Oakland, CA

Carmen Smith
Curriculum and Instruction
Elementary School Mathematics
Georgia Department of Education

Appendix C:

Protocol for Framework Comparisons (States)

In this task you will compare the NAEP framework—separately for grade 4 and grade 8—to state standards from CA, GA, IN, MA, TX, and WA. Your primary comparison will be to the state standards for the corresponding grade level[1]. However, we also have included state standards for some earlier grades (grades 3, 6, and 7), and you will use these in a more limited way, as described below.

As you will recall, the NAEP framework is organized into five *content areas*: Number Properties and Operations, Measurement, Geometry, Data Analysis and Probability, and Algebra.

Within grade level, and for each *NAEP content area*, compare the NAEP framework to the six states. (Note: the states may organize content differently, so you will have to find the corresponding state content where ever they put it.)

Answer two questions for each of the five NAEP content areas:

1. Is NAEP missing something in this content area?
    - Describe what is missing by citing text from the state framework that expresses it best.
    - Indicate where each of the 6 states includes this content in its standards (if the state includes it at all).
    - Rate how important you think it is that this content be included on NAEP. Rate the omission as of minor importance, moderate importance, or major importance.

2. Is NAEP overemphasizing something in this content area?
    - Describe what is overemphasized by citing the NAEP objective(s) in which the over-emphasized content appears.
    - For any topic that you consider over-emphasized in NAEP, rate its emphasis in each of the six states.

Details:

To ***"cite,"*** simply name the document (state or NAEP) and use the document's numbering scheme. E.g., "NAEP, Numbers 3 (a)" would serve as the citation for NAEP Numbers Properties and Operations, subtopic 3) (number operations), objective (a) (perform computations with rational numbers).

"Missing" means not mentioned or mentioned incidentally without being the focus of at least one objective.

"Overemphasis" means being the focus of an objective or several objectives whereas it was "missing" from most states.

Grade level issue: States have standards for each grade level, so a state's grade 8 standards usually focus on grade 8 content and exclude earlier grade content. NAEP, in

[1] For California, there are no grade 8 mathematics standards. Use a combination of grade 7 and Algebra standards for the comparisons.

contrast, tests only at grades 4, 8, and 12 and tends to include some content from earlier grades. (For example, ratios and proportions are in NAEP grade 8 but are usually covered in grades 6 or 7 for the states.) Thus, NAEP may “overemphasize” some important mathematics compared to grade 8 state standards because the content appears in earlier grade standards of the states. For any topic that you judge to be “overemphasized,” look at prior-grade state standards. If you find a reasonable match to NAEP, mark this in the appropriate column of the “emphasis” table.

Grain size issue: NAEP (or a state) might mention a concept, but the comparison state might treat the same concept with more detail. For example, NAEP grade 8 only mentions “roots” in Numbers 2 (d), where students are expected to “estimate” square or cubic roots between two whole numbers. Exponents are referenced by NAEP only indirectly in Numbers 1(f), which deals with scientific notation. In contrast, Indiana deals with roots and exponents together in 8.1.4, 8.1.5, 8.1.6, and 8.1.7. We want you to consider this as a real difference in emphasis between NAEP and Indiana. Thus, you should specify that NAEP is “missing” content on the laws of exponents and the relationship between roots and exponents. MA 8.N.7 expresses this concisely and can be cited in order to convey what is missing.

Clarity of framework: Finally, there will be many times when you will be struck by how well or poorly a framework/standards document is written. Try to get past your reaction to the writing and focus on the domain of mathematics being described. However, feel free to point out egregious failures of communication or unusual eloquence.

Appendix D:

List of Expert Reviewers
Alignment Analysis

List of Expert Reviewers
Alignment Analysis Meeting
February 21-24, 2007

Ann Bartosh
Mathematics Consultant
Kentucky Department of Education

Sybilla Beckmann
Department of Mathematics
University of Georgia

Robin Cohen
Principal
New York City Public Schools

Wade Ellis
Department of Mathematics
West Valley, California Community College

Lisa Emond
Mathematics Instructional Specialist
New York City Public Schools

Sharon Fiden
Principal
New York City Public Schools

Linda Fisher
Director of Assessment
Robert Noyce Foundation

Maia Fjelland
Math Coach
Hayward, California Unified School District

Michael P. Gallagher
Math Test Development
North Carolina Department of Education

Susan Hull
Director of Mathematics
The Charles Dana Center

Michele Watson
Director of Mathematics
Glastonbury Public Schools

Mark Johnson
Student Assessment Services
Massachusetts Department of Education

Lucy H. Michal
Director of Mathematics and Science
El Paso Collaborative for Academic Excellence

Lisa Mikus
Teacher
Horace Mann Elementary School, Newton, Massachusetts

Claire Pierce
Math Program Manager
Georgia Department of Education

Lynn Raith
Retired Teacher
Pittsburgh Public Schools

George Reese
College of Education
University of Illinois

Michael Roach
Mathematics Consultant
Indiana Department of Education

Monica P. Schlechter
Bilingual Educator
Oregon Public Schools

Barry Smith
Teacher
Tomlinson Middle School, Fairfield, Connecticut

Carmen Smith
Curriculum and Instruction
Elementary School Mathematics
Georgia Department of Education

Wiley Williams
Department of Mathematics
University of Louisville

Appendix E

Directions for Alignment Analysis

Directions for Alignment Analysis
2/21-2/24 2007

At this meeting, our primary focus is on the question: *How well does the NAEP assessment assess the learning specified in the NAEP Framework?*

Note that the NAEP mathematics framework identifies content at three levels:

- Content area (e.g., Number Properties and Operations),
- Subtopic (e.g., Number Sense), and
- Objective (e.g., Identify the place value and actual value of digits in whole numbers (grade 4).

We will implement a review procedure that cycles through each of the 5 content areas. For each content area, we will first make judgments at the sub-content area level and then evaluate the content area as a whole before moving on to the next content area. (We will not make judgments at the objectives level, but objectives serve to define the subtopic).

The work will also alternate between individual review and group (table-level) discussion. The final product will be a set of individual ratings from each participant. However, the group discussions should help to calibrate the individual ratings and enrich the information on which the ratings are based. Each table will have a designated table leader to keep the discussion process moving forward.

Here is an outline of the review process for **one content area,** with suggested timing:

I. Review the Framework description of the *content area* and discuss briefly (with your table) [5 minutes]

II. For each *subtopic* within that content area:

A. Individual work [5-15 minutes per subtopic]
1. Review the objectives that define the subtopic
2. Compare the item pool to the framework
3. Make notes related to each of the 3 subtopic-level study question and decide on initial ratings

B. Group work with your table [5 minutes per subtopic]
1. Flash vote on subtopic ratings[2]
2. Discuss and attempt consensus[3]
3. Table leader summarizes and closes the discussion

C. Individual work [5 minutes per subtopic]
1. Individuals record their **final ratings** for the subtopic
2. Individuals record **comments**[4]

[2] A flash vote enables the table to decide how much discussion is needed. Conserve time where possible.
[3] Table agrees or agrees that there is a disagreement. Disagreement can be a useful finding. However an attempt should be made to reach agreement so that only those disagreements that survive the attempt are registered in the subsequent individual ratings.
[4] We are seeking insights in the comments; therefore individuals should express their own insights rather than merely repeat the table leader's summary.

D. Repeat steps A-C for each subtopic in that content area

III. For the *content area* overall

A. Group work with your table [10 minutes]
1. Discuss the 2 content-area level study questions
2. Flash vote
3. Attempt consensus
4. Table leader summarizes

B. Individual work [10-15 minutes]
1. Individuals record their **final ratings** for the content area
2. Individuals **"write and cite"** their observations. Cite example items from other tests to illustrate what NAEP could do to improve assessment in this content area. Cite examples from NAEP test to illustrate what there is too much of or too little of. (Cite by naming test and giving item number.) Also identify any aspects of the NAEP Framework that are too ambiguous or faulty to use for evaluating the item pool in this content area.

Ratings and Questions:

- Consider the gross quality of the items; discount the contribution/value of low quality items.
- Your rating and comments should be based on your own analysis of the Framework and items, as well as your consideration of the discussion with others at your table. Attempt to join a consensus.
- All ratings use a 4 point scale, where

1= very well
2 = well enough
3 = not well enough
4 = poorly

Subtopic Ratings:

1. **FOCUS:** To what extent does the item pool **focus** on the most important knowledge and know-how in this subtopic rather than waste items on marginal knowledge? How well does the center of gravity of the items match the center of gravity of the subtopic as defined by the Framework?
    a. The NAEP objectives define the subtopic
    b. Knowledge on which other knowledge builds is more important than isolated bits of knowledge

2. **BALANCE:** How well does the item pool **balance** across the range of knowledge and know-how in the Framework subtopic?

3. **REACH:** How well does the item pool **reach** from less advanced and/or easier aspects of the subtopic to more advanced and/or harder aspects? To what extent does the item pool offer items allow all students to show what they learned?

Content Area Ratings and Questions:

1. **BALANCE:** How well does the item pool **balance** across the subtopics in this content area?

2. **COMPLEXITY:** Does the item pool for this content area exhibit a **complexity** distribution consistent with the framework specifications for complexity?

3. **WRITE AND CITE** for a given content area:

    a. Where NAEP is weak: find examples in the state/nation tests to illustrate the kind of items NAEP should have more of in this content area.

    b. Illustrate with NAEP items the kinds of problems there are too many of or too few of in this content area.

    c. Are there defects or ambiguities in the Framework that make it a poor tool for judging the item pool in this content area? Please specify.

Rating Form (sample)

Grade 4

Content Area 1	**Number Properties and Operations**	**NAEP does:**			
		Very well Rating: 1	**Well enough Rating: 2**	**Not well enough Rating: 3**	**Poorly Rating: 4**
	Content Area Balance	1	2	3	4
	Complexity				
	Low	1	2	3	4
	Moderate	1	2	3	4
	High	1	2	3	4
Subtopic 1	**Number Sense**				
	Balance	1	2	3	4
	Focus	1	2	3	4
	Reach	1	2	3	4
Subtopic 2	**Estimation**				
	Balance	1	2	3	4
	Focus	1	2	3	4
	Reach	1	2	3	4
Subtopic 3	**Number operations**				
	Balance	1	2	3	4
	Focus	1	2	3	4
	Reach	1	2	3	4
Subtopic 4	**Ratios and proportional reasoning**				
	Balance	1	2	3	4
	Focus	1	2	3	4
	Reach	1	2	3	4
Subtopic 5	**Properties of number and operations**				
	Balance	1	2	3	4
	Focus	1	2	3	4
	Reach	1	2	3	4

WRITE AND CITE

For Content Area 1:

d. Where is NAEP weak? Find examples in the state/nation tests to illustrate the kind of items NAEP should have more of in this content area.

e. Illustrate with NAEP items the kinds of problems there are too many of or too few of in this content area.

f. Are there defects or ambiguities in the Framework that make it a poor tool for judging the item pool in this content area? Please specify.

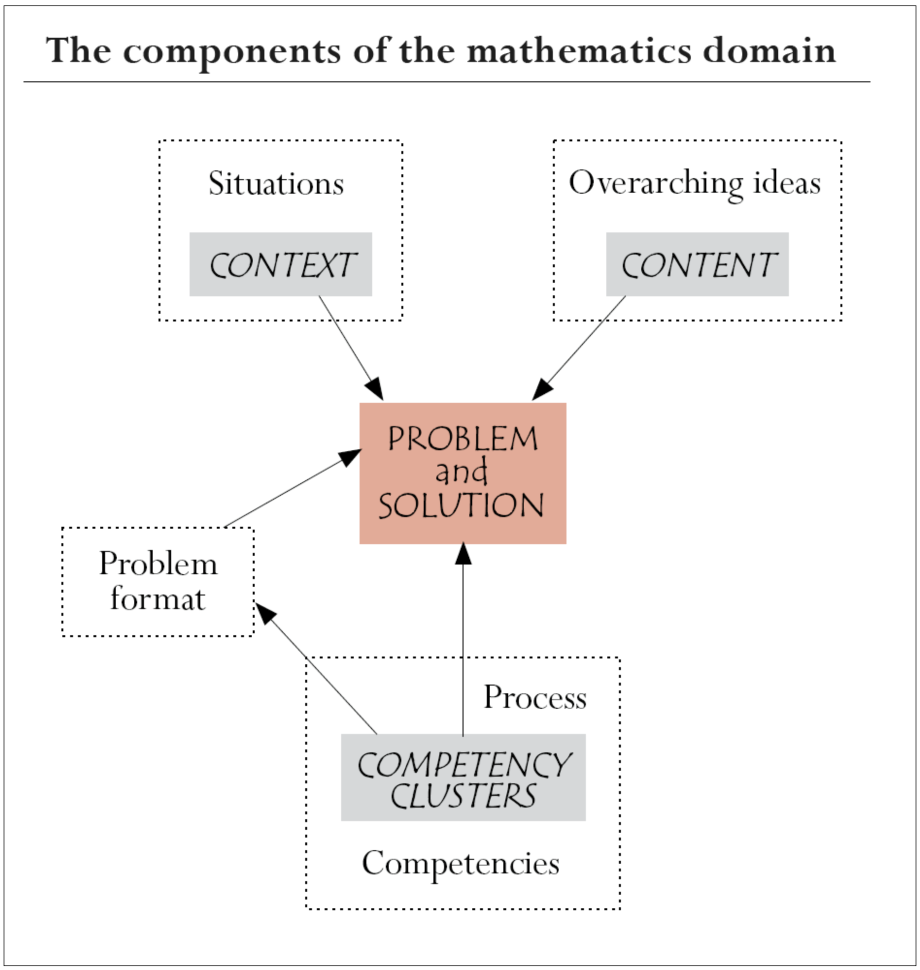
Appendix F:

On Design of Items

On design of items, especially more complex ones.

The 3C's schema

A good starting point for every designer for assessments in mathematics is the (basically) triangular schema used in the Framework of OECD/PISA for mathematical literacy. (see de Lange, 2007). The corners are the three C's: content, context and competencies. These three result in an item and feedback mechanism in the center of gravity of the triangle. An extra line of influence can be drawn from the competencies to the item: the competencies requested may influence the format of the item.



The Content

The content decision is the easiest one, in large scale assessments especially, as these are quite often laid out in a framework, and sometimes into some detail. Therefore, one just has to make sure that the essential content is covered in a satisfactory way. The bottom line: does the test as a whole have the proper balance in number, algebra, and geometry, items, for example?

The Context

The context domain is very complicated and quite often not well understood. If the context of the problem is mathematical, we may call it inner-mathematical and we are done with the context. Right? Wrong!

One can offer the students any problem in a variety of ways, with a lot of text, with visuals, either mathematically or from a real world, and so on. Example: you can ask 6 + 6 = ___, but you can also ask: what is six plus six, or what is the double of 6, or, ….. All of these count as different inner-mathematical items. Or you can show two number cubes with the six faces up and ask the total score. Is this still inner-mathematical?

Then you have the fake contexts: a basically inner-mathematical problem that has a context that has really nothing to do with the solution to the problem. Something like Jan is six years old, Mary is six years old, How old are they together?

Of course there are other contexts that are better, but only marginally: How many soda cans in two six packs? This all too real-world problem makes sense and has relevant mathematics in it. But mathematizing is definitely not an important aspect of this problem.

But the really nice problems are those where both the mathematics and the context interact nicely and make sense. On a very simple level: I need 20 cans of soda. How many six packs should I buy?

Another important factor in determining the quality of an item is how "near" to the student the item is. We can stay very close to the student's home, so to speak, or stray very far away. In an ideal test the "distances" to the student should be balanced also.

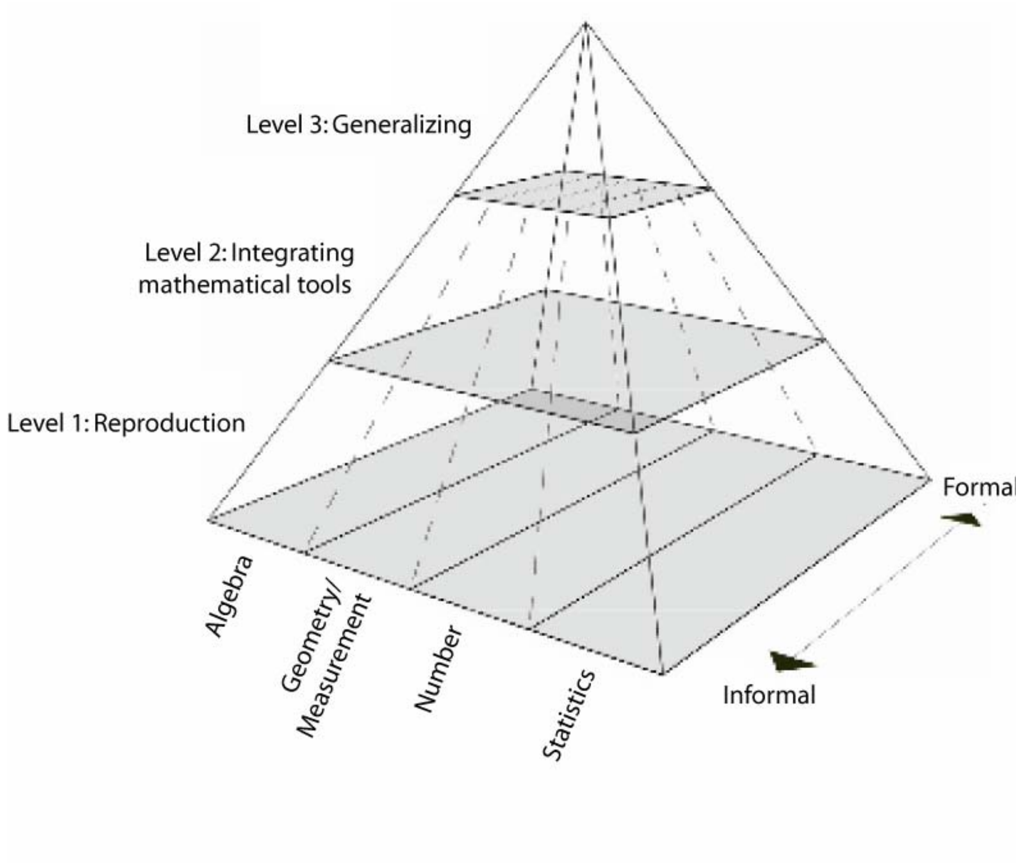
To make a shortcut:
In good assessments there should be a balance in

- inner-mathematical versus real-world context
- relevant context versus "motivational" context
- presentation of context (e.g., words, graphs, pictures)

One can have authentic contexts that are not real, but transferred from reality to some form of artificiality that is still authentic.

The Competencies

The competencies domain may be the most complex and difficult to deal with. Most frameworks nowadays have some kind of distribution of skills, following something that goes from “Reproduction and Definitions” to “Connections & Problem Solving” and “Reasoning, Thinking and Abstraction” or similar categories. Previously we have visualized some of the required design aspect in our “Assessment Pyramid.” This pyramid had been very successful for the design of tests, as well as for professional development of mathematics teachers.



It will be clear from the schema that only a limited number of variables can be presented. So the context is missing as a variable, but the content and competencies can be visualized quite dramatically: a good test will have items well spread over the whole pyramid.

Level of Formality

In this example, we have chosen, for the third variable in the pyramid, the level of formality of the mathematics involved. This seems an increasingly important aspect that measurers are interested in. For instance, the difference in scores for PISA between the real world champion Flanders and The Netherlands can be completely attributed to the competency cluster and the formal aspects of mathematics.

So for the test designer, this pyramid offers a lot of challenges. For NAEP, for instance, it is my opinion that the upper part of the pyramid is almost a vacuum. This is highly undesirable and makes the content validity of the whole NAEP test an important point to reflect on.

Balance Between Difficult and Easy

But what makes the task even more daunting is the fact that we also need a good balance between easy and difficult items. In PISA we have found out that indeed to construct easy complex items is an art by itself.

In short, the aspects that we need to take into account, apart from the usual suspects like validity and reliability, are:

1. Content appropriateness
2. Context balance: e.g., fake versus authentic; close to the student versus scientific; inner-math versus real world; mode of presentation (text versus pictures)
3. Competencies balance
4. Balance easy/difficult
5. Balance formal/informal

Another essential aspect of any test item is that the item really does measure what it is supposed to measure. It is nice to finally end up with an item that the designer is happy with from all of the abovementioned aspects. But the moment of truth comes when the item also does what it is supposed to do. In the “art” article I have given an example of a TIMSS item that clearly does not operationalize what it is said to do.

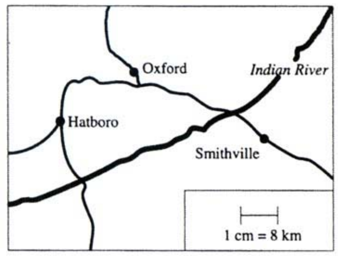
Linking Item with Framework

To provoke some discussion, and to invite the reader to reflect on the items and make a judgment, we start the examples with an old (mid-nineties) item from TIMSS. This item is from population 2, which indicates that the children who will have to solve the problem will be around 14 years of age.

EXAMPLE ITEM 3
FRACTIONS & NUMBER SENSE

Distance on map

One centimeter on the map represents 8 kilometers on the land.



About how far apart are Oxford and Smithville on the land?

A. 4 km

B. 16 km

C. 35 km

D. 50 km

Performance Category: Using Complex Procedures

The problem seems simple, actually too simple, for the intended age group. There is nothing wrong with the item as such; it requires the students to make an estimate of the distance from Oxford and Smithville, using the scale line. The multiple choices make it even easier for the students to find the correct answer, which is encircled to avoid embarrassing any reader.

The surprise lies in the "Intended Performance Category," which is "Using Complex Procedures." The interpretation is that, if a student has chosen answer C, he is able to use complex (mathematical) procedures. This seems somewhat far-fetched.

So the aspect that this example shows clearly is that we need to make sure that there is a reasonable relationship between the "expected performances" or "competencies" or "learning goals" and the item. This seems straightforward but is at the heart of the design of any coherent and consistent assessment design.

Jan de Lange
April 9, 2007

Appendix G:

List of Mathematician Reviewers

List of Mathematician Reviewers
Item Quality Meeting
February 17-18, 2007

Patrick Callahan
University of California, San Diego

Art Duval
University of Texas at El Paso

Wade Ellis
West Valley Community College

Roger Howe
Yale University

Wilfried Schmidt
Harvard University

Appendix H:

Directions for Review of Item Quality

Directions for Review of Item Quality
February 17-18 2007

Directions for Item Rating Task

You will receive packets of similar items. Each packet contains NAEP items intermingled with items sampled randomly from state assessments. Across the packets, all NAEP items that appeared on the 2005 or 2007 assessments are included.

Your task for today is to evaluate the adequacy of each NAEP and state item, using the attached Mathematical Quality rating scale. For each packet, a response sheet is provided on which you should enter your ratings and explain your concerns for items that you have judged to be less than adequate. You will also receive a pad of sticky notes for flagging items that you want to bring up in discussion. It will be especially valuable to illustrate points by contrasting two items that are similar but one is adequate and one isn't. This will help to focus the discussion and will also provide us with essential material for our report.

For most of the day we will work in two subgroups (Group A and Group B). Ratings will be done individually, but we will pause after each packet or set of packets to discuss emerging issues with your assigned subgroup. (Some packets have been batched to even out the work flow.) There will also be time for summative discussions across subgroups.

Rating Scale

Each NAEP item should assess mathematical content. In addition, items assess the student's ability to reason with the content. NAEP has defined three levels of complexity to capture reasoning with content. Items with high level of complexity ask students to "…engage in more abstract reasoning, planning, analysis, judgment, and creative thought…." (Mathematics Framework for the 2005 National Assessment of Educational Progress, 2004).Thus, it is consistent with the NAEP Framework to present a situation and ask students to identify relevant quantities and express their relationship symbolically, numerically or graphically; or to find an answer based on formulating a mathematical understanding of the situation.

Given the grade level of the students (their expected reading level and mathematical maturity), please rate each item as:

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.

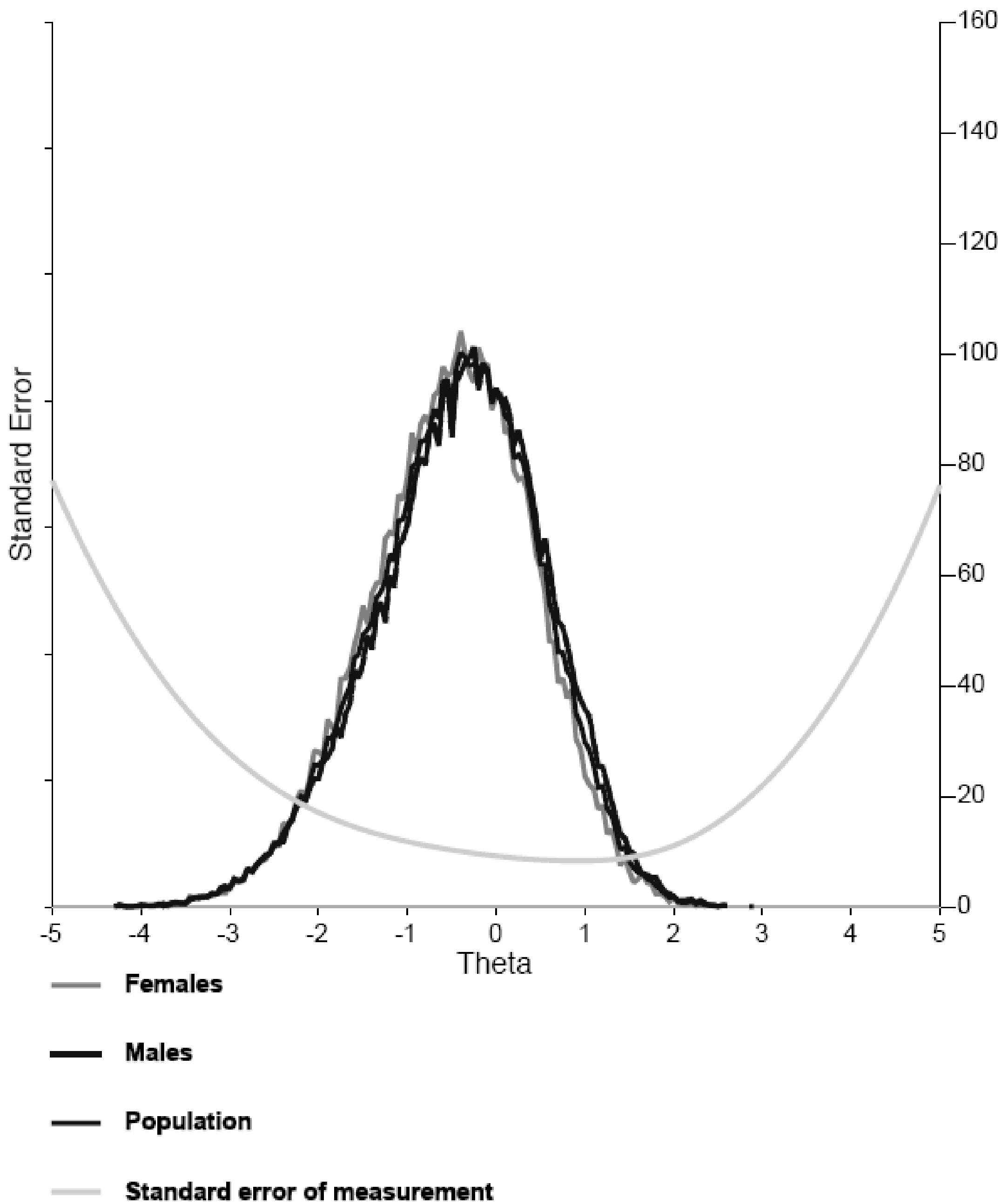
3. ***Seriously Flawed***

   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 graphics 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 rubric.

Appendix I:

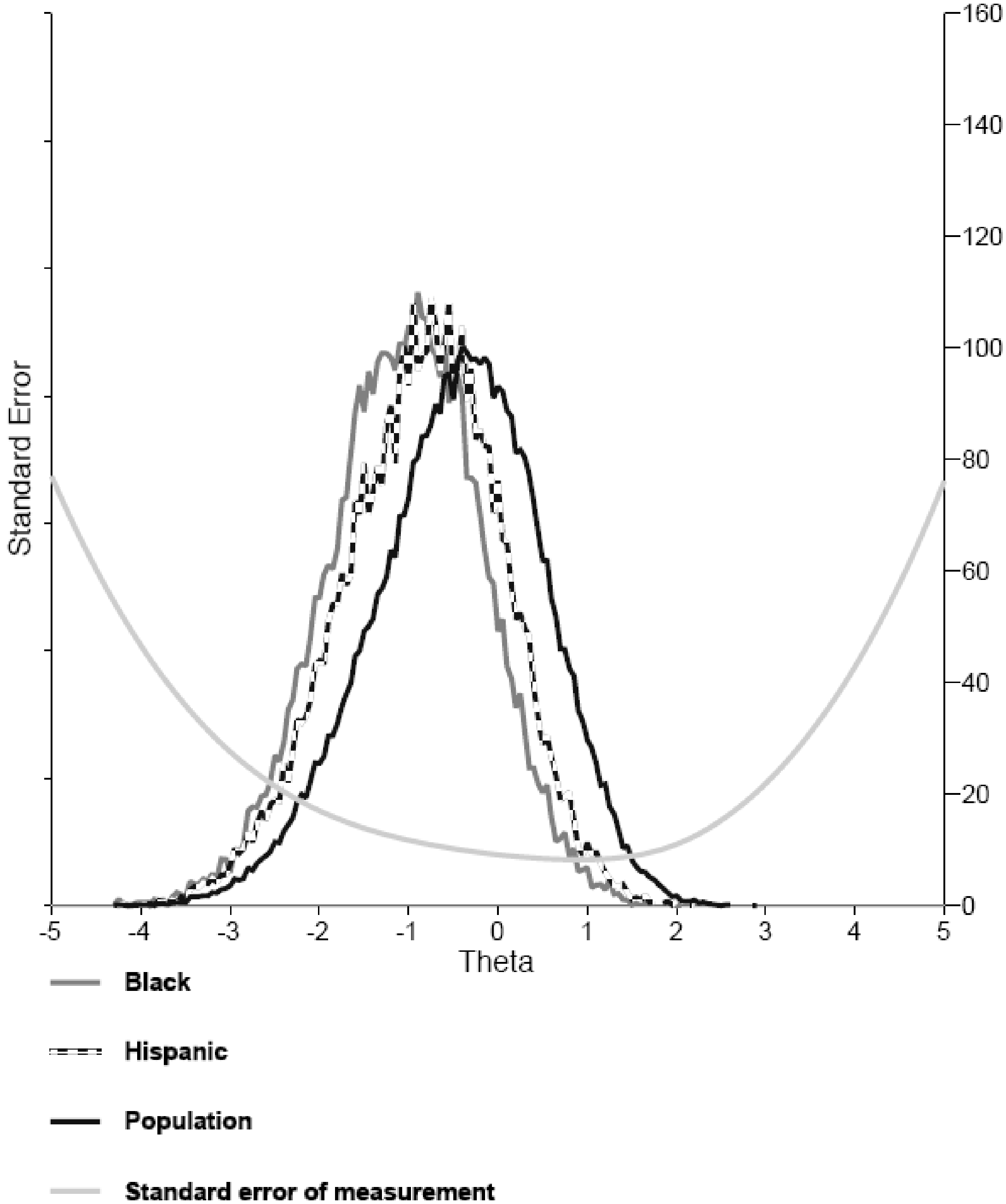
Plots of Standard Error of Measurement Relative to Population Performance, for Each Subscale and Each Mandated Reporting Group

Exhibit I-1. Grade 4 numbers and operations: Standard error of measurement and achievement distributions by gender



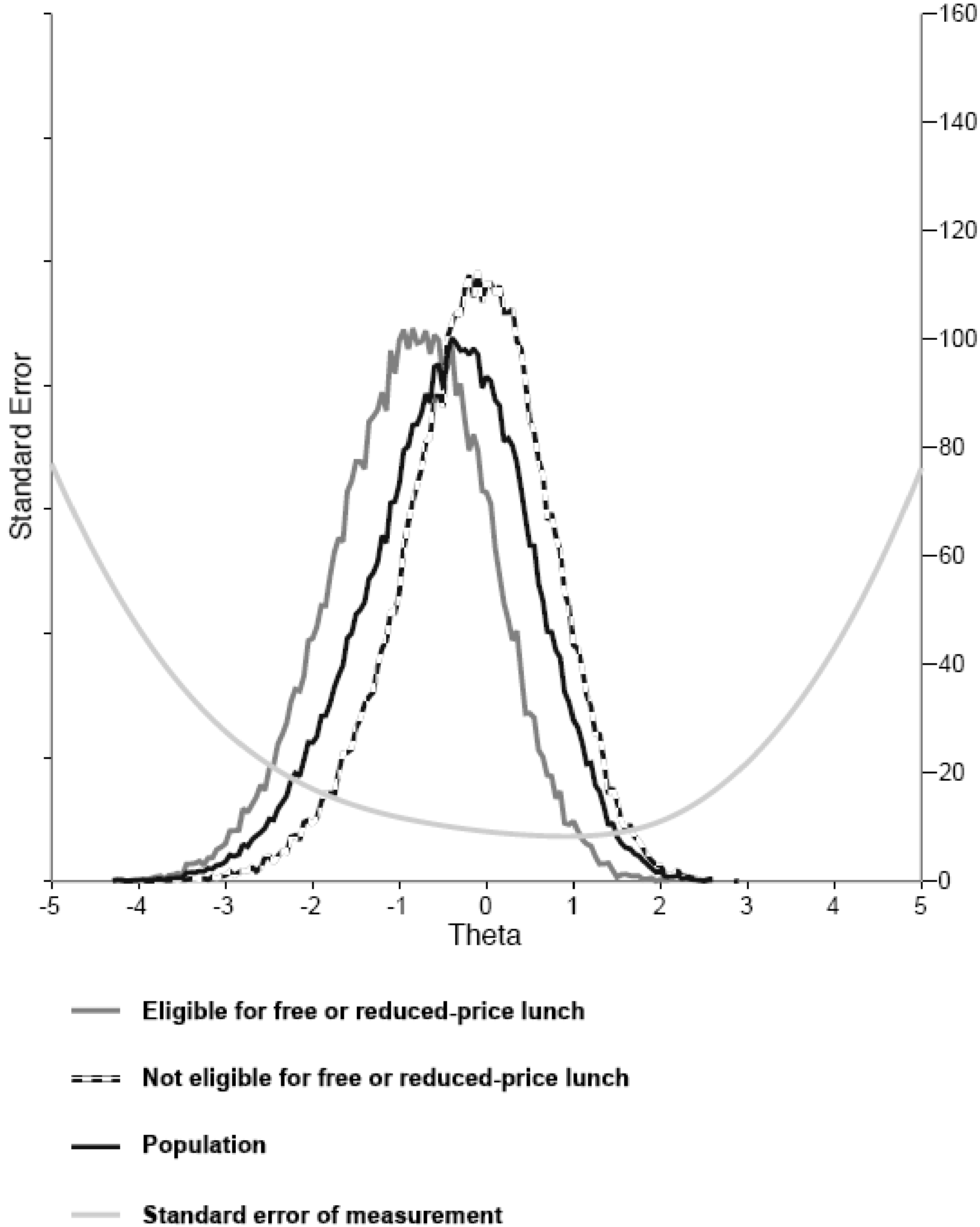
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-2. Grade 4 numbers and operations: Standard error of measurement and achievement distributions by race/ethnicity



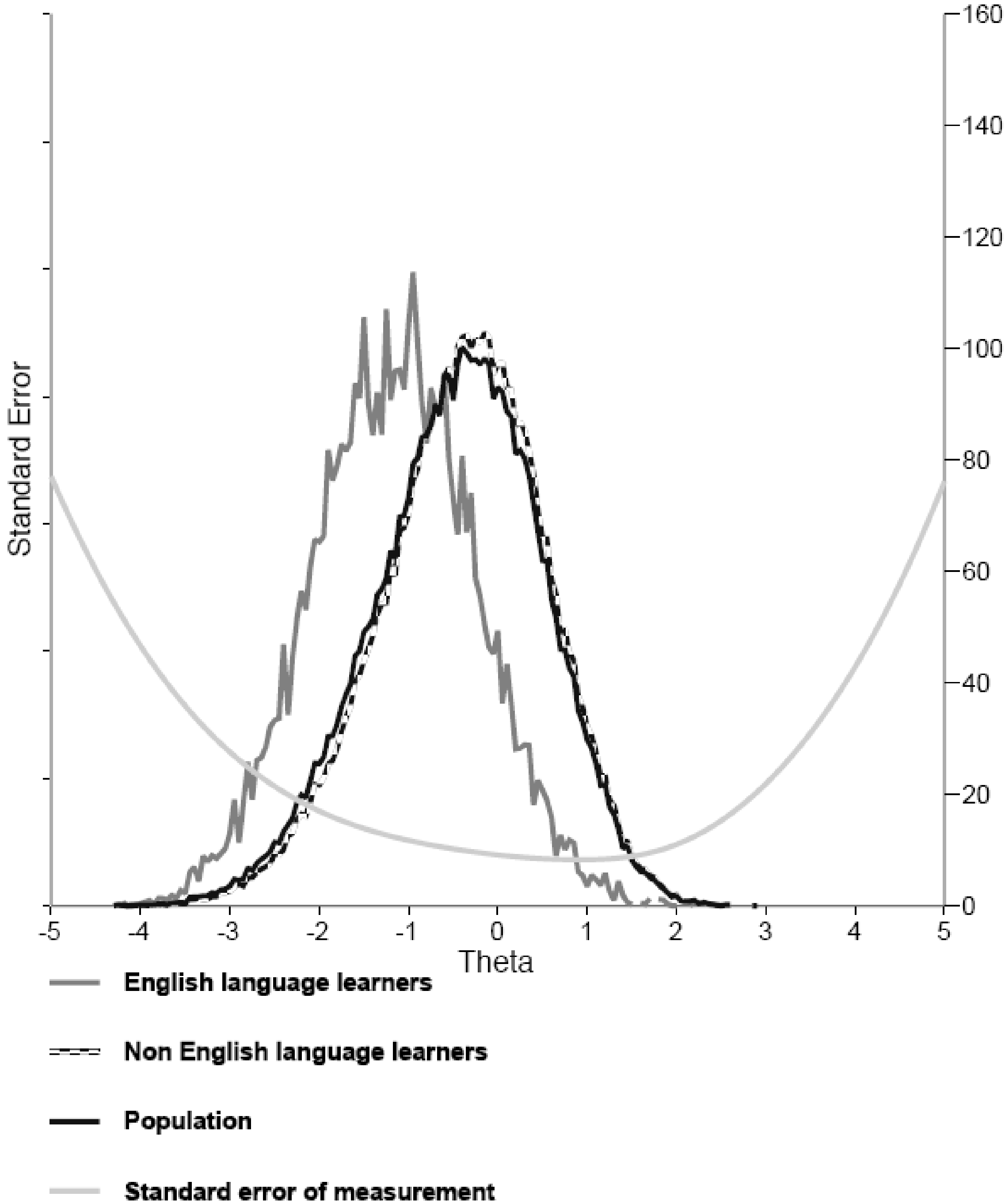
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-3. Grade 4 numbers and operations: Standard error of measurement and achievement distributions by eligibility for free/reduced price lunch



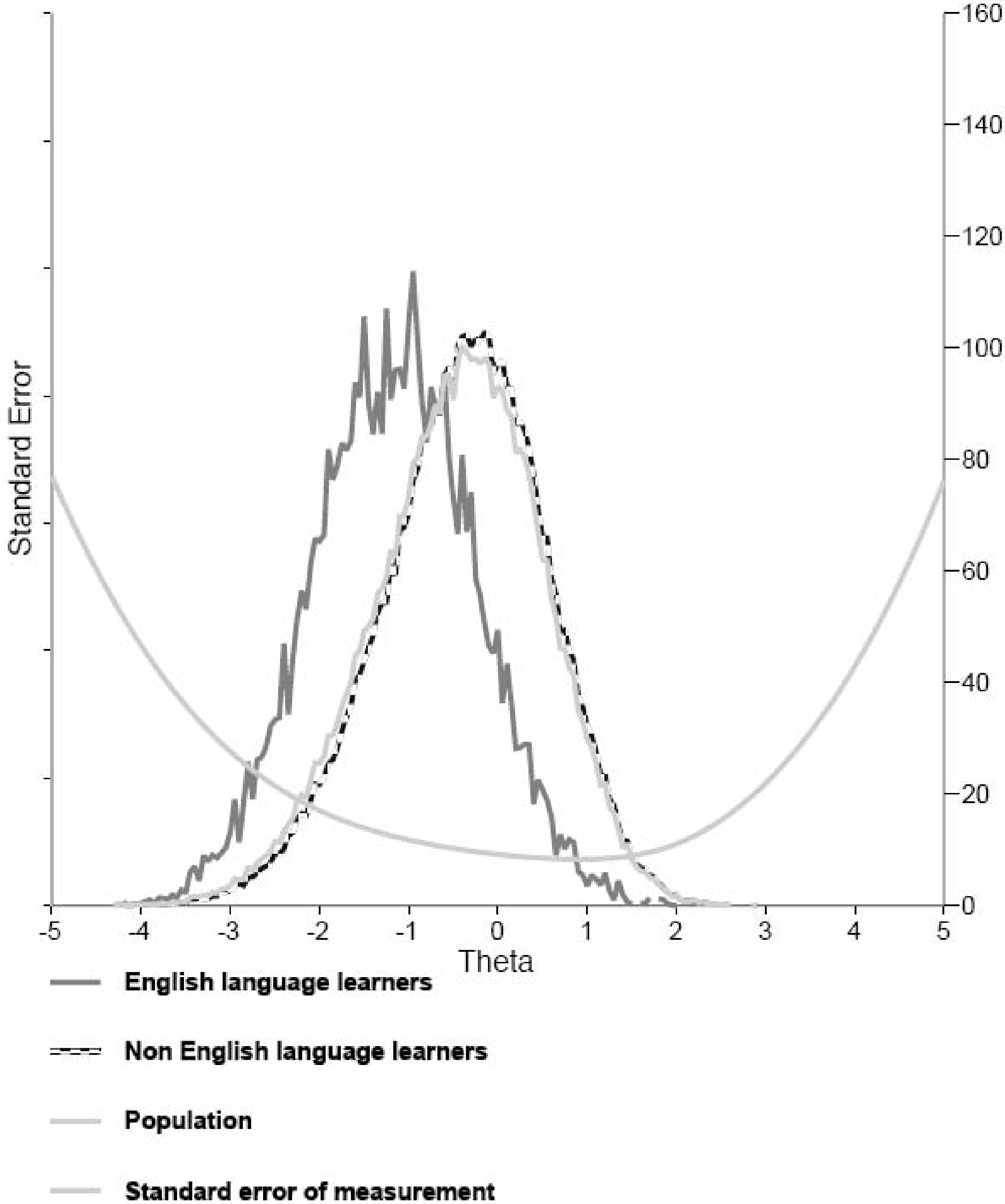
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-4. Grade 4 numbers and operations: Standard error of measurement and achievement distributions by student disability status



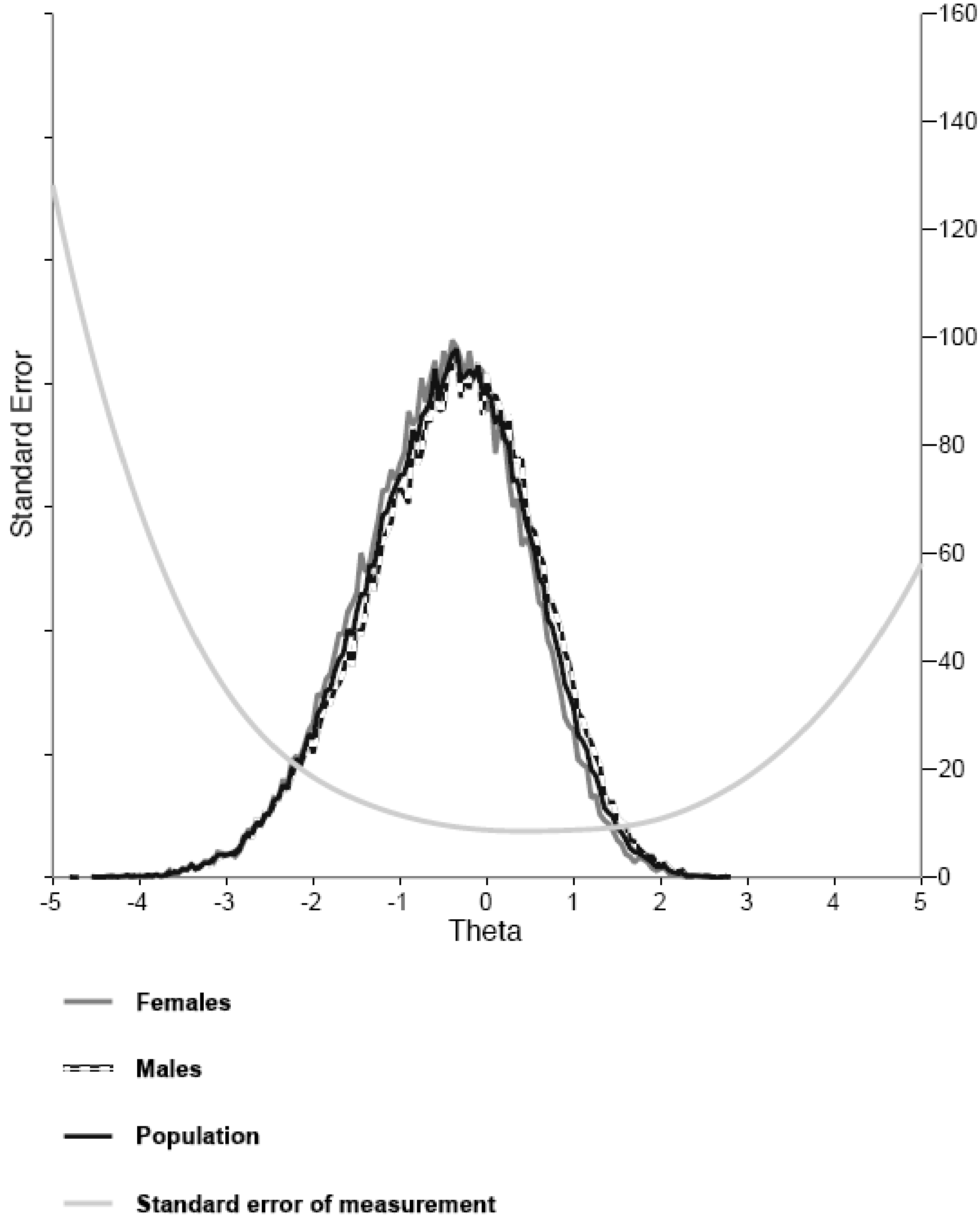
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-5. Grade 4 numbers and operations: Standard error of measurement and achievement distributions by ELL status



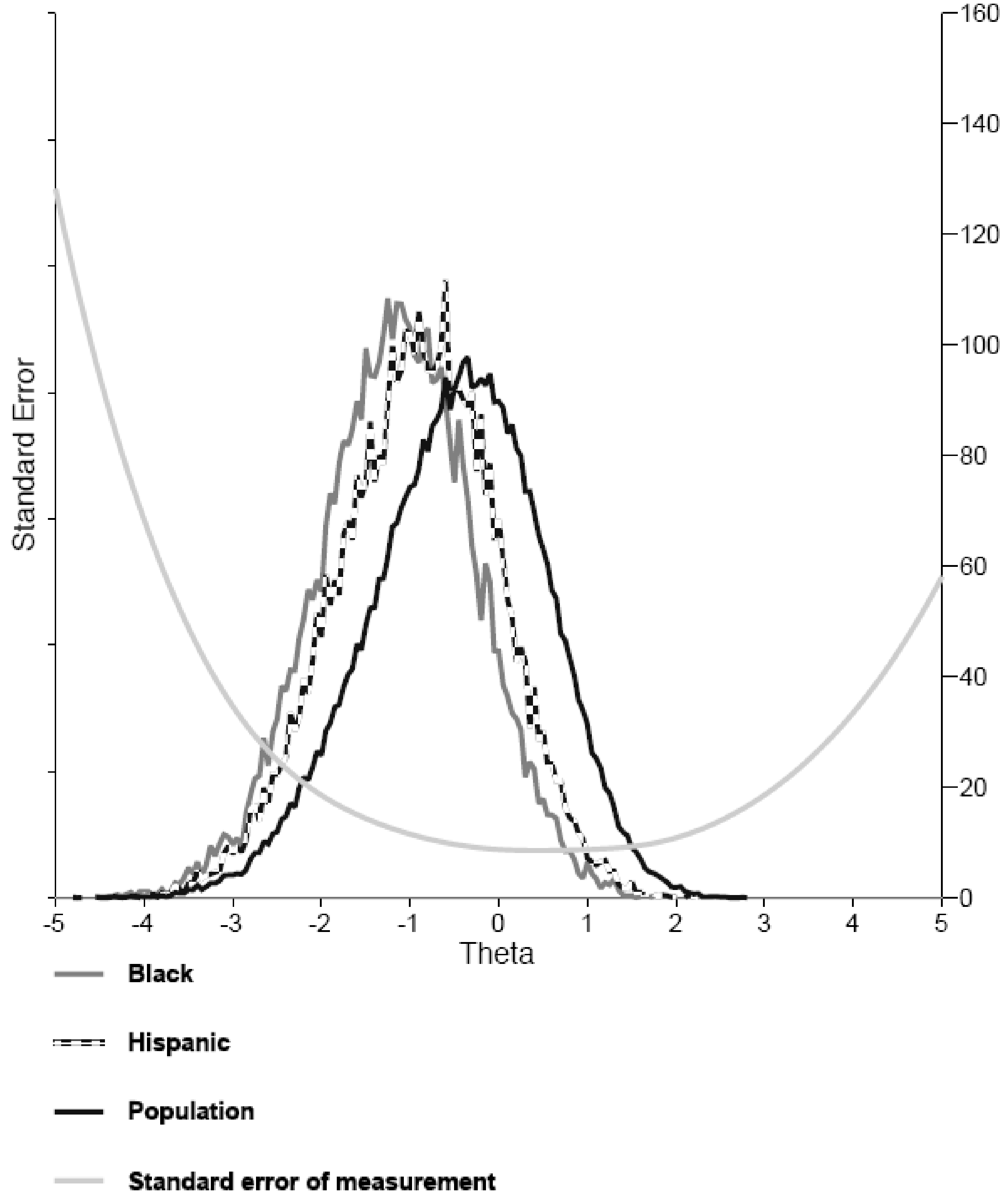
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-6. Grade 4 measurement: Standard error of measurement and achievement distributions by gender



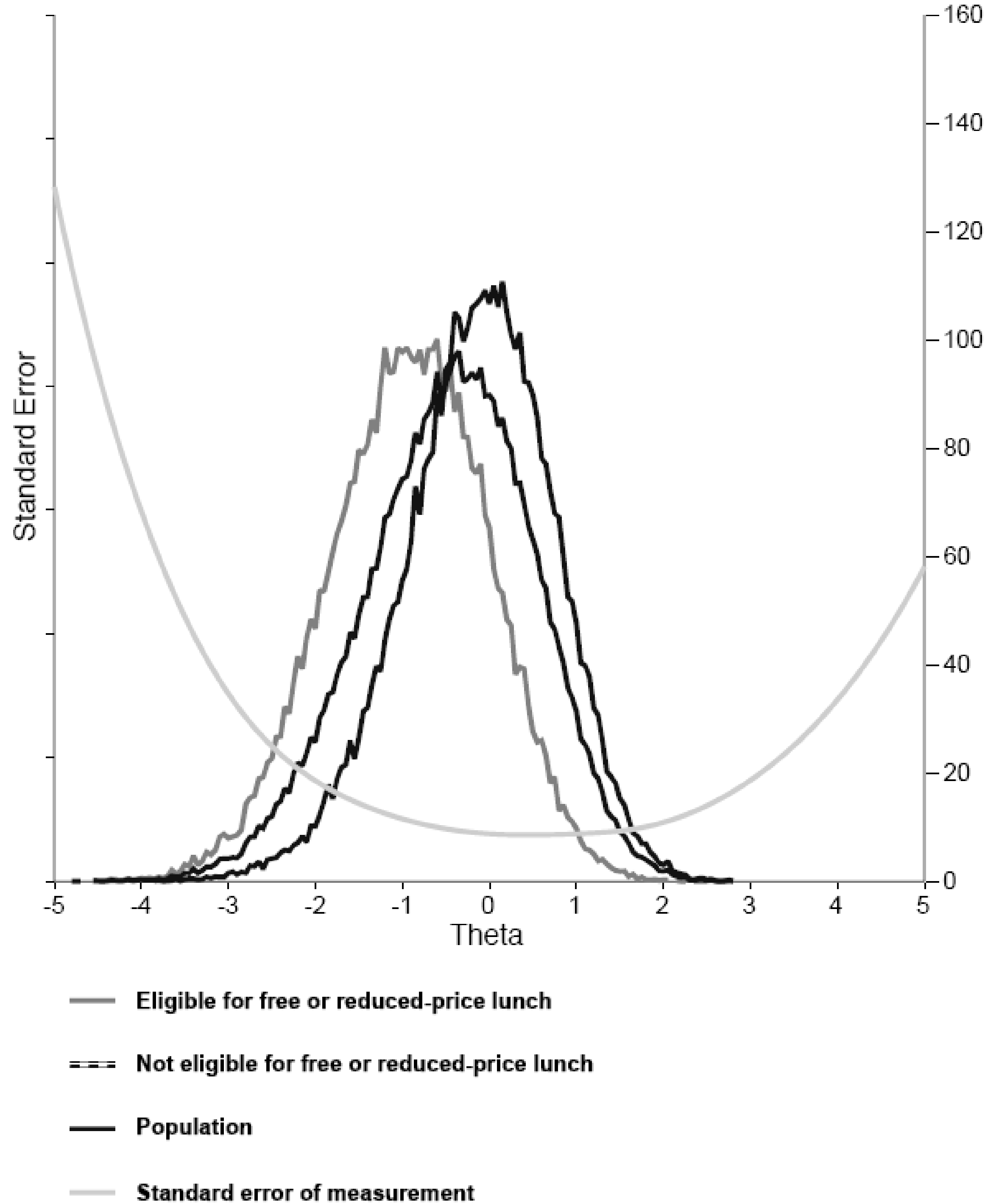
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-7. Grade 4 measurement: Standard error of measurement and achievement distributions by race/ethnicity



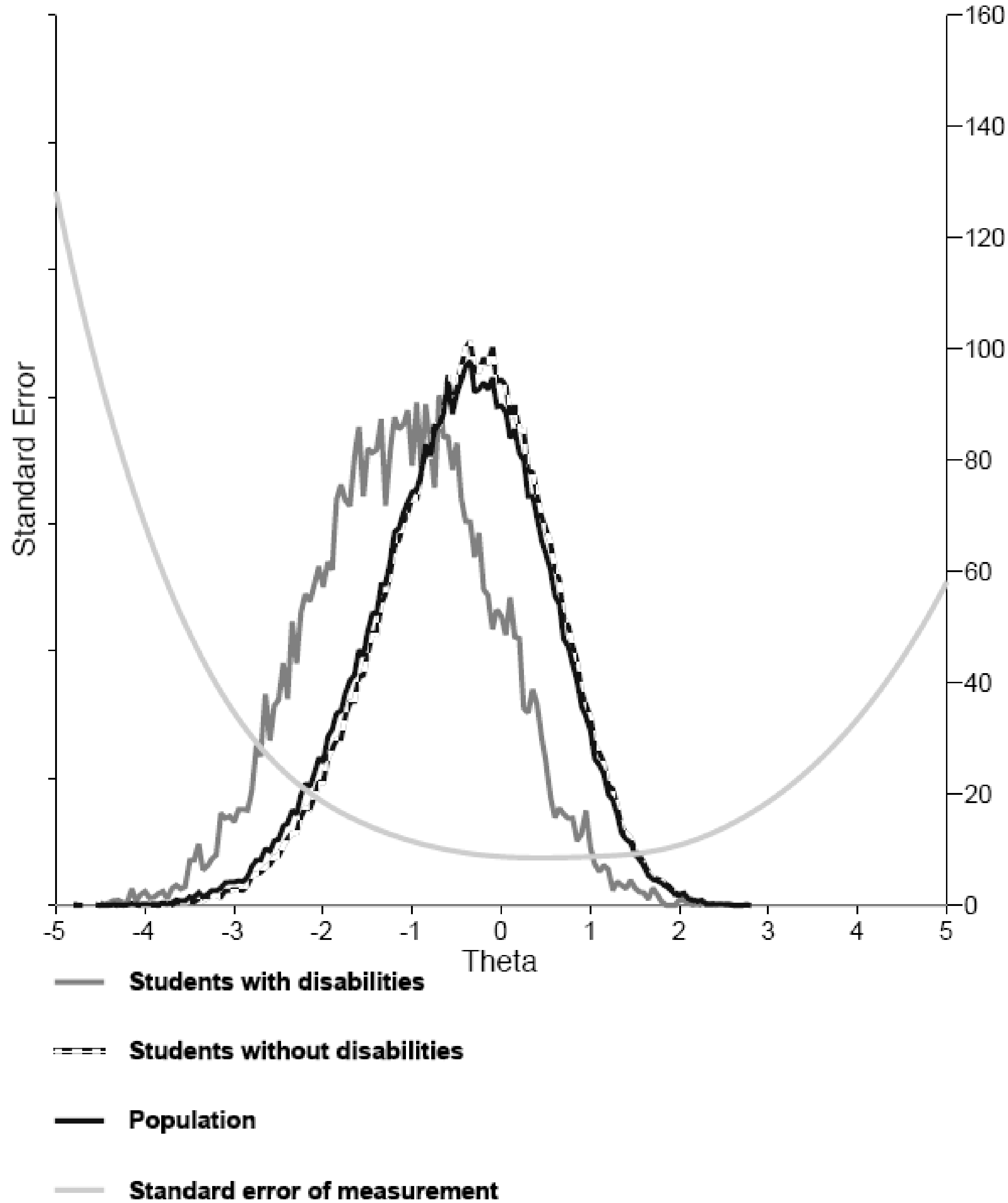
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-8. Grade 4 measurement: Standard error of measurement and achievement distributions by eligibility for free/reduced price lunch



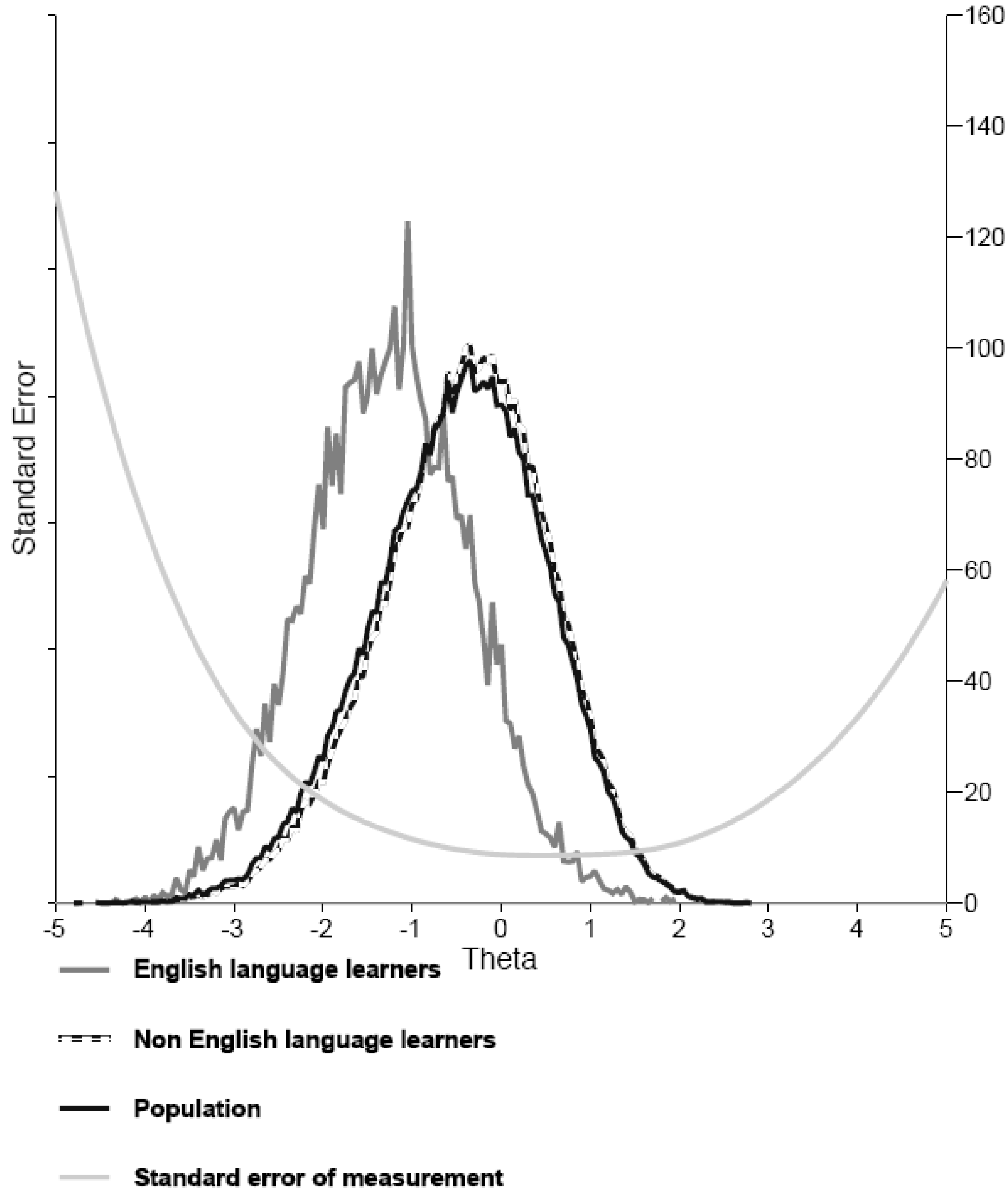
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-9. Grade 4 measurement: Standard error of measurement and achievement distributions by student disability status



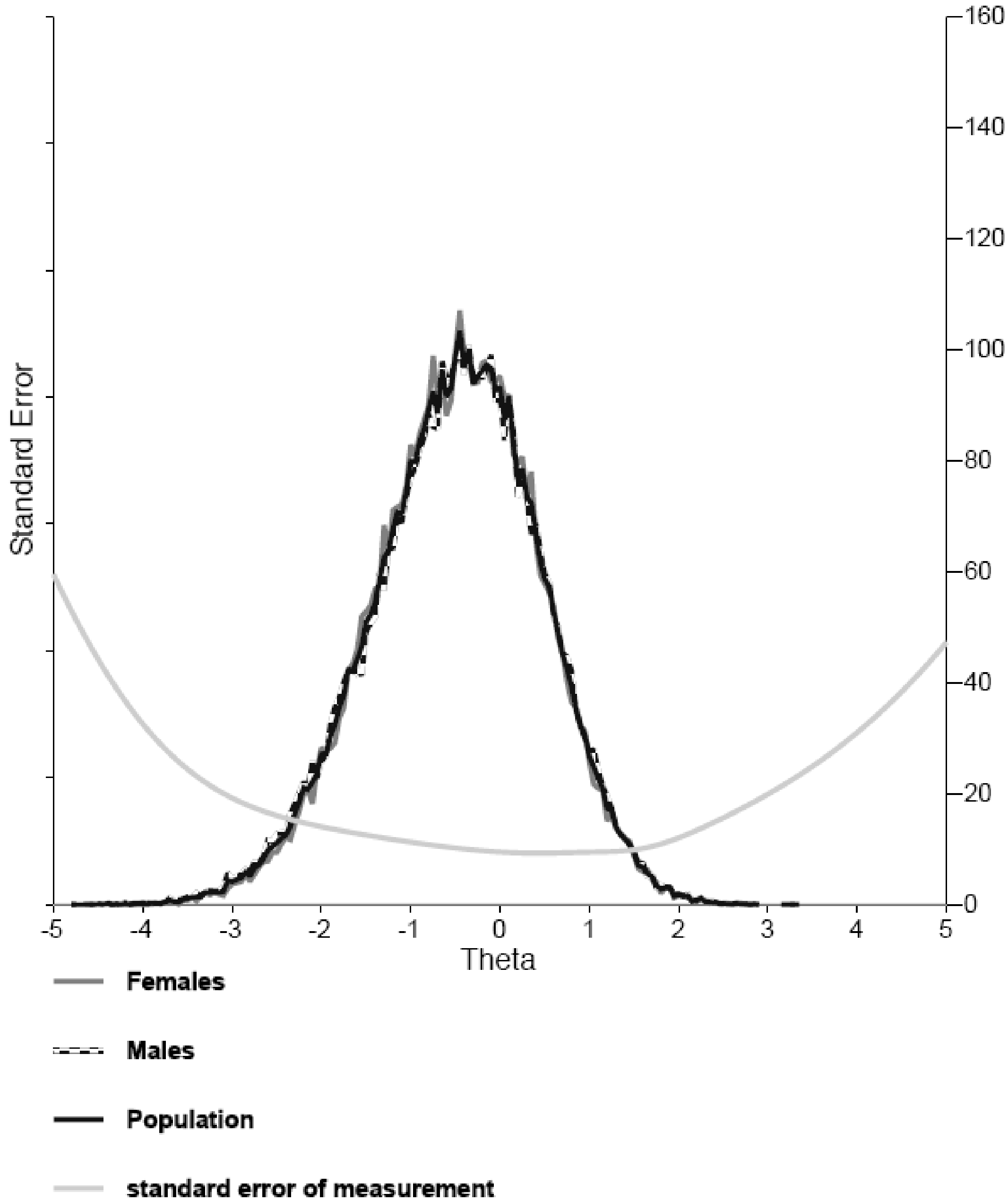
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-10. Grade 4 measurement: Standard error of measurement and achievement distributions by ELL status



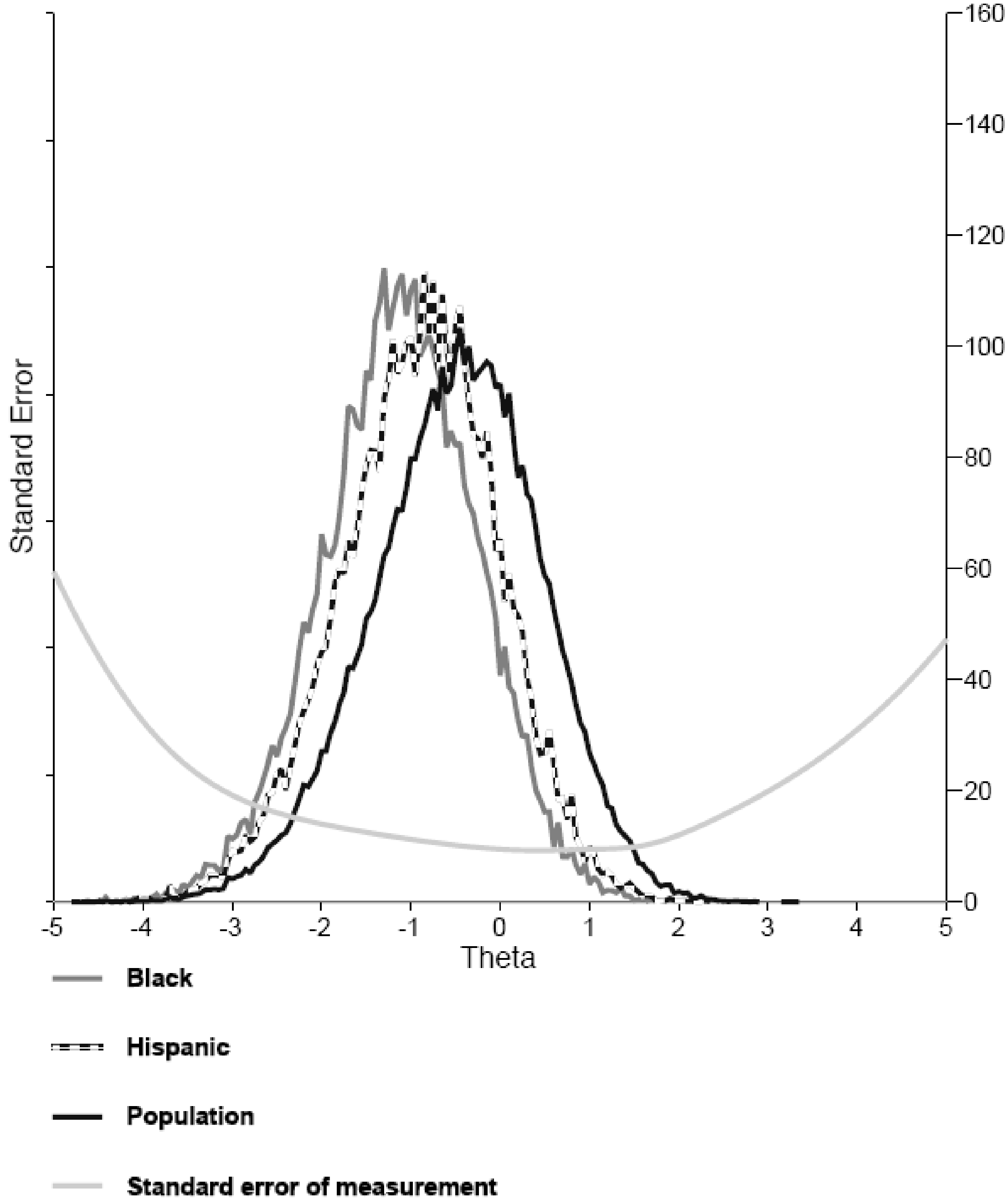
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-11. Grade 4 geometry: Standard error of measurement and achievement distributions by gender



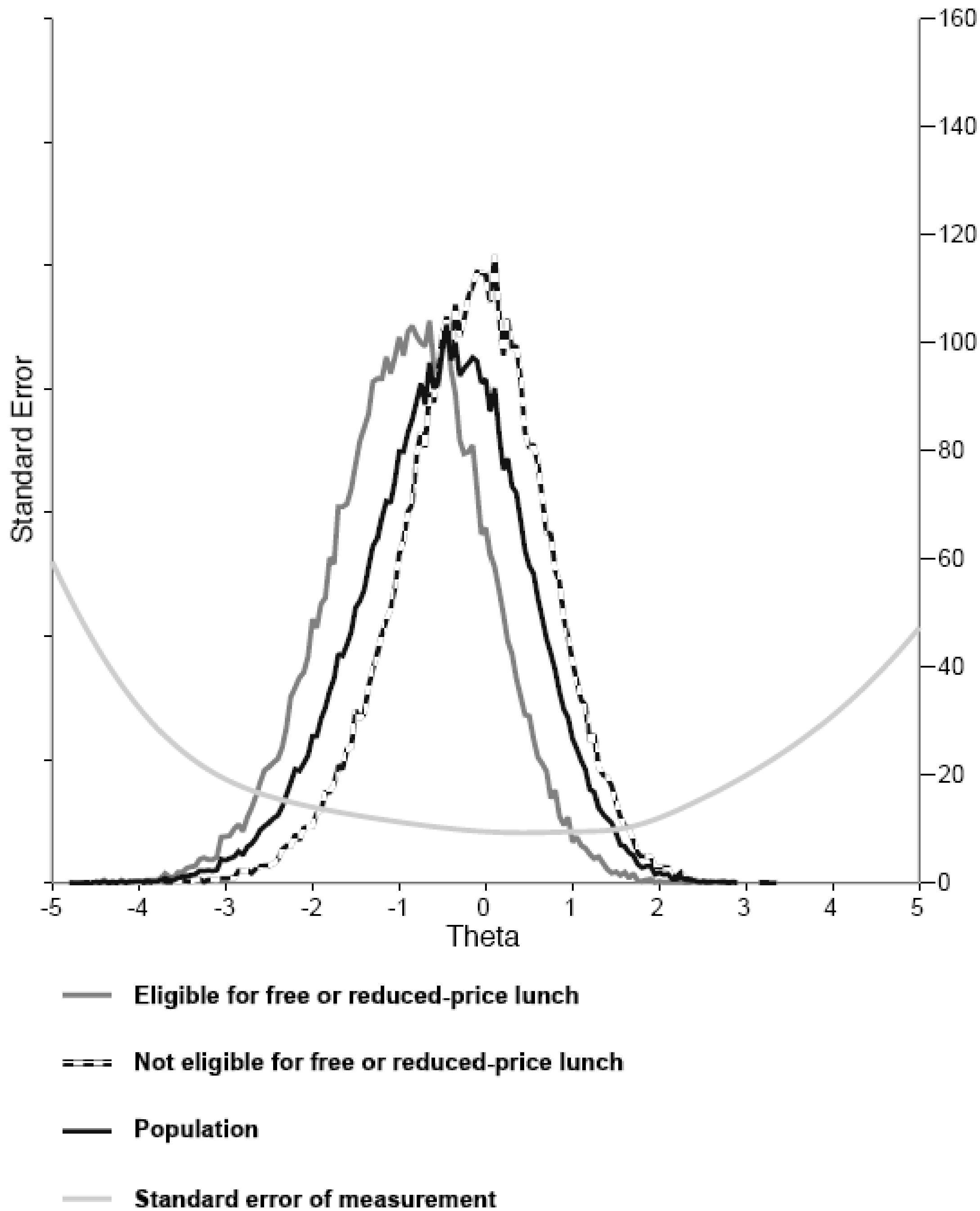
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-12. Grade 4 geometry: Standard error of measurement and achievement distributions by race/ethnicity



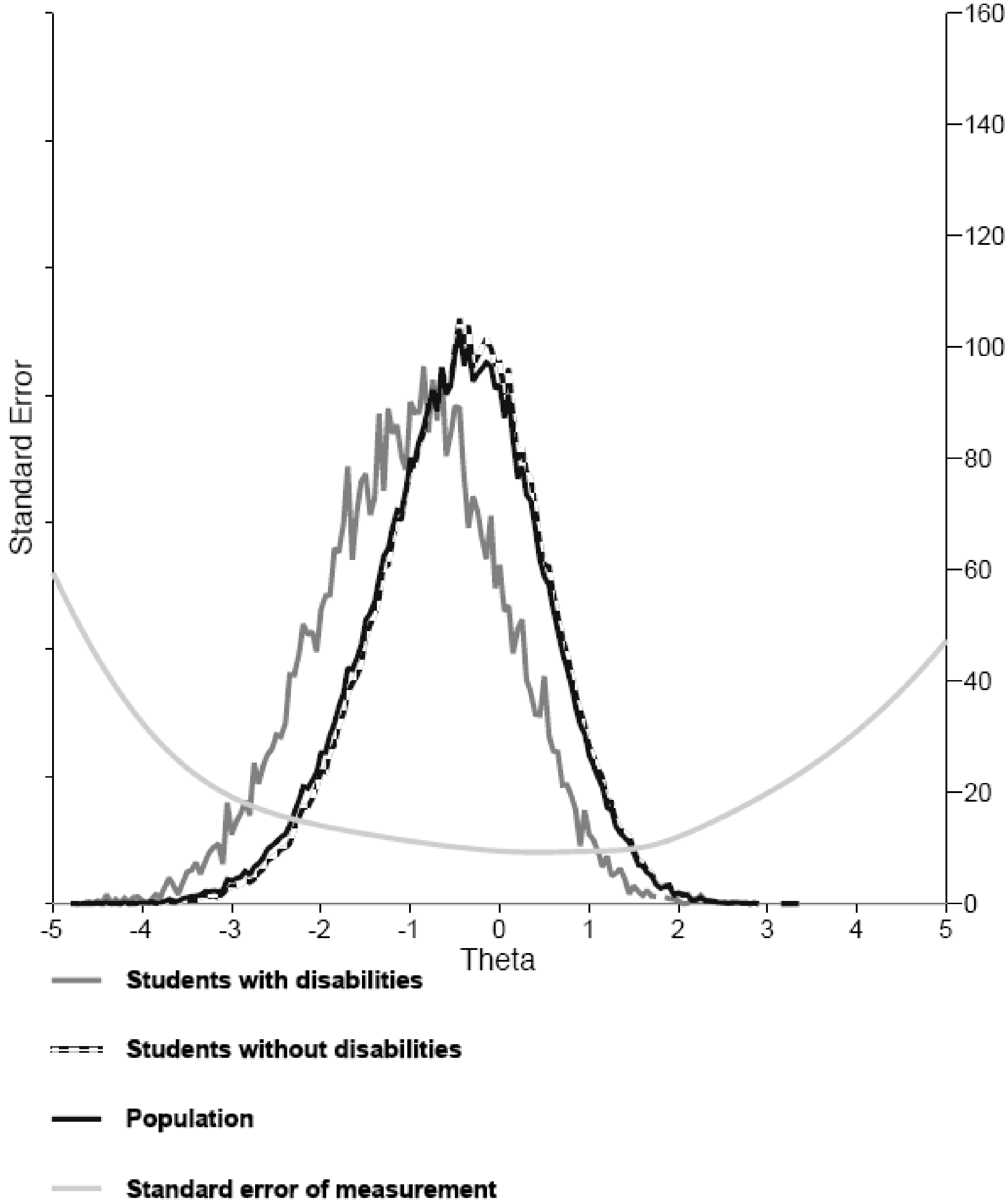
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-13. Grade 4 geometry: Standard error of measurement and achievement distributions by eligibility for free/reduced price lunch



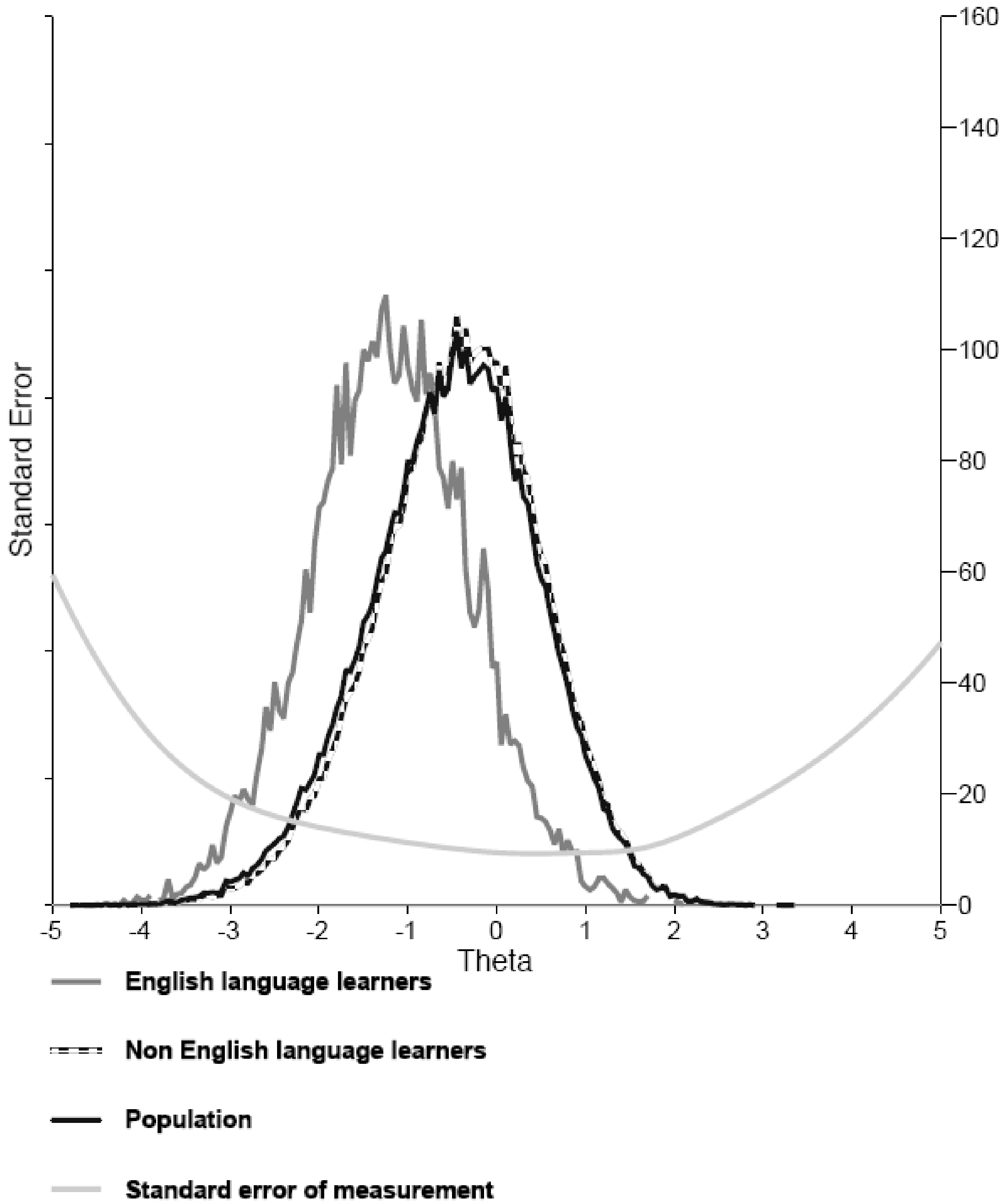
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-14. Grade 4 geometry: Standard error of measurement and achievement distributions by student disability status



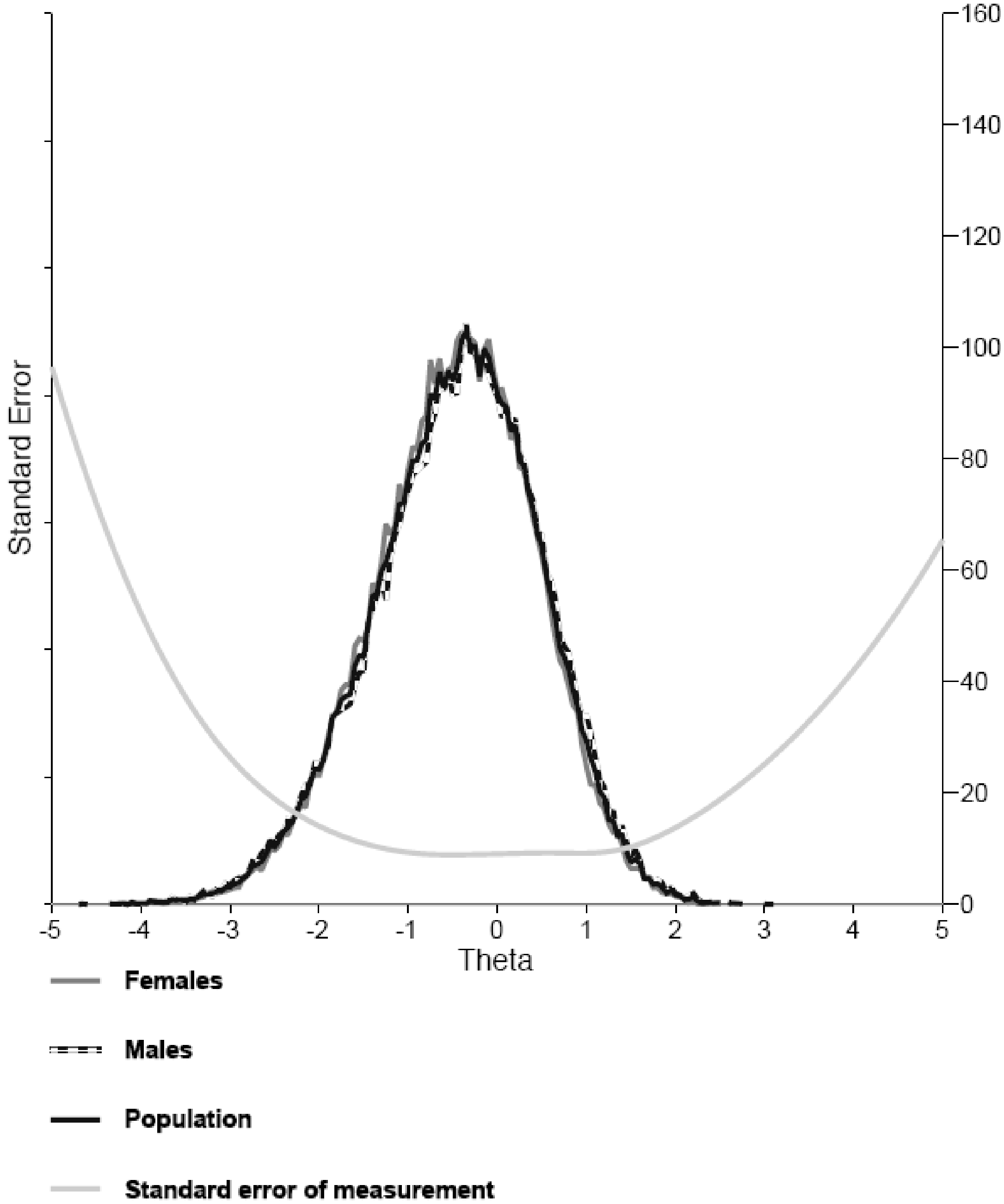
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-15. Grade 4 geometry: Standard error of measurement and achievement distributions by ELL status



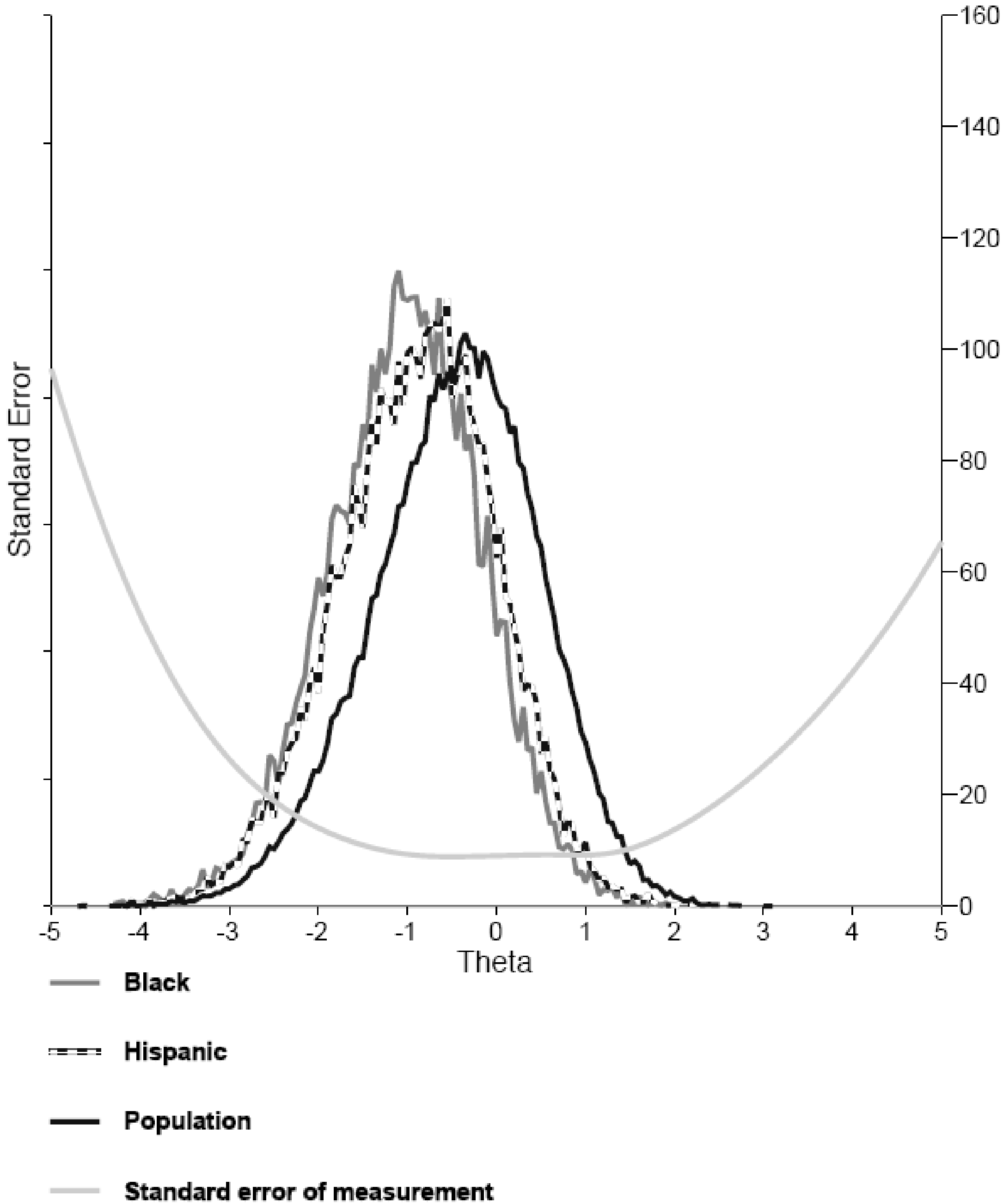
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-16. Grade 4 data analysis and probability: Standard error of measurement and achievement distributions by gender



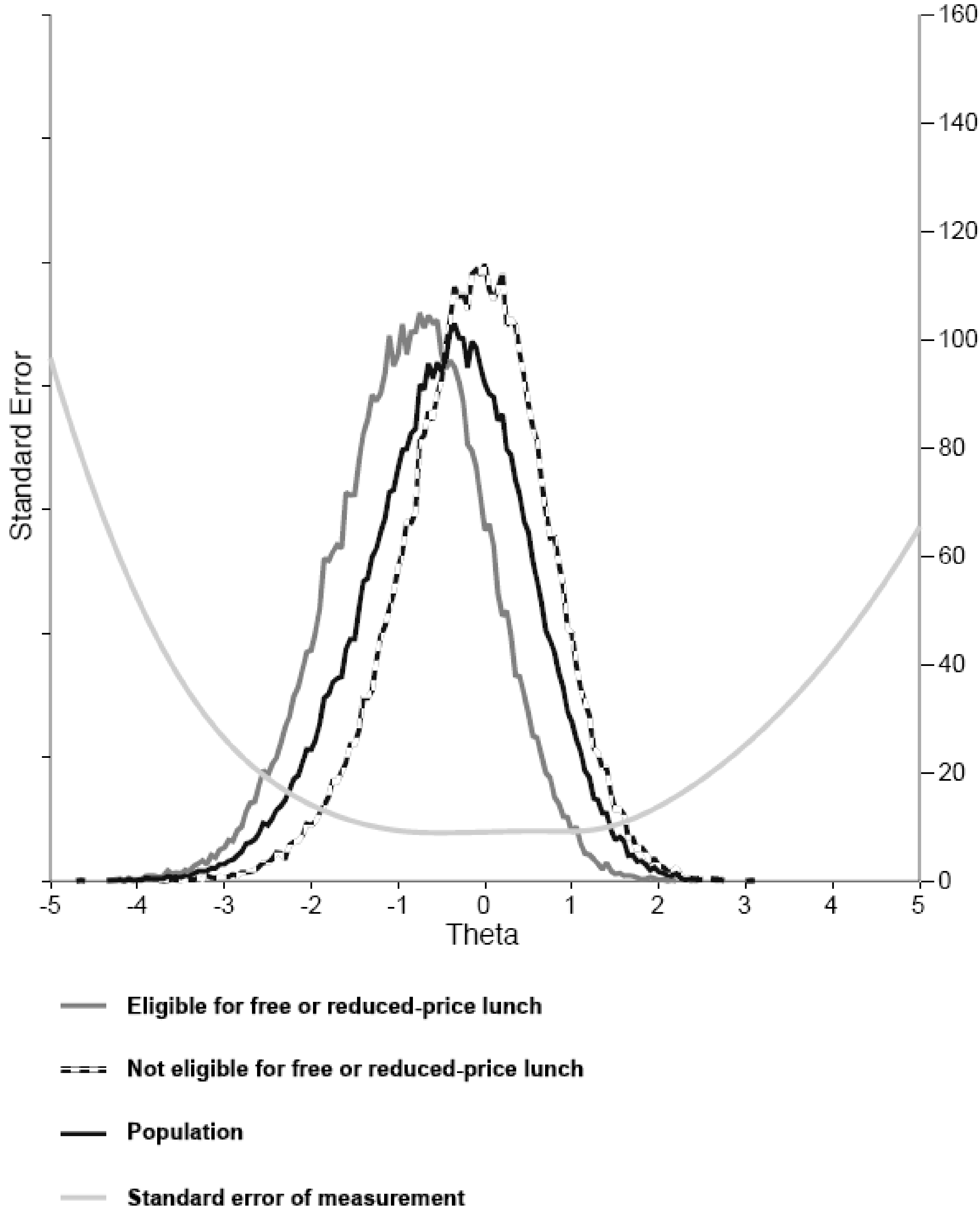
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-17. Grade 4 data analysis and probability: Standard error of measurement and achievement distributions by race/ethnicity



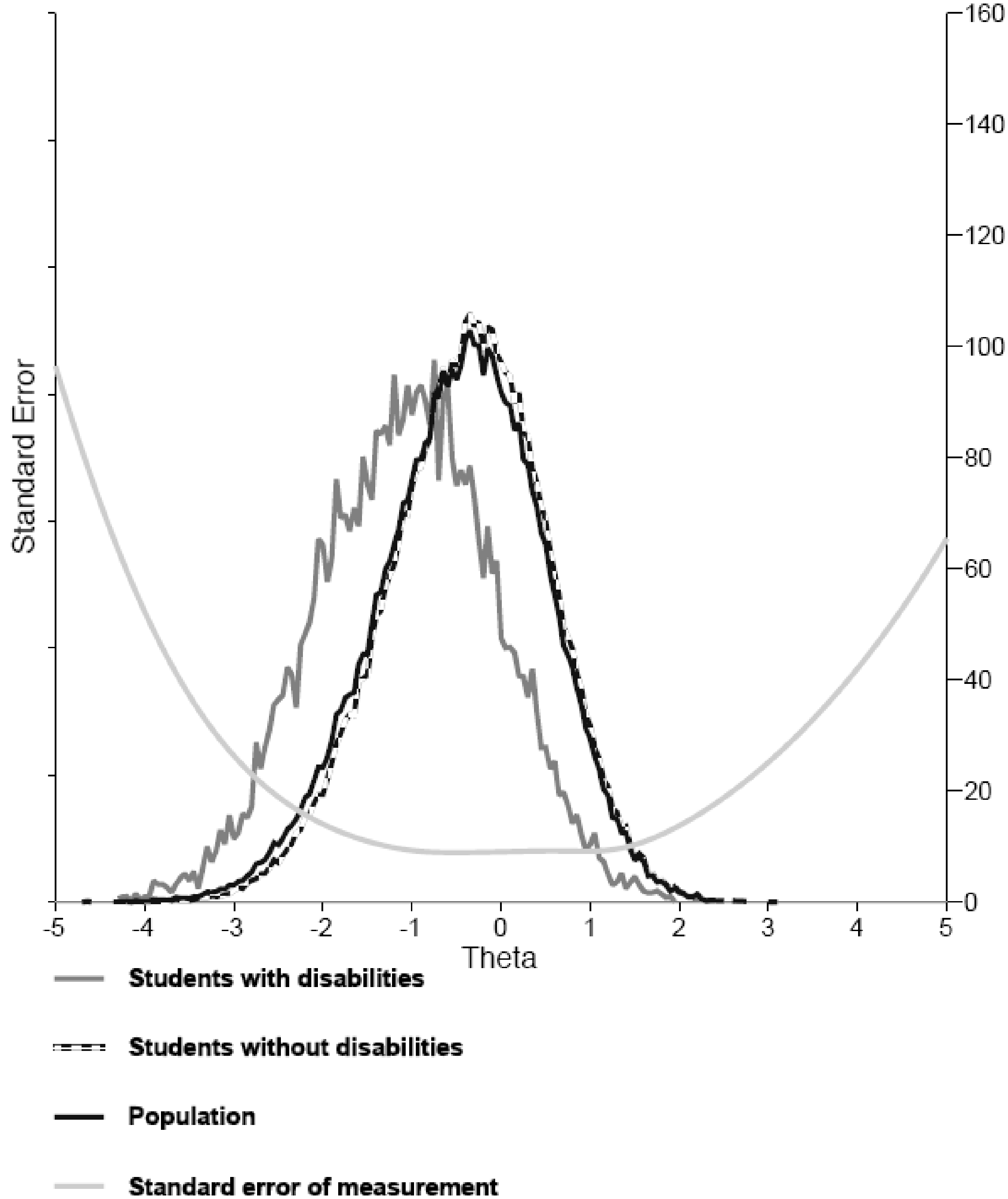
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-18. Grade 4 data analysis and probability: Standard error of measurement and achievement distributions by eligibility for free/reduced price lunch



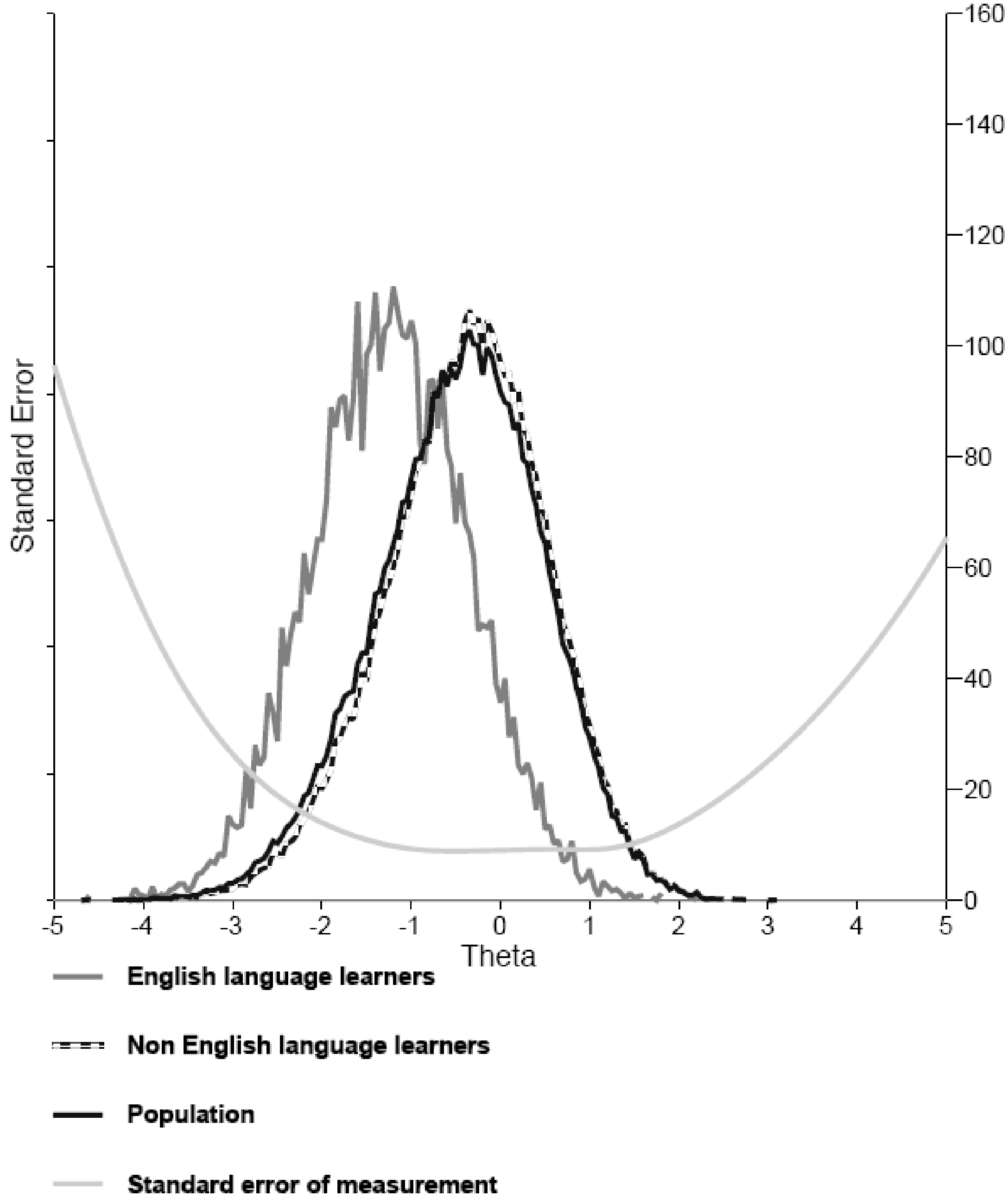
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-19. Grade 4 data analysis and probability: Standard error of measurement and achievement distributions by student disability status



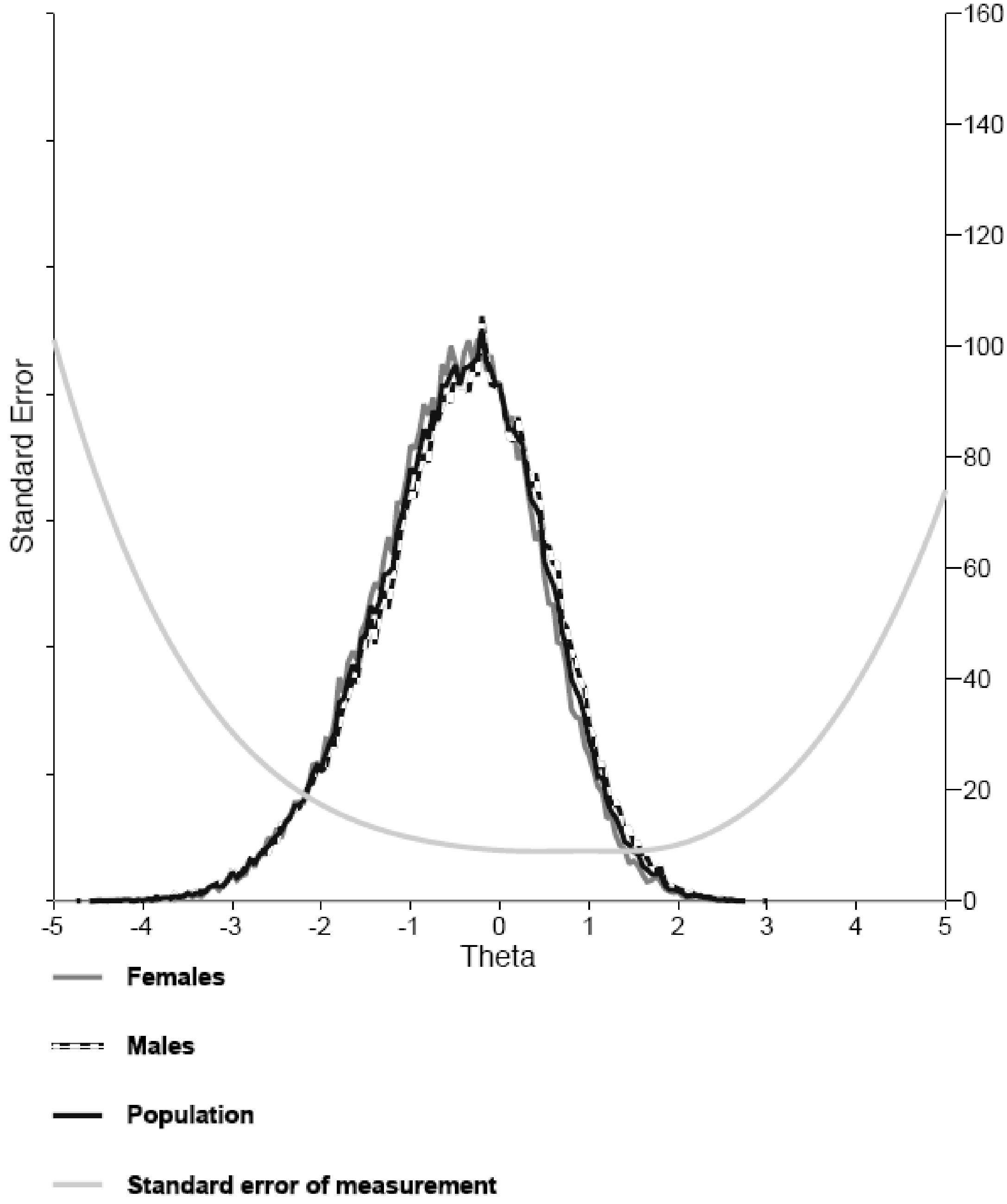
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-20. Grade 4 data analysis and probability: Standard error of measurement and achievement distributions by ELL status



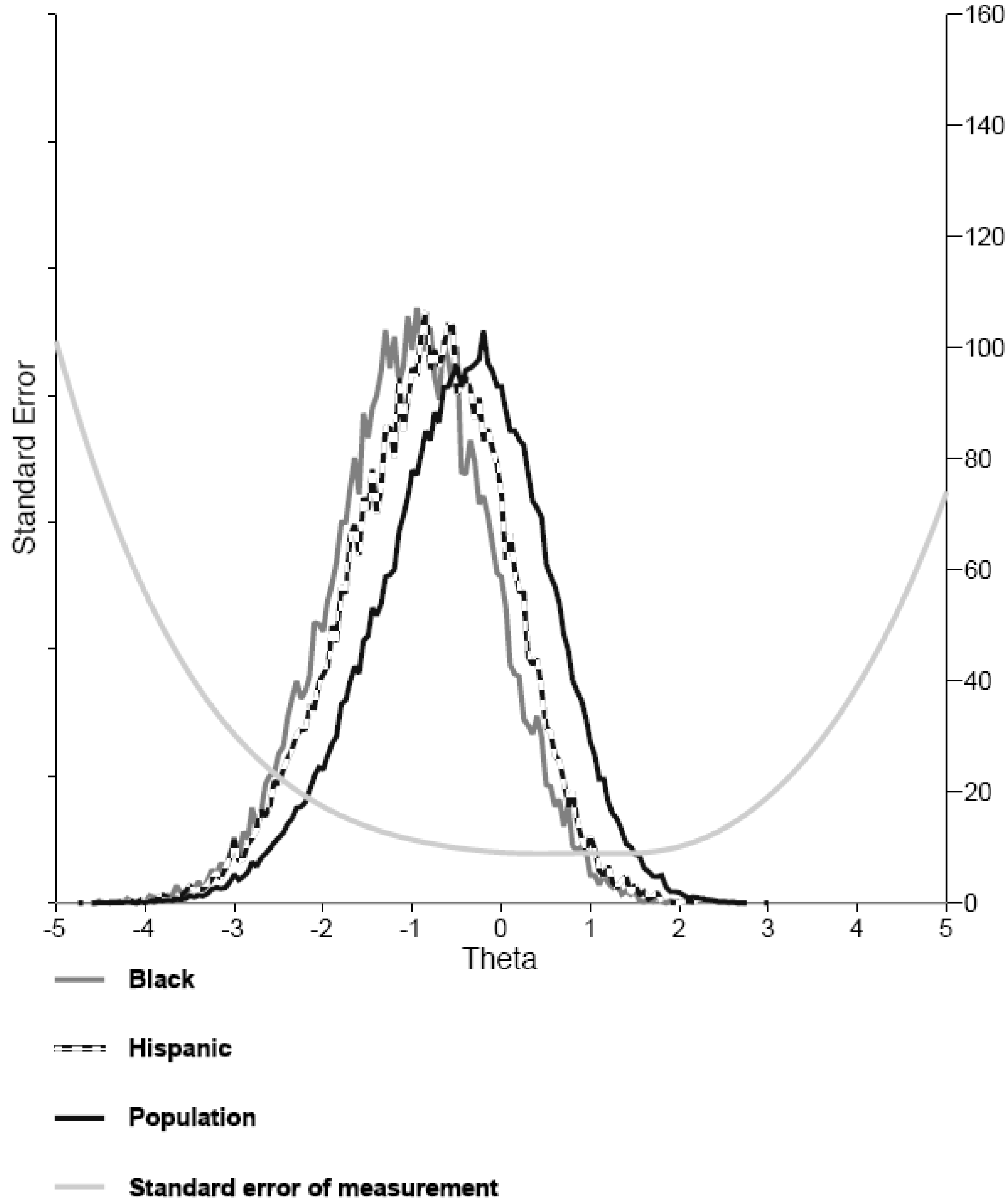
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-21. Grade 4 algebra: Standard error of measurement and achievement distributions by gender



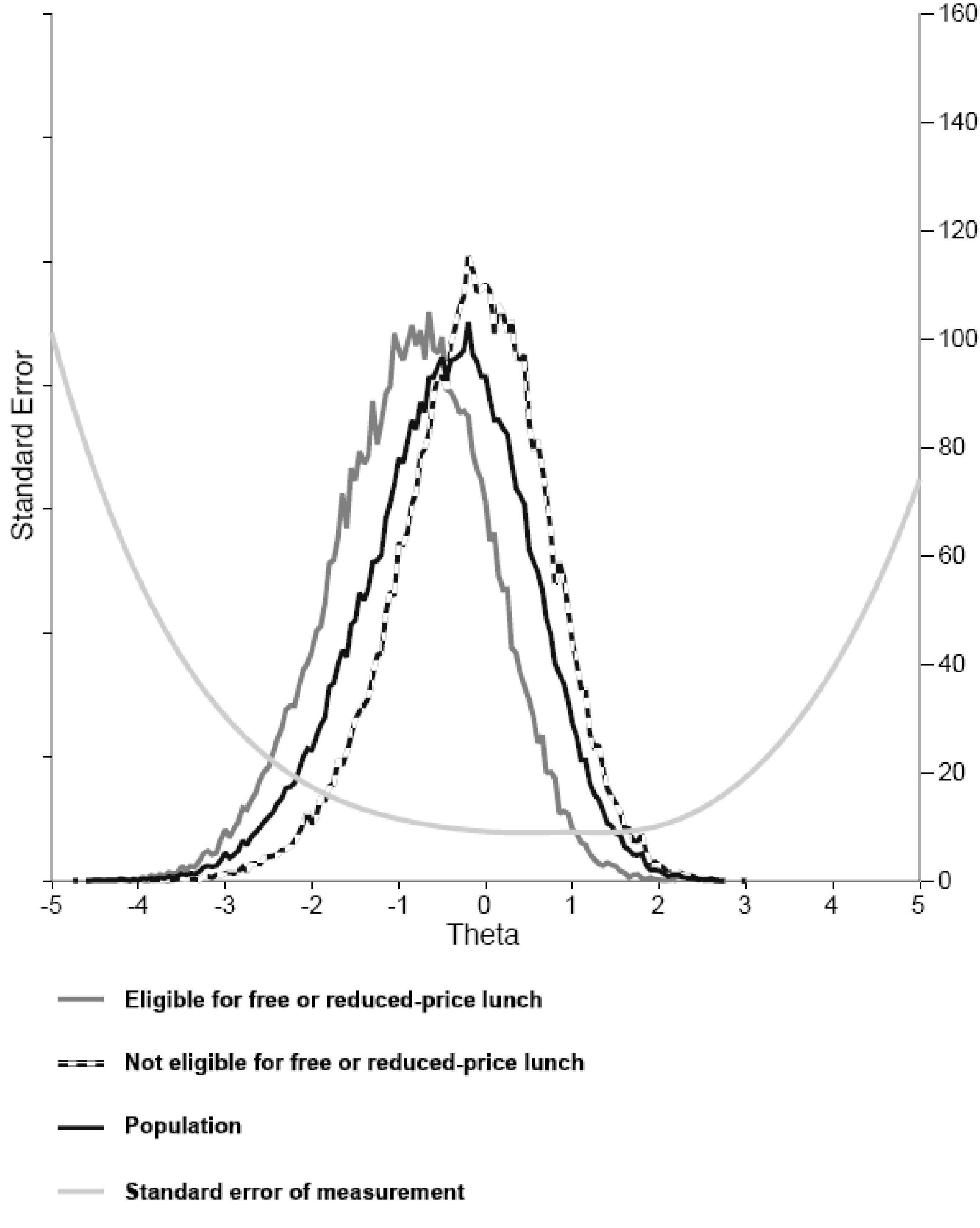
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-22. Grade 4 algebra: Standard error of measurement and achievement distributions by race/ethnicity



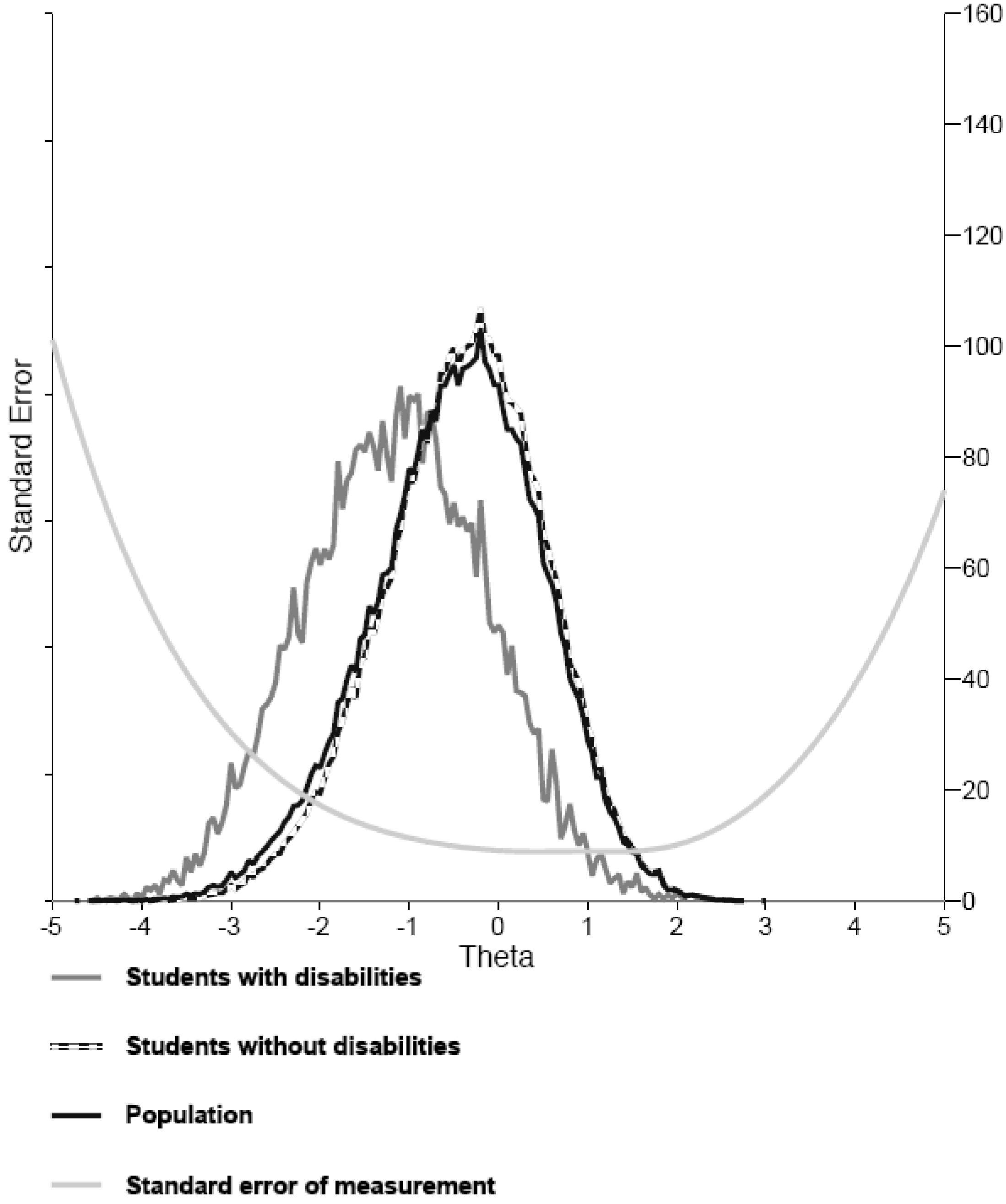
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-23. Grade 4 algebra: Standard error of measurement and achievement distributions by eligibility for free/reduced price lunch



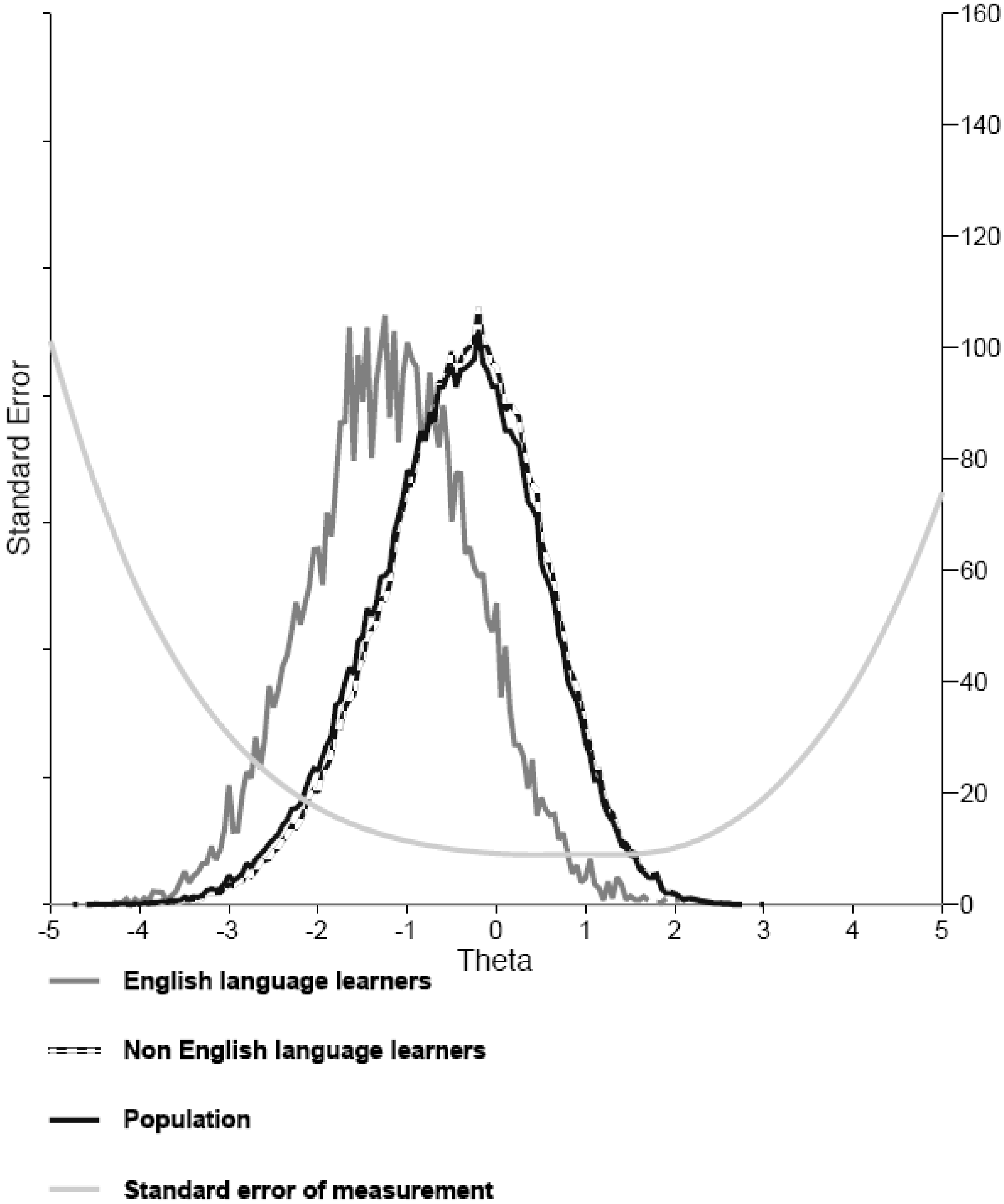
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-24. Grade 4 algebra: Standard error of measurement and achievement distributions by student disability status



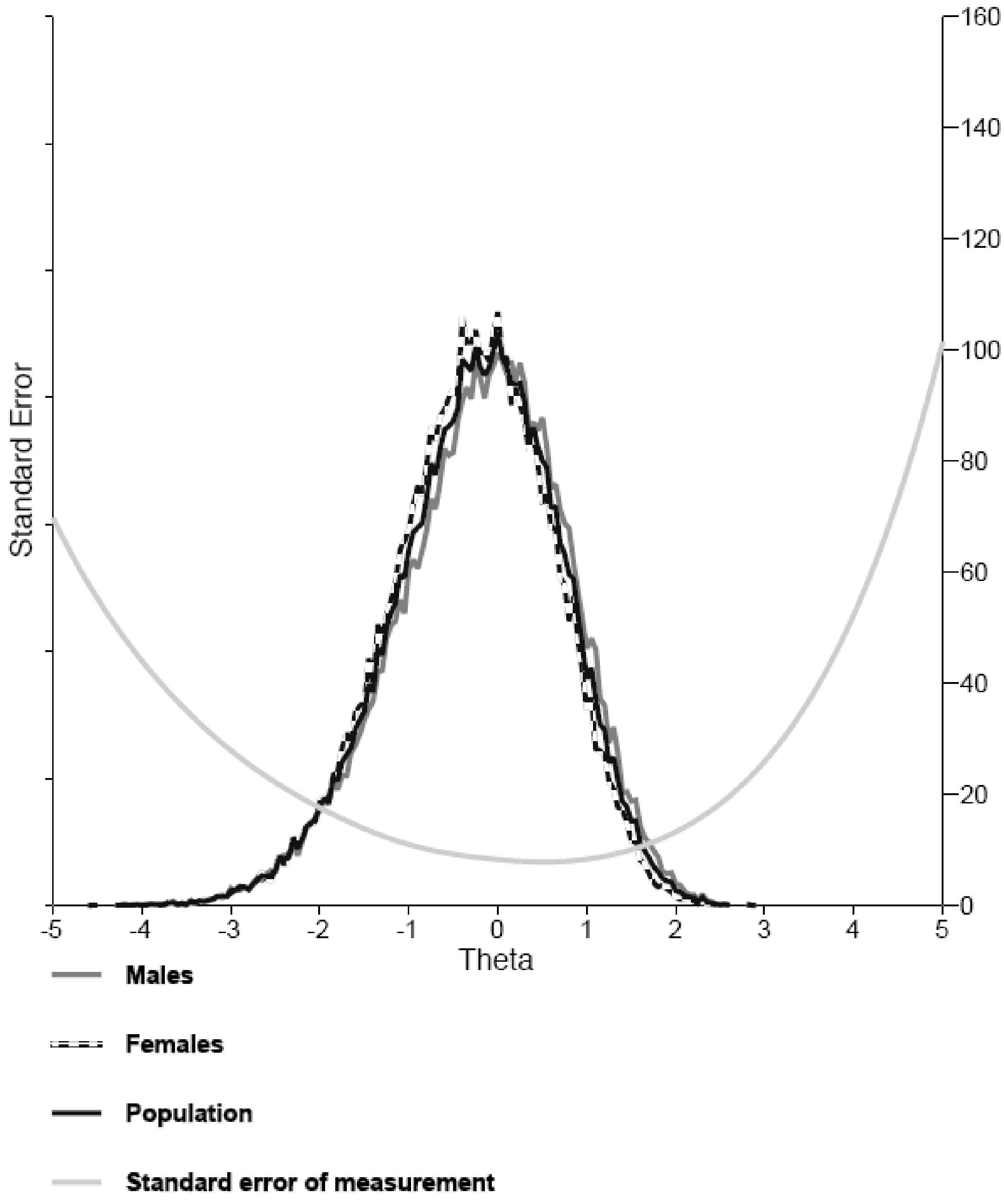
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-25. Grade 4 algebra: Standard error of measurement and achievement distributions by ELL status



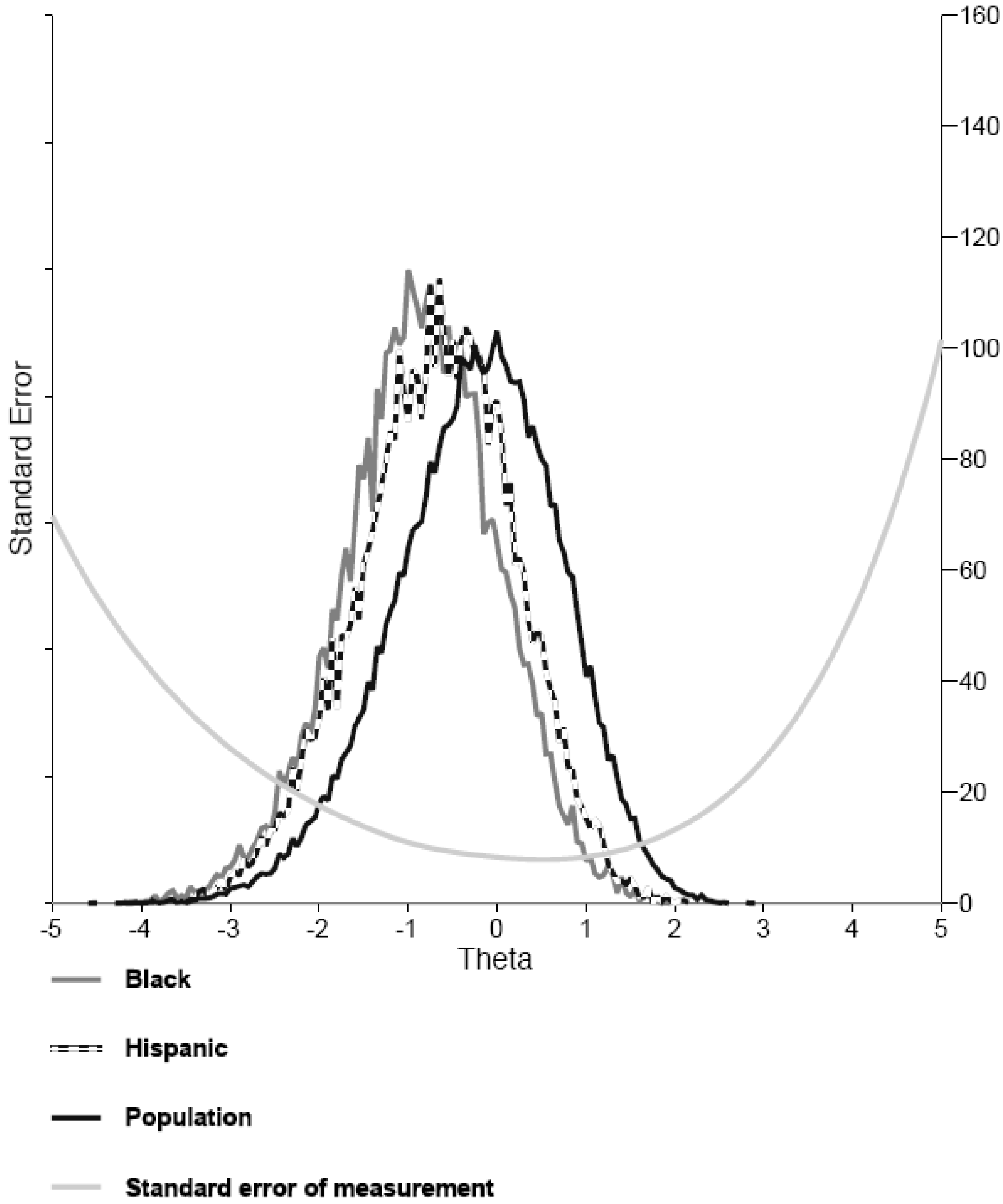
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-26. Grade 8 numbers and operations: Standard error of measurement and achievement distributions by gender



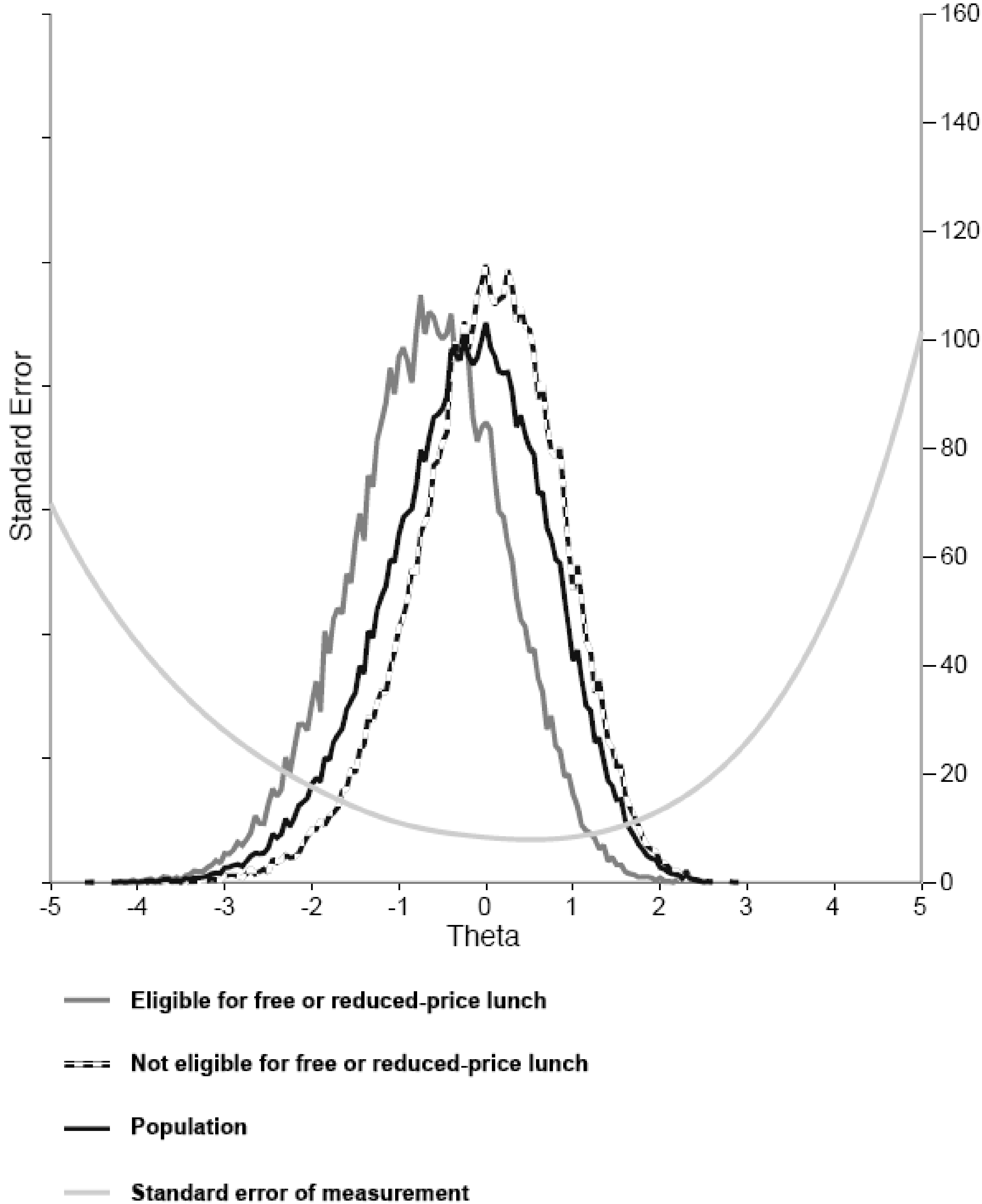
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-27. Grade 8 numbers and operations: Standard error of measurement and achievement distributions by race/ethnicity



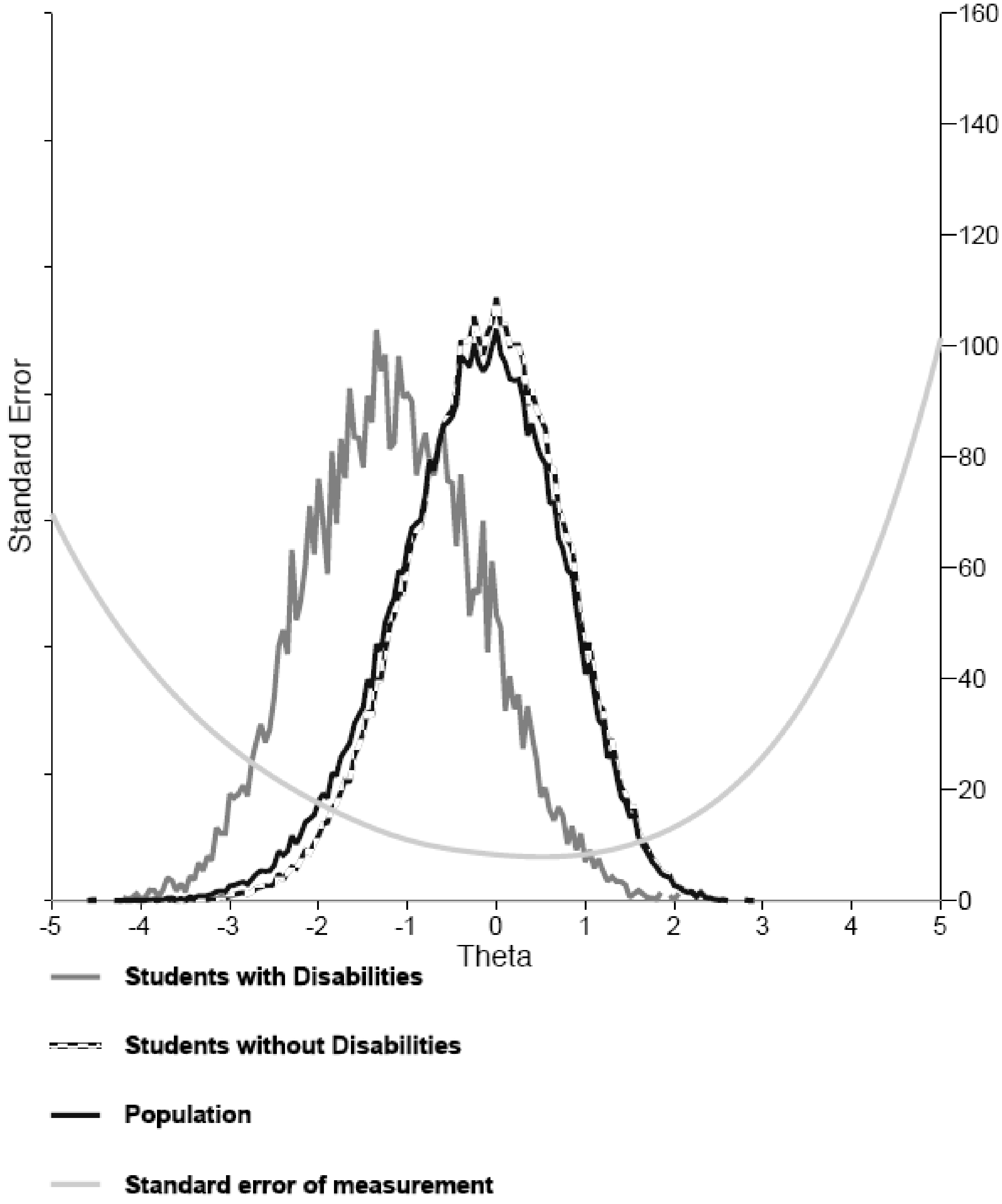
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-28. Grade 8 numbers and operations: Standard error of measurement and achievement distributions by eligibility for free/reduced price lunch



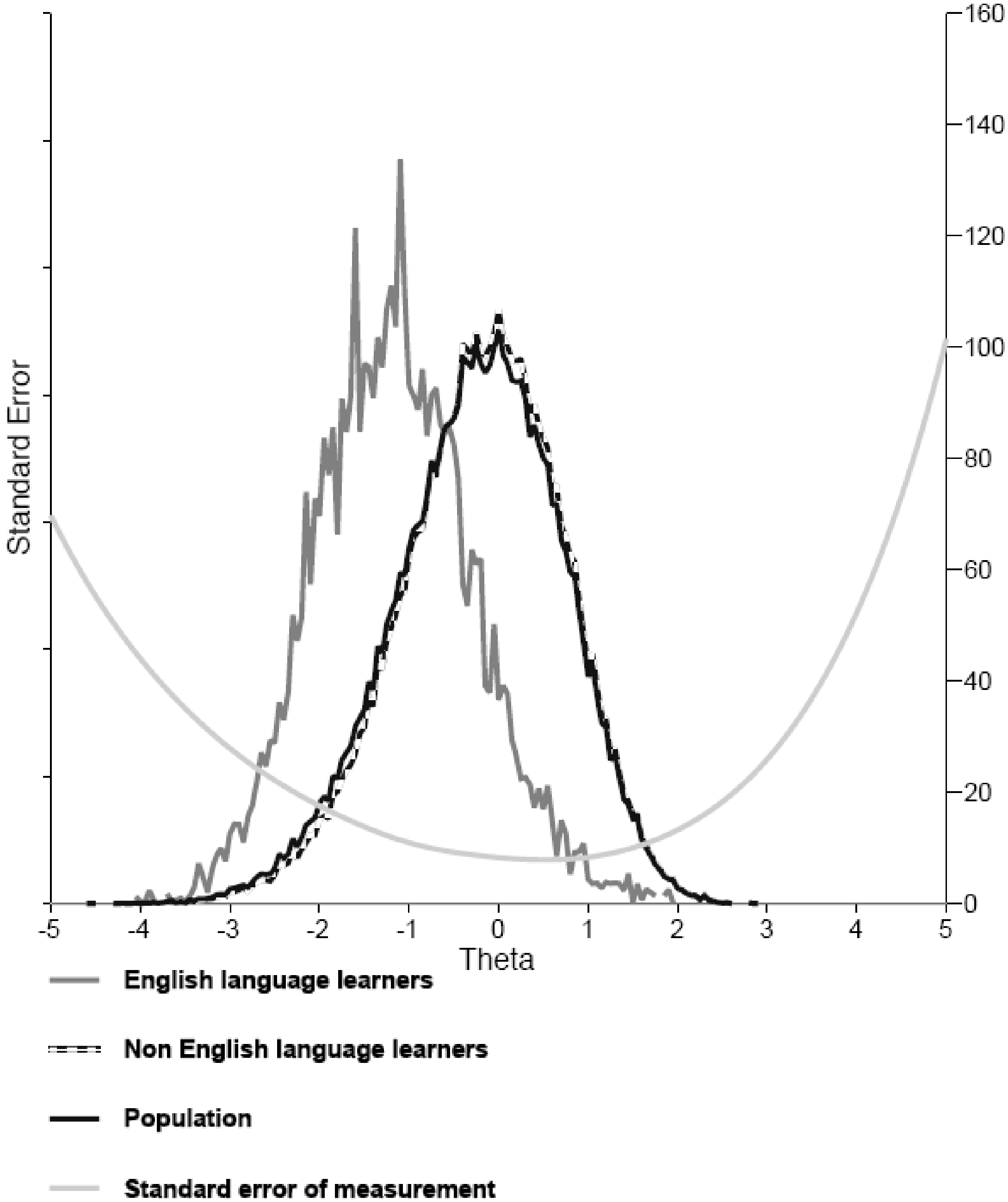
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-29. Grade 8 number and operations: Standard error of measurement and achievement distributions by student disability status



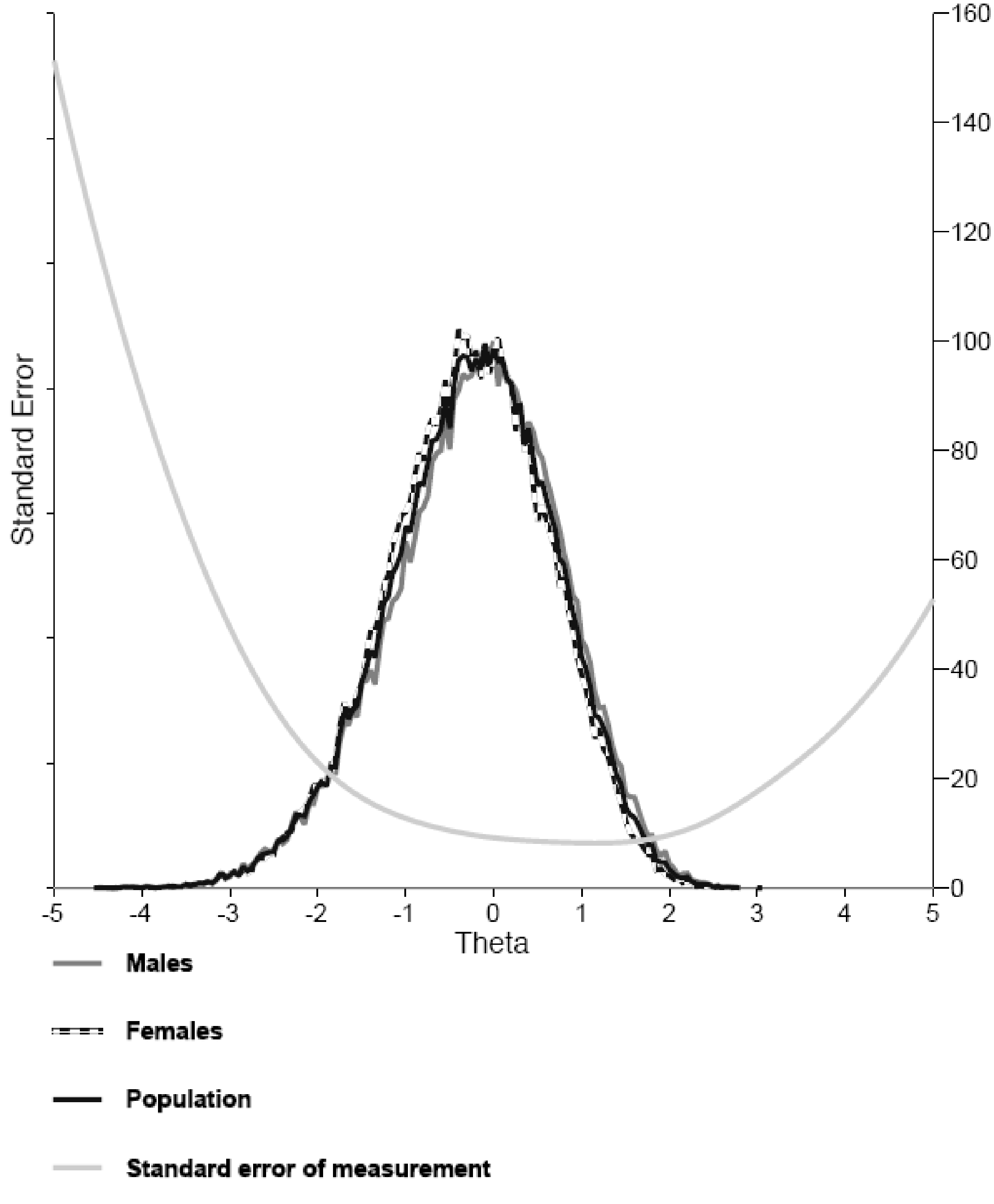
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-30. Grade 8 number and operations: Standard error of measurement and achievement distributions by ELL status



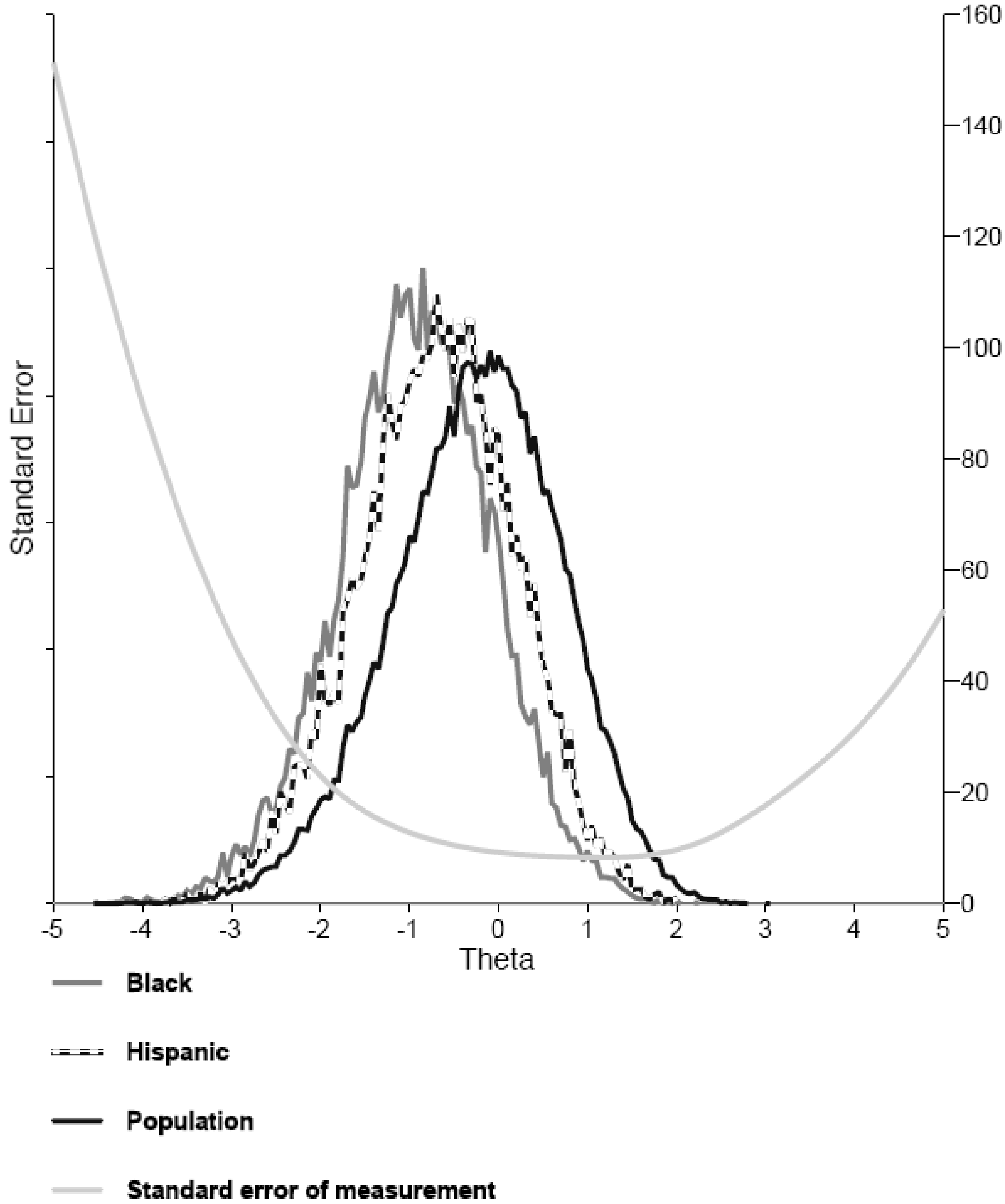
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-31. Grade 8 measurement: Standard error of measurement and achievement distributions by gender



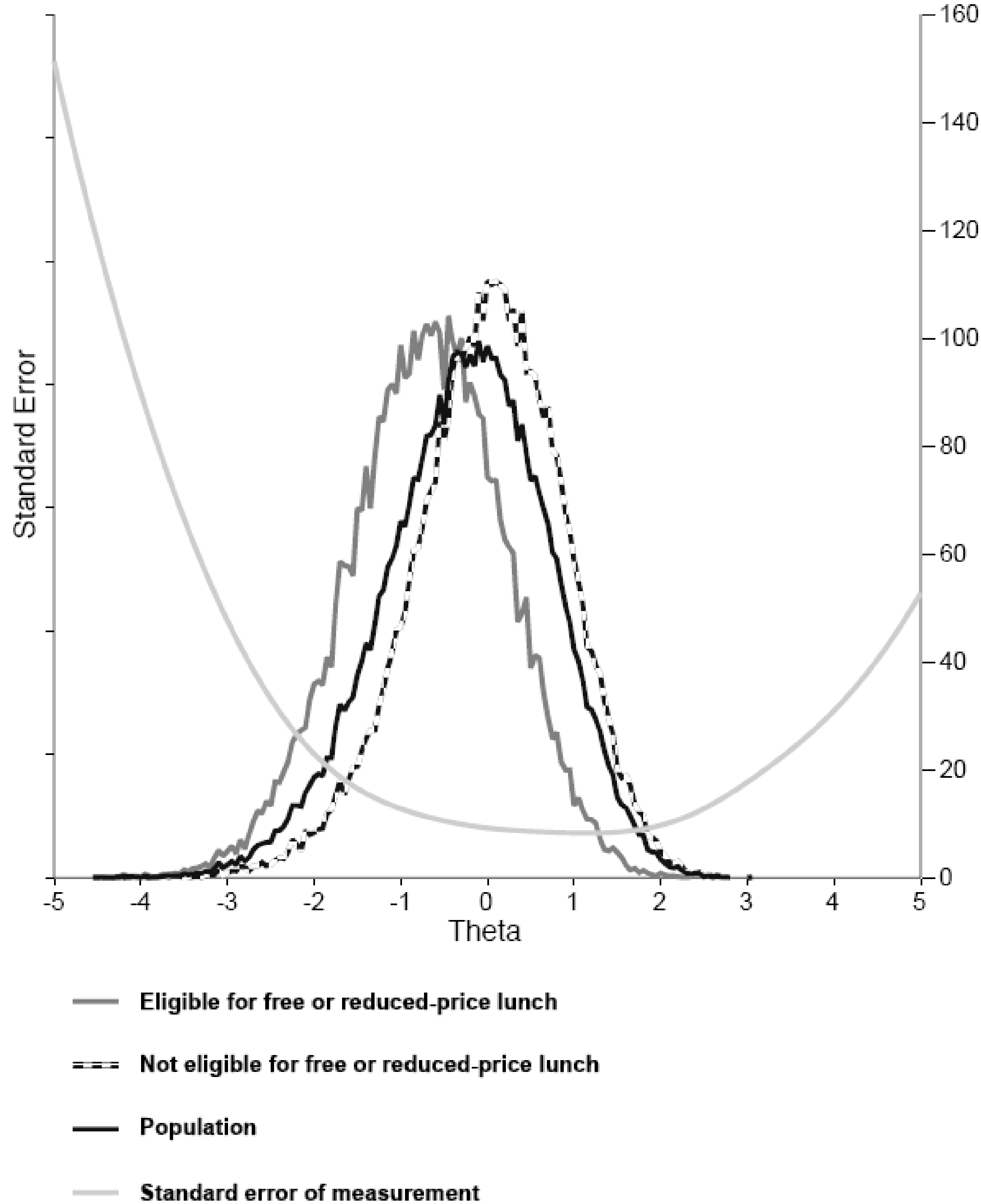
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-32. Grade 8 measurement: Standard error of measurement and achievement distributions by race/ethnicity



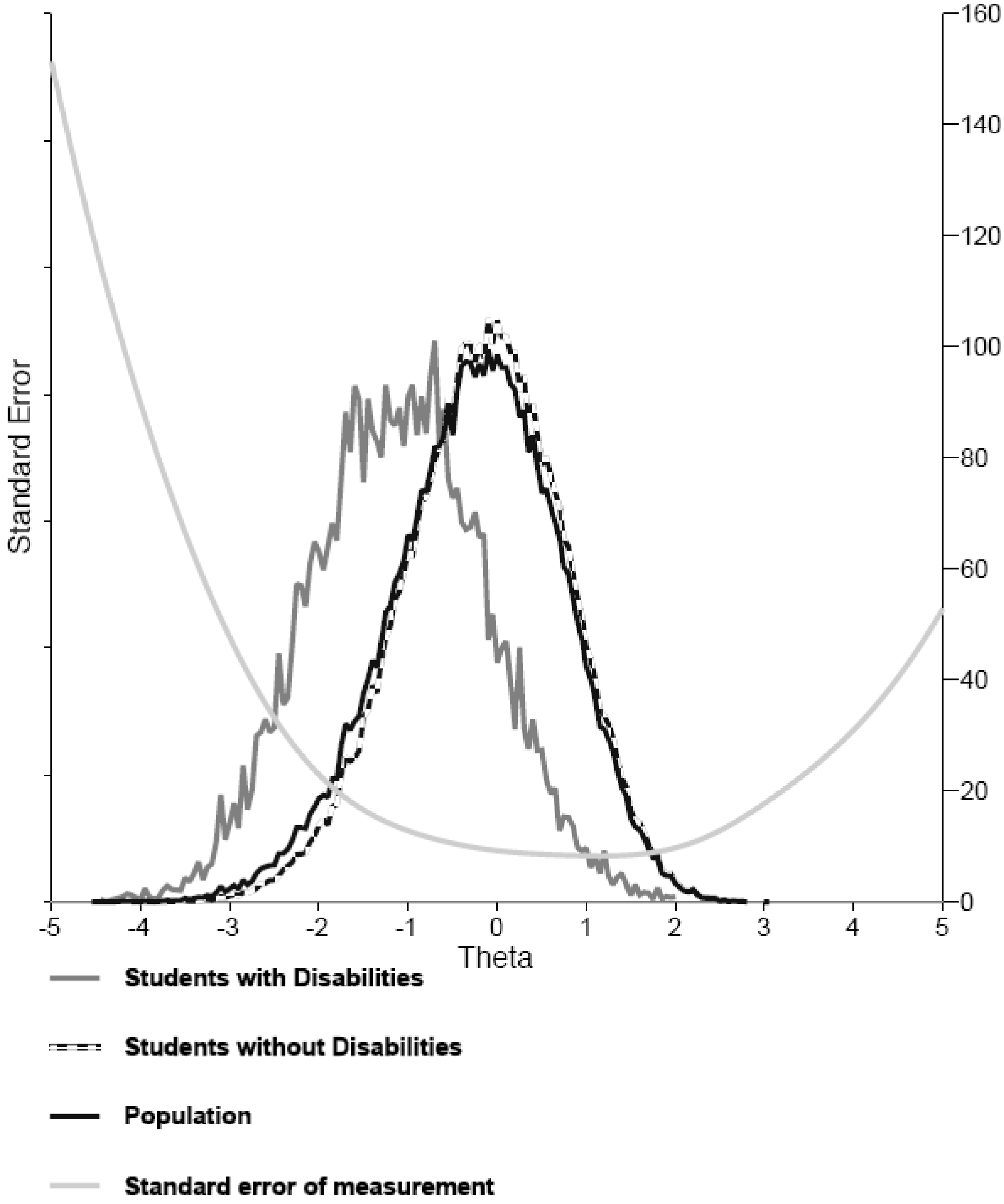
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-33. Grade 8 measurement: Standard error of measurement and achievement distributions by eligibility for free/reduced price lunch



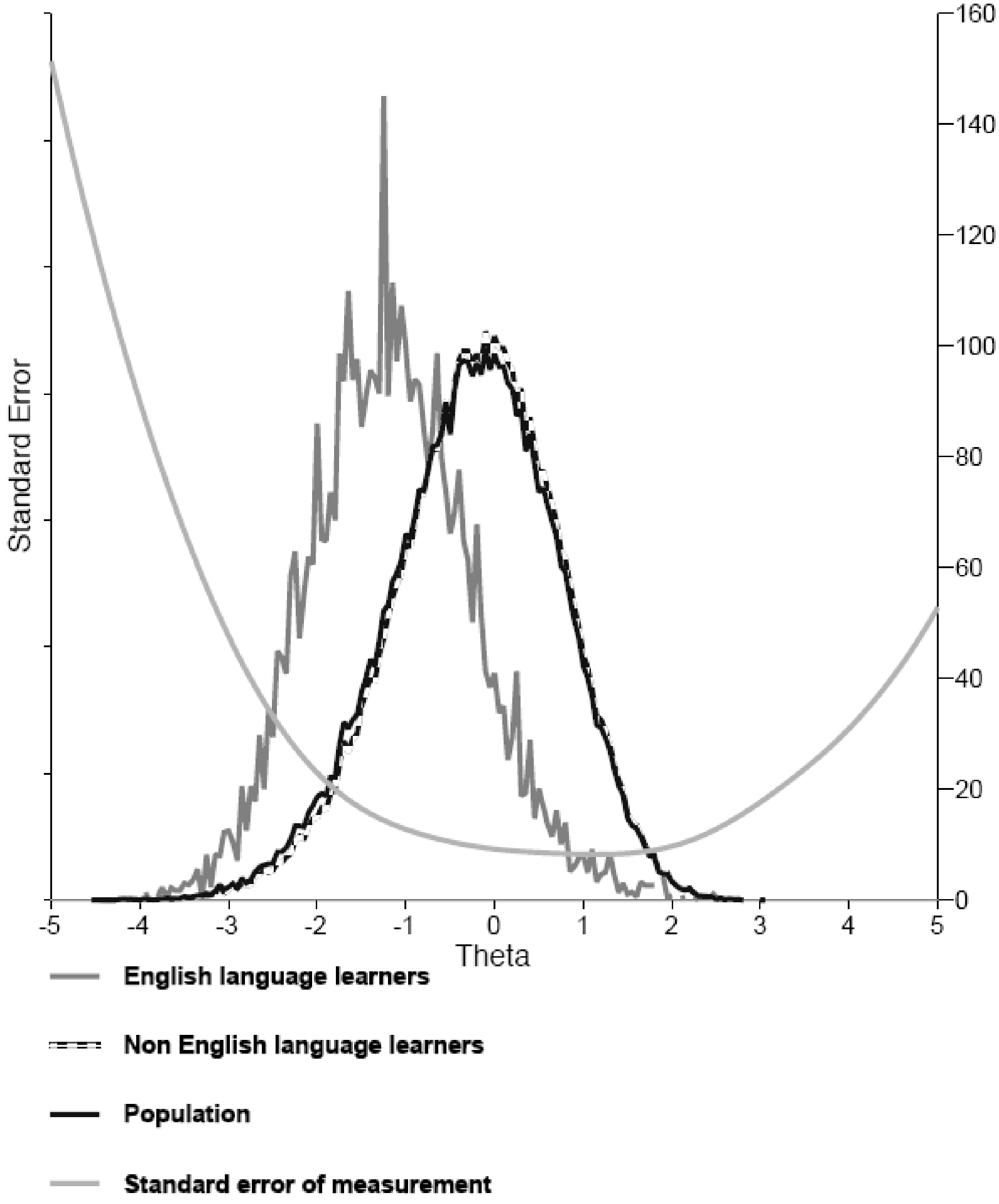
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-34. Grade 8 measurement: Standard error of measurement and achievement distributions by student disability status



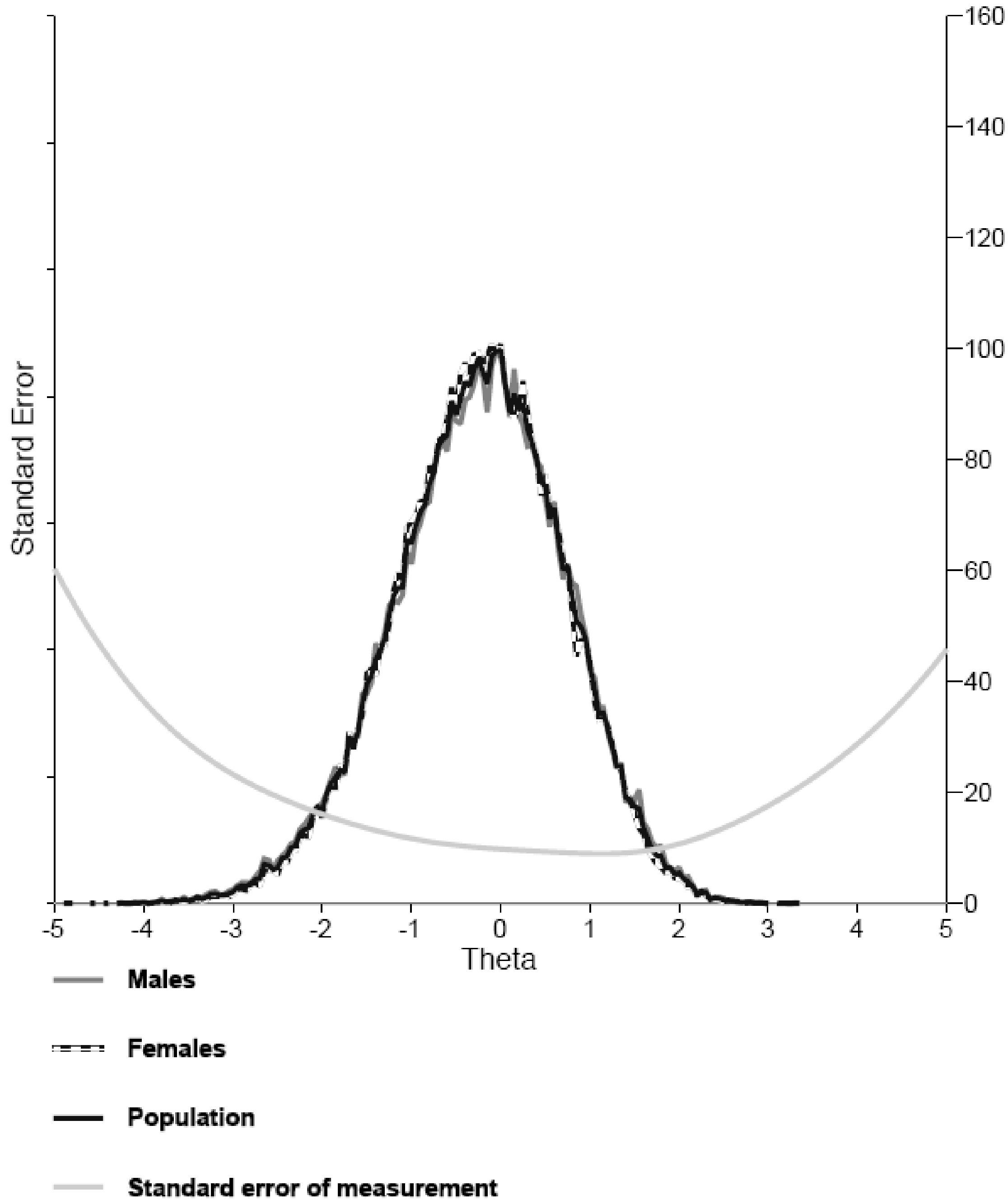
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-35. Grade 8 measurement: Standard error of measurement and achievement distributions by ELL status



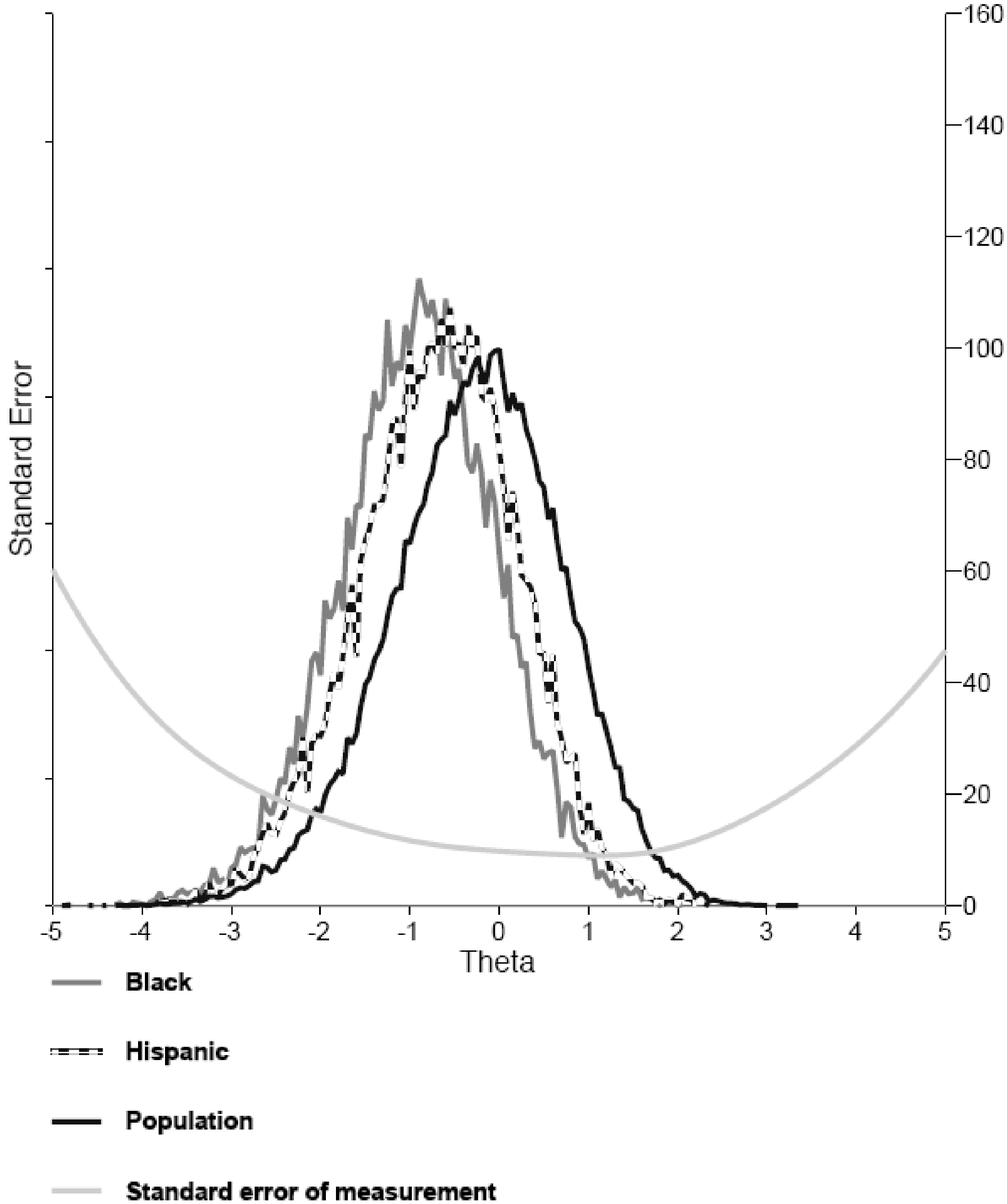
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-36. Grade 8 geometry: Standard error of measurement and achievement distributions by gender



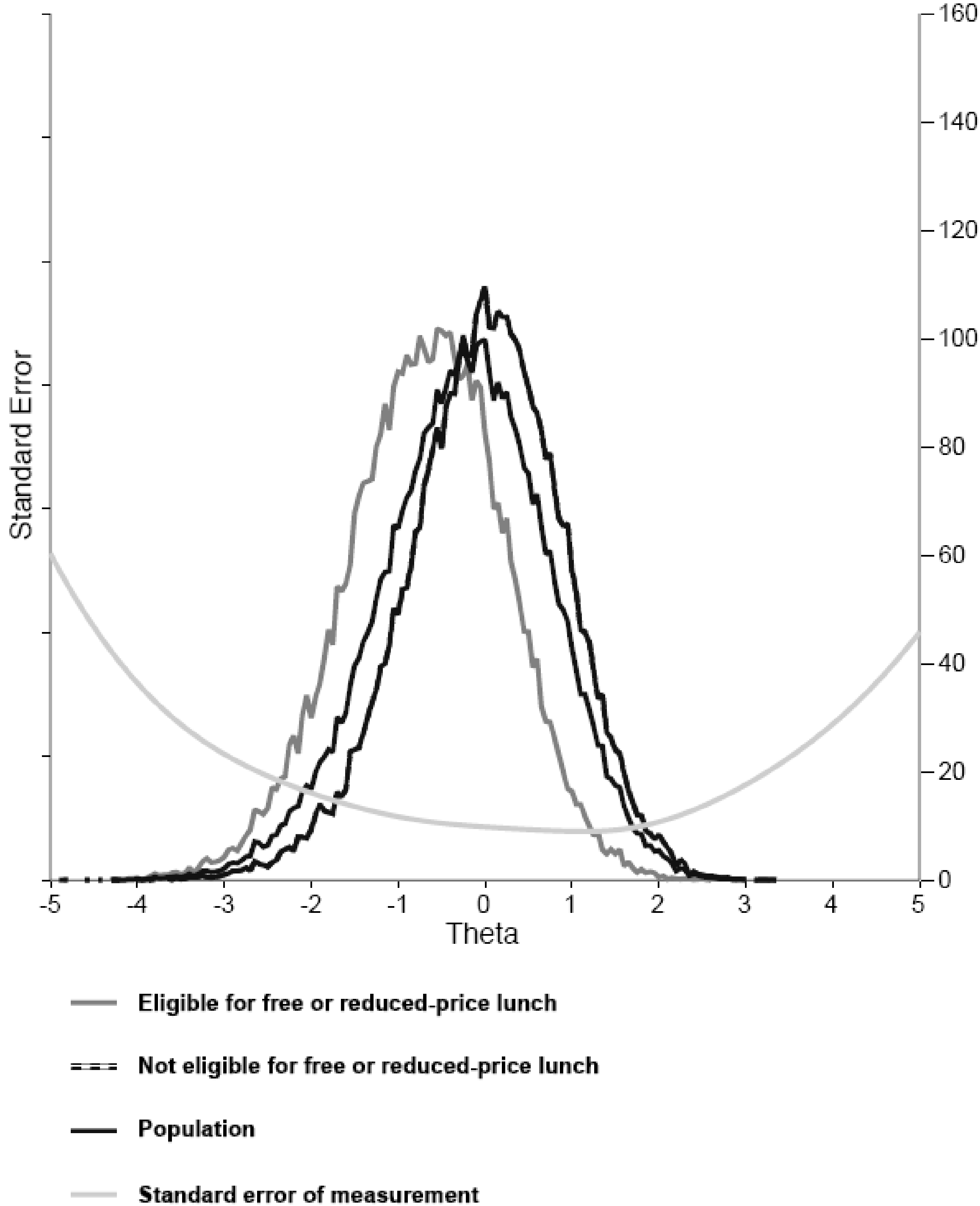
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-37. Grade 8 geometry: Standard error of measurement and achievement distributions by race/ethnicity



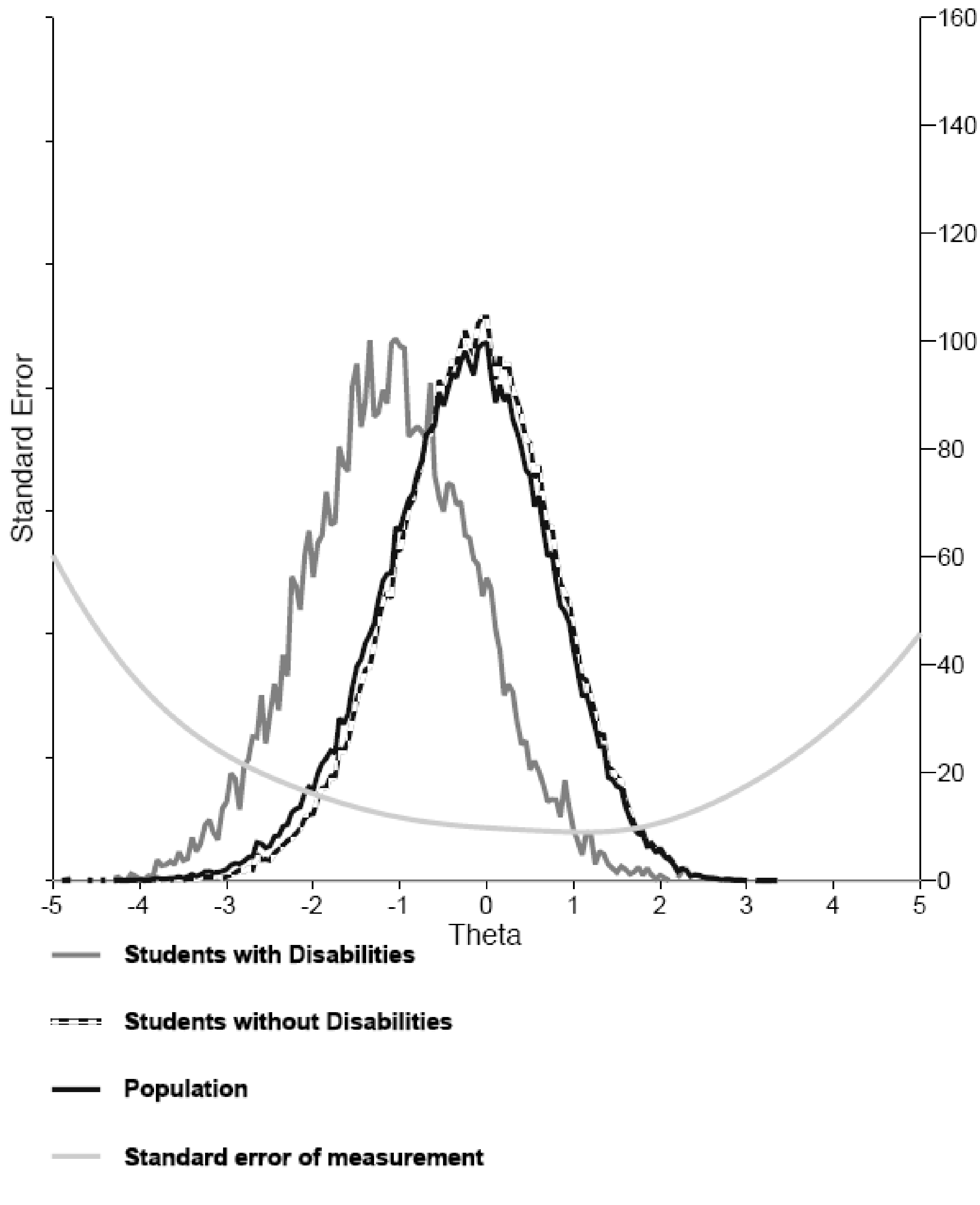
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-38. Grade 8 geometry: Standard error of measurement and achievement distributions by eligibility for free/reduced price lunch



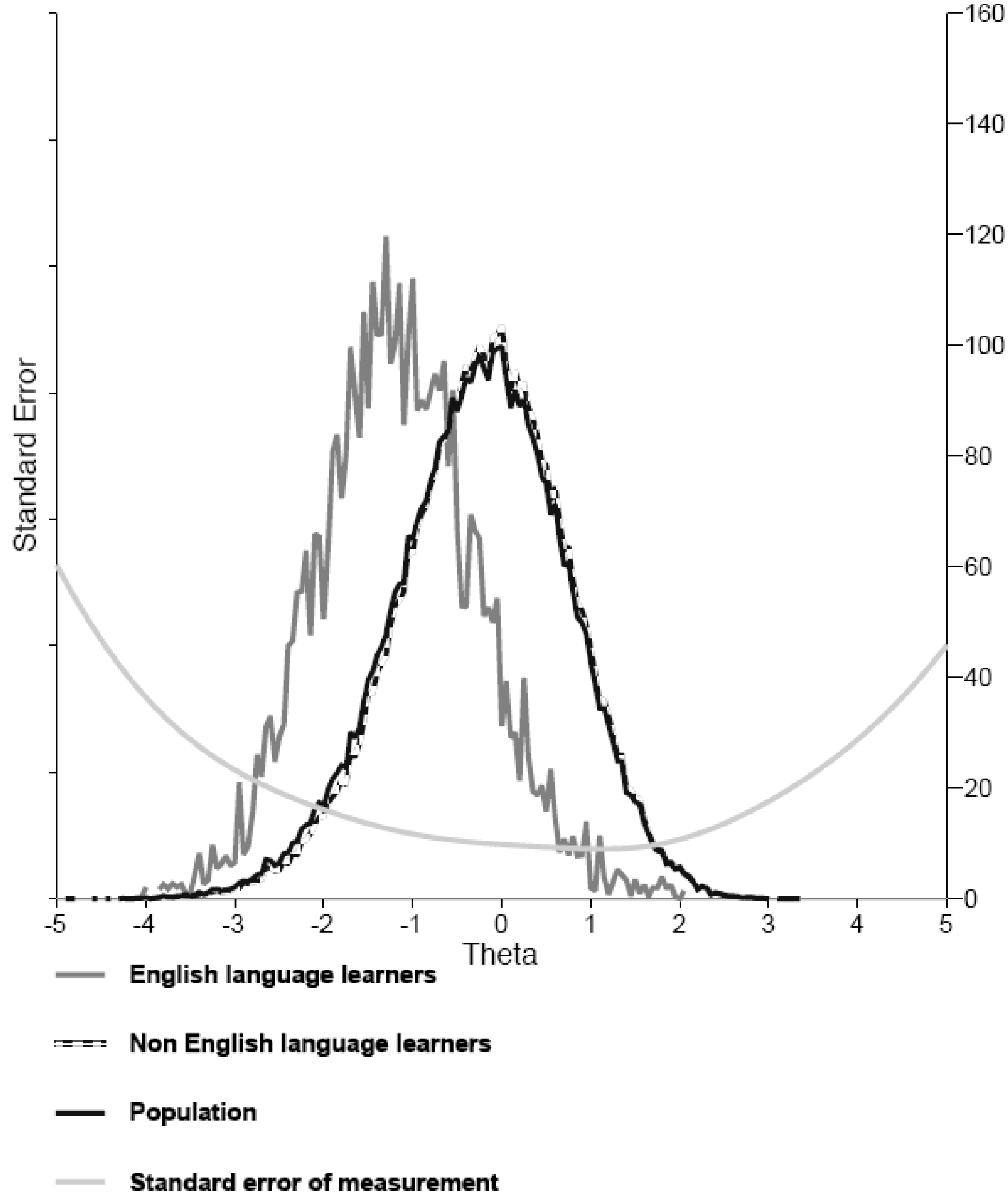
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-39. Grade 8 geometry: Standard error of measurement and achievement distributions by student disability status



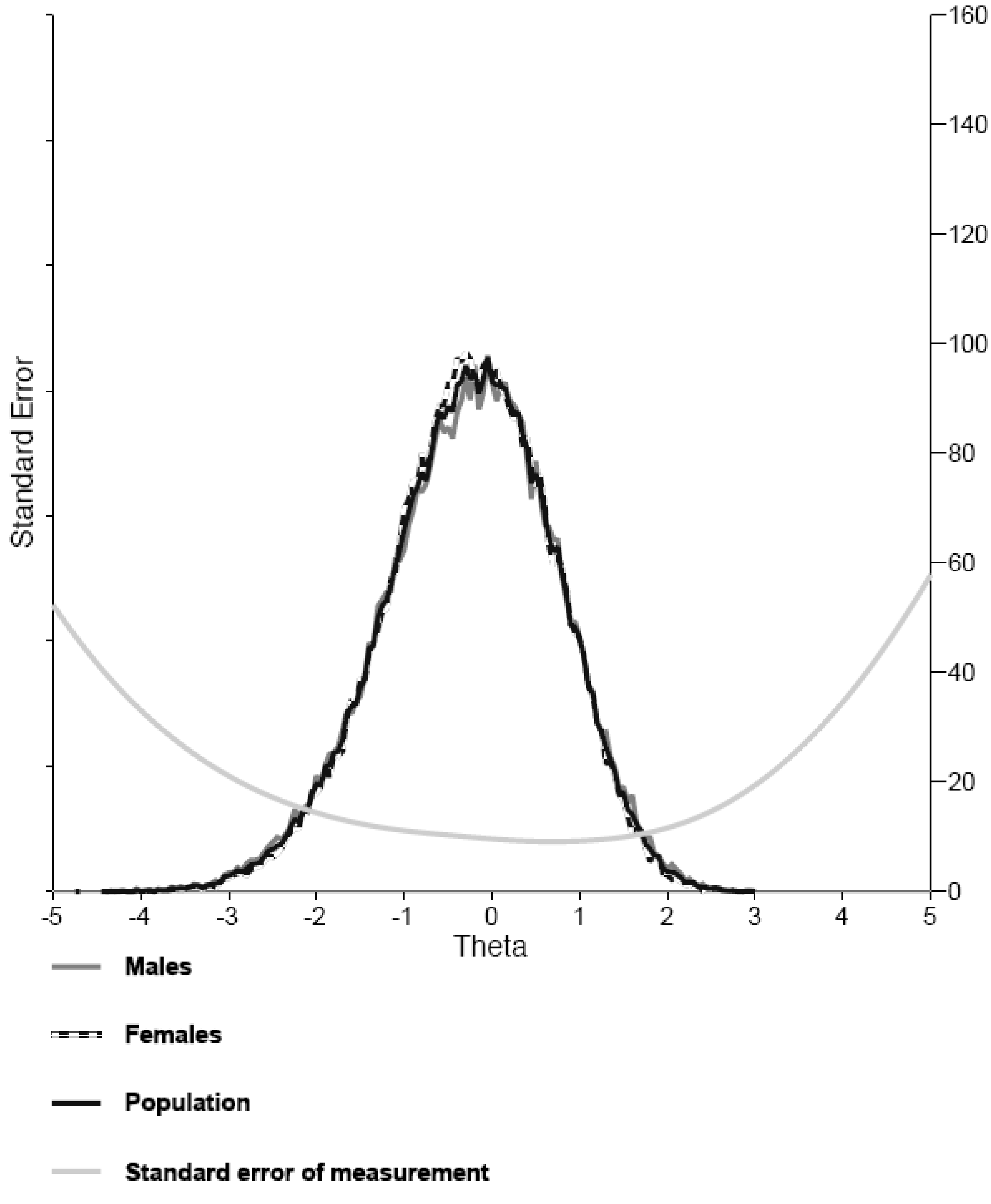
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-40. Grade 8 geometry: Standard error of measurement and achievement distributions by ELL status



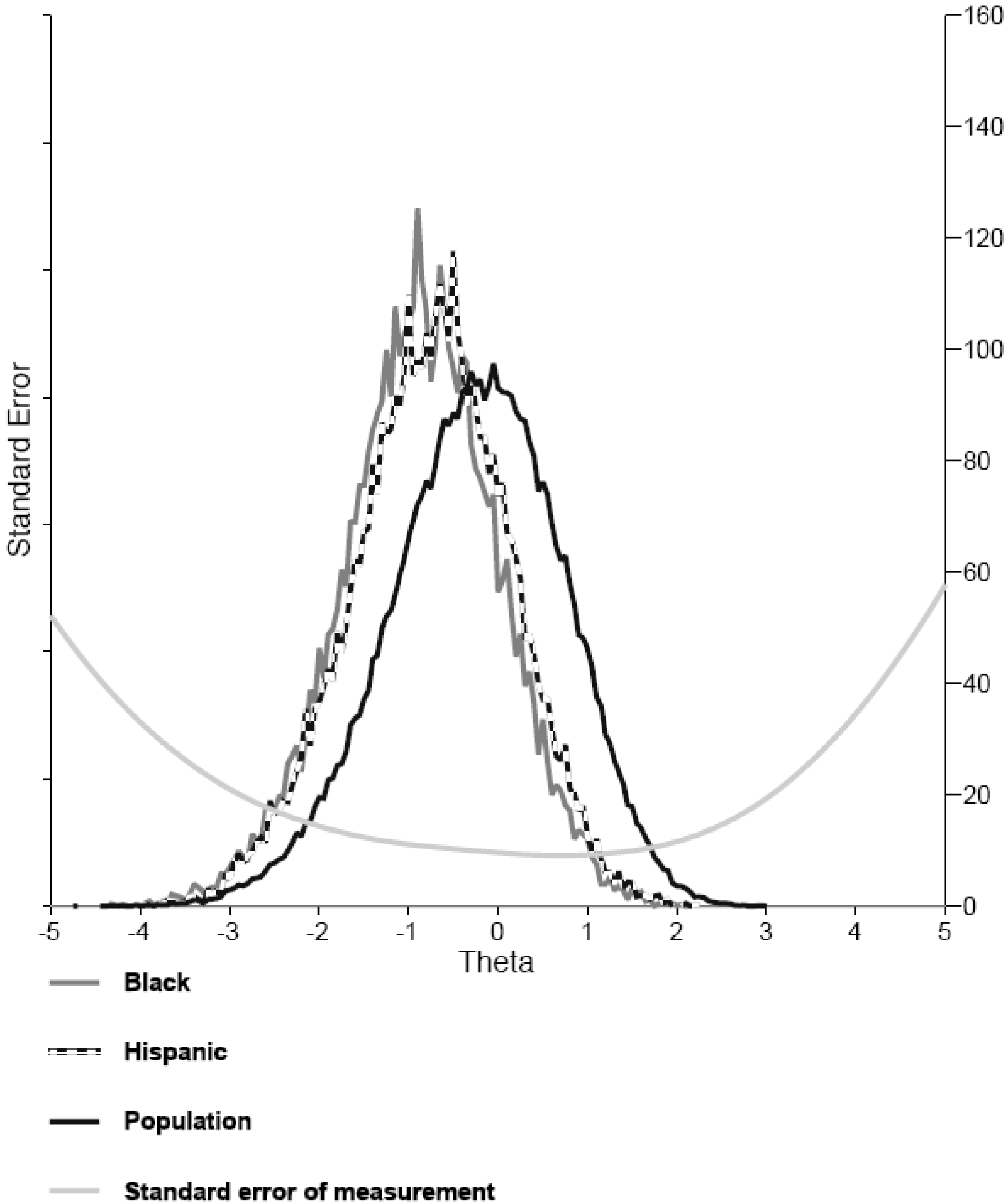
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-41. Grade 8 data analysis and probability: Standard error of measurement and achievement distributions by gender



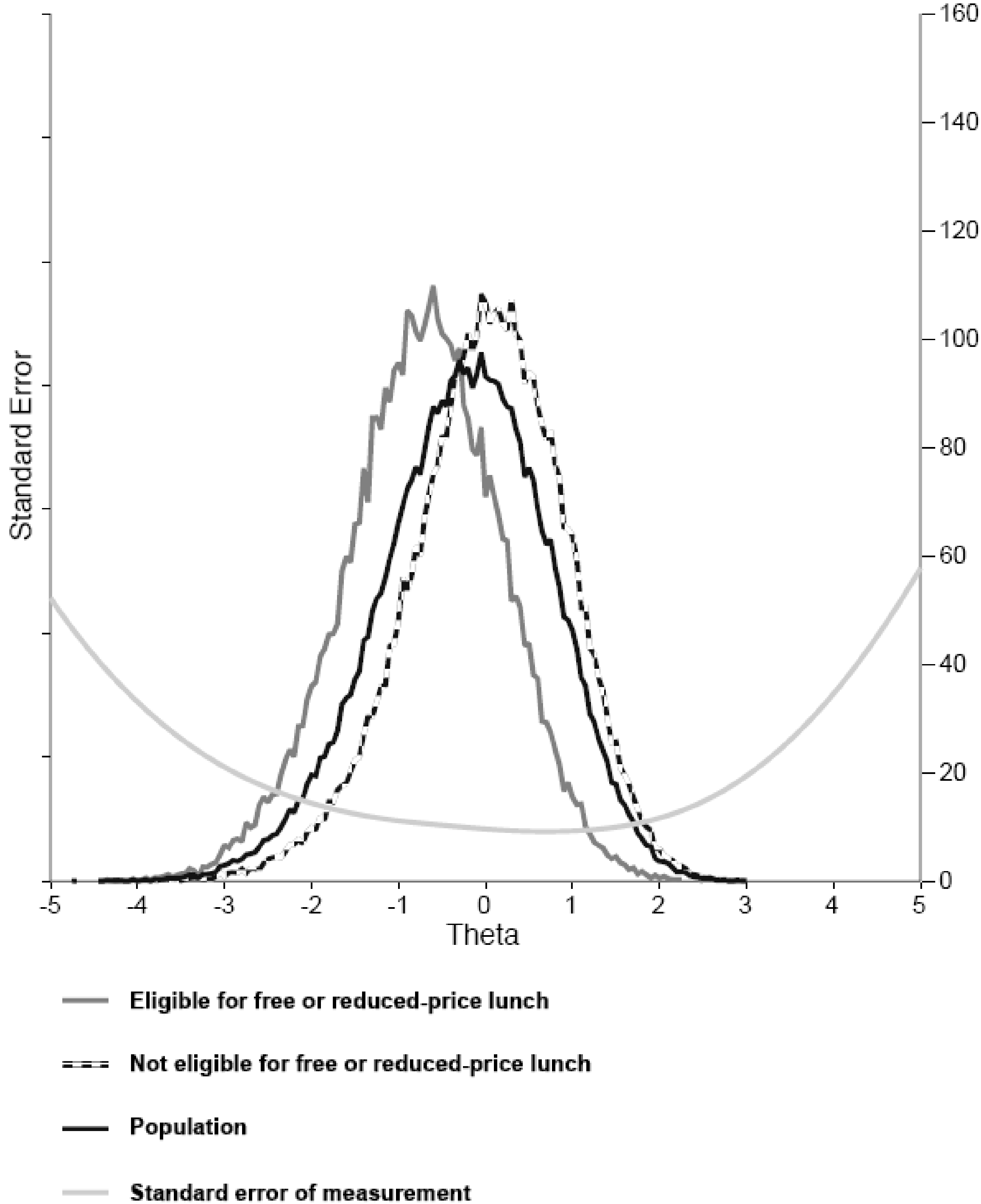
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-42. Grade 8 data analysis and probability: Standard error of measurement and achievement distributions by race/ethnicity



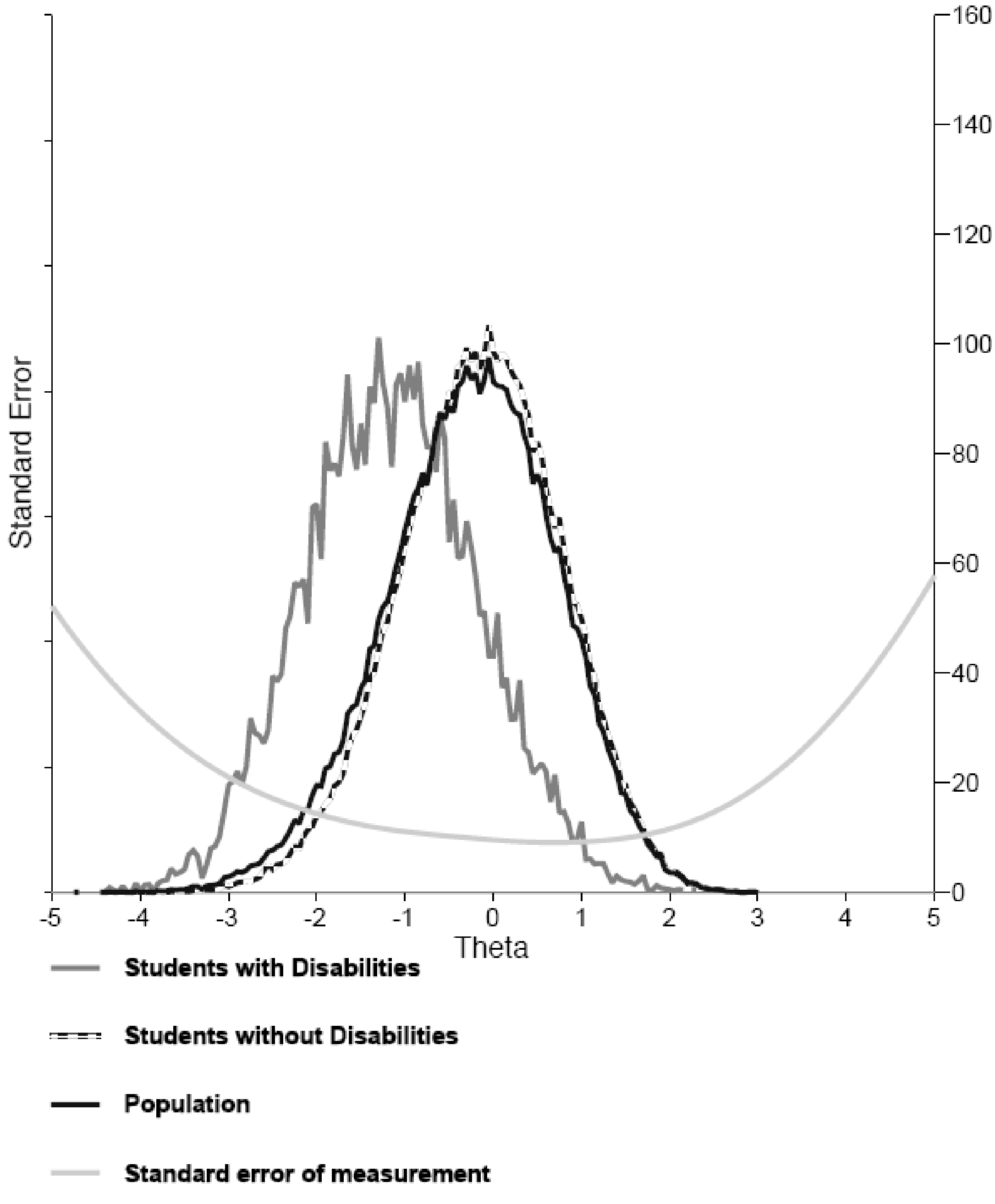
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-43. Grade 8 data analysis and probability: Standard error of measurement and achievement distributions by eligibility for free/reduced price lunch



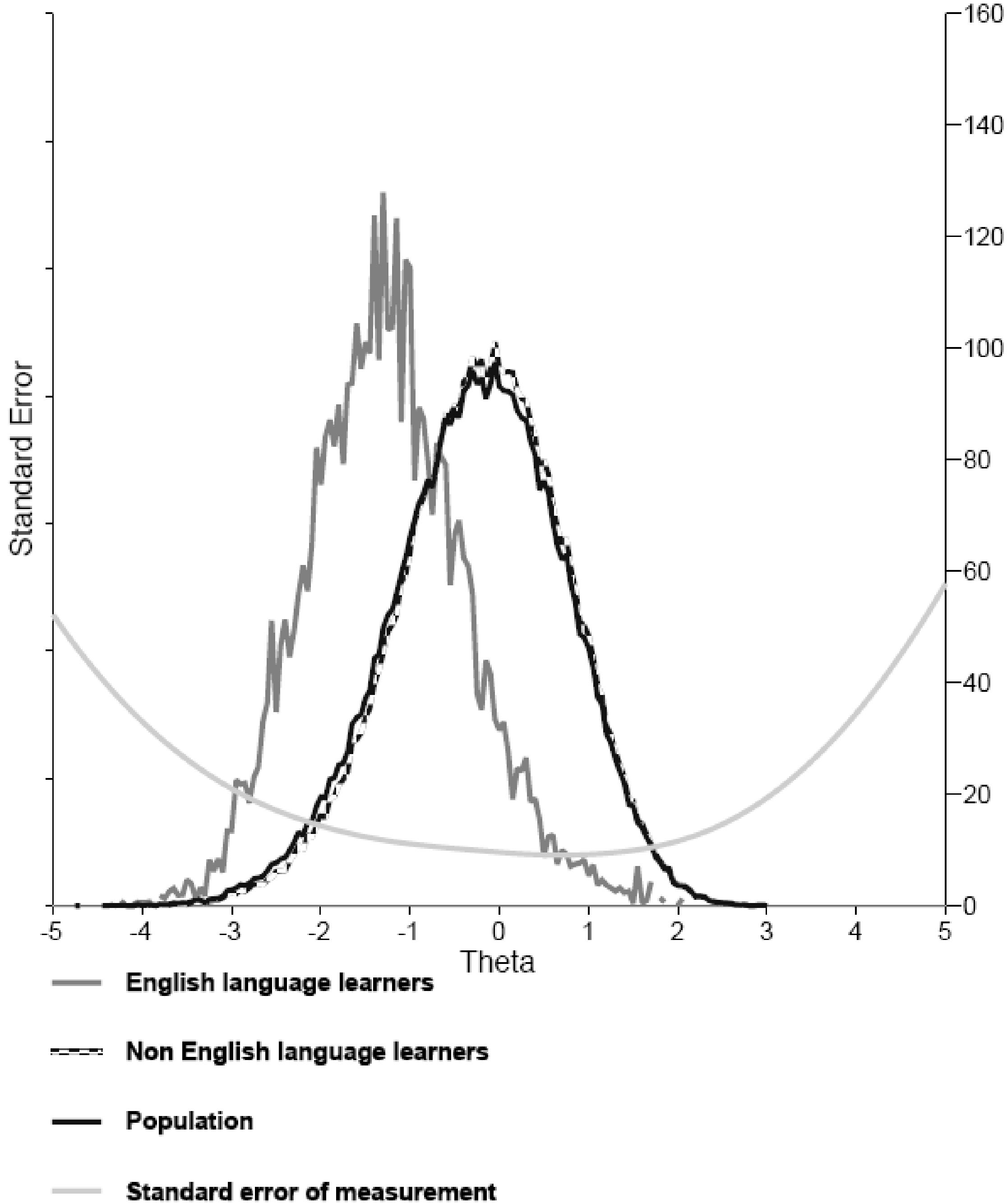
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-44. Grade 8 data analysis and probability: Standard error of measurement and achievement distributions by student disability status



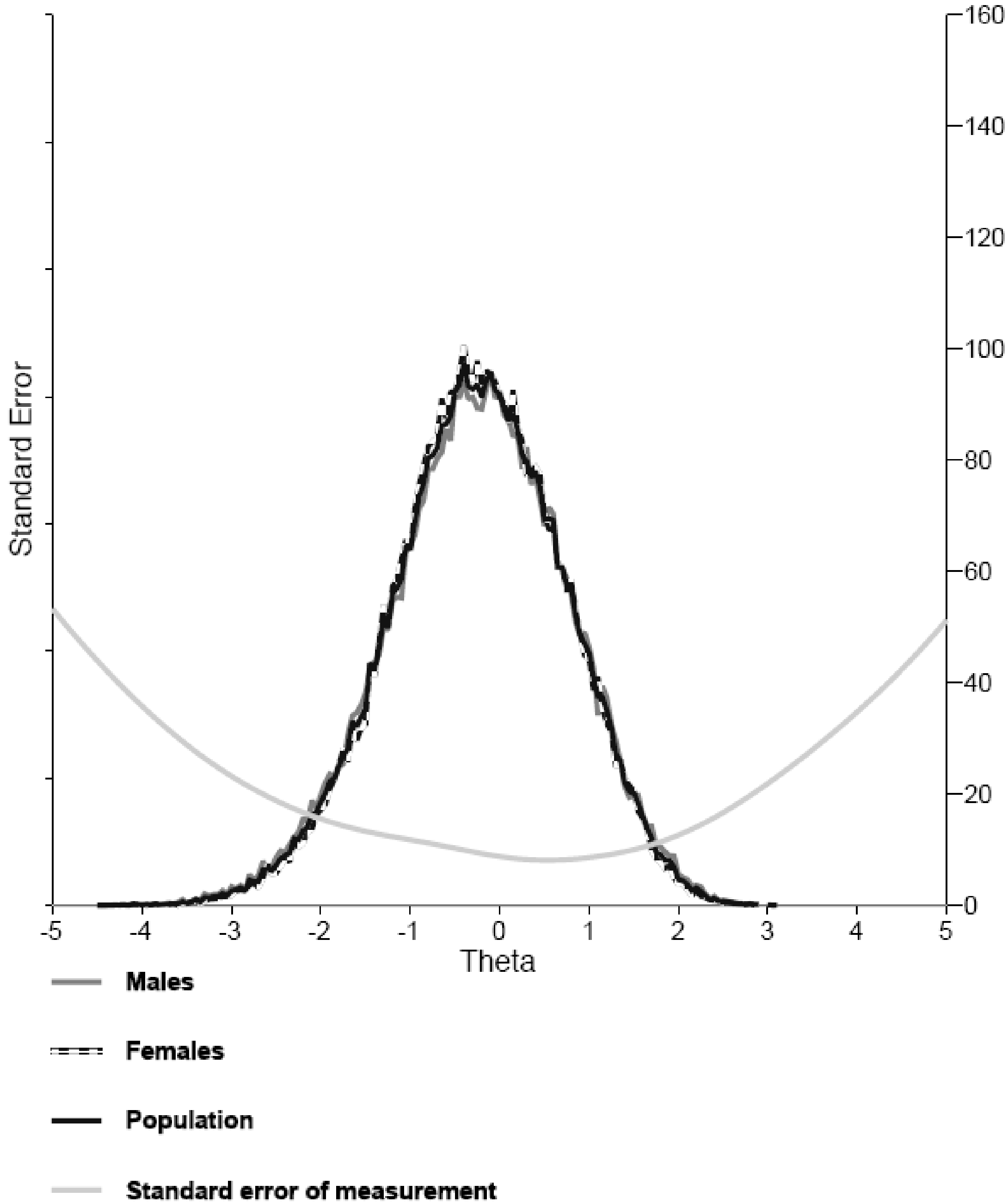
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-45. Grade 8 data analysis and probability: Standard error of measurement and achievement distributions by ELL status



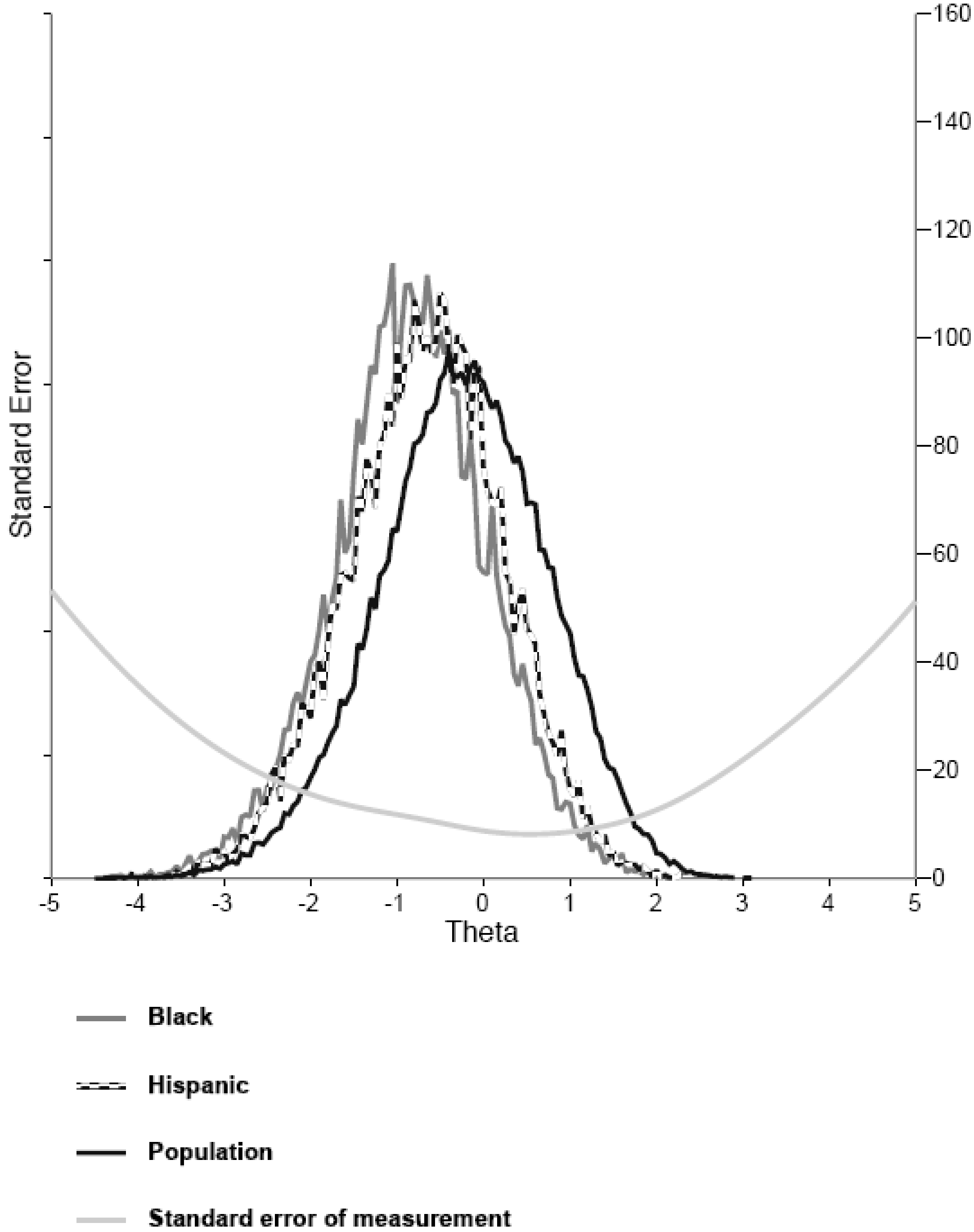
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-46. Grade 8 algebra: Standard error of measurement and achievement distributions by gender



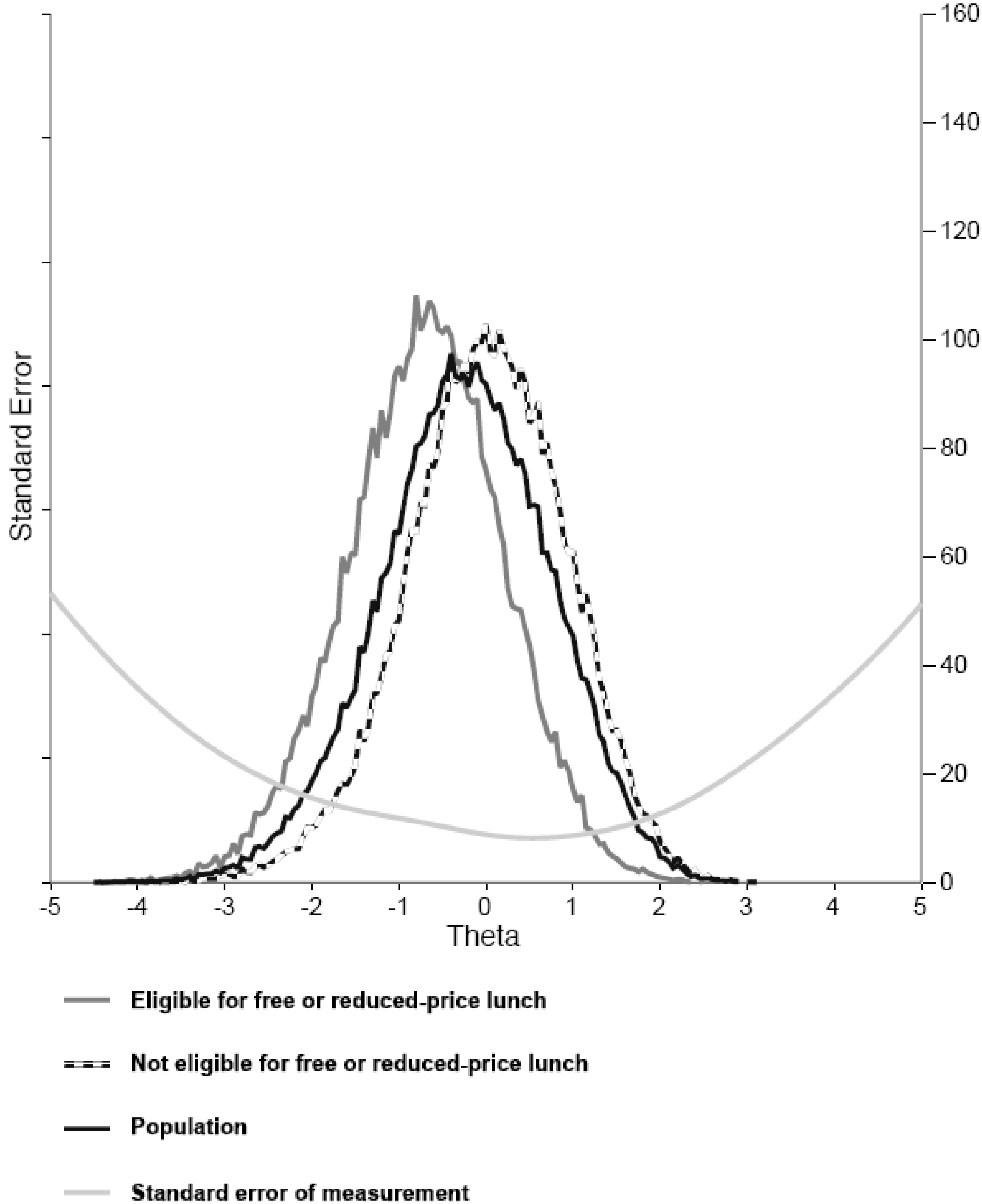
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-47. Grade 8 algebra: Standard error of measurement and achievement distributions by race/ethnicity



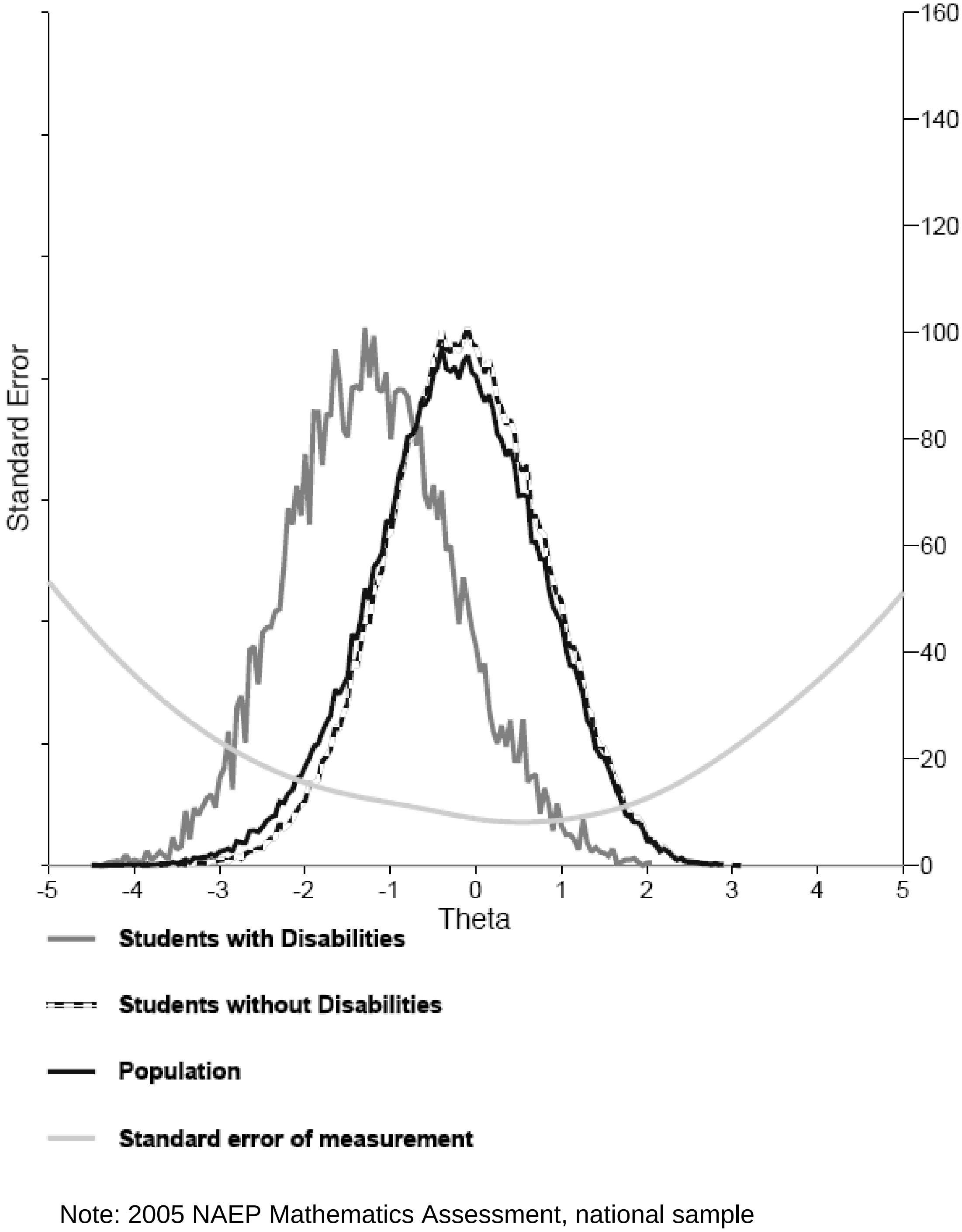
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-48. Grade 8 algebra: Standard error of measurement and achievement distributions by eligibility for free/reduced price lunch



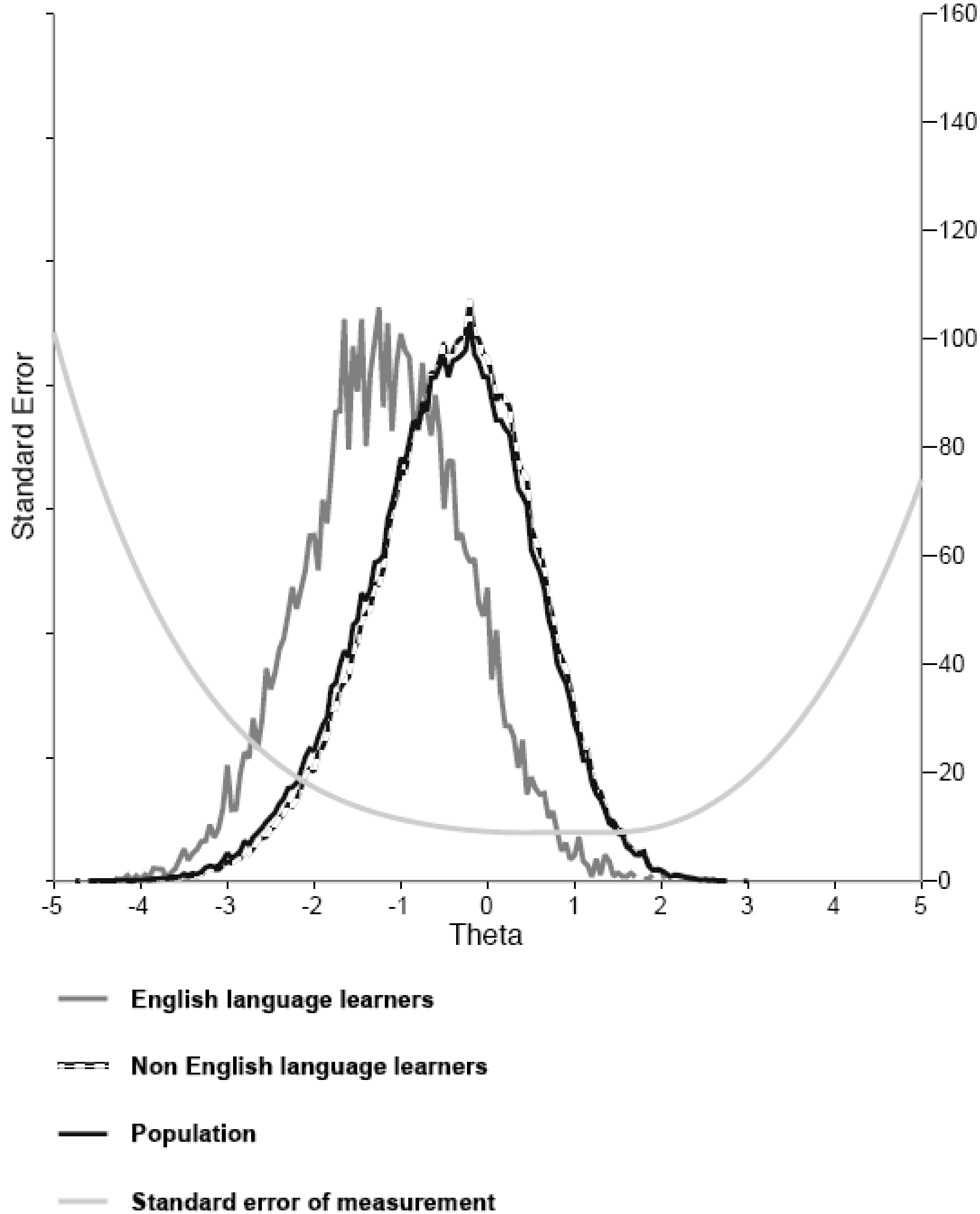
Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-49. Grade 8 algebra: Standard error of measurement and achievement distributions by student disability status



Note: 2005 NAEP Mathematics Assessment, national sample

Exhibit I-50. Grade 8 algebra: Standard error of measurement and achievement distributions by ELL status



Note: 2005 NAEP Mathematics Assessment, national sample

Appendix J

On Including All Item Response Skills in Framework

Why the NAEP mathematics assessment framework must include all skills required to respond successfully to the assessment and differentiate between those that are and are not in the target domain

Don McLaughlin
June 2007

Purpose of a test framework

It isn't sufficient to say that mathematics is what a mathematics test tests. Just as society writes down laws, in order to guide judges in rendering consistent judgments throughout the land, it is necessary to write down what a mathematics test is supposed to test, in order to guide test specialists in constructing and scoring tests in a manner that leads to consistent comparisons throughout the land. Any test whose scores are aggregated and reported is designed to differentiate between individuals or groups of individuals in a specified population in terms of the extent to which they have acquired target skills.

A test framework consists of a catalogue of skills in the target domain (e.g., mathematics skills taught in grades up to eight), a catalogue of item types allowed for the test, a catalogue of skills not in the target domain that are required for responding successfully to the allowable item types, and specification of methods for ensuring that variation in test scores in the target population accurately represents variation in the acquisition of the target skills.

In order to guide the writing of test items, a mathematics test framework must therefore specify the range of skill requirements for "auxiliary" skills not in the mathematics target domain (e.g., vision, general reading comprehension, attention) allowable for inclusion in the test items, along with accommodations and/or scoring rules that ensure that variation among students on those skills does not affect test results. The simplest example is "visual acuity," which is not part of the mathematics skill domain but which varies between students and would affect scores if not corrected. The usual accommodation is so well-known that most people consider it unimportant even to note: eyeglasses (or contact lenses or LASIK surgery).

Relation between test items and skills

In a testing context, each student (if he/she is engaged in the test) responds to a test item by applying a combination of skills he/she has acquired. Every test item requires multiple auxiliary skills outside the target domain, in addition to one or more skills in the target domain, and if the level of the auxiliary skills varies in the student population, then without some correction, part of the variation in test scores will be due to variation in auxiliary skills.

Even though an item invokes multiple skills, there is often a single skill that is most closely associated with performance on an item – the limiting skill for the item. Of the skills required by an item, it is the one on which there is greatest variation in the population being tested. The specifications for a test item typically specify the limiting skill for the item (e.g., "we want n items on addition of fractions with different denominators"). If a test is to be used to assess what is learned in a course, then all of the items should have skills taught in the course as their limiting skills. Unfortunately, such items are often not the most potent items for differentiating between students in a course because other items can have greater requirements for skills not taught in the course (e.g., general aptitude, reading, writing, task organization, attentiveness, math learned prior to the course). The common procedure of evaluating items in terms of their contribution to test "reliability" has the perverse effect of favoring such items.

The requirement that the test framework identify the target and auxiliary skills involved in each item is actually more complex than I've described. Different students may attack an exercise in different ways. To take a very simple mathematics example, some students may perform single digit additions as memory recall, while others may "count on their fingers." This dichotomy of recall versus sequential processing occurs for most problems, with resulting ambiguity in the meaning of test scores. Because both recall and sequential processing have a high, but not perfect, likelihood of success, the major impact of the difference is the time required for generating the correct answer. As long as there is a consensus to avoid measuring how fast a student can answer questions on a test, this ambiguity is difficult to overcome. However, careful design of items can increase the likelihood that either recall or sequential processing will account for substantial variation in performance; the test framework should specify that each item be constructed to elicit specific skills—that is, to be functionally determinate, to as great an extent as possible.

Measuring contributions of variation in target and auxiliary skills to variation in test scores

Once auxiliary skills are specified in the test framework, item writers can develop items that more carefully include and exclude these skills from particular items, and data analysts can compute the separate contributions of variation on these skills to variation in test outcomes. Without getting into details of the scoring methodology, one needs a psychometric model which allows for identifying multiple skills required for performance on individual items. For example, one model is to assume that a correct response requires both an auxiliary skill and a target skill. Then a logistic model (a la current IRT algorithms) would frame the probability of a correct response to item *i* by student *j* as the product of two components:

$$P_i(\theta_j, \psi_j) = \left(\frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}}\right)\left(\frac{e^{(\psi_j - h_j)}}{1 + e^{(\psi_j - h_j)}}\right) \quad ,$$

where θ refers to the target skill level and ψ refers to the auxiliary skill level of the student and *b* and *h* refer to the item requirements for these skills.

By including a small number of items which have virtually no target skill requirement (i.e., $b = -\infty$), an estimate for the auxiliary skill level of a student can easily be obtained. Note that even if the estimate of auxiliary skill levels has measurement error, that is a substantial qualitative improvement over the policy of ignoring the impact of variation in auxiliary skills on test score variation.

The two auxiliary factors involved in mathematics tests on which there is greatest variation in the student population are reading and writing. To the extent that students with better reading and writing skills score higher on a mathematics test than other students with the same mathematical skills, inferences about the effectiveness of mathematics education from scores on the test are suspect. One strategy is to reduce the reading and writing load of mathematics test items to a level that can be assumed to have been acquired by the vast majority of the student population. For some parts of the mathematics skill domain that may not be possible (e.g., assessing understanding or assessing mathematics modeling skills), because items with significant reading or writing requirements are needed to create problems that require those mathematics skills. However, efforts should be made to construct items that minimize auxiliary loads. For example, if one feels the need to assess "understanding" by asking a student to "explain" the reasoning behind a solution, it is possible to avoid a writing requirement by constructing a multiple choice item with short indicators of the competing explanations—indicators that would only be recognized through understanding.

The alternative strategy (and both can be used together) is to assess reading and writing levels in "auxiliary" test items. For example, a few items with substantial reading or writing requirements but very limited mathematics requirements would yield valuable information about the contribution of reading or writing skills to the mathematics scores of a student or group of students (or the impact of deficiency in those skills on scores).

Accommodations and test scoring

In 2007, it is common practice (a) to provide a range of testing accommodations for students who fall into certain categories (e.g., learning disabled, English language learners) in order to ensure that deficiencies in auxiliary skills known to be involved in the items do not impair performance, but also (b) not to take any steps to correct for any effects either of (1) residual deficiencies not removed or (2) advantages provided by removing part of the target skill requirement. This is a profound failing of current assessment methodology, and it can only be remedied by developing test frameworks that provide for the evaluation of the effects of non-standard test administrations (i.e., accommodated testing) on test outcomes.

A comprehensive program for the evaluation of the effects of accommodations on the skill levels required for successful performance on a test is urgently needed; inclusion of auxiliary, or "enabling," skills and accommodations in the test framework will bring this need out of the shadows and should make it apparent to educational policy makers who wish to formulate policy based on comparison of test outcomes between groups of students.

In any case, accommodation as currently practiced is a very blunt instrument for addressing the problem of auxiliary skill deficiencies because (1) it only provides for a single level of intervention, whereas the level of deficiency varies along a continuum, and (2) it is based on a categorical judgment (the student either does or does not have access to the accommodation) based on ill-defined criteria that are known to vary from state to state and even from school to school. If tests are constructed to operationalize a framework that includes assessment of auxiliary skill deficiencies, then it will be possible to phase out categorical, non-scaleable accommodations for cognitive disabilities.

Conclusion

Testing has many purposes, and for the purpose of determining within a classroom which students have failed to acquire the target skill of a week's lesson, a test framework need not be extensive. However, for tests that affect students' lives (e.g., high school exit exams and college entrance exams) or affect large scale educational policies (e.g., state assessments and NAEP), a valid test framework is essential. That test framework must go beyond a taxonomy of the target skill domain to include all the skills that are to be allowed to contribute to variation in test scores. The decision of what auxiliary skills to include in the framework should be determined by a cognitive analysis of the kinds of items that are candidates for inclusion in the test and which vary substantially in the population for which the test is designed.

If the purpose of a test is also to communicate effectively to students and teachers what the target skill domain is, then inclusion of items that require skills outside the target domain must be accompanied by communication that the test also assesses these auxiliary skills. Of course, the auxiliary skills are frequently of substantial importance for the integrated education of a student, so there need be no apology for including those skills in the test. However, if variation in the test results that are communicated back are purely in terms of the target domain (e.g., mathematics), then the production of scores from that test should remove the auxiliary skill contribution, insofar as possible. Including the auxiliary skills in the test framework is an initial step in accomplishing that.