Chapter 8: Report of the Task Group on Assessment

Camilla P. Benbow, Chair
Susan Embretson
Francis “Skip” Fennell
Bert Fristedt
Tom Loveless
Wilfried Schmid
Sandra Stotsky
Irma Arispe, Ex Officio
Ida Eblinger Kelley, U.S. Department of Education Staff

CONTENTS

Tables

Abbreviations

ACT	American College Testing
AP	Advanced Placement
AYP	Adequate Yearly Progress
CR	Constructed Response
CR-G	Constructed Response Grid In
CR-S	Constructed Response Short Answer
CR-EE	Constructed Response Extended
DIF	Differential Item Functioning
ELL	English Language Learners
ERIC	Education Resources Information Center
IRT	Item Response Theory
MC	Multiple Choice
MSPAP	Maryland State Performance Assessment Program
MTSA	Major Topics of School Algebra
NAEP	National Assessment of Educational Progress
NAGB	National Assessment Governing Board
NAS	National Academy of Science
NCES	National Center for Education Statistics
NCLB	No Child Left Behind
NCTM	National Council of Teachers of Mathematics
NELS	National Education Longitudinal Study
NVS	NAEP Validity Study
PISA	Programme for International Student Assessment
SAT	Scholastic Achievement Test
SAT-M	Math Scholastic Achievement Test
SES	Socioeconomic Status
SSCI	Social Sciences Citation Index
STPI	Institute for Defense Analysis Science and Technology Policy Institute
TIMSS	Trends in International Math and Science Study

Executive Summary

Introduction

Achievement tests are widely used to estimate what students know and can do in specific subject areas. Tests make visible to teachers, parents, and policymakers some of the outcomes of student learning. They also can drive instruction. Due to their important role in education today, especially after enactment of the No Child Left Behind Act, the Panel examined the quality of released items from the mathematics portions of the National Assessment of Educational Progress (NAEP) and six state tests, and reviewed the relevant scientific literature on the appropriate distribution of test content, the setting of performance categories, factors affecting measurement accuracy, and appropriate test design.

Key Questions Addressed by the Task Group

To address the charges in the Executive Order, the Assessment Task Group developed five primary questions:

Part One: Test Content and Performance Categories

1) What should the mathematics content of the NAEP and state tests be at Grades 4 and 8? How does the content of state tests compare with NAEP?
2) How are performance categories determined?

Part Two: Item and Test Design

3) How does item response format affect performance on multiple-choice and various kinds of constructed-response items?
4) What are some nonmathematical sources of difficulty or confusion in mathematics test items that could inappropriately affect performance? How prevalent are they on the NAEP and the six state tests examined?
5) How are calculators used in NAEP and state assessments and how does calculator use affect performance?

These questions are not independent of each other; they overlap because what one tests and how one chooses to test are intertwined. For example, the verbal context of the test items or calculator use could have bearing on what is actually measured.

Test Content

The content strands in most state mathematics tests are similar to the content strands in the NAEP mathematics test. Thus, the Task Group focused its investigation on the NAEP content strands, knowing that any suggestions for the NAEP would have implications for most state mathematics tests as well.

The Task Group presents in Table 1 possible recommendations that could flow from the general principles for organizing the content of the NAEP—that tests should measure what students should be learning. In preparation for algebra, students should become proficient in the critical foundations for algebra as described in the Conceptual Knowledge and Skills Task Group report.

Table 1: Suggested Reorganization of NAEP Content Strands

Grade 4	**Grade 8**
Number: Whole Numbers	Number: Integers
Number: Fractions and Decimals	Number: Fractions, Decimals, and Percents
Geometry and Measurement	Geometry and Measurement
Algebra	Algebra
Data Display	Data Analysis and Probability

The most critical skills to be developed before beginning algebra are extensive work with whole numbers, including whole number operations, and facility with fractions. The NAEP Validity Study (NVS), as well as others, have noted the relative paucity of items assessing fractions in both the fourth- and eighth-grade NAEP.

Moreover, the NVS indicates that half of the data analysis and probability section of the Grade 4 NAEP is probability-related. Given the importance of fractions for the conceptual understanding of probability, the Task Group questions whether probability can be measured appropriately in the fourth grade. Thus, the Task Group suggests that this strand at the fourth-grade level be limited to data analysis and titled as Data Display.

The review of NAEP content also led the Task Group to conclude that there needs to be a more appropriate balance in how algebra is defined and assessed at both the fourth- and eighth-grade levels of the NAEP. At the fourth-grade level, most of the NAEP algebra items relate to patterns or sequences. While the inclusion of patterns in textbooks or as state curriculum expectations may reflect a view of what constitutes algebra, patterns are not emphasized in the curricula of high-achieving countries. In the Major Topics of School Algebra set forth in the Task Group on Conceptual Knowledge and Skills report, patterns are not a topic of major importance. The prominence given to patterns at the preschool through Grade 8 level is not supported by comparative analysis of curricula or by mathematical considerations. Applying the general principle for selecting content for the NAEP and state tests, the Task Group strongly recommends that "algebra" problems involving patterns be greatly reduced in these tests.

It might be useful to note that the Trends in International Math and Science Study (TIMSS) content domains were changed at the time the Task Group was conducting its own work. Adopting the Task Group's recommendations would bring NAEP into greater alignment with TIMSS (Mullis et al., 2007).

Performance Categories

Once content is selected, decisions must be made about what the performance categories should be and how to assign student scores to them. The Task Group did not investigate what the cut scores for each category should be but, rather, how they should be determined. Although the states and NAEP varied in both process and method for such standard setting (setting cut scores), the six states studied in the NVS report and NAEP employed currently acceptable educational practices to quantify judgments of the standard-setting panelists and to map their judgments onto test scores. Limited research is available on standard-setting methods and processes. The Modified Angoff method requires the most plausible assumptions about raters and tests, but more research is needed comparing the outcomes based on alternative methods.

In the states the Task Group examined, classroom teachers, most of whom are not mathematics specialists, predominate in the standard-setting process. The expertise of mathematicians, as well as of mathematics educators, curriculum specialists, classroom teachers, and the general public, should be consistently used in the standard-setting process. The Task Group also found that the standard-setting panelists often do not take the test as examinees before attempting to set the performance categories, and that they are not consistently informed by international performance data. On the basis of international performance data, there are indications that the NAEP cut scores for performance categories are set too high. This does not mean, however, that the mathematical content of the test is too hard; it is simply a statement about the location of cut scores for qualitative categories such as "proficient" and "beyond proficient."

Recommendations for Test Content and Performance Categories

1) NAEP and state tests must ensure a focus on the mathematics that students should learn with achievement on critical mathematics content reported and tracked over time. NAEP should ensure that the Conceptual Knowledge and Skills' Critical Foundations and elements of the Major Topics of School Algebra are integral components of the mathematics assessed. The Task Group proposes reorganization, as well as possible title changes, of NAEP's current five content strands:

    a. Number Properties and Operations should be renamed and expanded into two separate categories—Grade 4 Number: Whole Numbers; and Fractions and Decimals; and Grade 8 Number: Integers; and Fractions, Decimals, and Percent.

        1. Whole Numbers will include emphasis on place value, comparing and ordering, and whole number operations at Grade 4. This will be expanded to include work with all integers, including operations with negative and positive integers at Grade 8.

        2. Fractions and Decimals will include recognition, representation and comparing and ordering at Grade 4. This will be expanded to include operations involving fractions, decimals, and percent at Grade 8.

b. Geometry and Measurement should be combined into one content area. Topics related to both Measurement and Geometry should serve as important contexts for problems in the Grade 4 and Grade 8 NAEP.

c. Within Algebra, a better balance is needed within the subtopic of patterns, relations, and functions at Grades 4 and 8. That is, there should be far fewer items on patterns.

d. Data Analysis and Probability should be renamed as Data Display at Grade 4 and expanded to include both data interpretation and probability at Grade 8.

2) Procedures should be employed to include a broader base for setting performance categories:

a. The Task Group recommends that standard-setting (setting cut scores) panels include individuals with high levels of expertise, such as mathematicians, mathematics educators, and high-level curriculum specialists, in addition to classroom teachers and the general public.

b. The standard-setting panelists should take the test as examinees before attempting to set the performance categories.

c. The standard setting should be informed by international performance data.

d. Research is needed on the impact of standard-setting procedures and methods (e.g., Bookmark Method, Modified Angoff procedure) in promoting the representation of a broad base of judgments.

Item and Test Design

It is important not only that appropriate content is measured and cut scores for student performance are set appropriately, but also that test scores reliably reflect the competencies intended to be measured. That is, the measurement itself must be carried out in a high-quality and appropriate manner.

Item Response Format

Many educators consider constructed-response items (e.g., short answer) as superior to multiple-choice (MC) items in measuring mathematical competencies and a more authentic measure of mathematical skill. They believe such items also offers the opportunity for students to explain principles and display a range of math skills including verbal explanations. The Task Group examined the literature on the psychometric properties of constructed-response items compared with multiple-choice items. The evidence found in the scientific literature does not support the notion that a constructed-response format, particularly the short-answer type, measures different aspects of mathematics competency compared with a multiple-choice format. While there are skills that may be measured only using a constructed-response format, concern about use of multiple-choice items in these tests at the fourth- and eighth-grade level is not warranted.

Nonmathematical Sources of Difficulty

The NVS panel found many examples of flawed items on NAEP and state assessments that could affect performance of all or some students and trend lines. The Task Group undertook its own examination of released items on state and NAEP tests, looking specifically for nonmathematical sources of difficulty (e.g., particular context portrayed within an item) and found many items on the NAEP and state tests affected by these sources of difficulty, resulting in too many flawed items. The Task Group presents seven types of flawed items illustrating nonmathematical sources of influence that could affect scores. Test developers should be sensitive to the presence of these types of flaws in the test development process.

Careful attention must be paid to exactly what mathematical knowledge is being assessed by a particular item and the extent to which the item is, in fact, focused on the intended mathematics. In other words, significant attention should be devoted to the actual design of individual mathematics items and to the evaluation of items for inclusion. More mathematicians should be involved in the process of designing and evaluating items, as should mathematics educators, curriculum specialists, linguistics experts, and cognitive psychologists.

The frequency of flawed items points to another possible gap in test development procedures that needs to be addressed. Psychometricians are trained to use highly sophisticated statistical models and data analysis methods for measurement, but are not as familiar with issues of item design with respect to measuring mathematical constructs. Item writers and item evaluators often do not have a college degree in the appropriate subject, and apparently do not have the kind of background in task and item design that would lead to a lower percentage of items that are flawed or marginal according to the mathematicians. Moreover, they receive limited feedback from psychometricians on how the items they develop end up functioning for students at varying levels of performance. That is, the feedback mechanism does not provide sufficient information to help pinpoint the sources of item deficiencies.

Calculators

Use of calculators in assessment is another frequently discussed design issue. While findings from the literature revealed that using calculators in assessment has no significant impact on performance overall or in problem solving, the research indicates that calculator use affects performance on computation-related items and also could change the nature of the competencies tested.

Recommendations on Item and Test Design

1) The focus in designing test items should be on the specified mathematical skills and concepts, not item response format. The important issue is how to most efficiently design items to measure content of the designated type and level of cognitive complexity.

2) Much more attention should be paid to what mathematical knowledge is being assessed by a particular item and the extent to which the item addresses that knowledge.

3) Calculators (the use of which constitutes a design feature) should not be used on test items that seek to measure computational skills. In particular, NAEP should not permit calculator use in Grade 4.

4) Mathematicians should be included in greater numbers, along with mathematics educators, and curriculum specialists (not just classroom teachers and the general public), in the standard-setting process and in the review and design of mathematical item content for state, NAEP, and commercial tests.

5) States and NAEP need to develop better quality control and oversight procedures to ensure that test items reflect the best item design features, are of the highest quality, and measure what is intended, with nonmathematical sources of variance in performance minimized.

6) Researchers need to examine whether the language in word problems is suitable for assessing their mathematical objectives before examining their impact in state assessments on student performance, especially the performance of special education students or English language learners.

7) More scientific research is needed on item and test design features.

I. Introduction

Achievement tests are widely used to estimate what students know and can do in specific subject areas. Tests make visible to teachers, parents, and policymakers some of the outcomes of student learning. Tests also can provide an efficient and fair way to assess student achievement. Finally, tests can drive both the content and format of classroom instruction.

Widespread, large-scale testing began in the 1960s after passage in 1965 of the Elementary and Secondary Education Act and the appropriation of Title I funds. Policymakers desired information to gauge the progress of education in the United States as a whole and, thus, the idea for a National Assessment of Educational Progress (NAEP) emerged. NAEP was implemented in 1969, and the long-term trend tests began during the 1972–73 school year. The main NAEP came later, in 1990.

With passage of the No Child Left Behind Act (NCLB) in 2001, the use of testing was expanded beyond its many uses at the time: end-of-course evaluations of learning by teachers; admission tests for college, graduate, or professional programs; loosely structured accountability systems at the school district level; and what had been required under ESEA as reauthorized in 1994 by IASA (the Improving America's Schools Act). In 1994, IASA required states to test students in all schools in reading and mathematics in three grade spans, and to use state assessments. NCLB expanded the number of grades required to be tested. NCLB also mandated, among other things, use of state assessments and other measures to hold schools and districts accountable for increasing all students' achievement, including the achievement of different subgroups of students. A few states had already created content standards on which their state tests, if developed, were based to ensure that students were learning the topics judged important for students to master. And some of these states had made passing state tests a requirement for high school graduation. NCLB, however, required all states to develop standards and state assessments in reading and mathematics in Grades 3–8 and once in high school, and set forth measures of teacher quality, as well. States could choose their own tests and set their own cut scores, but they had to demonstrate annual improvement for all subgroups of students through a measure called Adequate Yearly Progress (AYP).

A provision in NCLB also required all states to participate in NAEP beginning with the 2003 cycle. NAEP was to sample each state every 2 years so that the results on NAEP tests could be compared with the results of state tests. It is intended that NAEP and the state assessments, along with results from international assessments when they are available, inform the public and policymakers on the condition of education in the United States. Given the importance of the NAEP and state tests for measuring the outcomes of education, it is vital that the NAEP and state tests measure appropriately what is deemed important for children to learn in school. For more details on the history of NAEP and the two types of tests it gives, see Appendix A and the NAEP Web site. Descriptions of six state testing programs are provided in Tables 3 and 4.

In this context, the Assessment Task Group of the National Mathematics Advisory Panel (Panel) was formed to address the following charges in the Executive Order:

> (b) the role and appropriate design of standards and assessment in promoting mathematical competence;
>
> (f) the role and appropriate design of systems for delivering instruction in mathematics that combine the different elements of learning processes, curricula, instruction, teacher training and support, and standards, assessments, and accountability (Executive Order No. 13398).

II. Key Questions Addressed by the Task Group

To address the charges in the Executive Order, the Task Group developed five primary questions (divided into two areas):

Part One: Test Content and Performance Categories

1) What should the mathematics content of the NAEP and state tests be at Grades 4 and 8? How does the content of state tests compare with NAEP?
2) How are performance categories determined?

Part Two: Item and Test Design

3) How does item response format affect performance on multiple-choice and various kinds of constructed-response items?
4) What are some nonmathematical sources of difficulty or confusion in mathematics test items that could inappropriately affect performance? How prevalent are they on the NAEP and the six state tests examined?
5) How are calculators used in NAEP and state assessments, and how does calculator use affect performance?

These questions are not independent of each other. They overlap because what one tests and how one chooses to test are intertwined. For example, the verbal context of the test items or calculator use could have bearing on what is actually measured.

III. Background

Two studies were shared with the Task Group that offered a strong foundation for its work and began to answer the key questions. These were: 1) *Validity Study of the NAEP Mathematics Assessment: Grades 4 and 8* (Daro et al., 2007), 2) Response to the Validity Study of the NAEP Mathematics Assessment: Grades 4 and 8 (Schneider, 2007). These two documents, while addressing some of the Task Group's main concerns, led the group to probe some of the reports' findings in deeper and more specific ways.

A. NAEP Validity Study (NVS) Report

The NVS convened an expert panel involving mathematicians, mathematics educators, and an expert on state-based mathematics standards. They compared the NAEP mathematics framework with the standards and frameworks (test blueprints) of six states (California, Massachusetts, Indiana, Texas, Washington, and Georgia), two high-performing nations (Singapore and Japan), and standards outlined by the National Council of Teachers of Mathematics (NCTM) and Achieve, Inc.

The NVS examined the content areas of number properties and operations, algebra, geometry, measurement, and data analysis and probability strands in the 2005 NAEP mathematics framework to determine if NAEP was missing something or overemphasizing topics in a given content area.[1] The reviewers then described what was missing or being overemphasized, and rated the emphasis of each content topic as compared to each of the six states and Singapore and Japan. The panel of mathematicians also examined individual items from the NAEP tests and the six states' tests and found serious problems.

Quoting from the NVS Report:

> *Five percent of NAEP items were found to be 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 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 (Daro et al., 2007, p. 79–80).*

The NVS report showed that a high percentages of marginal and flawed items appeared in four major content areas in these tests: Algebra, Geometry, Measurement, and Data Analysis and Probability. The Number Properties and Operations section was better than the other four. As Exhibits IV-3 and IV-4 (pp. 82 and 83) of the NVS report show, for Grade 4, 16 of the 28 NAEP Algebra items and 8 of the 16 state Algebra items were classified as marginal or seriously flawed. As Exhibit IV-5 shows (p. 83) at Grade 8, 16 of 59 NAEP Algebra items were similarly classified. For Grade 4, 15 of 34 NAEP Geometry items, 15 of 42 NAEP Measurement items, and 12 of 19 state Measurement items were so classified. At Grade 8, 11 of 45 NAEP Geometry items, 10 of 37 NAEP Measurement items,

[1] The NVS was asked to address the following questions:
1. Does the NAEP framework offer reasonable content and skill-based coverage compared to the assessments of states (six were selected for study and are described in Table 1 and 2) 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? (Daro et al., 2007, p. i).

and 10 of 25 state Geometry items were so classified. For Grade 4, 5 of 12 state Data Analysis and Probability items were so classified. For Grade 8, 15 of 32 NAEP Data Analysis and Probability items, and 7 of 15 state Data Analysis and Probability items were so classified. In other words, more than one-fourth to more than one-half of the items in these four areas were rated by five mathematicians as not suitable for high-stakes tests.

The NVS report also was concerned about whether the percentage of marginal and flawed items in the NAEP tests might have influenced test results. Quoting from the report:

> *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 classified as marginal exhibited construct-irrelevant difficulties that could affect performance for some test takers (p. 82).*

Nonetheless, one central finding of the NVS was, "The NAEP mathematics assessment is sufficiently robust to support the main conclusions that have been drawn about United States and state progress in mathematics since 1990" (p. ii and 119). They noted, however, that while the framework was reasonable, the specifications communicated to test developers were not detailed enough. In addition, while they thought item quality was typical of large-scale assessments, it could be improved.

The NVS made recommendations for improving the NAEP that flowed from its study. Two of these recommendations are of particular importance to the Task Group. First was its recommendation to sharpen the focus of the current NAEP framework. Specifically, it recommended, "[F]ocus: don't worry about leaving things out; worry about targeting the most important things…Explicitly address high priority issues that cut across content areas" (p. vi).

The second recommendation in the NVS report of importance to the Task Group involved item quality and the provision of exemplars of good items for future NAEP tests. NVS recommended improved quality assurance, with particular attention focused on the following: mathematical quality of the items, quality of the situated mathematics problems (e.g., word problems), measurement of complexity, non-construct relevant sources of item difficulty (i.e., nonmathematical sources of difficulty; e.g., verbiage; complex graphical displays; vocabulary), item performance and construction (e.g., response format such as multiple choice versus constructed response), and the range of item difficulty and curricular reach.

B. NCES Response to the NVS Report

In its response to the NVS report, the National Center for Education Statistics (NCES) claimed that student gains on the main NAEP have not been underestimated by the number of poor-quality items. The NCES statistical analysis of marginal and flawed items revealed that their mean discrimination indices (mean r-biserial for each category) did not differ from the items judged as adequate (Schneider, 2007). NCES was also less critical of NAEP's quality control process for item development, but NCES concurred with the NVS recommendation about the importance of item quality and the provision of exemplars of good items for future NAEP tests.

IV. Methodology

The Task Group determined a strategy to probe the quality of the state tests given their particular charge. The Task Group needed to determine if the recommendations for the NAEP also applied to the state tests they could inspect and if there potentially were other issues. Thus, the Task Group undertook a review of released test items from six state tests and NAEP. Moreover, it wanted to explore more deeply the validity of the process for setting performance categories, especially given the recent NCES report, *Mapping 2005 State Proficiency Standards onto the NAEP Scales* (National Center for Education Statistics, 2007). In that study, NCES mapped state performance categories in reading and mathematics onto the appropriate NAEP scale using data from fourth and eighth grades in the 2004–05 school year. Finally, the Task Group took some of the recommendations for the NAEP a step further, explored appropriate content, and posed additional questions, such as the impact of calculator usage.

With the assistance of Abt Associates Inc. (Abt), the Task Group conducted searches of the scientific literature with respect to the questions posed. The Institute for Defense Analyses Science and Technology Policy Institute (henceforth STPI) also assisted, in particular, with the review of released test items from the NAEP and state assessments. See Appendix B for more detail on the Task Group's methodology.

The Task Group conducted its work during a three-month time period. Consequently, it was limited in its ability to collect, examine, and analyze an extensive amount of information. For example, the identification of relevant literature was limited to what could be identified and reviewed in that time period. Furthermore, there was insufficient time to field a survey. The analysis was, thus, based on information readily available on state and NAEP Web sites, in publications, and in prior analysis and research.

V. Part I: Test Content and Performance Categories

A. Question 1: What Should the Mathematics Content of NAEP and State Tests Be at Grades 4 and 8? How Does the Content of State Tests Compare with NAEP?

1. Background

Currently, NAEP assesses mathematics organized by the following five content strands:

- Number Properties and Operations
- Geometry
- Algebra
- Measurement
- Data Analysis and Probability

These strands were developed to meet the requirement that large-scale achievement tests of this nature must measure competencies reflecting all areas of mathematics taught and considered most important at a given developmental level. To assess the appropriateness of the content strands, the Task Group examined five sources: 1) the Critical Foundations, the skills and knowledge essential for success in algebra as described in the Conceptual Knowledge and Skills Task Group report; 2) the NAEP Validity Study (Daro et al., 2007); 3) findings from the Task Group's literature review; 4) the Task Group-generated review of released items from NAEP and selected state tests; and 5) the Panel's National Survey of Algebra Teachers (Hoffer, Venkataraman, Hedberg, & Shagle, 2007).[2] In the Panel's survey, teachers identified particular aspects of mathematical content areas (e.g., fractions and success with word problems) as both critically important to the preparation for algebra and insufficiently acquired by students in introductory algebra courses. These findings added to the Task Group's basis for reviewing these topics. In addition, the Task Group's review of some state tests and their released items provided further information on how one might reset the focus of test content frameworks.

2. Reorganizing the Content Strands of NAEP and Implications for State Assessments

Based on the review of the five sources described earlier in this section, the Task Group proposes several principles for reorganization, as well as possible title changes, for the five content strands of the NAEP and, potentially, for state tests. The suggested reorganization is presented in Table 1 and represents the possible outcome of employing the principles for organizing the content of the NAEP. This possible reorganization has implications for state mathematics tests, as well.

[2] The order in which the sources are listed bears no significance of the importance of each source.

Table 1: Suggested Reorganization of NAEP Content Strands

Grade 4	Grade 8
Number: Whole Numbers	Number: Integers
Number: Fractions and Decimals	Number: Fractions, Decimals, and Percent
Geometry and Measurement	Geometry and Measurement
Algebra	Algebra
Data Display	Data Analysis and Probability

The suggested principles are as follows:

1) NAEP and state tests must ensure a focus on the mathematics that students should learn with achievement on critical mathematics content reported and tracked over time. NAEP should ensure that the Conceptual Knowledge and Skills' Critical Foundations and elements of the Major Topics of School Algebra are integral components of the mathematics assessed. The Task Group proposes for reorganization, as well as possible title changes, for NAEP's current five content strands:

    a. Number Properties and Operations should be renamed and expanded into two separate categories—Grade 4 Number: Whole Numbers; and Fractions and Decimals; and Grade 8 Number: Integers; and Fractions, Decimals, and Percent.

        1. Whole Numbers will include emphasis on place value, comparing and ordering, and whole number operations at Grade 4. This will be expanded to include work with all integers, including operations with negative and positive integers at Grade 8.

        2. Fractions and Decimals will include recognition, representation and comparing and ordering at Grade 4. This will be expanded to include operations involving fractions, decimals, and percent at Grade 8.

    b. Geometry and Measurement should be combined into one content area. Topics related to both Measurement and Geometry should serve as important contexts for problems in the Grade 4 and Grade 8 NAEP.

    c. Within Algebra, a better balance is needed within the subtopic of patterns, relations, and functions at Grades 4 and 8. That is, there should be many fewer items on patterns.

    d. Data Analysis and Probability should be renamed as Data Display at Grade 4 and expanded to include both data interpretation and probability at Grade 8.

    These principles and their possible implications are now explained in greater detail.

Number Properties and Operations

The Task Group suggests that Number Properties and Operations be expanded and renamed as Number. It should include a focus on whole numbers, including place value, comparing and ordering, and whole number operations (i.e., addition, subtraction, multiplication, division—arithmetic) at Grade 4 and then be expanded to include extensive work with all integers (negative and positive) at Grade 8. A proposed additional content area involving number would focus on fractions. At the Grade 4 level, this would involve beginning work with fractions and decimals including, recognition, representation, and comparing and ordering. This would be expanded to include operations with fractions, decimals, and percent at Grade 8. Similarly, the focus of work with whole numbers and fractions on state tests should expand as concepts and operations are developed from year to year, particularly at Grades 5, 6, and 7, which are grade levels when the NAEP test is not offered.

A review of the eighth-grade NAEP Number Properties and Operations content area (Daro et al., 2007) found an emphasis on topics from number theory—factorization, multiples, and divisibility. Their review suggests, however, the need to ensure that eighth-grade students have developed proficiency with whole numbers, positive and negative integers, fractions, decimals, and percent given their importance as prerequisites for algebra. Because this content area stood out in the NVS review as under-sampling grade-level content, "It is possible that students are making gains in this content area that are not being detected by NAEP" (p. 123). In the Panel's judgment, it is also possible that students are losing ground that goes undetected. Indeed, because the NAEP minimizes this area, this could be a driving force for reduced attention to it within the school curriculum.

One of the Task Group's greatest concerns is that fractions (defined here as fractions, decimals, and related percent) are underrepresented on NAEP. The NVS, as well as others, have noted the relative paucity of items assessing fractions in both the fourth- and eighth-grade NAEP. The validity study identified fewer than 20% of items as involving fractions and decimals in Grade 8. It also was noted that, while Number Properties and Operations should be the most emphasized content area at the fourth-grade level, the NAEP provides a very limited assessment of fractions at this level. Implementation of the Task Group's recommendations would result in a more appropriate balance of content and address the issue of underrepresentation of fractions on the NAEP.

Geometry and Measurement

As seen in Table 1, the Task Group also suggests that Geometry and Measurement be combined into one content area, which would make the Grade 4 and 8 test frameworks consistent with that of Grade 12 NAEP (2005).The proposed merging of these content areas also would address the concern that there is a "need for a close look at how the NAEP measurement objectives compare to the treatment of measurement elsewhere" (Daro et al., 2007, p. 9). Such an examination was deemed important given that measurement is second only to number properties and operations in the fourth-grade NAEP in terms of the number of items assessed in a particular content area. The larger number of measurement items within NAEP however, is "not well leveraged to include fractions or decimals used in realistic situations" (p. 126). It is also noted that while there is considerable overlap in the

NAEP and Trends in International Math and Science Study (TIMSS) assessments involving measurement, there is greater emphasis in NAEP on using measurement instruments and units of measurement. As a result, NAEP may be underestimating achievement in measurement. TIMSS includes a higher percentage of items on estimating, calculating, or comparing perimeter, areas, and surface area.

The review of the Geometry items indicated wide variation across six states. For example, the NAEP at Grade 8 includes more geometry than the comparison states or nations. Also, the eighth-grade NAEP Geometry items assess symmetry and transformations more than those of the states and emphasize parallel lines and angles less than the comparison states. Finally, it should be noted that topics related to both measurement and geometry (e.g., perimeter, area, and circumference) could serve as important contexts for problems within the Grade 4 and Grade 8 NAEP. This constitutes another principle for organizing the content of the NAEP not previously noted.

Algebra

Algebra is the most heavily weighted content topic on the eighth-grade NAEP, with 30% of the assessment targeting algebra objectives (NAEP 2007 framework). Fifteen percent of the fourth-grade NAEP is dedicated to algebra. At the fourth-grade level most of the NAEP algebra items relate to patterns or sequences (Daro et al., 2007). Chazan et al. (2007) also note that released Grade 4 NAEP items place a heavy emphasis on pattern completion at the expense of other types of algebraic reasoning. This is cause for concern. While states' inclusion of patterns in textbooks or as curriculum expectations may reflect their views of what constitutes algebra, patterns are not emphasized in high-achieving countries (Schmidt, 2007). The NVS (Daro et al., 2007) recommended better item balance within the algebra subtopic of patterns, relations, and functions at the Grade 4 level. In the Conceptual Knowledge and Skills Task Group's Major Topics of School Algebra (MTSA), patterns are not a topic of major importance. The prominence given to patterns at the preschool to eighth-grade level is not supported by comparative analysis of curricula or by mathematical considerations (Wu, 2007). In addition, this has broad implications for interpreting student performance. For example, although student performance on the eighth-grade NAEP Algebra strand has increased, reviewers note the underrepresentation of high-complexity items in algebra (Daro et al., 2007). Thus, one cannot be clear on what this increased performance means.

Data Analysis and Probability

While recognizing that data analysis provides the context for many interesting problems in mathematics, the Task Group notes that the work of the NVS indicates that half of the Data Analysis and Probability section of the Grade 4 NAEP is probability related, whereas TIMSS has a greater proportion of items than NAEP that emphasize reading, interpreting, and making predictions from tables and graphs, and data representation, especially at the fourth-grade level. Given the importance of fractions for the conceptual understanding of probability, the Task Group questions whether probability can be taught and measured appropriately at the fourth-grade level. As students begin work with fractions, probability becomes a more viable mathematics topic and thus should come later in elementary and middle school. The Task Group, therefore, suggests that the Grade 4 content

area be renamed as Data Display, consistent with TIMSS 2007, and at Grade 8, as Data Analysis and Probability. The focus at the eighth-grade level would be expanded to include both data interpretation and probability.

Comparison to TIMSS

It is useful to note here that the TIMSS content domains were changed (Mullis et al., 2007) at the time the Task Group was conducting its own study. The Grade 4 content domains are now identified as Number, Geometric Shapes, and Measures and Data Displays. At this level, TIMSS has merged Geometry and Measurement, as the Task Group also suggests, and deleted the domain formerly called Patterns, Equations, and Relationships, again consistent with the concerns the Task Group raises about patterns and algebra at the fourth-grade level. The Grade 8 content domains are Number, Algebra, Geometry, and Data and Chance. At this level, TIMSS has infused Measurement within Geometry and expanded Data to include Probability. The Task Group's suggested principles for reorganizing the NAEP would bring it into greater alignment with TIMSS.

3. A Comparison of State Test Content with NAEP Content

How does the content of State Tests Compare with the Content of the NAEP Tests?

A comparison by strand of the percentage of items on state tests for Grades 4 and 8 with the percentage of items on the Grade 4 and Grade 8 NAEP test in 2007 yields the following results in Table 2:

Table 2: A Comparison of State Test Content with NAEP Content

Grade 4			**Grade 8**		
	NAEP (2007)	*State Tests*		*NAEP (2007)*	*State Tests*
Numbers	40%	15–48%	Numbers	20%	10–26%
Measurement and Geometry (combined)	35%	18–34%	Measurement and Geometry (combined)	35%	20–28%
Algebra	15%	12–28%	Algebra	30%	20–28%
Data Analysis, Statistics and Probability	10%	6–20%	Data Analysis, Statistics and Probability	15%	12–20%

Source: Daro et al., 2007.

What do these comparisons indicate? First, they show considerable differences in content distribution across these six states for most strands at both Grades 4 and 8, as well as differences from the weight given a strand on the corresponding NAEP test. The percentages or content emphasis for each strand at each grade level for each of the six states can be seen in Table 3. Table 4 shows the percentages for each strand on the 2003 and 2007 NAEP tests. In Grade 4, NAEP has a greater emphasis on Numbers and in Measurement and Geometry combined than all six states but a much lower percentage in Algebra and Data Analysis, Statistics, and Probability. In Grade 8, NAEP still has a greater emphasis on Measurement and Geometry combined than all six states, but (because of a change in the weight assigned Algebra from 2003 to 2007) now has a higher percentage of its 2007 test in Algebra than all six states. The NAEP tests tend to be lower in the percentage of items in Data Analysis, Statistics, and Probability at both grade levels than most of the six states.

More information from a literature review offers additional support for organizing NAEP and state assessments in the areas of content frameworks can be found in Appendix C.

Table 3: Summary of State Mathematics Test Content, Grade 4 and 8

State	Grade	Content Strand Weights - Number of Questions or Points Possible (% of total)						Total Questions or Points	Item Formats	Use of testing aids		
		Number	Measurement	Geometry	Algebra and Functions	Statistics, Data Analysis, and Probability	Other		Constructed Response	Calculators	Manipulatives	Formula Sheets
California	4th	31 (48%)	12 (18%)		18 (28%)	4 (6%)		65 (100%)	Multiple Choice	No	—	—
	8th	100%						65	Multiple Choice	No	No	No
Georgia (QCC)	4th	— (20%)	— (23%)		— (12%)	— (10%)	— (35%)	— (100%)	Multiple Choice	No	—	—
	8th	— (17%)	— (22%)		— (23%)	— (12%)	— (26%)	— (100%)		No	—	—
Indiana	5th	13* (15%)	13* (15%)	11* (13%)	14* (17%)	10* (12%)	23* (27%)	84* (100%)	Multiple Choice, Constructed Response	No	—	No
	9th	9* (10%)	16* (17%)	10* (11%)	25* (27%)	11* (12%)	22* (24%)	93* (100%)		Yes	—	Yes
Massachusetts	4th	19* (35%)	13* (25%)		11* (20%)	11* (20%)		54* (100%)	Multiple Choice, Short Answer, Open Response	No	Yes	No
	8th	14* (26%)	14* (26%)		15* (28%)	11* (20%)		54* (100%)		by section	No	Yes
Texas	4th	11 (26%)	6 (14%)	6 (14%)	7 (17%)	4 (10%)	8 (19%)	42 (100%)	Multiple Choice, some Gridded	No	ruler	measurement conversions
	8th	10 (20%)	5 (10%)	7 (14%)	10 (20%)	8 (16%)	10 (20%)	50 (100%)		No	ruler	measurement conversions
Washington	4th	3–6 (8–17%)	3–6 (8–17%)	3–6 (8–17%)	3–6 (8–17%)	3–6 (8–17%)		35 (100%)	Multiple Choice, Short Answer, Extended Response	by section	by section	High School only
	8th	4–7 (9–20%)	4–7 (9–20%)	4–7 (9–20%)	4–7 (9–20%)	4–7 (9–20%)		50 (100%)		by section	by section	High School only

Notes:
* Content strand weight based on number or points possible instead of number of items in strand.
— Information not available

California: The expectation for 8th grade is that students will take CST Algebra 1 test. However, only about half the cohort takes that test. The others take a general math test as they are not ready for algebra.

Indiana's ISTEP+ is administered in the fall of each academic year and draws from the curricula of all previous grades.

Other strands are Computation and Problem Solving (Georgia and Indiana) and Mathematical Processes and Tools (Texas).

Source: This table was created for the Task Group by STPI using publicly available data from state Web sites. Data on California from S. Valenzuela (personal communications, February 1, 2008).

Table 4: Summary of 2003 and 2005 NAEP Mathematics Test Content, Grade 4 and 8

Year	Grade	Content Strand Weights - (% of total)					Cognitive Dimension			Item Formats	Use of testing aids		
		Number	Measurement	Geometry	Algebra and Functions	Statistics, Data Analysis, and Probability					Calculators	Manipulatives	Formula Sheets
2003	4th	40%	20%	15%	15%	10%	Conceptual: at least 1/3 of items	Procedural: at least 1/3 of items	Problem Solving: at least 1/3 of items	Multiple Choice (50%), Short and Extended Constructed Response (50%)	four function calculators provided for approximately 1/3 of items	students are provided rulers	Selected formulas and conversion factors (ones students are not necessarily expected to have memorized) are given on a per-item basis (e.g., volume of a cylinder, number of feet in a mile).
	8th	25%	15%	20%	25%	15%					scientific calculators provided for approximately 1/3 of items	students are provided rulers and protractors	
2005	4th	40%	20%	15%	15%	10%	Low Complexity 25% of score	Moderate Complexity 50% of score	High Complexity: 25% of score	Multiple Choice (64%), Short Constructed Response (32%), Extended Constructed Response (4%)	four function calculators provided for approximately 1/3 of items	students are provided rulers	
	8th	20%	15%	20%	30%	15%				Multiple Choice (69%), Short Constructed Response (28%), Extended Constructed Response (4%)	scientific calculators provided for approximately 1/3 of items	students are provided rulers and protractors	

Notes:
Various populations, rather than individual students, are the targets of the NAEP assessments. In particular, the assessment administered to any given student does not follow all the strict NAEP guidelines for mathematics assessment composition. Instead, the guidelines apply to the entire set of items for a given year and grade. The entire set of items consists of ten 25-minute blocks. The booklets administered to students participating in the mathematics assessment contain only two 25-minute blocks, in part to minimize the burden on students participating in the assessment. In effect, each student takes one-fifth of an assessment.

Assessments in 2003 and earlier classified the "cognitive dimension" of an item according to the "mathematical ability" required of a student responding to the item (conceptual understanding, procedural knowledge, and problem solving). The 2005 assessment changed the focus to the item itself; it classified the cognitive dimension of an item according to its complexity (low, moderate, high). On the 2003 assessments, a single item may be assigned to more than one mathematical ability level. Thus, this rule means that at least one-third of the items must have a major element of conceptual understanding. For 2005 Item Format Percentages see http://www.ed.state.nh.us/education/doe/organization/Curriculum/NAEP/2005/NAEPReport4MathWCoverRecoveredCorrect.pdf.

Source: STPI compiled this table using information from 1) National Assessment Governing Board, U.S. Department of Education, Mathematics Framework for the 2005 National Assessment of Educational Progress, September 2004, retrieved on October 1, 2007 from http://www.nagb.org/pubs/m_framework_05/toc.html and 2) National Assessment Governing Board, U.S. Department of Education, Mathematics Framework for the 2003 National Assessment of Educational Progress, September 2002, retrieved on October 1, 2007 from http://www.nagb.org/pubs/math_framework/toc.html.

B. Question 2: How Are Performance Categories Determined?

Question 1 concerned the nature and the weighting of the content that should appear on the assessment of mathematics. Question 2 examines how students' scores on mathematics tests are assigned to a particular performance category, e.g., Basic or Proficient. Of foremost concern is the minimum performance level on a test required for a student to be placed in a certain category. Performance level categories appear on both NAEP and the state tests, but the labels and underlying procedures may differ.

1. Background

Establishing performance categories involves a set of procedures currently known in educational measurement as standard setting (or setting cut scores). Judgments about performance categories are made by a panel of persons selected for their expertise or educational perspective. The exact procedures for determining performance categories can range from the panel's judgment about the test as a whole (i.e., the minimum percent of items passed at the various levels) to quantified judgments of individual items with respect to expected performance of students in the categories.

Several procedures and methods for combining judgments in standard setting have been developed. These procedures typically involve training panelists on the definitions of the standards and the nature of performance within the categories, soliciting judgments about the relationship of the test to the performance categories, and providing successive feedback to the panelists about their judgments. Various methods to combine judgments have been developed. Variants of the Bookmark method and the Modified Angoff method involve panelists judging how students at varying levels of competency will respond to representative test items. In these two methods, the cut score for competency classifications is determined by linking the judgments to empirical indices of item difficulty. In contrast, the Body of Work method requires the panelist to classify representative students into competency categories by examining their full pattern of item responses. While the methods are all scientifically acceptable, they may differ in effectiveness. The Bookmark method may involve the most assumptions about the data, while the Body of Work method may demand the highest level of rater judgment. While more research is needed in this area, the Modified Angoff method performs well against several criteria for psychometric adequacy (Reckase, 2006).

The Task Group was interested in determining the nature of the performance categories and the standard-setting procedures and methods for NAEP and the six states.

2. A Review of State Assessments and NAEP

The standard-setting procedures of NAEP and six states were examined with respect to the following seven questions. Not all information was fully available on all questions for each state.

1) What are the performance standards of NAEP and the states?
2) How were the NAEP and state performance standards established?
3) Are they based on procedures in which experts inspect actual item content or on global definitions?
4) Are empirical procedures used to combine individual expert opinions?
5) What is the background of the experts?
6) What descriptions or instructions are given, if any, about the nature of performance at different levels?
7) Do the experts receive the items in an examination under the same conditions as the students?

To answer these questions, documents available from Web sites of NAEP (National Assessment Governing Board) and six states (California, Georgia, Indiana, Massachusetts, Texas and Washington) were retrieved by STPI and provided to the Task Group. These documents were reviewed for relevant data by the Task Group members.

For the first question, NAEP employs a three-category system, Basic, Proficient and Advanced. The six states employed similar categories, although some made more distinctions than others. California's performance categories are labeled as Far Below Basic, Below Basic, Basic, Proficient, Advanced; Georgia, Does Not Meet, Meets, Exceeds Standard; Indiana, Did Not Pass, Pass, Pass+; Massachusetts, Warning, Needs Improvement, Proficient, Advanced; Texas, Basic, Proficient, Advanced; and Washington, Basic, Proficient, Advanced. For NAEP and all six states, global definitions of the performance categories are available. The definitions are all characterized as "global" because the definitions were fairly abstract characterizations of behavior that would require high degrees of judgment to determine the categorization of student performance.

For the other six questions, we draw on data in Table 5. First, although there is variability in the methods, all states use a contemporary method for standard setting or setting cut scores. Second, the Bookmark method was most frequently applied in standard setting. Second, individual item content is judged in NAEP and in all states except Massachusetts, where whole tests from students at various performance levels are examined. Third, empirical combination of judgments is implemented in all states. Fourth, the background of the experts varies within panels and probably somewhat across states. For example, Georgia uses primarily classroom teachers as experts while Texas represents broader contingencies, including curriculum experts from higher education and non-educators. However, in both NAEP and the six states, classroom teachers generally predominate as standard-setting panelists. Fourth, all six states train the panelists prior to eliciting their ratings. Fifth, although all six states applied some training procedures for panelists, the Task Group cannot judge the quality without having access to exact content. Sixth, only two states have the panelists experience the items in the same way as the test-takers.

Table 5: Information on Features of Standard-Setting Procedures (Setting Cut Scores) for NAEP and the Six States

	1. How Established?	2. Item Content Judgments?	3. Combination Procedures	4. Background of Experts	5. Instructions & Definitions	6. Test Taken?
NAEP	Modified Angoff Method	Yes	Empirical with successive feedback.	55% teachers, 15% non-teacher educators, and 30% members of the general public. Panelists should be knowledgeable in mathematics. Panelists should be familiar with students at the target grade levels. Panelists should be representative of the nation's population in terms of gender, race and ethnicity, and region.	Yes	N/A
California	Bookmark Method	Yes	N/A	N/A	Yes	Yes
Georgia	Modified Angoff Method	Yes	Empirical with successive feedback.	Primarily the panelists selected were educators currently teaching in the grade and content area for which they were selected to participate.	Yes	Yes
Indiana	Bookmark Method	Yes	Empirical preliminary followed by feedback & consensus.	Not specifically given, but appears to be classroom teachers.	Yes	None specified. Probably first viewed in panel setting.
Massachusetts	Expert Opinion – Body of Work Method	No	Empirical aggregation of first judgments. Details not available about feedback & consensus.	The panel consists primary of classroom teachers, school administrators or college and university faculty, but also non-educators including scientists, engineers, writers, attorneys, and government officials.	Yes	None specified. Probably first viewed in panel setting.
Texas	Item-mapping	Yes	Empirical preliminary followed by feedback & consensus.	The majority of the panelists on each committee were active educators—either classroom teachers at or adjacent to the grade level for which the standards were being set, or campus or district administrative staff. All panels included representatives of the community "at large."	Yes	None specified. Items probably first viewed in panel setting.
Washington	Bookmark Method	Yes	Empirical preliminary followed by feedback & consensus.	The majority of the panelists on each committee were active educators—either classroom teachers with some representation of higher education.	Yes	Yes

Source: This table was created by the Task Group using publicly available information from state Web sites. Data on California from S. Valenzuela (personal communications, February 1, 2008).

3. Conclusion

Although NAEP and the six states the Task Group examined varied in both process and method for standard setting or setting cut scores, NAEP and all states for which information was available employed currently acceptable educational practices. The methods may differ in effectiveness; however, scant evidence on their efforts is available. The Bookmark method may involve the most assumptions about the data while the Body of Work method may demand the highest level of judgment from the raters. The Modified Angoff method is preferred (Reckase, 2006) because the assumptions of the Bookmark method (e.g., unidimensionality) are probably not met in practice. The Body of Work method is less often applied to year-end tests because it requires higher levels of judgments from the experts. More research is needed on standard-setting methods and processes.

It was found that classroom teachers, most of whom are not mathematics specialists, predominate in the standard-setting process. Higher levels of expertise, including the expertise of mathematicians, as well as mathematics educators, curriculum specialists, classroom teachers and the general public, should be consistently used in the standard-setting process. The Task Group also found that the standard-setting panelists often do not take the complete test as examinees before attempting to set the performance categories, and that they are not consistently informed by international performance data. On the basis of international performance data, there are indications that the NAEP cut scores for performance categories are set too high. This does not mean that the test content is too hard; it is simply a statement about the location of cut scores for qualitative categories such as "proficient" and "beyond proficient." Additional information on this literature review can be found in Appendix D.

C. Part I: Recommendations on Test Content and Performance Categories

1) NAEP and state tests must ensure a focus on the mathematics that students should learn with achievement on critical mathematics content reported and tracked over time. NAEP should ensure that the Conceptual Knowledge and Skills' Critical Foundations and elements of the Major Topics of School Algebra are integral components of the mathematics assessed. The Task Group proposes reorganization, as well as possible title changes, of NAEP's current five content strands:

    a. Number Properties and Operations should be renamed and expanded into two separate categories—Grade 4 Number: Whole Numbers; and Fractions and Decimals; and Grade 8 Number: Integers; and Fractions, Decimals, and Percent.

        1. Whole Numbers will include emphasis on place value, comparing and ordering, and whole number operations at Grade 4. This will be expanded to include work with all integers, including operations with negative and positive integers at Grade 8.

2. Fractions and Decimals will include recognition, representation and comparing and ordering at Grade 4. This will be expanded to include operations involving fractions, decimals, and percent at Grade 8.

b. Geometry and Measurement should be combined into one content area. Topics related to both Measurement and Geometry should serve as important contexts for problems within the Grade 4 and Grade 8 NAEP.

c. Within Algebra, a better balance is needed within the algebra subtopic of patterns, relations, and functions at this level. That is, there should be many fewer items on patterns.

d. Data Analysis and Probability should be renamed as Data Display at Grade 4 and expanded to include both data interpretation and probability at Grade 8.

2) Procedures should be employed to include a broader base for setting performance level categories:

a. The Task Group recommends that standard-setting (setting cut scores) panels include high levels of expertise, such as mathematicians, mathematics educators, and high-level curriculum specialists, in addition to classroom teachers and the general public.

b. The standard-setting panelists should take the complete test as examinees before attempting to set the performance categories.

c. The standard setting should be informed by international performance data.

d. Research is needed on the impact of standard-setting procedures and methods (e.g., Bookmark Method, Modified Angoff procedure) in promoting the representation of a broad base of judgments.

VI. Part II: Item and Test Design

It is important not only that appropriate content is measured and cut scores for student performance are set appropriately but also that test scores accurately reflect the competencies intended to be measured. That is, the measurement itself must be carried out in a high-quality and appropriate manner. Test specifications that dictate the content of mathematics are not sufficient to ensure that valid assessments will be obtained. Thus, the Task Group reviewed the area of item and test design.

A. Question 3. How Does Item Response Format Affect Performance on Multiple-Choice and Various Kinds of Constructed-Response Items?

1. Background

Constructed-response formats, in which the examinee must produce a response rather than select one, are increasingly utilized in standardized tests. One motivation to use the constructed-response format arises from its presumed ecological validity, the belief that it reflects tasks in academic and work settings, and stresses the importance of "real-world" tasks. Constructed-response formats also are believed to have potential to assess dynamic cognitive processes (Bennett, Ward, Rock, & Lahart, 1990), and systematic problem solving and reasoning at a deeper level of understanding (Webb, 2001), as well as to diagnose the sources of mathematics difficulties (Birenbaum & Tatsuoka, 1987). Finally, constructed-response formats also may encourage classroom activities that involve skills in problem solving, graphing, and verbal explanations of principles (Pollack, Rock, & Jenkins, 1992). However, these purported advantages can incur a cost. The more extended constructed-response formats require raters to score them. Hence, they are more expensive and create delays in test reporting, as well as possibly introducing subjectivity in scoring.

In contrast, multiple-choice items have been the traditional type used on standardized tests of achievement and ability for over a century. Multiple-choice items can be inexpensively and reliably scored by machines or computers; they may require relatively little testing time and they have a successful history for psychometric adequacy.

Constructed-response (CR) items vary substantially in the amount of material that an examinee must produce. There are three basic types of CR items:

- The *grid-in* constructed-response format (CR-G) requires the examinee to obtain the answer to the item stem and then translate the answer to the grid by filling in the appropriate bubble for each digit.

- The *short answer* constructed-response format (CR-S) varies somewhat. The examinee may be required to write down just a numerical answer or the examinee may need to produce a couple of words to indicate relationships in the problem. The

CR-S format potentially can be scored by machine or computer, given a computerized algorithm that accurately recognizes the varying forms of numerical and verbal answers. Further, an intelligent algorithm also may provide for alternative answers (e.g., slightly misspelled words, synonyms).

- The *extended* constructed-response format (CR-EE) requires the examinee to provide the reasoning behind the problem solution. Thus, the CR-EE format would include worked problems or explanations. This format readily permits partial credit scoring; however, human raters are usually required. Use of human raters, however, can lead to problems with consistency and reliability of scoring.

The stems of CR-G and CR-S items and multiple-choice (MC) items can be identical, especially if the correct answer is a number. It is not clear how the stems of CR-EE can be identical to MC items, although this possibility cannot be excluded.

The NMP Assessment Task Group examined the literature on the psychometric properties of constructed-response items as compared to multiple-choice items. The original focus was to address the following three questions:

1) Do the contrasting item types (e.g., multiple choice, constructed response) capture the same skills in these tests equally well?
2) What does the scientific literature reveal?
3) What are the implications for NAEP and state tests?

2. A Review of the Literature

Impact of response format on mathematical skills, knowledge, and strategies. Potentially the most pressing issue about response format is the extent to which the same skills, knowledge, and strategies can be measured by the MC and CR item formats. The research generally does not support major differences in the nature of the construct that is measured by CR and MC items, nor in the strategies that are applied. However, much more data on this issue are potentially available because many state accountability, graduation, and year-end tests employ both item formats.

Impact of response format on psychometric properties. The evidence about the psychometric properties of constructed-response items as compared to multiple-choice items is inconsistent and depends on the source and the design of the comparison. If the studies utilize operational test data, comparisons of MC and CR items have indicated greater omit rates and greater difficulty for the CR items. This pattern is probably repeated on many state tests and would be a strong finding if such data were available for study by the methods employed in the Task Group's study. It should be noted, however, that studies on operational test items were not designed to isolate the impact of format by controlling or measuring other properties of items. If the studies utilized stem-equivalent versions of MC and CR items, the difference in psychometric properties depended on other design features of the items, such as the nature of the distractors and the use of grid-in responses. For example, some studies have found the CR format to be more difficult, which is consistent with the operational test studies. Other studies, however, have found the MC items to be more difficult when the distractors are constructed to

represent common error patterns. Moreover, little evidence from any design is available to support differences between MC and CR items on item discrimination levels, differential item functioning (DIF), strategy use, and the nature of the construct that is measured.

Impact of response format on differences between groups. The results on the interaction of the magnitude of gender-related differences in performance and item format are inconsistent and depend on the design of the specific study. However, the evidence suggests either no impact of response format on gender-related differences or that the relatively lower scores of girls than boys on mathematics items may be lessened in the constructed-response format.

The interaction of racial-ethnic differences with item format also has been examined in several studies. The research provides some evidence that Black-White differences in performance in mathematics are lessened on CR item format as compared to the MC item format.

Other results on item format that are potentially interesting include Hastedt and Sibberns (2005) finding on TIMSS data that scores based on MC versus CR items produced only slight differences in the relative ranking of the various participating countries. And, DeMars (2000) found that the difference between MC and CR items in difficulty depended on the test context. The two item formats differed less in difficulty on high-stakes tests than on low-stakes tests.

Additional information from this literature review can be found in Appendix E.

3. Conclusion

The available evidence on comparing the psychometric properties of MC items and CR items must be interpreted in the context of several factors. These factors include the following limitations: 1) the limited scope of the available scientific literature, 2) the uncontrolled design features for comparisons based on operational tests, 3) the design strategy in available controlled comparisons of MC and CR items, 4) the limited scope of the controlled comparisons, and 5) the impact of test context on the relative performance on MC and CR items. These limitations and the methodology for this review are discussed in more length in Appendix E.

Given the limitations of the research, there is little or weak evidence to support the CR format as providing much different information than the MC format. For example, the available evidence provides little or no support for the claim that different constructs are measured by the two formats or that item discrimination varies across formats. Although some evidence suggests that CR items are more difficult, especially for the more extended CR formats, there is some contrary evidence that indicates that more difficult MC items can be constructed for their stem-equivalent CR items. Finally, the impact data do not support much difference between the two item formats. That is, the impact of response format on gender differences is inconsistent, while the impact on racial-ethnic differences is weak. Suggestions to guide the evaluation of assessment item design are listed in Appendix F.

B. Question 4: What are Some Nonmathematical Sources of Difficulty or Confusion in Mathematics Test Items That Could Negatively Affect Performance?

Because flawed and marginal items on NAEP and state assessments could affect performance of students and could affect trend lines, the Task Group probed this issue.

1. Background

A crucial skill in learning mathematics is gaining the ability to understand what mathematical relationships and operations are intended by the language of word problems. Word problems are very common in most if not all state assessments, as well as in school curriculum materials. Nevertheless, it is clear that several nonmathematical aspects of word problems can adversely influence performance on tests of mathematical competence. These include misleading language and confusing visual displays. Problems also can emerge when reading, writing, and other skills that overlap with mathematical competence have an undue influence on performance.

The chapter of the NVS report on "Language that is unclear, inconsiderate, or misleading" provides only three examples of test items that show language or wording defects, even though many test items exhibited difficulties of this nature, as suggested by the comments on pages 94 and 95 (Daro et al., 1997). In addition, only two of the examples in this section were mathematical story problems, or, as they may also be labeled, situated mathematics problems.

How prevalent are poorly worded problems on high-stakes assessments? The Task Group wanted to find out if there was evidence on the frequency of language or wording issues from other analyses of test items on state, NAEP, or commercial mathematics assessments. Have any researchers systematically analyzed state, national, or commercial tests to determine the number of problems with poorly chosen, or developmentally inappropriate, unnecessary, or misleading language? Have any researchers found empirical evidence on the difficulties that students in general or various subgroups of students have with items that could be judged as linguistically defective? Are there research-based recommendations on language or wording issues to avoid, not only in abstract mathematics problems (problems not contextualized in real life) but especially in applied, or situated, problems that typically use everyday language to describe the givens of a mathematical problem?

2. A Review of the Literature

The Task Group undertook a review of the literature by examining 28 studies that met the Panel's criteria for quantitative empirical studies. The methodology for this review can be found in Appendix F. The Task Group was able to group most of these 28 studies into three general areas of interest in mathematics assessments. Seven looked at gender-related issues, two of which (Sappington, Larsen, Martin, & Murphy, 1991; McLarty, Noble, & Huntley, 1989) examined whether gender-related wording in mathematics word problems could lead

to a difference in scores between boys and girls. A third (Low & Over, 1993) reported on whether girls more often incorporated irrelevant information into constructed responses than did boys. The other four examined the differences in boys' and girls' scores on mathematics word problems with respect to their format, i.e., whether they provided MC options for response or required CR (Reiss & Zhang, 2006; Pomplun & Capps, 1999; Ryan & Fan, 1996; Wilson & Zhang, 1998).

The studies that examined differential responses by gender to the format of a test item found that, as also noted in the section comparing MC and CR formats, that for the most part, females do better on CR formats, while boys do better on MC formats. However, complexities appear when these differences are explored in greater depth. Reiss and Zhang (2006) found, when they controlled for language skills, that girls did less well than boys on both types of formats, more so on MC than on CR. The researchers observed that "the advantage females have in reading and writing improves their mathematics scores" while "males' lower reading and writing scores negatively impact their mathematics performance" (p. 13). In no way, however, do the researchers suggest that reading and writing skills are not important in mathematics; at issue was the role these skills play in mathematics assessments. It was not clear to the researchers whether raters rated responses by females more favorably because their responses were more complete (and of a higher quality) or because the females wrote more words, a dilemma in interpretation that has been found in writing assessments.

Another four studies examined, in differing ways, mathematical problem-solving difficulties for students who have learning disabilities, mathematics disabilities, or low reading skills (Fuchs & Fuchs, 2002; Moyer, Moyer, Sowder, & Threadgill-Sowder, 1984; Moyer, Sowder, Threadgill-Sowder, & Moyer, 1984; Larsen, Parker, & Trenholme, 1978). These students have difficulty reading and understanding how to solve mathematics word problems, sometimes because of the syntactic complexity of the language. Indeed, three other studies (Ketterlin-Geller, Yovanoff, & Tindal, 2007; Bolt & Thurlow, 2006; Johnstone, Bottsford-Miller, & Thompson, 2006) used a "read-aloud" method to explore what these kinds of students find difficult, in part to determine how items might be altered to remove what these students verbalized as difficulties. But in none of these studies did the researchers explore what the tests were actually measuring or whether the test items were defective or abnormal in any way. As a result, they did not explore the effects of erroneous, misstated, or poorly constructed items on student performance.

Five other studies examined assessment issues for English language learners (ELLs) (Brown, 2005; Butler & Stevens, 1997; Abedi & Lord, 2001; Abedi, Lord, Hofstetter, & Baker, 2000; Abedi, 2003). The researchers were interested in the effects of these students' English language limitations on test performance, in ways to accommodate their English language limitations on test items, or in the effects of accommodations in test items on them. In none of these studies, however, did the researchers examine how appropriate the language in test items was for assessing their mathematical objectives, whether the studies examined the effects of the original language or of altered language.

The remaining studies examined a variety of other issues, ranging from correlations of socioeconomic status (SES), race, and ethnicity with achievement (Lubienski, 2001; Lubienski, 2002; Lubienski & Shelley, 2003), the influence of scoring quality on assessment

reliability (Myerberg, 1999), and various issues in mathematics education (Romberg, 1992), to the features of mathematics and language arts tests constructed in various ways (Perkhounkova & Dunbar, 1999), a processing model to predict difficulty (Kintsch & Greeno, 1985), and the effects of personalized and reworded mathematics word problems on problem-solving skill (DeCorte, Verschaffel, & De Win, 1985; Davis-Dorsey, Ross, & Morrison, 1991). However, none of these studies addressed the relationship between flawed items and individual or group differences in performance. That is, this research did not examine the suitability of the language in a word problem for assessing a mathematical objective.

The Task Group examined other research that focused on content validity, reliability, and item performance. They reviewed the bibliography of the NVS for references to studies addressing the language or wording of test items. The five-page NVS bibliography revealed no published studies on language-related factors in test items influencing mathematical performance. The Task Group also examined the table of contents of a newly published volume on mathematics assessment. Not one of the titles of the 22 chapters in *Assessing Mathematical Proficiency*, edited by Alan Schoenfeld, published by Cambridge University Press in May 2007, hints at a discussion of language or wording issues in test items. Nor does a recent article by Lane and Stone (2006). None of these works addressed the Task Group's specific interest in language or wording issues.

The Task Group's review of research on content frameworks also noted no studies examining mathematical item quality, aside from the NVS itself. Most of the studies describe content validity, and examine scope of content and depth of treatment within content areas in relation to national and international tests. Only a few comment extensively on item difficulty and complexity, which may be considered aspects of item quality or test content or both, as seems to be the case in a 2004 analysis of the contents of six state exit tests by Achieve, Inc.

In sum, while the Task Group found many studies on other aspects of mathematics assessments, including item performance and item difficulty, they did not locate any studies that examined how suitable the wording of a test item may be for its mathematical objectives or the effects of wording-related issues in test items on student performance. Therefore, the Task Group proceeded to examine an array of test items from NAEP and state tests to see what kinds of language or wording flaws could be found.

3. Seven Types of Flaws in Released Items from State Assessments and NAEP

The Task Group determined first the extent to which quality is an issue as it relates to the language or wording of an item. The Task Group did an initial cursory reading of the word problems in the 2005 NAEP assessments for Grade 4 and Grade 8, and in assessments for Grade 4 and Grade 8 from six states: California, Georgia, Indiana, Massachusetts, Texas, and Washington. No special significance should be attached to these particular states, except that they were included in the NVS report. The Task Group simply wanted a sufficiently varied pool of items. In all cases, it used only released items that were supplied to them.

The Task Group did not look for all kinds of defects in test items. It focused only on defects in the language or visual displays for word problems. It did not try to determine 1) correctness of answers, 2) appropriateness of constructed-response items, 3) the quality of the rubrics given for grading constructed responses, or 4) the quality of wrong options in multiple-choice items. A test item was judged unsatisfactory if its language or visual display seemed to be distracting, confusing, or misleading, or if its wording or context made the test item too difficult for some students with grade-level mathematics skills. That is, the Task Group focused on the wording or the context for the problem that might, in its judgment, lead some students to give wrong responses independent of their mathematical skills.

A sufficient number of unsatisfactory items were found to warrant a detailed review of released items, with the goal of pinpointing various types of flaws. An array of released test items from NAEP and state tests were then examined. The Task Group stresses that examples could have equally been selected from other states to illustrate the types of flaws found. Enough flawed items were found to support a recommendation that, in future test development, careful attention should be paid to exactly what mathematical knowledge is being assessed by a particular item and the extent to which the item is, in fact, focused on that mathematics.

Below are seven types of flaws that the Task Group identified. Some of the graphics below have been reduced in size for ease of presentation. The Task Group also found many examples of satisfactory word problems in which the nonmathematical knowledge is minimal and for which the student is expected, as appropriate for a mathematics test, to convert relationships described verbally into mathematical symbolism or calculations. See Appendix G for examples of satisfactory word problems.

1) Use of nonmathematical knowledge in a word problem that might not be equally available to all students, or use of terms whose meaning might not be equally available to all students.

   For example: Grade 8: Block 8, M12-Item 11 on NAEP 2005

   Ms. Thierry and 3 friends ate dinner at a restaurant. The bill was $67. In addition, they left a $13 tip. Approximately what percent of the total bill did they leave as a tip?

   A) 10 % B) 13 % C) 15 % D) 20 % E) 25 %

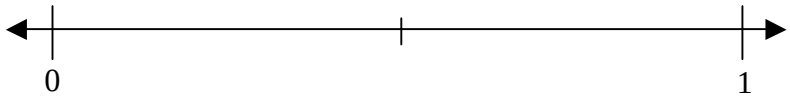
   Comment: This problem assesses conversion of a relationship described verbally into appropriate mathematical symbolism. But there are terminology issues that might trip up some students who would otherwise be able to understand the relationship described. They might not know what a tip is. More importantly, the use of ‘bill’ in one place and ‘total bill’ in another place clouds the relationship: Which is correct: 13/67 or 13/80? Some students will have the nonmathematical knowledge needed for this problem. For others, it will be unfamiliar or vague. It is this feature that makes this question flawed.

2) Use of a "real-world" setting for an essentially mathematical problem, a use that seems to serve only as a distraction because there is no apparent mathematical purpose for that setting.

For example: Grade 4, Massachusetts, 2006
Question 16: Multiple Choice
Reporting Category: Number Sense and Operations
Standard(s): 4.N.4 (No calculator permitted)

The picture below shows four fractions and a number line. Wilson's homework is to place a point on the number line for the location of each of the fractions.

$\frac{1}{6}$ $\frac{1}{3}$ $\frac{1}{12}$ $\frac{1}{4}$



If Wilson places the fractions correctly, which fraction will be closest to 0 *on the number line?*

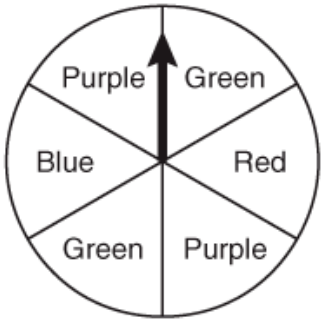
A. $\frac{1}{6}$ *B.* $\frac{1}{3}$ *C.* $\frac{1}{12}$ *D.* $\frac{1}{4}$

Comment: The content of this problem is strictly mathematical. The test-taker must identify which of four given fractions is closest to 0 on the number line. Students who know that $\frac{1}{12}$ is the smallest of the four fractions and understand the relationship between smallness and closeness to 0 should choose the correct answer. But some fourth-graders might be confused by seeing the fractions listed twice or be distracted by the story about Wilson. Straightforward mathematical questions should not be turned into questions about what someone else might do.

3) A focus on logical reasoning in what is essentially a nonmathematical problem.

   For example: Grade 4, Texas, Problem 32 in probability and statistics

   Kyle will spin the arrow on a spinner like the one shown below.



   If Kyle spins the arrow twice, which of these is NOT a possible outcome?

   F. Green, green G. Purple, green H. Blue, blue J. Red, orange

   Comment: There is no mathematics involved in selecting Option J, the right answer. The student who did not choose Option J did not read the problem with care. While careful reading is part of solving a mathematics problem, a problem involving logical reasoning on a broadly given mathematics assessment also should have a mathematical component.

4) An unnatural sequence of sentences in the word problem, probably created to make the problem "suitable" for assessing mathematical reasoning.

   For example: Grade 8, Washington, Problem 33

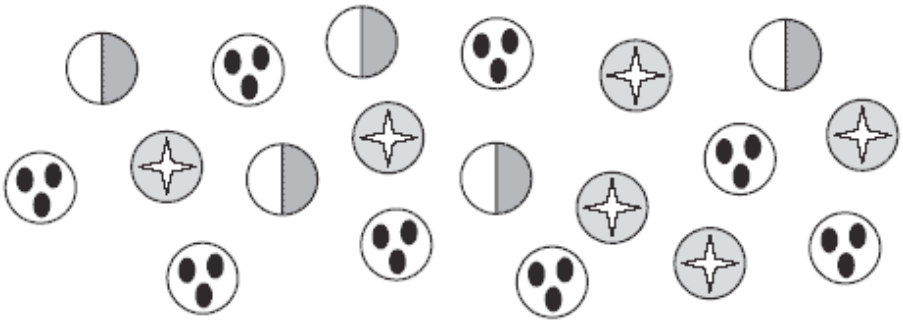
   Barb's class is conducting a walkathon. Her mother pledges $15.00. Her father pledges $3.50 per mile. Barb says she can determine the amount of money she will earn using the equation p = *3.5* m + *15. Explain the meaning of* m *in the equation. Explain the meaning of* p *in the equation.*

   Comment: The natural question has been convoluted so as to permit the test-maker to ask for explanations.

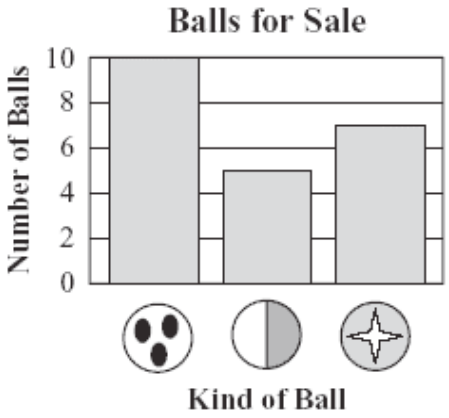
5) Use of a visual display having little connection with mathematics.

For example**:** Grade 4, Massachusetts, 2006, Question 20: Multiple Choice
Reporting Category: Data Analysis, Statistics, and Probability
Standard(s): 4.D.1 (No calculator permitted)

The picture below shows the balls that are for sale at a store.



Which of the following graphs shows the correct number of each kind of ball?

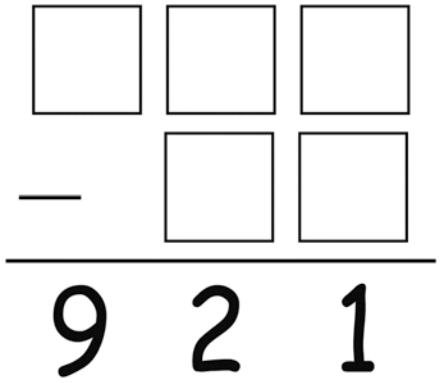


Comment: Solving this problem requires good eyesight as well as the ability to point and count with one hand while covering already counted items with the other.

6) A reliance on general understanding or ingenuity beyond the level of the actual mathematics involved.

For example: Grade 4, NAEP 2005, Problem 11, Page 3–16

Audrey used only the number tiles with the digits 2, 3, 4, 6, and 9. She placed one tile in each box below so that the difference was 921. Write the numbers in the boxes below to show where Audrey placed the tiles.

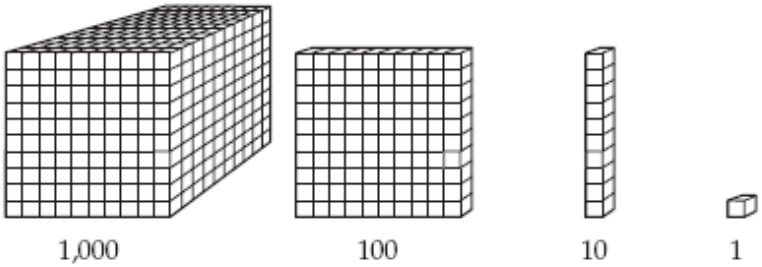


Comment: The mathematics involved is an understanding of the subtraction algorithm. But some students who are proficient with the subtraction algorithm might get this problem wrong because of its puzzle format. While the skills for doing this puzzle can be taught, they are not critical skills in mathematics, and large-scale assessments should not, in effect, be saying that it is important that every teacher teach these skills. Although it might be argued that this problem also involves mathematical reasoning and that only students who can reason at this level will do the problem correctly, this particular type of mathematical reasoning is not central to mathematics.

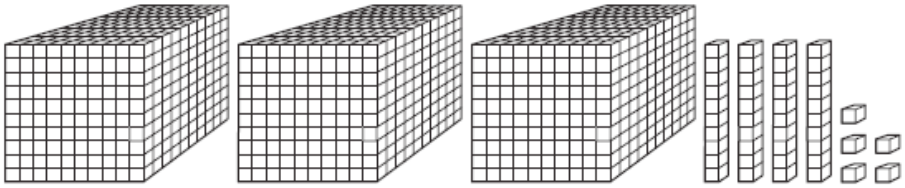
7) A required understanding of a pedagogical technique or tool that might be used for teaching mathematics but is not a part of its content.

For example: Grade 4, Indiana, Problem 3, Page 2–3

Look at the place-value blocks below.



What number does the following place-value model represent?



Answer _____

Comment: Place-value blocks are a tool for teaching, but one should not expect all students to be familiar with them. A student could figure out how to do the problem without ever having heard of a place-value block, but this makes the item more difficult for such a student than for a student who used one in class.

4. Discussion

A crucial skill in mathematics is the ability to understand what mathematical relationships and operations are intended by the language of word problems. But, flawed items that contain misleading language or confusing visual displays could affect performance of students and could affect trend lines and comparisons from NAEP and state assessments. Because the NVS report indicated that NAEP and state assessments include many items with misleading language and confusing visual displays, the Task Group searched the literature for relevant studies.

No directly relevant studies were identified on how suitable the wording of a test item may be for its mathematical objectives or the effects of wording-related issues on student performance. Thus, research on how aspects of mathematical problems and their item context (e.g., item format, problem scenario, wording, visual displays) are related to the construct that is measured, psychometric properties, and adverse impact should be supported, conducted, and reported. Furthermore, positive examples of well-designed items are needed to guide test development.

To begin this process, the Task Group examined state tests to provide examples of both undesirable and desirable content in mathematics word problems. In approaching this task, the Task Group's premise was that the major purpose of word problems on broadly given assessments should be to assess skill in converting relationships described verbally into mathematical symbolism or calculations. Many flawed items were found on the state tests in sufficient quantity to raise further concerns about item quality. The examples given above illustrate seven types of flaws that were found. Our findings, when combined with NVS findings on the large percentage of flawed and marginal items, point to possible gaps in test development procedures that need to be addressed. Developers of NAEP and state tests use sophisticated psychometric models and methods to select items and yet, according to NCES, these statistics are unable to detect the type of flaws noted in the NVS study.

Several aspects of the item and test development process may contribute to the large numbers of undetected flawed and marginal items.

First, there is a gap in the educational background of psychometricians and item writers. Psychometricians are trained to use highly sophisticated statistical models and data analysis methods for measurement but are not as familiar with issues of item design with respect to measuring mathematical constructs. Typical item writers and item evaluators often do not have a college degree in the appropriate subject and typically have little or no training in task and item design.

Second, item writers receive limited feedback from psychometricians on how the items they develop end up functioning for students at varying levels of performance. That is, the feedback mechanism does not provide sufficient information to help pinpoint the sources of item deficiencies.

Third, traditional psychometric indices of item quality are not sufficient indicators of item quality. According to the NCES report, the flawed and marginal items differed little from the adequate items in the average biserial correlations with total score, which is a classical test theory indicator of item quality. On other achievement tests, such as the state tests, the statistical criteria for evaluating item quality may be set much lower than the indices reported by NCES for NAEP. The lowered statistical criteria may be necessary to accommodate the inherent heterogeneity of educational achievement tests. Requiring high item discrimination may counter efforts to broadly represent an item domain. But an unintended consequence of broad representation is that it can allow even more items with marginal features to meet the low standard.

Fourth, it is increasingly maintained in some educational circles that ensuring that test items fulfill blueprints, along with traditional psychometric indices of item quality, provides sufficient evidence for test validity (e.g., Lissitz & Samuelson, 2007). As the findings of NVS suggest, these criteria do not provide the necessary assurance that students are responding to the items in the manner assumed by the test developers. Further, relying only on content specifications contrasts sharply with current standards for constructing tests (Myerberg, 1999), which expect multiple kinds of evidence for the construct validity of any test. While content specifications are part of the required evidence to support educational test validity, other kinds of evidence are also needed, including evidence based on theory, logical analysis, and scientific research (Embretson, 2007). Specifically, they include the current theory of the domain structure (e.g., the Conceptual Knowledge and Skills Task Group's view of how content "strands" relate to performance in algebra) and item design features. For the latter, the Task Group cannot assume without empirical evidence that students do indeed apply the knowledge, processes, and strategies that are intended for an item classified in a blueprint.

These several factors, taken together, work against ensuring that the items used to assess mathematical competencies are of the highest quality. Better procedures in item development, quality control, and oversight appear needed to counter this problem.

5. Conclusion

The Task Group examined state tests to provide examples of desirable content in mathematics word problems. In approaching this task, the Task Group's premise was that the major purpose of word problems on broadly given assessments should be to assess skill in converting relationships described verbally into mathematical symbolism or calculations. However, word problems also should satisfy the following conditions:

a) be written in a way that reflects natural and well-written English prose at the grade level assessed;
b) assess mathematics knowledge and skills for the grade level of the assessment, as judged by agreed-upon benchmarks, while restricting nonmathematical knowledge to what would be general knowledge for most students.

c) clearly assess skill in converting relationships described verbally into mathematical symbolism or calculations or, if a "real-world" setting is used, the problem uses a setting that aids in solving a problem that would be more cumbersome to state in strictly mathematical language.

Appendix G offers examples of word problems that follow these guidelines.

C. Question 5: How Are Calculators Used in NAEP and State Assessments and How Does Calculator Use Affect Performance?

1. Background

Tests that assess achievement in mathematics are administered under a variety of conditions, and using a variety of procedures, instructions, and technologies. For example, some tests are administered in large groups using paper and pencil booklets while other tests are administered in small groups in which each student is seated at a computer. These conditions may affect performance. Taken together, the diverse conditions under which tests are administered constitute an area of test design.

A very salient aspect of test design is the use of calculators. On some tests, calculators are made available for all items while, on other tests, calculators are available for only some items or for no items at all. Calculator use may affect performance in several ways, including total time on test, the strategies that students apply, the skills that are measured, and might result in differences between diverse groups.

Abt Associates Inc. conducted a review of the scientific literature on the effects of calculator usage on mathematics achievement test scores, using selection criteria described in the Assessment Task Group methodology statement. Below is a description of the studies identified, followed by a synthesis of the results of that literature search.

2. A Review of the Literature

Loyd (1991) noted, in a study involving eighth-graders completing a summer enrichment program (45% of the 160 students), that there was no evidence that use of calculators increased or decreased the speed with which examinees performed on four different types of items on a test. Calculator use was found to be advantageous with some item types (computation-based items), but less so with others.

Loveless (2004a) investigated the extent to which the use of calculators on NAEP computation items at the fourth-grade level produced significantly different results compared to student performance when calculators were not used. He also analyzed the impact of using calculators on performance gaps among black, white, and Hispanic students. Findings indicated that large differences in performance on computation items occurred when students used calculators on the fourth-grade NAEP. In 1999, students averaged 85% correct on whole number computation items when using calculators. On the same items, students who did not have access to calculators averaged only 57% correct on whole-number computation items.

Deeper analyses showed differences in achievement within the whole number operations of addition, subtraction, multiplication, and division. Interestingly, when comparing white, black, and Hispanic students, the gaps relative to achievement in computation narrow when black and Hispanic students have access to and use calculators. A conclusion drawn from this work is that, when young children have access to calculators on test items that focus on computations involving whole numbers, the results will not be indicative of their computational fluency with the operations assessed. In support of Loveless (2004a), Carpenter et al. (1981) also found increased performance on the long-term NAEP computation items for all three age groups if calculators were allowed but not on the problem-solving items.

Dye (1981) assessed eighth-graders using one of the forms used in a prior state mathematics assessment. One student group had access to calculators, one group was told that they could bring calculators and use them if they wished, and one group did not use calculators. The results indicated that the use of a calculator did not make any significant difference on final test scores; however, it was found that, if a mathematics test included many computation items, using a calculator would increase scores. It needs to be noted, however, that some design problems in this study lessened the Task Group's confidence in the conclusions drawn.

Hanson et al. (2001) studied 50 eighth-grade students completing a set of NAEP problems and a set of computation tests with their own calculator and comparable sets of problems with a scientific calculator provided to them. The researchers found no performance advantages associated with calculator type, nor was there an advantage related to student background characteristics (gender, race, math ability, socioeconomic status). Hanson et al. did find that calculator preference depended on the complexity of the student's own calculator relative to the standard one provided. The researchers concluded that there was no compelling reason to prohibit students from bringing their own calculators to a testing situation. On the other hand, the work of Chazan et al. (2007) seems to indicate that experience with calculators matters. They discovered, on the 2003 eighth-grade NAEP, that students who use calculators on a regular basis in their schooling scored higher on algebra and functions items than students who reported little use of calculators. Among all eighth-graders, regardless of socioeconomic status, the average scale scores of students who reported that they used calculators was 6 to 11 points higher on algebra and functions items than those who reported that they did not use calculators.

Brooks et al. (2003) analyzed calculator use on the Stanford Series Achievement Tests. They found that the score differences between calculator users and nonusers on the Stanford 10, which is the latest edition of the Stanford Achievement Series, were not large enough to warrant development of separate score conversion tables. This decision is consistent with findings on recent prior editions of the Stanford Series. The American College Testing Program (ACT) conducted a study in 1996 to assess effects of using a calculator on ACT's mathematics tests. The main purpose was to determine the effect of calculator use on the ACT's PLAN-ACT score scale. This study found that calculator use did not affect scores on either the PLAN or ACT tests. Additionally, the study revealed no differences related to gender and ethnicity with regard to calculator use on the PLAN and ACT tests. On the College Board Scholastic Assessment Test (SAT), however, Lawrence and Dorans (1994) did

find that, while most of the items on experimental versions of a pretest taken by thousands of students were unaffected, calculator usage affected item difficulties of those test items that had a heavy computational load. Such items became less difficult with calculator usage.

Long et al. (1989) looked at the role of calculators on the performance on the Missouri state test in 1987 for the tested 8th- and 10th-graders. In these two grades, the use of calculators was allowed but calculators were not provided. Students who used calculators did perform better but the advantage decreased as problem complexity increased. Ansley et al. (1989) and Forsyth and Ansley (1982) reported similar results with two samples of Iowa high school students. Use of calculators did not affect scores on the quantitative test of the Iowa Test of Educational Development, which is purported to be a test of problem solving. In the Ansley et al. (1989) study, all students in the 10th grade of a single high school were tested and randomly assigned to calculator or no-calculator condition. In the Forsyth and Ansley (1982) study several high schools that were matched on student characteristics participated, with each school being randomly assigned to calculator or no-calculator condition.

3. Conclusion

Based on the literature review conducted by the Task Group, it does not appear that using a calculator has a significant impact on test scores overall. However, the use of a calculator does seem to increase scores on computation-related items. Tables 3 and 4 capture key features of the NAEP and the six state tests, including information on calculator and tool use in assessment. Calculators are permitted for use in solving 35–40% of the fourth-grade NAEP test items. This is not the case for the six state tests reviewed. (One of the six states allowed calculators but only on certain sections.)

Thus, care must be taken to ensure that computational proficiency is not assessed using calculators. Additionally, the Task Group highlights one more aspect of this issue. It appears as if students who are comfortable with the calculator may have an advantage in knowing how and when the calculator may be used profitably in problem solving. If there are differences, therefore, in comfort level, the use of calculators might add nonmathematical sources of difficulty to test scores. This should be avoided.

D. Part II: Recommendations on Item and Test Design

1) The focus in designing test items should be on the specified mathematical skills and concepts, not item response format. The important issue is how to most efficiently design items to measure content of the designated type and level of cognitive complexity.

2) Much more attention should be paid to what mathematical knowledge is being assessed by a particular item and the extent to which the item addresses that knowledge.

3) Calculators (the use of which constitutes a design feature) should not be used on test items that seek to measure computation skills. In particular, NAEP should not permit calculator use in Grade 4.

4) Mathematicians should be included in greater numbers, along with mathematics educators, and curriculum specialists (not just classroom teachers and the general public), in the standard-setting process and in the review and design of mathematical item content for state, NAEP, and commercial tests.

5) States and NAEP need to develop better quality control and oversight procedures to ensure that test items reflect the best item design features, are of the highest quality, and measure what is intended, with nonmathematical sources of variance in performance minimized.

6) Researchers first need to examine whether the language in word problems is suitable for assessing their mathematical objectives before examining their impact in state assessments on student performance, especially the performance of special education students or English language learners.

7) More scientific research is needed on item and test design features.

BIBLIOGRAPHY

Abedi, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data*. CSE Report, CSE-R-603.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16–26.

Achieve, Inc. (2004). *Do graduation tests measure up?: A closer look at state high school exit exams*. Washington, DC: Achieve, Inc.

American College Testing Program. (1996). *Using calculators on the PLAN and ACT assessment mathematics tests.* Iowa City, IA: American College Testing Program.

Ansley, T.N., Spratt, K.F., & Forsyth, R.A. (1989). The effects of using calculators to reduce the computational burden on a standardized test of mathematics problem solving. *Educational and Psychological Measurement, 49*(1), 277–286.

Behuniak, P., Rogers, J.B., & Dirir, M.A. (1996). Item function characteristics and dimensionality for alternative response formats in mathematics. *Applied Measurement in Education, 9*(3), 257–275.

Bennett, R.E., Ward, W.C., Rock, D.A., & Lahart, C. (1990). Toward a framework for constructed-response items (ETS Research Report No. 90-7). Princeton, NJ: Educational Testing Service.

Birenbaum, M., & Tatsuoka, K.K. (1987). Open-ended versus multiple-choice response formats—It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*(4), 385–395.

Birenbaum, M., Tatsuoka, K.K., & Gurtvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement, 16*(4), 353–363.

Bolt, S.E., & Thurlow, M.L. (2006). *Item-level effects of the read-aloud accommodation for students with reading disabilities*. (Synthesis Report 65). Minneapolis, MN: University of Minnesota.

Brooks, T.E., Case, B.J., Cerrillo, T., Severance, N., Wall, N., & Young, M.J. (2003). *Calculator use on Stanford series mathematics tests*. Austin, TX: Harcourt Assessment, Inc.

Brown, C.L. (2005). Equity of literacy-based math performance assessments for English language learners. *Bilingual Research Journal. 29*(2), 337–363.

Burton, N.W. (1996). How have changes in the SAT affected women's math scores? *Educational Measurement: Issues and Practice, 15*(4), 5–9.

Butler, F.A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations.* No. 448 CSE Technical Report. Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Carpenter, T.P., Corbitt, M.K., Kepner, H.S., Lindquist, M.M., & Reys, R.E. (1981). Calculators in testing situations: Results and implications from national assessment. *Arithmetic Teacher, 28*(5), 34–37.

Chazan, D., Leavy, A.M., Birky, G., Clark, K., Lueke, M., McCoy, W., et al. (2007). What NAEP can (and cannot) tell us about performance in algebra. In P. Kloosterman & F. Lester, Jr. (Eds.), *Results and interpretations of the 2003 mathematics assessment of the national assessment of educational progress* (pp.186–188). Reston, VA: National Council of Teachers of Mathematics.

Daro, P., Stancavage, F., Ortega, M., DeStefano, L., & Linn, R. (2007). *Validity study of the NAEP mathematics assessment: Grades 4 and 8.* (Chapters 2 and 3). Washington, DC: National Center for Education Statistics and American Institutes for Research.

Davis-Dorsey, J., Ross, S.M., & Morrison, G.R. (1991). The role of rewording and context personalization in the solving of mathematical word problems. *Journal of Educational Psychology, 83*, 61–68.

De Corte, E., Verschaffel, L., & De Win, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology, 77,* 460–470.

DeMars, C.E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11*(3), 279–299.

DeMars, C.E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55–77.

Dossey, J.A., Mullis, I.V.S., & Jones, C.O. (1993). *Can students do mathematical problem solving? Results from constructed-response questions in NAEP's 1992 mathematics assessment* (No. 23-FR01). Princeton, NJ: Educational Testing Service.

Dye, D.L. (1981). *The use and non-use of calculators on assessment testing*. St. Paul, MN: Minnesota State Department of Education.

Embretson, S.E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher, 36,* 8, 449–456.

Executive Order No. 13398, National Mathematics Advisory Panel, *Federal Register,* Vol. 71, No. 77, p. 20519, April 21, 2006.

Forsyth, R.A., & Ansley, T. N. (1982). The importance of computational skill for answering items in a mathematics problem-solving test: Implications for construct validity. *Educational and Psychological Measurement, 42*(1), 257–263.

Fuchs, L.S., & Fuchs, D. (2002). Curriculum-based measurement: Describing competence, enhancing outcomes, evaluating treatment effects, and identifying treatment nonresponders. *Peabody Journal of Education, 77,* 64–84.

Fuchs, L.S., & Fuchs, D. (2002). Mathematical problem solving profiles of students with mathematics disabilities with and without comorbid reading disabilities. *Journal of Learning Disabilities, 35*(6), 563–573.

Gallagher, A. (1992). *Strategy use on multiple-choice and free-response items: An examination of differences among high scoring examinees on the SAT-M* (No. ETS Research Report No. 92-54). Princeton, NJ: Educational Testing Service.

Garner, M., & Engelhard Jr., G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12*(1), 29–51.

Hanson, K., Brown, B., Levine, R., & Garcia, T. (2001). Should standard calculators be provided in testing situations?: An investigation of performance and preference differences. *Applied Measurement in Education, 14*(1), 59–72.

Hastedt, D., & Sibberns, H. (2005). Differences between multiple choice items and constructed response items in the IEA TIMSS Surveys. *Studies in Educational Evaluation*, *31*, 145–161.

Hoffer, T., Venkataraman, L., Hedberg, E.C., & Shagle, S. (2007). *Final report on the national survey of algebra teachers for the National Math Panel*. Chicago: University of Chicago, National Opinion Research Center.

Hombo, C.M., Pashley, K., & Jenkins, F. (2001). *Are grid-in response format items usable in secondary classrooms?* Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Johnstone, C.J., Bottsford-Miller, N.A., & Thompson, S.J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners (Technical Report 44)*. Minneapolis, MN: National Center on Educational Outcomes.

Karantonis, A., & Sireci, S.G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1) pp. 4–12.

Katz, I.R., Bennett, R.E., & Berger, A.E. (2000). Effects of response format on difficulty of SAT-Mathematics items: It's not the strategy. *Journal of Educational Measurement, 37*(1), 39–57.

Kenney, P. (2000). Families of items in the NAEP mathematics assessment. In N.S. Raju, J.W. Pellegrino, M.W. Bertenthal, K.J. Mitchell, & L.R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP*. Washington, DC: National Academy Press.

Kenney, P., Silver, E., Alacaci, C., & Zawojewski, J. (1998). *Content analysis project: State and NAEP mathematics assessments. Report of results from the Maryland-NAEP Study*. Conducted under contract with the National Assessment Governing Board. Retrieved on September 1, 2007 from http://www.marces.org/mdarch/pdf/M031915.pdf.

Ketterlin-Geller, L., Yovanoff, P., & Tindall, G. (2007). Developing a new paradigm for conducting research on accommodations in mathematics testing. *Exceptional Children, 73*(3), 331–347.

Kintsch, W., & Greeno, G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92*(1), 109–129.

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress* (No. CSE Technical Report 357). Los Angeles, CA: Center for Research on Evaluation, Standards and Student Testing (CRESST) University of California Los Angeles.

Lane, S., & Stone, C.A. (2006). Performance assessments. In B. Brennan (Ed.), *Educational measurement*. Westport, CT: Praeger.

Larsen, S., Parker, R., & Trenholme, B. (1978). The effects of syntactic complexity upon arithmetic performance. *Learning Disability Quarterly, 1*(4), 80–85.

Lawrence, I.M., & Dorans, N.J. (1994). *Optional use of calculators on a mathematical test: Effect on item difficulty and score equating.* Princeton, NJ: Educational Testing Service.

Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring.* Symposium conducted at the Council of Chief State School Officers National Conference on Large-scale Assessment, Phoenix, AZ.

Linn, R., & Kiplinger, V. (1994). *Linking statewide tests to the national assessment of educational progress: Stability of results.* National Center for Research on Evaluation, Standards, and Student Testing Technical Report 375. Los Angeles, CA: University of California–Los Angeles.

Lissitz, R.W., & Samuelson, K. (2007). Suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36,* 8, 437–448.

Long, V.M., Reys, B., & Osterlind, S.J. (1989). Using calculators on achievement tests. *Mathematics Teacher, 82,* 318–325.

Loveless, T. (2004a). *Computation skills, calculators, and achievement gaps: An analysis of NAEP items.* American Educational Research Association Conference, San Diego, CA.

Loveless, T. (2004b). *The 2004 Brown Center Report on American education: How well are American students learning?* Washington, DC: The Brookings Institution Press.

Low, R., & Over, R. (1993). Gender differences in solution of algebraic word problems containing irrelevant information. *Journal of Educational Psychology, 85*(2), 331–339.

Loyd, B.H. (1991). Mathematics test performance: The effects of item type and calculator use. *Applied Measurement in Education, 4*(1), 11–22.

Lubienski, S.T. (2001). *A second look at mathematics achievement gaps: Intersections of race, class, and gender in NAEP data*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Lubienski, S.T. (2002). A closer look at Black-White mathematics gaps: Intersection of race and SES in NAEP achievement and instructional practices data. *The Journal of Negro Education, 71*(4), 269–288.

Lubienski, S.T., & Shelley II, M.C. (2003, April). *A closer look at U.S. mathematics instruction and achievement: Examinations of race and SES in a decade of NAEP data.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

McLarty, J.R., Noble, A.C., & Huntley, R.M. (1989). Effects of item wording on sex bias. *Journal of Educational Measurement 26*(3), 285–293.

Moyer, J.C., Sowder, L., Threadgill-Sowder, J., & Moyer, M.B. (1984). Story problem formats: Verbal versus telegraphic. *Journal for Research in Mathematics Education, 15*(1) 64–68.

Moyer, J.C., Moyer, M.B., Sowder, L., & Threadgill-Sowder, J. (1984). Story problem formats: Drawn versus verbal versus telegraphic. *Journal for Research in Mathematics Education, 15*(5) 342–351.

Mullis, I. V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y., Arora, A., & Erberber, E. (2007). TIMSS 2007 assessment frameworks. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Myerberg, N.J. (1999, April). *The relationship between scoring quality and assessment reliability*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

National Center for Education Statistics (2007). *Mapping 2005 state proficiency standards onto the NAEP scales* (NCES 2007-482). U.S. Department of Education. Washington, DC: Author.

National Mathematics Advisory Panel (2008). Reports of the task groups and subcommittees. Washington, DC: Author.

Neidorf, T., Binkley, M., Gattis, K., & Nohora, D. (2006). *Comparing mathematics content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 assessments* (NCES 2006-029). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

O'Neil Jr., H.F., & Brown, R.S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education, 11*(4), 331–351.

Perkhounkova, Y., & Dunbar, S.B. (1999, April). *Influences of item content and format on the dimensionality of tests combining multiple-choice and open response items: An application of the Poly-DIMTEST procedure.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Pollack, J.M., Rock, D.A., & Jenkins, F. (1992, April). *Advantages and disadvantages of constructed-response item formats in large-scale surveys.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Pollack, J.M., & Rock, D.A. (1997). *Constructed response tests in the NELS: 88 high school effectiveness study. National Education Longitudinal Study of 198, second follow up* (No. NCES 97-804). Chicago, IL and Princeton, NJ: National Opinion Research Center and Educational Testing Service.

Pomplun, M., & Capps, L. (1999). Gender differences for constructed-response mathematics items. *Educational and Psychological Measurement, 59*(4), 597–614.

Reckase, M.D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice, 25*(2), pp. 4–18.

Reiss, P.P., & Zhang, S. (2006). Why girls do better in mathematics in Hawai'i: A causal model of gender differences on selected and constructed response items. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

Romberg, T.A. (1992). Mathematics assessment and evaluation: Imperatives for mathematics education. New York: SUNY Press.

Romberg, T.A. (1992). Problematic features of the school mathematics curriculum. In P.W. Jackson (Ed.). Handbook of Research on Curriculum. New York: Macmillan.

Romberg, T.A. (1992). Perspectives on scholarship and research methods. In D.A. Grouws (Ed.) Handbook of Research on Mathematics Teaching and Learning. New York: Macmillan.

Ryan, K.E., & Fan, M. (1996). Examining gender DIF on a multiple-choice test of mathematics: a confirmatory approach. *Educational Measurement: Issues and Practices, 15*(4), 15–20.

Sappington, J., Larsen, C., Martin, J., & Murphy, K. (1991). Sex differences in math problem solving as a function of gender-specific item content. *Educational and Psychological Measurement, 51*(4), 1041–1048.

Schmidt, W., & Houang, R. (2007). Lack of focus in mathematics: Symptom or cause? In T. Loveless (Ed.), *Lessons learned: What international assessments tell us about math achievement.* Washington, DC: Brookings Institution Press.

Schneider, M. (2007). *Response to the "Validity study of the NAEP mathematics assessment: Grades 4 and 8*." U.S. Department of Education, National Center for Education Statistics. Retrieved November 26, 2007, from http://nces.ed.gov/whatsnew/commissioner/remarks2007/11_23_2007.asp.

Schoenfeld, A. (Ed.). (2007). Assessing mathematical proficiency. New York: Cambridge University Press.

Silver, E., & Kenney, P. (1993). An examination of relationships between the 1990 NAEP mathematics items for grade 8 and selected themes from the NCTM standards. *Journal for Research in Mathematics Education, 24*(2), 159–167.

Traub, R.E., & Fisher, C.W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1*(3), 355–369.

U.S. Department of Education (2002). *National Assessment Governing Board. Mathematics framework for the 2003 national assessment of educational progress*. Retrieved on October 1, 2007 from http://www.nagb.org/pubs/math_framework/toc.html.

U.S. Department of Education (2004). *National Assessment Governing Board. Mathematics framework for the 2005 national assessment of educational progress*. Retrieved on October 1, 2007 from http://www.nagb.org/pubs/m_framework_05/toc.html.

Webb, D.C. (2001, August). *Classroom assessment strategies to help all students master rigorous mathematics*. Co-presenter for a four-day workshop for District of Columbia Public School mathematics teachers, American Association for the Advancement of Science, Washington, DC.

Wilson, L.D., & Zhang, L. (1998, April). *A cognitive analysis of gender differences on constructed-response and multiple-choice assessments in mathematics*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

Wu, H. (2007). *Fractions, decimals, and rational numbers.* Retrieved on November 1, 2007 from http://math.berkeley.edu/~wu/.

Zieleskiewicz, J. (2000). Subject-matter experts' perceptions of the relevance of the NAEP long-term trend items in science and math. In N. S. Raju, J.W. Pellegrino, M.W. Bertenthal, K.J. Mitchell, and L.R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP*. Washington, DC: National Academy Press.

APPENDIX A: National Assessment of Educational Progress (NAEP)

Background

Since 1969 the National Assessment of Educational Progress (NAEP) has been regularly conducting assessments of samples of the nation's students attending public and private schools at the elementary, junior high, and high school levels. NAEP's goal, since its inception, has been to make available reliable information about the academic performance of U.S. students in various learning areas. To this end, NAEP has produced more than 200 reports in 11 instructional areas.

Teachers, administrators, and researchers from across the United States have helped propel NAEP into the valuable informational source it is today. As a result, members of the educational community are able to make use of NAEP's findings on students' learning experiences to inform policymakers and to improve students' educational experiences.

NAEP is an indicator of what students know and can do. Only group statistics are reported, no individual student or teacher data are ever released.

NAEP is conducted under congressional mandate and is directed by National Center for Educational Statistics (NCES) of the U.S. Department of Education. NCES currently contracts with the Educational Testing Service (ETS) to design instruments, and conduct data analysis and reporting; Westat, Inc., to conduct sampling and data collection activities; and National Computer Systems to manage materials distribution, scoring, and data processing.

Who Is Sampled?

Every 2 years, NAEP assesses nationally representative samples of more than 120,000 students in public and private schools in Grades 4, 8, and 12. The NAEP state assessment samples also include students from both public and private schools to be representative of schools in the participating state. Scientific sampling procedures are used to ensure reliable national, regional, and state samples.

Schools

Schools are randomly selected for NAEP based on demographic variables representative of the nation's schools. Trained NAEP staff members administer the assessment. In NAEP state assessments, the participating schools work with a coordinator designated by the respective state department of education to collect information on a statewide level.

Students

Students are selected randomly; their names are not collected. Confidentiality of all participants is ensured and their names do not leave the school.

What Subjects Are Assessed?

The academic subject areas assessed vary from year to year. According to the law authorizing NAEP, all subjects listed under National Educational Goal 3 are to be tested periodically in the national assessment. Reading, writing, mathematics, and science are the most frequently assessed subjects. To minimize the burden on students and schools, no student takes the entire assessment. Instead, assessment sessions are limited to 1 $\frac{1}{2}$ to 2 hours. Questionnaires are also given to students, teachers, and principals to obtain current information about school and instructional practices that may influence learning and student performance.

When Do Assessments Take Place?

Assessments occur throughout the school year; however, most are conducted January through March. State assessments occur in February.

Source: http://www.nagb.org/about/abt_naep.html.

APPENDIX B: Methodology of the Assessment Task Group

The Assessment Task Group addressed several different kinds of questions related to the influence of different item types and test administration procedures on student responses; the content validity, item types, and item difficulties of the National Assessment of Educational Progress (NAEP) and state tests; and the way NAEP and state performance categories are established. Several different sources of information contributed to the resulting report, each involving somewhat different methodological considerations. These included a review of relevant research literature, elaboration of findings from a recently completed report by the NAEP Validity Studies (NVS) Panel, and an analysis of the content and performance categories of NAEP and selected state mathematics achievement tests.

Literature Review

Literature searches were conducted by Abt Associates Inc. (Abt) to locate studies on mathematics assessment that included the content validity of NAEP, the effect of test administration procedures, the influence of item wording, and the skills and concepts captured by various item types. The criteria for selecting relevant studies required that they a) be published between 1970 and 2007 in a journal, government or national report, book, or book chapter; b) involve K–12 mathematics assessments; c) be available in English; and d) use quantitative methods for analyzing data. Because of the diversity of pertinent topics and associated forms of research, no other general methodological criteria were imposed but, rather, the Task Group made individual judgments about the appropriateness and quality of each candidate study located in the search.

Electronic searches were made in Education Resource Information Center (ERIC), PsycInfo, and the Social Sciences Citation Index (SSCI) using the search terms identified by the Assessment Group, shown below.

Assessment or testing *and* math *and* each of the following:

item language	verbiage	short answer
item wording	excessive language	true-false
question language	validity	administration procedure
question wording	reliability	manipulatives
item wordiness	item type	calculators
language bias	item structure	formulas
linguistic simplification	multiple choice	accommodations
language density	constructed response	Bloom’s taxonomy
essential language	open response	bias

Additional studies were identified through manual searches of relevant journals, Internet search engines, and reference lists and recommendations from experts. Abstracts from these searches were screened for relevance to research questions and appropriate study design. For studies deemed relevant, the full study report was obtained. Citations from those articles and research reviews were also examined to identify additional relevant studies. Abt extracted summary information from the qualifying studies and provided it to the Task Group along with the complete articles. The Task Group then further screened the studies to narrow those included in their reviews to the most relevant and highest quality available.

The Task Group drew on the studies located and screened through this procedure for its reviews of the content validity of NAEP mathematics assessment, the influence of item format on test performance, and calculator use during mathematics achievement assessments.

Analysis of NAEP and State Test Mathematics Items

The Task Group's assessment of the mathematics items used in NAEP and state achievement tests initially drew on a report of a validity study released in the early fall of 2007 (Daro et al., 2007). The Task Group was briefed on the NAEP Validity Study by the authors, given access to an embargoed early version of the report, and shown the response of National Center for Education Statistics (NCES) to that report. The Task Group then conducted its own further analysis of the items in the six state tests represented in the NVS sample.

The IDA STPI (Institute for Defense Analyses Science and Technology Policy Institute) provided the Task Group with information on the test frameworks, testing procedures, and test items for the six state mathematics tests used in the NVS report: California, Georgia, Indiana, Massachusetts, Texas, and Washington. STPI collected the state assessment information it provided to the Task Group from each state's Department of Education Web site and the NAEP information from the U.S. Department of Education's National Center for Education Statistics Web site. STPI assembled the relevant material it located in response to the Task Groups request but did not conduct any analyses of that material.

One of the mathematicians on the Task Group then analyzed the released items for the Grade 4 and Grade 8 state tests and NAEP provided in the STPI material. The results of that analysis were then further reviewed by the other Task Group members and incorporated in the Assessment Report to supplement the analysis by five mathematicians that was reported by the NVS.

Analysis of the Content and Performance Categories of NAEP and State Mathematics Tests

The material provided to the Task Group by STPI on the test frameworks, testing procedures, and test items for the six state mathematics tests used in the NVS report included descriptions of the content of each test, the performance categories, and the procedures for establishing the performance categories. STPI collected this information from each state's department of education Web site and the NAEP information from the NCES Web site. The reports and descriptive summaries they provided were reviewed by the members of the Task Group and used, along with studies from the literature review, as the basis for their analysis of these topics.

APPENDIX C:
Test Content Frameworks and Items: A Review

Ten studies that assessed, in various ways, the content validity of the mathematics portion of the National Assessment of Educational Progress (NAEP) for Grades 4 and 8 were reviewed for this portion of the report, most of which are reviewed in this section. Two of the studies (Daro et al., 2007; Kenney et al., 1998) compared the NAEP framework with the framework of other mathematics assessments (among other topics). Five of the studies (Kenney et al., 1998; Loveless, 2004b; Neidorf et al., 2006; Silver & Kenney, 1993; Zieleskiewicz, 2000) compare NAEP items to frameworks from other mathematics assessments. Three studies (Daro et al., 2007; Silver & Kenney, 1994; Silver et al., 1992) compared NAEP items to the NAEP framework. The remaining three studies used a variety of methods to explore NAEP content and items. Kenney (2000) discussed the rationale for creating families of items and demonstrates the creation of families with released NAEP items. Linn and Kipplinger (1994) tested whether an equating function could be developed to equate standardized achievement test scores to NAEP scores.

Test Content Frameworks

Daro et al. (2007) convened an expert panel involving mathematicians, mathematics educators, and an expert on state-based mathematics standards. They compared the NAEP mathematics framework with the standards and frameworks (test blueprints) of six states (California, Massachusetts, Indiana, Texas, Washington, and Georgia), two high-performing nations (Singapore and Japan), and standards outlined by the National Council of Teachers of Mathematics (NCTM) and Achieve, Inc. In examining the content areas of Number Properties and Operations, Algebra, Geometry, Measurement, and Data Analysis and Probability in the 2005 NAEP mathematics framework, the reviewers attempted to determine if NAEP was missing something or overemphasizing topics in a given content area. The reviewers then described what was being overemphasized and rated the emphasis of each content topic as compared to each of the six states and Singapore.

Item Comparisons within Content Frameworks

Daro et al. (2007) indicated that, at the fourth-grade level, the only area where NAEP has a higher percentage of items than the other frameworks was Measurement. It also was noted that, while Number Properties and Operations is the most emphasized content area at the fourth-grade level, the NAEP provides a very limited assessment of fractions at this level. The NAEP Geometry items assess symmetry and transformations more than the other states, and emphasize parallel lines and angles less than the comparison states. Moreover, the fourth-grade NAEP Algebra content area appears to be especially problematic. The pattern items overemphasize sequences of numbers that grow in a regular way; and, this type of pattern is used in NAEP more than in the other frameworks. Mathematics reviewers

suggested that NAEP consider pattern items based on the relationship between two quantities. The review panel recommended better item balance within the algebra subtopic of patterns, relations, and functions at this level.

The review of the eighth-grade NAEP Number Properties and Operations content area found an emphasis on topics from number theory—factorization, multiples, and divisibility. Given this review, a focus dedicated to ensuring that eighth-grade students have developed proficiency with whole numbers, negative integers, and fractions, decimals, and percent may be considered given their importance as prerequisites for algebra. Because this content area stood out in the review as undersampling grade level content, "It is possible that students are making gains in this content area that are not being detected by NAEP" (p. 123). In the Panel's judgment, it is also possible that students are losing ground that goes undetected. Indeed, because the NAEP minimizes this area, this could be a driving force for reduced attention to it within the school curriculum.

The eighth-grade Measurement content area appears to be assessing lower-level concepts and skills and, as a result, NAEP may be underestimating achievement. It also was noted that the larger number of measurement items is "not well leveraged to include fractions or decimals used in realistic situations" (p. 126). The review of the Geometry items indicated wide variation across the six states. The NAEP at Grade 8 includes more geometry than the comparison states or nations. A consensus does not appear to exist on what is important in geometry at Grade 8.

Loveless (2004b) found that the majority of the fourth- and eighth-grade NAEP items assessing problem solving, algebra and numbers sense involve whole numbers. While this is understandable at the fourth-grade level, it is cause for serious concern at the eighth-grade level. Fractions, decimals, and percent are under-assessed. In items assessing problem solving, whole numbers make up approximately 72% of the fourth-grade items and approximately 70% of the eighth-grade items. The possible overemphasis regarding whole numbers continues for the eighth-grade NAEP algebra items as well. This suggests again raising the level of arithmetic to include more direct assessment of fractions, decimals, and percents within the number and algebra content areas and that the confinement of arithmetic to whole numbers is largely responsible for the low grade-level demands of many of the items Loveless also questions the identification of some of the items as algebra.

Neidorf et al. (2006) compared the mathematics content in NAEP, Trends in International Mathematics and Science Study (TIMSS), and the Program for International Student Assessment (PISA). They note that the NAEP and TIMSS content frameworks are quite consistent with regard to their basic organization of mathematics content. They both have five main content areas: Number, Measurement, Geometry, Data, and Algebra. They did note different emphases within topics and subtopics and in some grade level expectations. PISA differs from both NAEP and TIMSS in that it samples 15-year-olds rather than specific grades and that it focuses on problem-solving using what are called, in the world of K–12 education, "real-world" problems, rather than curriculum content areas. However, the mathematics content assessed by PISA is consistent with the NAEP eighth-grade mathematics framework.

Neidorf et al. (2006) noted that PISA has more items classified as Data Analysis and fewer as Algebra and Number Sense than the other two assessments. The NAEP and TIMSS comparison indicates that there is a greater emphasis by NAEP on applications. Moreover, TIMSS includes a higher proportion of items involving ratio and proportion and, thus, has a more appropriate balance for assessing number using fractions, decimals, and percent. While there is considerable overlap in the NAEP and TIMSS assessments involving measurement, there is greater emphasis in NAEP on using measurement instruments and units of measurement. TIMSS included a higher percentage of items involving estimation, calculation, or comparing perimeter, area, volume, and surface area. With regard to data, NAEP has a greater proportion of probability items, whereas TIMSS has a greater proportion of items than NAEP that emphasize reading, interpreting, and making predictions from tables and graphs and data representation, especially at the fourth-grade level. Finally, in NAEP, "mathematical reasoning" is included in making conjectures and other related subtopics. This is not the case in TIMSS.

The Task Group notes that the TIMSS content domains were recently changed (Mullis et al., 2007). The Grade 4 content domains are now identified as Number, Geometric Shapes, and Measures and Data Displays. At this level, TIMSS has merged Geometry and Measurement and deleted the domain formerly called Patterns, Equations, and Relationships. The Grade 8 content domains are Number, Algebra, Geometry, and Data and Chance. At this level, TIMSS has infused Measurement within Geometry and expanded Data to include Probability.

Kenney et al. (1998) compared the mathematics portion of the 1996 NAEP and Maryland State Performance Assessment Program (MSPAP) at the eighth-grade level as part of the Content Analysis Project supported by National Assessment Governing Board (NAGB). It should be noted that the MSPAP is no longer used as Maryland's eighth-grade assessment due to a host of problems. Nonetheless, based on a comparison of the content frameworks, there was moderate congruence regarding the content characteristics of the MSPAP and NAEP Grade 8 tests. Content areas and topics were similar; however, the similarity was more evident in some content areas than in others. For instance, the Measurement items were nearly identical. The differences between the two tests are not sufficient to account for the magnitude of the difference between proficient performance on the MSPAP (48%), a high-stakes assessment, and on the NAEP (24%). This is likely the result of different performance categories.

Zieleskiewicz (2000) completed a study that involved 30 raters who were selected to evaluate math items on the long-term trend NAEP and the main NAEP. The reviewers felt that both the long-term trend and main NAEP frameworks assess important mathematics, with little variation across the types of raters, which included classroom teachers and mathematics specialists (e.g., university professors, leaders in professional organizations, assessment specialists).

Linn and Kiplinger (1994), moreover, in their work linking statewide tests to NAEP, found substantial content differences between the state tests and NAEP, with the majority of the statewide test items falling into one of the NAEP content areas—Number and Operations. They note that, if linking state assessments and NAEP is a goal, tests should be developed with a common framework.

Finally, Kenney (2000) reviewed the rationale for creating a family of items about a specific topic. She suggests that the ideal method for creating an item family for the NAEP would be to begin with the topic (e.g., algebra) and information based on research about students' understanding of the topic. A family of items would be built based on theoretical grounds and validated by examining results from tests. It was proposed that item families would increase NAEP's potential to provide important information about the depth of students' knowledge in a particular content strand or across content strands. It was suggested that research could support creating item families on fractions, decimals, probability, with vertical item families assessing depth in these content areas. Proportionality in measurement, geometry, and number would be a horizontal item family that would assess an important concept (proportionality) across content areas.

These studies guided the Task Group's thinking when developing the principles for organizing the content of the NAEP and state tests. Together, they form the rationale for any recommendations drawn from the general principles.

APPENDIX D: Establishing Performance Categories

Establishing performance categories involves a set of procedures currently known in educational measurement as standard setting (or setting cut scores). Judgments about performance categories are made by a panel of persons selected for their expertise or educational perspective. The exact procedures to classify students' test scores into performance categories can range from a panel consensus global judgment about the test as a whole (i.e., the minimum percentage of items passed at the various levels) to quantified judgments of individual items with respect to expected performance of students in the categories.

Several procedures and methods for combining judgments in standard setting have been developed. These procedures typically involve training panelists on the definitions of the standards and the nature of performance within the categories, soliciting judgments about the relationship of the test to the performance categories and providing successive feedback to the panelists about their judgments. Various methods to combine judgments have been developed. Variants of the Bookmark method and the Modified Angoff method involve panelists judging how students at varying levels of competency will respond to representative test items. In these two methods, the cut score for competency classifications is determined by linking the judgments to empirical indices of item difficulty. In contrast, the Body of Work method requires the panelist to classify representative students into competency categories by examining their full pattern of item responses. While the methods were all scientifically acceptable, they may differ in effectiveness. The Bookmark method may involve the most assumptions about the data, while the Body of Work method may demand the highest level of rater judgment. While more research is needed in this area, the Modified Angoff method performs well against several criteria for psychometric adequacy (Reckase, 2006).

The Task Group was interested in the following questions about standard setting in NAEP and the six states:

1) What are the performance categories of NAEP and the states?
2) How were the NAEP and state performance categories established?
3) Are they based on procedures in which experts inspect actual item content or on global definitions? (Definitions are characterized as "global" when fairly abstract characterizations of behavior necessitate high degrees of judgment to determine the categorization of student performance.)
4) Are empirical procedures used to combine individual expert opinions?
5) What is the background of the experts?
6) What descriptions or instructions are given, if any, about the nature of performance at different levels?
7) Do the experts receive the items in an examination under the same conditions as the students?

Method

To answer these questions, documents available from Web sites of NAEP (National Assessment Governing Board) and six states (California, Georgia, Indiana, Massachusetts, Texas and Washington) were retrieved by Institute for Defense Analyses Science and Technology Policy Institute (STPI) and provided to the Task Group. These documents were reviewed for relevant data by the Task Group members.

Results

Table D-1 shows the performance categories and definitions given by the NAEP and six states that were studied. Information was not fully available on all questions for each state.

Table D-1: Standard-Setting Procedures of NAEP and Six States

	Performance Categories	**Definitions**
NAEP	Basic, Proficient, Advanced	Global
California	Far Below Basic, Below Basic, Basic, Proficient, Advanced	Global and by Area
Georgia	Does Not Meet, Meets, Exceeds Standard	Global
Indiana	Did Not Pass, Pass, Pass+	Global, brief
Massachusetts	Warning, Needs Improvement, Proficient, Advanced	Global
Texas	Basic, Proficient, Advanced	Global
Washington	Basic, Proficient, Advanced	Global

The first question that was examined was the definitions of performance categories on NAEP and the six states. NAEP and all six states employed a three category system, although the labels varied somewhat. NAEP's performance categories are Basic, Proficient, Advanced. California's performance categories are labeled as Below Basic, Basic, Proficient, Advanced; Georgia, Does Not Meet, Meets, Exceeds Standard; Indiana, Did Not Pass, Pass, Pass+; Massachusetts, Warning, Needs Improvement, Proficient, Advanced; Texas, Basic, Proficient, Advanced; and Washington, Basic, Proficient, Advanced. For NAEP and all states, global definitions of the performance categories are available. Data on the NAEP and six states are tabulated in Table D-2.

For question number 2, several different standard-setting (or setting cut scores) methods have been developed over the last decade. The most widely used methods involve a generally similar standard-setting process. That is, the standard-setting process begins with a training session for the panelists, focusing on the definitions of the standards and the relevant behaviors. Then, the panelist was asked to rate, categorize, or set cutlines, depending on the exact standard-setting method. The process is iterative, with feedback about the results given and opportunities to revise judgments.

Three different standard-setting methods were employed in the states for which information was available. Historically, the Bookmark method (Lewis, Mitzel, & Green, 1996) is the most widely used method. Prior to the standard-setting process, items are ordered by their empirical difficulty in the item response theory metric. Then, the panelist sets marks in the ordered set of items to designate the points at which the minimally competent student in

a category (e.g., Basic, Proficient, Advanced) is more likely to pass than to fail the item (often defined as a probability of .67). Although panelists are instructed to examine all items, items near the marks probably receive the most scrutiny. Item mapping is a modified Bookmark method, which differs somewhat in the standard-setting process as compared to the standard Bookmark method. The Modified Angoff method requires each panelist to consider each test item and to estimate for each item what percentage of students who minimally qualify for a category (e.g., "meets standards") would answer the item correctly (this is also referred to as assigning a p-value). This method involves empirically aggregating ratings and giving feedback to panelists, followed by opportunities to revise ratings. Because each item must be rated, close scrutiny of each item is required. The Body of Work method is more holistic, because panelists examine test protocols for students at varying score levels. Their material includes item content, actual item responses, and the scoring rubrics. The panelist's task is to determine which students fall in the various categories.

As summarized by Karantonis and Sireci (2006), scant research is available on how the popular Bookmark method compares to other methods. Thus, insufficient empirical evidence is available to recommend it over the other methods. Further research should be conducted, and variables such as reliability across panelists, exact item content, domain multidimensionality, as well as resulting levels set for the standards, should be examined.

For question number 3, the standard-setting (or setting cut scores) methods reviewed by the Task Group all involve the actual inspection of item content by the panelists. However, some methods involved more intensive consideration of item content than others. In particular, the Modified Angoff method requires judgments for each item. The Bookmark method involves discussion of items, but the quantified judgments are for the category distinctions. Items with more extreme difficulties may be not considered extensively. In the Body of Work method, items are given but they are not judged.

For question number 4, judgments that are elicited from the panelist may be combined empirically. In practice with the various methods, judgments are often taken repeatedly and combined, thus allowing feedback and possible revision of judgments.

For question number 5, the background of the experts used to set standards varies within panels and possibly between states. Classroom teachers may be predominantly represented, but other experts, such as curriculum experts from higher education, may be present. Further, community representatives also may be panelists.

For question number 6, the standard-setting process for the methods described above typically involve extensive instruction about the definitions of the standards and the procedures used to set standards. Such instructions are expected to have substantial impact on the judgments. This question was scored separately because states may deviate from typical procedures or methods.

For question number 7, the experience of actually taking test items not only serves to establish the panelist's understanding of the subject area test items but also to have the experience of the students who take the tests. Judgments of items that are viewed under operational conditions are based more on individual information than on panel consensus.

All states for which information was available applied one of the current standard-setting (or setting cut scores) methods. The data in Table D-2 can be summarized as follows. First, although there is variability in the methods, all states use a contemporary method for standard setting. The Bookmark method was most frequently applied in standard setting. Second, item content is judged in all states except Massachusetts. Third, empirical combination of judgments is implemented in all states. Fourth, the background of the experts varies within panels and probably somewhat across states. For example, Georgia uses primarily classroom teachers as experts while Texas represents broader contingencies, includes curriculum experts from higher education and non-educators. Fifth, all states train the panelists prior to eliciting their ratings. Finally, only two states have the panelists experience the items in the same way as the test-takers.

Table D-2: Information on Features of Standard-Setting Procedures (Setting Cut Scores) for NAEP and the Six States

	1. How Established?	2. Item Content Judgments?	3. Combination Procedures	4. Background of Experts	5. Instructions & Definitions	6. Test Taken?
NAEP	Modified Angoff Method	Yes	Empirical with successive feedback.	55% teachers, 15% non-teacher educators, and 30% members of the general public. Panelists should be knowledgeable in mathematics. Panelists should be familiar with students at the target grade levels. Panelists should be representative of the nation's population in terms of gender, race and ethnicity, and region.	Yes	N/A
California	Bookmark Method	Yes	N/A	N/A	Yes	Yes
Georgia	Modified Angoff Method	Yes	Empirical with successive feedback.	Primarily the panelists selected were educators currently teaching in the grade and content area for which they were selected to participate.	Yes	Yes
Indiana	Bookmark Method	Yes	Empirical preliminary followed by feedback & consensus.	Not specifically given, but appears to be classroom teachers.	Yes	None specified. Probably first viewed in panel setting.
Massachusetts	Expert Opinion – Body of Work Method	No	Empirical aggregation of first judgments. Details not available about feedback & consensus.	The panel consists primary of classroom teachers, school administrators, or college and university faculty, but also non-educators including scientists, engineers, writers, attorneys, and government officials.	Yes	None specified. Probably first viewed in panel setting.

Continued on p. 8-57

Table D-2, continued

	1. How Established?	2. Item Content Judgments?	3. Combination Procedures	4. Background of Experts	5. Instructions & Definitions	6. Test Taken?
Texas	Item-mapping	Yes	Empirical preliminary followed by feedback & consensus.	The majority of the panelists on each committee were active educators—either classroom teachers at or adjacent to the grade level for which the standards were being set, or campus or district administrative staff. All panels included representatives of the community "at large."	Yes	None specified. Items probably first viewed in panel setting.
Washington	Bookmark Method	Yes	Empirical preliminary followed by feedback & consensus.	The majority of the panelists on each committee were active educators—either classroom teachers with some representation of higher education.	Yes	Yes

Source: This table was created by the Task Group using publicly available data from state Web sites. Data on California is from S. Valenzuela (personal communication, February 1, 2008).

Discussion

Although the NAEP and states varied in both process and method for standard setting (or setting cut scores), all states for which information was available employed currently acceptable educational practice. The methods may differ in effectiveness; however, scant evidence is available. The Bookmark method may involve the most assumptions about the data, while the Body of Work method may demand the highest level of judgment from the raters. The Modified Angoff method is preferred (Reckase, 2006) because the assumptions of the Bookmark method (e.g., unidimensionality) are probably not met in practice. The Body of Work method is less often applied to year-end tests because it requires higher levels of judgments from the experts. More research is needed on the standard-setting process.

It was found that classroom teachers, most of whom are not mathematics specialists, predominate in the standard-setting process. Higher levels of expertise, including the expertise of mathematicians, as well as mathematics educators, high-level curriculum specialists, classroom teachers and the general public, should be consistently used in the standard-setting process. The Task Group also found that the standard-setting panelists often do not take the complete test as examinees before attempting to set the performance categories, and that they are not consistently informed by international performance data. On the basis of international performance data, there are indications that the NAEP cut score for performance categories are set too high. This does not mean that the test content is too hard (sufficient mathematical item complexity).

APPENDIX E: Item Response Format and Performance on Multiple-Choice and Various Kinds of Constructed-Response Items

Introduction

Constructed-response (CR) item formats, in which the examinee must produce a response rather than select one, are increasingly utilized in standardized tests. One motivation to use the CR format arises from its presumed ecological validity by more faithfully reflecting tasks in academic and work settings, and stressing the importance of "real-world" tasks. CR formats also are believed to have potential to assess dynamic cognitive processes (Bennett, Ward, Rock & Lahart, 1990) and principled problem solving and reasoning at a deeper level of understanding (Webb, 2001), as well as to diagnose the sources of mathematics difficulties (Birenbaum & Tatsuoka, 1987). Finally, CR formats also may encourage classroom activities that involve skills in demonstrating problem-solving methods, graphing, and verbal explanations of principles (Pollack, Rock & Jenkins, 1992). However, these purported advantages can incur a cost. The more extended CR formats require raters to score them. Hence, they are more expensive and create delays in test reporting.

In contrast, multiple-choice (MC) items have been the traditional type used on standardized tests of achievement and ability for over a century. MC items can be inexpensively and reliably scored by machines or computers, they may require relatively little testing time, and they have a successful history for psychometric adequacy.

The Assessment Task Group examined the literature on the psychometric properties of constructed-response items as compared to multiple-choice items. The original focus was to address the following three questions:

1) Do the contrasting item types (e.g., MC, CR) capture the same skills in these tests equally well?
2) What does the scientific literature reveal?
3) What are the implications for National Assessment of Educational Progress (NAEP) and state tests?

Methodology

Constructed-response formats. CR items vary substantially in the amount of material that an examinee must produce. There are three basic types of CR items:

- The *grid-in* constructed-response format (CR-G) requires the examinee to obtain the answer to the item stem and then translate the answer to the grid by filling in the appropriate bubble for each digit.

- The *short answer* constructed-response format (CR-S) varies somewhat. The examinee may be required to write down just a numerical answer or the examinee may need to produce a couple of words to indicate relationships in the problem. The CR-S format potentially can be scored by machine or computer, given a computerized algorithm that accurately recognizes the varying forms of numerical and verbal answers. Further, an intelligent algorithm also may provide for alternative answers (e.g., slightly misspelled words, synonyms).

- The *extended* constructed-response format (CR-EE) requires the examinee to provide the reasoning behind the problem solution. Thus, the CR-EE format would include worked problems or explanations. This format readily permits partial credit scoring; however, human raters are usually required. Use of human raters, however, can lead to problems with consistency and reliability of scoring.

The stems of CR-G and CR-S items and MC items can be identical, especially if the correct answer is a number. It is not clear how the stems of CR-EE can be identical to MC items, although this possibility cannot be excluded.

Coding. With the Task Group's guidance, Abt Associates Inc. (Abt), a research group hired to assist the National Mathematics Advisory Panel, developed a list of variables to code for identified studies on this topic. The coding scheme is as follows:

Table E-1: List of Variables for Coding Studies

Category	Description
Citation	Reference citation.
Purpose of Study	Brief summary of the purpose/focus of the publication.
Content area	Studies that were not about math were excluded. Give more information on type of math if available.
Assessment	Assessment being investigated, if available.
Grade level	Grade being investigated. If not provided, age or other grouping characteristic (e.g., high school).
Item type(s) investigated	CR-G = constructed response, grid format CR-SR = constructed response, short response CR-EE = constructed response, extended essay CR-O = constructed response, other—provide details MC = multiple choice Other = other—provide details
General description of design	Brief summary of study design.
Number of items	Number of items for each type included in the study.
Description of Sample	Sample size, sampling technique (e.g., random, matrix, stratified, purposive), source of sample (e.g., national, region, locale, school district), specific characteristics (e.g., college-bound, general population, special population).
Appropriateness of sample	Enter Y if sample is representative of test taking population. Enter N if sample is not representative of test taking population and describe gap.
Source of items	Operational test, specially constructed for study, etc.
Summary of findings	Brief summary of findings. Possible reference to more detail in original text.
Subgroup performance	English language learners, gender, race/ethnicity, etc.
Psychometric properties	Item/test reliability, item/test difficulty (p-value), differential item functioning (DIF) results.
Information on scoring	Information on scoring of test (e.g., rubrics, criterion-referenced, norm-referenced).
Design flaws	Describe any obvious design flaws.

Search procedures. Abt used key words to search the literature and identify a broad band of potentially relevant research to all research questions addressed by the Task Group. Abt identified 161 articles that were potentially relevant for the specific research question on item format. They were then screened for several criteria: 1) inclusion of comparisons based on mathematics items, 2) presentation of empirical evidence, 3) published as a document other than a conference paper, 4) relevancy to the research question, 5) had the appropriate grade level or assessment [i.e., nothing higher than Advanced Placement (AP) or SAT, and 6] availability of the article.

Abt then extracted information from the 31 articles that remained and provided it to the Assessment Task Group. The full articles also were provided. An examination of the 31 articles by the Task Group led to further restriction of the set for the following reasons: 1) technical reports that were superseded by a published version, 2) irrelevant purpose for this specific research question, 3) inappropriate sample, and 4) inclusion of only MC or CR items, but not both. Ten additional articles were excluded; thus, 21 relevant articles were available to address the question on item format. To analyze the results, the 21 articles were examined in detail by the Task Group for relevant data on the several issues of concern.

Results

Impact of response format on mathematical skills, knowledge, and strategies. Potentially the most pressing issue about response format is the extent to which the same skills, knowledge, and strategies can be measured by the MC and CR item formats. Traub and Fisher (1977) found that stem-equivalent MC and CR mathematics items measured the same construct on a national math achievement test. That is, items from the two formats loaded on the same factor. Further, the skills and abilities measured by separate tests (i.e., ability to follow directions, recall and recognition memory, and risk-taking) had similar correlations with mathematics scores based on the two formats. Behuniak et al. (1996) also found that items in MC and CR format loaded on the same factor, but CR items were significantly more difficult. In contrast, Birenbaum et al. (1992) found a format effect in which there were larger performance differences on stem-equivalent MC and CR items than on parallel (i.e., superficially different) items in the same task format. Pollock and Rock (1997) examined National Education Longitudinal Study (NELS) data and found that MC items loaded on a different factor than CR items, although the factors correlated highly (r = .86). DeMars (1998) found that competencies on a state achievement test that were calculated from MC items versus CR items correlated in the .90s when the measures were statistically corrected for unreliability.

Katz, Bennett, and Berger (2000) studied strategy choice for stem-equivalent MC and CR items by analyzing verbal reports of problem solving processes during item solving, using "talk-aloud" procedures. Katz et al. (2000) found that the "plug-in strategy," which is usually associated with MC items, was used nearly as commonly with CR items. For MC items, examinees "plug-in" numbers from the response alternatives to identify the key. Students were observed adapting the plug-in strategy to CR items by estimating potential solutions and plugging-in numbers. O'Neil and Brown (1998) administered a questionnaire following the administration of a state standardized achievement test that contained both CR and MC items. Students reported greater use of systematic cognitive strategies for CR items than for MC items. However, they reported greater self-checking activity for MC items.

Thus, these studies generally do not support major differences in the nature of the construct that is measured by CR and MC items, nor in the strategies that are applied. However, much more data on this issue is potentially available because many state accountability, graduation, and year-end tests employ both item formats.

Impact of response format on psychometric properties. Several studies have results that are relevant to the psychometric properties of CR and MC items. Specifically, the psychometric properties that have been examined include item difficulty, item discrimination, omission rates, and differential item functioning (DIF) by diverse subgroups. The results vary somewhat over the exact type of CR format that is used.

The psychometric properties of MC items to their equivalent CR-G format were compared in three studies. Behuniak, Rogers, and Dirir (1996) found a moderate effect size (eta^2 of .057), which indicated that the CR-G items were harder. However, item format was unrelated to item discriminations and to gender-related DIF. Burton (1996) was interested primarily in the impact of item format on gender differences with mathematics items. She

reports only DIF as item-level statistics and found that MC items exhibited no gender-related DIF, while the CR-G format had very small and inconsistent DIF. Hombo, Pashley, and Jenkins (2001) found that CR-G items were more difficult than the MC stem-equivalent items for most items. They also found sizable differences between the grid-in responses and their accompanying written responses, suggesting that examinees have difficulty translating their answer to the grid. Hombo et al. (2001) did not report results on item discrimination or DIF.

In summary, these studies are consistent in supporting the conclusion that the CR-G format leads to higher levels of item difficulty as compared to stem-equivalent MC items. Yet, the results suggest that some, but not all, of the increased difficulty may be attributable to examinee difficulties in translating answers to grids. These studies do not provide support for format differences in item discrimination or gender-related DIF. Again, it should be noted that the number of studies yielded by the search procedures is very small.

Other studies comparing the psychometric properties of MC and CR items use the CR-S response format, which may be either in written or verbal format. Birenbaum, Tatsuoka, and Gutvirtz (1992) and Birenbaum and Tatsuoka (1987) found that stem-equivalent MC items were more difficult and less discriminating than CR-S items. However, because the MC items were constructed from previously administered CR-S items, they were able to contain common CR errors as distractors. This item design presumably minimizes the feedback received by examinees when their calculated answer does not appear as an alternative. Katz, Bennett, and Berger (2000) find MC items somewhat more difficult (proportion correct of .78 and .75 versus .66 and .74, with difference between .78 and .66 being statistically significant) than their stem-equivalent CR-S items. However, they note that large differences between item formats occurred when the MC stem-equivalent item did not have distractors representing common errors. Thus, MC items may be substantially easier than their CR-S counterparts because feedback about incorrect answers may have been provided.

In other studies, operational test data are used to examine the psychometric properties of MC and CR items. These comparisons do not involve specially constructed items (e.g., no stem-equivalent items) to control for other design differences. The nature of the MC items and the CR items that were compared is typically less well specified. In fact, the CR items may have been expressly constructed to represent other aspects of performance. DeMars (1998) found CR items on a low-stakes form of a state high school proficiency test more difficult than MC items. DeMars (2000) also found a similar format effect in a study that included both low-stakes and high-stakes high school proficiency tests. Koretz, Lewis, Skewes-Cox, and Burstein (1993) found that the omit rates are higher for CR items than for MC items on the National Assessment of Educational Progress (NAEP), which can lead to greater apparent difficulty for the CR items. Dossey et al. (1993) reported that, although NAEP CR-S items have proportions correct in the target range of .40 to .60, CR-EE items are very difficult. Garner and Engelhard (1999) also reported that the CR items on a state achievement test are more difficult than MC items, but the CR items exhibited less gender-related DIF.

In summary, the evidence about the psychometric properties of CR items as compared to MC items is inconsistent and depends on the source and the design of the comparison. If the studies utilize operational test data, comparisons of MC and CR items have indicated greater

omit rates and greater difficulty for the CR items. This pattern is probably repeated on many state tests and would be a strong finding if such data were available for study by the methods employed in this study. It should be noted, however, that studies on operational test items were not designed to isolate the impact of format by controlling or measuring other properties of items. If the studies utilized stem-equivalent versions of MC and CR items, the difference in psychometric properties depended on other design features of the items, such as the nature of the distractors and the use of grid-in responses. For example, some studies have found the CR format to be more difficult, which is consistent with the operational test studies. Other studies, however, have found the MC items to be more difficult when the distractors are constructed to represent common error patterns. Moreover, little evidence from any design is available to support differences between MC and CR items on item discrimination levels, DIF, strategy use and the nature of the construct that is measured.

Impact of response format on differences between groups. Finally, the impact of response format on differences between diverse groups has been examined in several studies. The most information is available on how item format might differentially affect mathematics performance of males and females. Historically, boys score higher than girls on many tests of mathematical competency (see the Learning Processes Task Group report). Hastedt and Sibberns' (2005) analysis of Trends in International Math and Science Study (TIMSS) data indicated that the magnitude of gender-related differences depended on item format, with girls scoring relatively higher on the CR item format. DeMars (1998) found that the interaction of gender with response format differed between two forms of a low-stakes state high school proficiency test, which included both CR-S and CR-EE as well as MC items. On one form, no significant interaction was found, while on the other form a small significant interaction was observed, indicating that girls scored relatively higher on the CR items, but still not as high as boys. DeMars (2000) examined both low-stakes and high-stakes high school proficiency tests and found that gender interacted with item format; namely, girls scored relatively higher on the CR format while the MC format favored boys. Although Gallagher (1992) also found that gender differences were greater on CR items (i.e., boys performing better), her comparison was based on high-ability students only. Garner and Engelhard (2000), moreover, examined a state high school graduation test. They found small gender-related differences on MC items, favoring boys, and smaller gender-related differences on CR items, but again favoring boys. Pollock and Rock's (1996) analysis of NELS data found that performance of males and females did not vary as a function of MC versus CR items. Thus, while girls scored lower, this was not due to item format. Burton (1996), moreover, found that the changes in item content on math section of the Math Scholastic Aptitude Test (SAT-M), among which was the inclusion of CR-G items, did not impact gender-related differences in quantitative scores, which traditionally has favored boys. Finally, Koretz, Lewis, Skewes-Cox, and Burstein (1993) report that gender-related differences in omitting either MC items or CR items were infrequent on NAEP.

Thus, the results on the interaction of the magnitude of gender-related differences in performance and item format are inconsistent and depended on the design of the specific study. However, the evidence suggests either no impact of response format on gender-related differences or that the relatively lower scores of girls than boys on mathematics items may be lessened in the constructed-response format.

The interaction of racial-ethnic differences with item format also has also been examined in several studies. Historically, minority groups score lower on tests of mathematical performance (see the Learning Processes Task Group report). In DeMars's (2000) study, proportion minority interacted with format, indicating that black-white differences were greater on MC items. Pollock and Rock's (1996) analyses of NELS data, moreover, indicated that demographic variables based on race-ethnicity (black versus white, Hispanic versus white) correlated more highly with the MC item factor than the CR item factor, indicating greater adverse impact on MC items. Finally, Koretz, Lewis, Skewes-Cox, and Burstein (1993) found that racial-ethnic groups differed on the relative rate of omitted items for MC and CR items on NAEP. Thus, as a set, these studies provide some evidence that black-white differences in performance in mathematics are lessened on CR item format as compared to the MC item format.

Other results on item format that are potentially interesting include Hastedt and Sibberns (2005) finding on TIMSS data that scores based on MC versus CR items produced only slight differences in the relative ranking of the various participating countries. And, DeMars (2000) found that the difference between MC and CR items in difficulty depended on the test context. The two item formats differed less in difficulty on high-stakes tests than on low-stakes tests.

Discussion

The available evidence on comparing the psychometric properties of MC items and CR items must be interpreted in the context of several factors. These factors include the following limitations: 1) the limited scope of the available scientific literature, 2) the uncontrolled design features for comparisons based on operational tests, 3) the design strategy in available controlled comparisons of MC and CR items, 4) the limited scope of the controlled comparisons, and 5) the impact of test context on the relative performance on MC and CR items.

First, the literature that could be retrieved by the methods in this study did not yield many journal articles and widely circulated technical reports. Yet, analyzing the psychometric properties of test items is routine test development procedure for many state and national tests, many of which contain both MC and CR item formats and most of which contain demographic information on examinees. It is unclear if these results are unavailable due to lack of appropriate publication outlets, lack of incentives to provide results, or some combination of these features. Despite some limitations in these comparisons (discussed further in this section), it would be useful to know if the CR items on current tests measured the same construct, had greater difficulty but equal discrimination and DIF, and result in lessened adverse impact on some groups of test takers as compared to MC items. Given the lack of evidence from the wider sphere of operational tests, the best conclusion about these issues from the studies that are available is that the evidence is weak or inconsistent.

Second, even if comparison data from operational tests were more available, the evidence is limited by the design of the items that appear on operational tests. That is, the goal of operational tests is to assess mathematical competency broadly, not to compare the MC and CR item formats. Thus, MC and CR items differ on a number of features, only one of which is format. Thus, more carefully controlled comparisons are desirable to isolate the impact of response format.

Third, however, the available comparisons between MC and CR items under controlled designs (i.e., stem-equivalent items) have yielded inconsistent results. Moreover, there is another design issue that has emerged—namely, the strategy for constructing stem-equivalent items. In some studies, CR items are created by removing the distractors from operational MC items. Evidence from these studies suggests that CR items are more difficult. In other studies, MC items are created to correspond to CR items by using common errors as distractors. Evidence from these studies suggests that MC items are more difficult.

Fourth, the controlled designs for comparing MC items to CR items have had limited scope. The items that were compared involved short numerical responses. Items at a higher level of complexity, those that involve understanding principles or showing steps, have not been compared between formats. Perhaps MC items that involve higher levels of complexity cannot be constructed; but then again, maybe they can but appropriate studies have not been undertaken.

Fifth, test context may interact with comparisons of MC and CR item formats. For example, one study found that a high-stakes context versus a low-stakes context of the same operational test was associated with decreased differences between the item formats. One interpretation is that the high-stakes test context may evoke sufficient levels of motivation in the examinees to complete the more time-consuming constructed-response formats. Moreover, tests with mixed item formats may lead to reduced format differences, due to item-solving strategies carrying over from one format to another. That is, in one study reviewed in this section, the plug-in strategy that can be effectively applied to the MC format may be extended to the CR format if the MC format precedes it.

Given the limitations described in this section, there is little or weak evidence to support the CR format as providing much different information than the MC format. For example, the available evidence provides little or no support for the possible claim that different constructs are measured by the two formats or that item discrimination varies across formats. Although some evidence suggests that CR items are more difficult, especially for the more extended CR formats, there is some contrary evidence that indicates that more difficult MC items can be constructed for their stem-equivalent CR items. Finally, the impact data does not support much difference between the two item formats. That is, the impact of response format on gender differences is inconsistent, while the impact on racial-ethnic differences is weak.

Item response format is one of the several design features that may impact item complexity. The evidence found in the scientific literature did not support the notion, however, that CR format, particularly the short answer type, measures different aspects of mathematics competency compared to MC. The impact of item format may interact with other design features, such as test context or strategy for developing controlled comparisons items. Thus, the important issue is not whether to select MC versus CR format, but rather how to most efficiently design items to measure content of the designated type and level of cognitive complexity.

APPENDIX F: Factors to Evaluate the Quality of Item Design Principles

Ensuring that high-stakes tests, such as the NAEP and various state tests, are of the highest quality psychometrically is critical. Measurement instruments need to be accurate and unbiased. The Task Group presents suggestions, or factors, for how quality control might be carried out. These factors, posed in the form of questions to be addressed, are relevant to principles such as item format and problem context (word problems), as well as item administration methods, such as including calculators and manipulatives.

1) Are the items generated from the principles appropriate for the targeted construct, or are they more likely to have nonmathematical sources of difficulty? What additional factors can reduce this vulnerability?

    a. For example, do items created in constructed-response (CR) formats rely more heavily on verbal skills than mathematical skills? Which CR formats or rubrics are less likely to involve these skills?

    b. Are items generated with "real-world" context more likely to contain confusing or irrelevant verbal material, visual displays, or practical knowledge? What mechanisms reduce this confounding?

2) Are the items generated from the principles generally appropriate to provide maximal information for the competency levels targeted by the assessment?

    a. Several factors reduce information, including unreliability in the scoring mechanisms, inappropriate item difficulty for the targeted levels, low item discrimination, and high vulnerability to guessing.

    b. What mechanisms reduce this source of confounding (e.g., machine scoring of CR)?

3) Are the items generated from the principles more likely to be vulnerable to differential item functioning (DIF)?

4) Are the items generated from the principles appropriate for model-based approaches to measurement?

    a. Current state and national assessment typically apply item response theory (IRT) approaches to scaling items. These approaches allow equating of tests across forms and time, which is necessary to examine trend and maintain comparable standards.

Are there special mechanisms to adapt diverse item design principles to IRT models? IRT easily accommodates binary and polytomous formats, including partial credit scoring. Formats known to produce local dependence (a violation of IRT assumptions) can sometimes be accommodated by special mechanisms, such as testlet scoring.

APPENDIX G: Descriptors Used in the Literature Search and Exemplars of Satisfactory Word Problems

To help the Assessment Task Group locate any previous work that might have been done with respect to language or wording defects in test items used in mathematics assessments, Abt Associates Inc. (Abt) conducted an extensive search of the research literature and related items. The descriptors used by Abt appear in the following chart.

Table G-1: Descriptors Used in Literature Search

Math* and	
Assessment*	
Test*	
Testing*	
	Item language
	Item wording
	Question language
	Question wording
	Item wordiness
	Language bias
	Linguistic simplification
	Language density
	Essential language
	Non-essential language

*__Note:__ All of the terms in the list were searched simultaneously with math, and assessment or test or testing.

Examples of Satisfactory Word Problems

The Task Group examined state tests to provide examples of desirable content in mathematics word problems. The major purpose of word problems on broadly given assessments should be to assess skill in converting relationships described verbally into mathematical symbolism or calculations. Moreover, they should satisfy the following conditions:

a) be written in a way that reflects natural and well-written English prose at the grade-level assessed;

b) assess mathematics knowledge and skills for the grade level of the assessment as judged by international benchmarks while restricting nonmathematical knowledge to what would be general knowledge for most students;

c) clearly assess skill in converting relationships described verbally into mathematical symbolism or calculations, or, if they use a "real-world" setting, they use a setting that aids in solving a problem that would be more cumbersome to state in strictly mathematical language.

The following five items illustrate some features that are relevant for quality word problems.

1) The following is an appropriate example of a satisfactory item for which the nonmathematical knowledge is minimal and the student is expected (as appropriate in a mathematics test) to convert relationships described verbally into mathematical symbolism or calculations:

   For example: Grade 4, Georgia, No. 1, Page 2–5

   The local park is having a game day. There are 5 teams, with 3 boys and 4 girls on each team. How many children are there in all?

   ☐ *12* ☐ *15* ☐ *20* ☐ *35*

   Comment: This problem assesses whether the student can convert a relationship described verbally into appropriate mathematical symbolism. Moreover, the nonmathematical nouns are at an appropriately low level of vocabulary.

2) Here is an appropriate word problem in which the real-world setting is an aid to the student in solving a problem that could have been expressed in strictly mathematical language:

   For example: Grade 8, Massachusetts, N.12 on Page 2–20

   Mona counted a total of 56 ducks on the pond in Town Park. The ratio of female ducks to male ducks that Mona counted was 5:3. What was the total number of female ducks Mona counted on the pond?

   A. 15 *B. 19* *C. 21* *D. 35*

   Comment**:** A student has to decide which fractions are relevant. Moreover, any statement of a ratio problem similar to this problem becomes harder to read if the context is removed. [However, the problem could be improved by making the total number of ducks equal to 120, so that the total could be divided by any of 3, 5, and 8 without giving a remainder.]

3) Here is an appropriate example of a test item requiring logical reasoning with appropriate mathematical aspects:

   For example: Grade 8, NAEP 2005, Problem 5, page 5–28

   Ravi has more tapes than magazines. He has fewer tapes than books. Which of the following lists these items from the greatest in number to the least in number?

   A) Books, magazines, tapes *B) Books, tapes, magazines*
   C) Magazines, books, tapes *D) Tapes, magazines, books*

   Comment: Although this problem could also be appropriate for a language arts assessment, it is also appropriate for a mathematics assessment because the language of inequalities is so closely related to the terminology in the item.

4) Here is an appropriate item with a natural sequence of sentences in which disciplined mathematical reasoning is the cornerstone:

   For example: Grade 8, Washington, No. 20

   Mrs. Bartiletta's class has 7 girls and 3 boys. She will randomly choose two students to do a problem in front of the class. What is the probability that she will choose 2 boys?

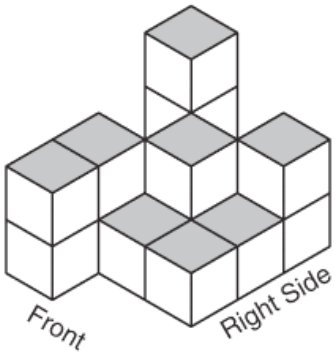
   A. $\frac{1}{15}$ *B.* $\frac{2}{5}$ *C.* $\frac{3}{7}$ *D.* $\frac{5}{19}$

   Comment: The student must first realize that this is a "without replacement" problem. Then the student is free to choose either permutations or combinations for the denominator. After that, however, the student must be consistent for the numerator. Finally, a fraction reduction is needed to match Option A. [This problem could reasonably be viewed as beyond the level required for performance at Grade 8.]

5) Here is an appropriate problem focused on one of the three methods of making planar pictures to represent 3-dimensional objects:

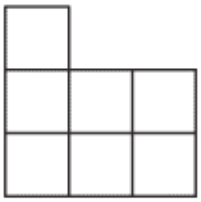
For example: Grade 8, Texas, No. 50

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

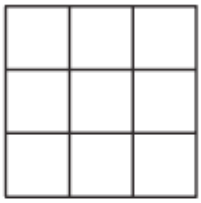


Which drawing best represents a front view of this solid figure?

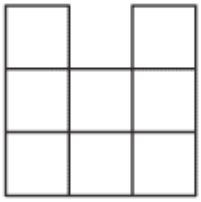
F



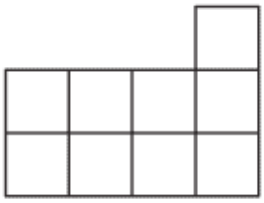
H



G



J



Comment: The problem is stated nicely. In particular, the phrase “by stacking” and the attribution to Melody make it clear to the student that he/she is being faced with a static situation.