Examining the Content and Context of the Common Core State Standards: A First Look at Implications for the National Assessment of Educational Progress

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

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

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

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

Panel Members:

Peter Behuniak
University of Connecticut

George W. Bohrnstedt
American Institutes for Research

James R. Chromy
Research Triangle Institute

Phil Daro
Strategic Education Research Partnership (SERP) Institute

Lizanne DeStefano
University of Illinois

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Robert Linn
University of Colorado Boulder

Ina V.S. Mullis
Boston College

Scott Norton
Council of Chief State School Officers

Gary Phillips
American Institutes for Research

Lorrie Shepard
University of Colorado Boulder

David Thissen
University of North Carolina, Chapel Hill

Karen Wixson
University of North Carolina, Greensboro

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Janis Brown
National Center for Education Statistics

For Information:

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

CONTENTS

Introduction

Since its inception more than four decades ago, the National Assessment of Educational Progress (NAEP) has served as a key indicator of what the nation's students know and can do in academic subjects. NAEP's role has evolved over time in response to the changing educational landscape. As the states became more invested in using assessments for educational accountability, Congress responded by expanding NAEP's mandate to include state (as well as national) estimates of student achievement. Eventually, under the No Child Left Behind Act, state NAEP assessments became more frequent and more comprehensive, with state-level participation in Grades 4 and 8 reading and mathematics assessments required by law. With every state developing its own assessments for accountability and setting its own benchmarks for proficiency, NAEP assessments provided a mechanism for putting the achievements of students in all states on a common scale. In addition, NAEP assessments have served as independent monitors of progress because they have no high-stakes consequences for schools or students. NAEP's frameworks also are not aligned with any one curriculum, but are intended to capture the achievements of students schooled under different curricula.

In addition to reflecting the different curricula in the states, NAEP also must embody emerging themes in education in a manner that contributes to the educational dialogue and positions NAEP assessments to measure new aspects of student learning when they occur. That is, to fulfill its mission, NAEP must both lead and reflect.

Now, the educational landscape is changing again. Under the leadership of the National Governors Association Center for Best Practices (NGA Center) and the Council of Chief State School Officers (CCSSO), a new set of common standards, the Common Core State Standards (CCSS), have been developed and widely adopted by the states. These new standards, and the assessments being built to measure them, offer the possibility of far greater uniformity in curriculum and assessment across the nation than has characterized U.S. education in the past. In addition, the CCSS embody many emerging themes of education reform, including a focus on college and career readiness for all students by Grade 12, a more coherent set of learning expectations across grades that builds on contemporary research into learning progressions, and an acknowledgment of the greatly expanded role of technology in teaching and learning.

In this context, the NAEP Validity Studies Panel (NVS Panel) determined to devote a substantial portion of its annual validity research agenda in 2011 and 2012 to exploring the relationship between NAEP and the CCSS, and to considering how NAEP can work synergistically with the CCSS assessments to provide the nation with the most useful information about educational progress. This is a very early look at a changing landscape. States are just beginning to roll out their CCSS-based curricula, and the federally funded consortia that are developing assessments for the CCSS will not begin operational testing until the 2014–15 school year. Nevertheless, it is clear that the CCSS, and the larger education trends that they embody, will be a major factor in shaping NAEP's future. By acting proactively, but deliberatively, the National Assessment Governing Board (Governing Board) and the National Center for Educational Statistics (NCES) can support NAEP's continued validity and enhance its utility over the coming decades.

The Studies

Included in this volume are two substantial studies exploring the relationship between the content of the NAEP mathematics, reading, and writing assessments and the CCSS in mathematics and English language arts (ELA). In part, because the assessments being developed by the two federally funded consortia to measure the CCSS (the Partnership for Assessment of Readiness for College and Careers [PARCC] and the Smarter Balanced Assessment Consortium [Smarter Balanced]) were at a very nascent stage when this work was being done, the studies focus on the standards themselves, while acknowledging that a comprehensive analysis will eventually require an examination of the consortia assessments at the item level. These two content studies are complemented by two shorter white papers that explore, respectively, the potential for incorporating learning progressions into NAEP assessments and the implications for the NAEP program of coming changes in psychometric approaches to statewide testing.

Following are brief descriptions of the major findings from each study.

The National Assessment of Educational Progress and the Common Core State Standards: A Study of the Alignment Between the NAEP Mathematics Framework and the Common Core State Standards for Mathematics (CCSS-M)

Gerunda Hughes, Phil Daro, Deborah Holtzman, and Kyndra Middleton

This study by Dr. Hughes and colleagues convened a panel of mathematicians and mathematics educators to compare the Grades 4 and 8 NAEP mathematics frameworks with the CCSS in mathematics (CCSS-M). For the CCSS-M, adjacent grades were included in the analyses.

This study found the preponderance of content in the CCSS-M also is found in the NAEP Mathematics Framework, but with some differences. The differences are potentially important and should receive attention in the normal revision of the framework and the assessments. Four types of discrepancies were observed. Compared to the NAEP framework, the CCSS-M have:

1. More rigorous content in eighth-grade algebra and geometry.
2. More extensive and systematic treatment of mathematical expertise (found in the Standards for Mathematical Practice).
3. A more conceptual perspective on many mathematical topics, explicitly stating the mathematics to be understood rather than the type of problem to be solved.
4. Some content taught at higher grades than is assessed in the fourth-grade NAEP assessment. For example, the study of proportional relationships is concentrated in Grades 6 and 7, and data sets and probability are taught in Grades 6 and 7, respectively.

These are important differences and these areas should be considered a priority in the normal revision of the NAEP Mathematics Framework.

The study also found that the CCSS-M include a preponderance of content included in the NAEP framework by the grade level assessed, with several important exceptions as noted in the results reported above. As implementation of the CCSS continues, an analysis should be conducted to estimate the effect on overall NAEP scores that follows from dropping content from the curriculum that is assessed by NAEP but not included in the CCSS-M. This should be done to avoid misinterpreting this effect as a general decline in mathematics achievement, when it may be due to a specific decline in a subdomain that has been intentionally deemphasized in the CCSS-M.

Study of NAEP Reading and Writing Frameworks and Assessments in Relation to the Common Core State Standards in English Language Arts

Karen K. Wixson, Sheila W. Valencia, Sandra Murphy, and Gary Phillips

This study by Dr. Wixson and colleagues convened a panel of reading experts, and a separate panel of writing experts, to compare the Grades 4, 8, and 12 NAEP reading and writing assessments with the CCSS in English language arts (CCSS-ELA). In addition to the NAEP frameworks, assessment materials from the 2009 and 2011 NAEP reading and writing assessments were used in the analysis.

Overall, the study found that *there is sufficient alignment between NAEP reading and writing assessments and the CCSS-ELA documents to make panelists cautiously optimistic about NAEP's continuing relevance and viability*. With attention to the specific issues identified in this report and a systematic program of special studies and probe studies to inform future assessments, the panelists concluded that NAEP could continue to serve not just as an independent monitor of student achievement in an era of CCSS, but also as an intellectual tool to promote the design and use of quality assessments apart from CCSS.

Reading: Many aspects of the current NAEP reading assessment reflect conceptualizations of the reading process found in CCSS-ELA documents, including a cognitive focus aligned with research, a broad range of text types, high-quality and appropriate length of texts used in assessment, attention to literary and informational comprehension, use of text pairs, attention to reader-text interactions in item development, inclusion of writing in response to reading, parsimony and elegance in crafting questions to align with specific texts, and thoughtful, meaningful items that are well sequenced and crafted.

Furthermore, the Governing Board's policy of aligning Grade 12 NAEP with standards for preparedness for postsecondary education and training is consistent with the intention of the CCSS-ELA standards to assure that students achieve college and career readiness no later than the end of high school.

Some specific similarities and differences include the following:

1. NAEP reading selections at Grades 4 and 8 generally fall within the quantitative ranges called for in the CCSS-ELA, while the Grade 12 passages examined are consistently less difficult than what is called for in the CCSS-ELA quantitative indexes.
2. The cognitive targets specified in the NAEP Reading Framework are compatible with the CCSS-ELA Anchor Standards.
3. An important area of difference between CCSS-ELA and the NAEP Reading Framework is the manner in which disciplinary reading is addressed. The conceptual framing for the CCSS-ELA positions disciplinary reading for the purposes of knowledge building. In contrast, the NAEP Reading Framework subsumes disciplinary texts under "informational texts," sampled from varied content areas and assessing general comprehension.
4. There are differences in how the NAEP Reading Framework and CCSS-ELA address vocabulary, with the CCSS-ELA placing a heavy emphasis on academic vocabulary.
5. The CCSS-ELA include K–5 standards for foundational skills, while NAEP reading assessments target comprehension beginning at Grade 4. Because foundational skills are not part of the NAEP reading assessments, comparisons of fourth-grade performance between NAEP and assessments built to reflect the CCSS may need to be carefully mapped and analyzed.

Writing: There are also broad similarities between the current NAEP Writing Assessment and the CCSS-ELA. Both the NAEP Writing Framework and CCSS present writing as a social, communicative activity; emphasize the importance of audience, purpose, and task; and treat rhetorical flexibility as an important component of skilled writing performance. The NAEP Writing Framework and the CCSS are aligned in other important ways as well: They address similar broad domains of writing, and identify and discuss essentially the same valued characteristics of effective writing—development of ideas, organization, and language facility and conventions. The NAEP scoring guides for writing emphasize adapting writing to purpose, task, and audience and the types of writing found in the CCSS-ELA, and the pool of NAEP writing prompts contains a broad range of audiences and forms, an aspect of range described in the CCSS-ELA.

The writing panel identified several gaps in alignment between the NAEP Writing Framework and the CCSS-ELA that should be considered as well:

6. The CCSS-ELA clearly emphasize integration of the language arts, while the NAEP Writing Framework does not. In particular, the CCSS-ELA stress writing about reading and writing from sources (writing based on research). NAEP writing tasks rely primarily on background knowledge and personal experience.
7. The CCSS-ELA are explicit in acknowledging that the teaching of writing is a shared responsibility across disciplines, and writing activities within the disciplines are integrated with content learning. Although the NAEP Writing Framework acknowledges the situated nature of writing and its importance in all

disciplines, the NAEP writing assessment deals with generic writing skills and general and academic vocabulary.

8. The NAEP writing assessment limits the role that technology plays in assessment to students' use of a computer to compose and edit with a limited set of commonly available tools. On the other hand, the CCSS-ELA convey a portrait of college and career-ready students who use technology and digital media strategically and capably.
9. The NAEP writing assessment assesses on-demand writing in an abbreviated time frame, while the CCSS-ELA emphasize writing under a variety of conditions, conveying expectations for students' use of writing processes.

The Relevance of Learning Progressions for NAEP

Lorrie Shepard, Phil Daro, and Fran Stancavage

This paper discusses the history and use of learning progressions, including their use in the CCSS. It considers the potential for using learning progressions in NAEP, either as a guide to assessment development or as a reporting device.

The paper notes that learning progressions are a highly popular innovation in assessment and instructional design. The core principles that undergird them have strong theoretical and research grounding, although specific, practical applications are rare, at least in U.S. contexts. Given the salience of hypothesized learning progressions in the design of the CCSS and the Next Generation Science Standards (NGSS), it is important to consider the relevance of formally developed learning progressions for the future design of NAEP assessments.

Because NAEP assessments must be sufficiently robust to assess progress toward the standards across multiple curricula, it is highly unlikely that formal learning progressions (which require detailed development of instructional activities and corresponding assessment tasks tied to the frameworks) could be the main building blocks of a newly design NAEP. Nonetheless, NAEP assessments must be designed in such a way as to be able to monitor the success of deeper curricular reforms where they occur. For NAEP to continue to be an independent monitor, the Governing Board and NCES must have a strategic vision that attends to both breadth and depth in representing subject-matter expertise. In a recent white paper on the future of NAEP (National Center for Education Statistics, 2012), an expert panel recommended that the NAEP domain specifications be broadened such that the NAEP reporting framework as historically conceived would be situated within a larger, "super-assessment" domain. In this context, assessment tasks tied to learning progressions in mathematics, science, or literacy could be embedded within an extended or enhanced NAEP framework, and both performance outcomes and psychometric functioning of the assessment tasks could be compared for students with and without instructional opportunities tied directly to learning progressions curricula.

In addition to considering the possibility of testing learning progressions by embedding them within the NAEP sampling frame or administering them in special probe studies, the authors also considered the feasibility of building example learning progressions into the NAEP item pool to enable their use as a reporting strategy. The authors constructed four quasi-learning progressions using existing NAEP items in combination with Balanced Assessment of Mathematics items but concluded, based on this exercise, that such an approach is infeasible and likely to be misleading until there is more widespread implementation of the CCSS and thereby greater congruence between a hoped-for and the actual empirical ordering of items.

What Might Changes in Psychometric Approaches to Statewide Testing Mean for NAEP?

David Thissen and Scott Norton

The authors explored two psychometric features of statewide testing that, mediated through the CCSS consortia tests, are likely to have significant implications for NAEP assessments. The first is the move toward computerization of testing and the second is the greatly decreased number of unique state tests. The latter creates new challenges and opportunities for NAEP to serve as a common metric across states.

With regard to the widespread movement toward computerized testing, the authors conclude that computerization of NAEP assessments is inevitable. There are several reasons for computerization. NAEP assessments may be computerized so that technology-enhanced item types can be delivered when required by the frameworks, as has already happened with the science interactive computer tasks in 2009 and is planned for the technology and engineering literacy (TEL) assessment in 2014. NAEP assessments may be computerized so that they appear more comparable with the statewide assessments being developed by the consortia, or to facilitate linking with those assessments. They may be computerized simply because computer administration has become more cost effective—this will ultimately happen for all assessments as the cost of computing equipment decreases and the costs of printing and physical distribution and scoring of paper response sheets grow. Finally, all assessments will gradually become computerized as computer use becomes ubiquitous for real-world tasks, both within and outside schools.

The literature review conducted by Rosenberg and Townsend and included as an appendix to the white paper concluded that comparability of results can often be maintained as a test makes the transition from paper-and-pencil to computerized administration. At the same time, aspect of computerization often have an effect on results for some subgroups of the population. This suggests that the computerization of NAEP is best approached in the way that all other changes made to NAEP assessments since the advent of the "new design" in 1983 have been approached: Careful consideration should be given to the design of the computerized administration, and a bridge study should be carried out to ensure the comparability of results across the transition (unless an a priori decision is made to "break trend").

With regard to the anticipated decrease in the number of state tests, the authors note that assessments developed by the two major consortia, Smarter Balanced and PARCC, may reduce the number of statewide tests in Grades 4 and 8 from nearly 50 to the low single digits, starting in the 2014–15 academic year.[1] With such a small set of tests to work with, linkage may become feasible, permitting close quantitative comparison between NAEP results and those obtained with the consortia tests, and providing a mechanism to link the consortia tests' scales with each other across the two groups of states.

Because correspondence between the results of disparate educational assessments tend to change over time, any linkage between the NAEP scale and the consortia statewide tests will need to be maintained regularly over the years of their use. However, a singular opportunity exists in a short window of time—essentially right now—to design data collection for linkage between the NAEP scale and the consortia assessments while the latter are under development. At this time, central control remains possible, and cooperative agreements to collect suitable linking data may be more easily obtained than will be the case after the consortia tests branch and fork into two dozen statewide assessments.

Conclusion

In general, the study authors, and the NVS Panel as a whole, were unanimous in recommending that NAEP continue to play its historical role as an independent monitor. In the short run, while the states are transitioning to the CCSS, NAEP assessments can provide a stable measure of trends in a shifting landscape of state assessments. In the longer run, the independent monitoring role for NAEP assessments is likely to remain important, in part because of the less biased perspective on achievement offered by NAEP's low-stakes administration, and also because there will still be a need to bring achievement for students in all states onto a common metric. Nevertheless, the NVS Panel cautioned that if NAEP is to remain viable as a credible independent monitor, it will need to evolve as instruction and assessment change around it. Furthermore, NAEP assessments must anticipate change in order to be able to measure it, and, as a result, the NAEP program should continue its tradition as a leader in assessment innovation. The consortia have high aspirations to deliver ground-breaking assessments based on the most current research. However, they are bound to be constrained by the cost and logistical requirements of providing individual student scores for all students in Grades 3–8 and high school. Freed of these constraints, the NAEP program can be more nimble and should use its competitive advantage to advance the art and science of assessment for the nation.

The NVS Panel also agreed with the following conclusions of the two white papers:

- Learning progressions are an important development that can increase the coherence between instruction and assessment, but they are unlikely to find a place in NAEP's design, given the fact that NAEP assessments must remain curriculum neutral and learning progressions are inherently curriculum-based.

---

[1] There are some states that have chosen not to join either consortium and will presumably continue to develop their own tests, at least for the foreseeable future.

- Computerization of NAEP assessments is inevitable and will offer the opportunity for a number of innovations and efficiencies. Bridge studies will be important to maintain trend during the shift to computerization.
- With the goal of providing a common metric against which the results of the PARCC and Smarter Balanced assessments can be compared, NCES should aggressively pursue the goal of a formal linking study to be carried out in concert with the field testing of the CCSS assessments.

As NCES looks to the future, examining areas of alignment and nonalignment between NAEP assessments and CCSS assessments is a first step. A next step might be to launch special studies within the NAEP program that could investigate the penetration of some of the more advanced skills espoused by the CCSS in contexts where these skills are being taught. Any changes to the main NAEP frameworks should be made gradually and deliberately, as uptake of CCSS-based curricula expands. This would ensure that NAEP maintains the appropriate balance between leading and reflecting.

It is our intention that the set of studies reported here will help NCES and the NAEP program begin their journey.

A Study of the Alignment Between the NAEP Mathematics Framework and the Common Core State Standards for Mathematics (CCSS-M)

Gerunda B. Hughes
Howard University

Phil Daro
Strategic Education Research Partnership (SERP) Institute

Deborah Holtzman
American Institutes for Research

Kyndra Middleton
Howard University

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

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

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

Executive Summary

Introduction

For decades, prior to the inception of the Common Core State Standards (CCSS), the National Assessment of Educational Progress (NAEP) was the only vehicle through which states could assess the progress of their students using a common metric. Now, 45 states, 4 U.S. territories, and the District of Columbia have adopted the CCSS to provide a clear and consistent curriculum framework to prepare students for college and the workplace. But because NAEP is a critical monitor for comparing results of student achievement across states, it is imperative that the newer CCSS standards and the NAEP frameworks be examined to determine the degree of alignment. The results will allow policymakers to make decisions about what changes, if any, should be made to the NAEP frameworks.

Methodology

This alignment study focuses primarily on the conceptual match between the subtopics and objectives in the NAEP Mathematics Framework and the content standards in the Common Core State Standards for Mathematics (CCSS-M) in Grades K–8. While an item-to-framework study is also critical when inquiring about alignment, items from the CCSS assessment consortia were not available at the time of this study.

Two criteria were used to describe the degree of alignment between the CCSS-M and the NAEP Mathematics Framework: the extent of content coverage and the grade at which the content was covered. To obtain the necessary data, two mappings were conducted: (a) CCSS-M to NAEP Mathematics Framework; and (b) NAEP Mathematics Framework to CCSS-M.

Findings

The study's findings relied on the judgment of four panels of experts who identified the specific CCSS-M content that was not covered well in the NAEP mathematics subtopics and objectives for Grade 4 and Grade 8 and the specific NAEP mathematics content that was not covered well in the CCSS-M at or before the grade level of the NAEP assessment.

The study did not find wide areas of content in the NAEP Mathematics Framework that were not covered in the CCSS-M. Similarly, the study did not find wide areas of content in the CCSS-M that were not covered by the NAEP Mathematics Framework. Nevertheless, there were differences in specificity and conceptual understandings between the CCSS-M and the NAEP Mathematics Framework that are important to note: (1) the CCSS-M have more rigorous content in eighth-grade algebra and geometry; (2) the CCSS-M infuse and distribute the development of mathematical expertise, such as the ability to estimate accurately, throughout the standards for mathematical content, whereas the NAEP Mathematics Framework assesses estimation as a skill in isolation from the vast majority of the content; (3) the CCSS-M attend to developing conceptual understandings of a greater number of mathematical topics (such as unit fractions, patterns, and functions) than does the

NAEP Mathematics Framework; and (4) the CCSS-M introduce some mathematics content, such as probability, at higher grades than does the NAEP Mathematics Framework.

Conclusions, Recommendations, and Next Steps

Certainly, there are differences between the NAEP Mathematics Framework and the CCSS-M. For example, the NAEP Mathematics Framework is an *assessment* framework that prescribes what should be tested on NAEP. The CCSS-M, on the other hand, provide a *curriculum* framework that prescribes what should be taught in classrooms. In those few areas where content is covered by the NAEP Mathematics Framework, but not included in the CCSS-M, and vice versa, studies should be conducted to determine how estimates of students' achievement status and growth are affected by the degree of alignment between what is taught and what is tested.

Historically, the NAEP frameworks have aspired to represent the union of all the various state curricula while reaching beyond these curricula to lead as well as reflect. As a result, NAEP often has pushed on the leading edge of what the nation's children know and should able to do. The introduction of the CCSS-M provides both new opportunities and challenges for NAEP. As the nation moves toward widespread implementation of instruction and assessment based on the CCSS-M, NAEP must balance the goals of comparability over time (i.e., maintaining trend) with current relevance.

CONTENTS

Background

Since its founding in 1963, the National Assessment of Educational Progress (NAEP) has made a unique contribution to American education. Since 1990, when state NAEP was authorized by Congress, NAEP—also referred to as "the Nation's Report Card"—has been the only vehicle through which states can compare the progress of their students against a common standard. Originally, only some states participated in state NAEP, but with the passage of No Child Left Behind, every state receiving Title I funds was required to take state NAEP in reading and mathematics. In 2010, however, the Common Core State Standards (CCSS) for English language arts and mathematics were released, and soon thereafter adopted by 45 states, 4 U.S. territories, and the District of Columbia.

The CCSS Initiative is a state-led effort coordinated by the National Governors Association Center for Best Practices and the Council of Chief State School Officers. The initiative, which includes the development of educational standards, is a collaboration among teachers, school administrators, and experts that was formed to provide a clear and consistent framework of what is needed to prepare American children for college and the workforce. Specifically, the initiative defines the knowledge and skills students should gain during their K–12 education so that they graduate from high school ready to succeed in entry-level, credit-bearing academic college courses or in meaningful workforce training programs. As of this writing, two federally funded state consortia are developing assessments aligned with the CCSS for general education students in Grades 3–8 and high school: the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (Smarter Balanced). In addition, two other state consortia are developing English language arts and mathematics assessments linked to the CCSS for students with severe cognitive disabilities: the Dynamic Learning Maps Alternate Assessment System Consortium and the National Center and State Collaborative consortium. Finally, the World-Class Instructional Design and Assessment consortium, as well as a second consortium led by WestEd, are developing English language proficiency assessments for English learners.

The Charge

In spring 2011, the National Center for Education Statistics (NCES) asked the NAEP Validity Studies Panel (NVS Panel) to undertake a study of the validity and utility of NAEP in the context of the CCSS. NCES asked that the study address the following questions:

1. What is the conceptual match between NAEP and the CCSS?
2. How should the content in the assessment frameworks and the standards be compared?
3. What could be learned from this comparison?

Two interrelated studies were commissioned: one in reading and writing and the other in mathematics. The purposes of these studies were twofold: (1) to compare the content of the current NAEP reading and mathematics frameworks in grades assessed by NAEP with the content standards of the CCSS in English language arts and mathematics; and (2) to make recommendations to NCES regarding broad issues related to the content comparison of NAEP subtopics and objectives and the CCSS, including the extent of alignment that is appropriate to support NAEP's continuing role as an independent monitor. In the current study, only mathematics is addressed.

NAEP and the Common Core State Standards for Mathematics (CCSS-M): Different Types of Mathematics Frameworks

NAEP began assessing mathematics in 1973, and the long-term trend component of NAEP, which reports on achievement among 9-, 13-, and 17-year-olds, has continued unbroken since that time. A second mathematics trend line, known as "main NAEP," began in 1990 using a new assessment instrument.

The main NAEP mathematics assessment is administered at the national and state levels and in selected urban districts. Results are reported on student achievement at Grades 4, 8, and 12 at the national level and at Grades 4 and 8 at the state level and in large urban districts that volunteer to participate. The main NAEP assessment is based on a framework that is updated periodically, but it has nevertheless been possible to continue the main NAEP trend lines from 1990 through the 2013 assessment for all grade levels. (The greatest changes were introduced in the Grade 12 content objectives in 2009, but special analyses were conducted and confirmed that the Grade 12 trend line could be maintained.)

The NAEP Mathematics Framework is an assessment framework, not a curriculum framework. Because it must fairly assess students from across the country, it spans the full range of mathematics that *could be* taught in America's classrooms. What *is* taught and learned in American classrooms depends on individual state or district mathematics curricula coupled with the educational preparation and instructional practices of teachers and the attentiveness and engagement of students.

The absence of an "official" national curriculum allows for a certain level of flexibility and freedom of choice as to the breadth of content and the depth of coverage in classrooms. This has led to the criticism that the U.S. mathematics curriculum is "a mile wide and an inch deep." The challenge for the CCSS Initiative, then, was to be able to answer the question: What essential mathematical knowledge and skills do students in Grades 4, 8, and 12 need to possess to be equipped to take full advantage of two important postsecondary opportunities—college and careers?

To address this challenge for grades K–12, the CCSS Initiative solicited input, advice, and guidance from professional educators, subject-area experts, policy groups, and the public on how to frame the standards. After reviewing the comments received, the initiative developed the CCSS standards, which were announced in 2009, were released in 2010, and will be assessed across all states in the respective consortia in 2015 when the PARCC and Smarter Balanced assessments are available. The standards were designed to be robust and relevant and based on a careful study of what (1) *is being taught* in countries with whom the United States has to compete and (2) *needs to be taught* to adequately prepare America's young people for successful postsecondary experiences and opportunities. Specifically, for the latter component, the standards' objective was to define a more focused, coherent curriculum framework. In the area of mathematics, the CCSS Initiative developed the CCSS for Mathematics (CCSS-M) content standards to delineate what mathematical content should be taught and learned and what mathematical expertise students should develop.

For more information on the NAEP Mathematics Framework and the CCSS-M, see Appendix A.

Comparing Standards to Standards

Comparing standards to standards can present many challenges and may result in many errors. The purpose of these comparisons is to determine what is substantially the same and what is different about the two sets of standards. One must remember that what is being compared is text. The text is written in a genre that is highly structured, almost in outline form. The authors of the text have choices to make about their structure: what should be superordinate, what should be subordinate, how precisely each topic should be described, and so on. A major goal in comparing standards to standards is to minimize the occurrence of interpretive errors such as pseudo-discrepancies, pseudo-matches, and pseudo-precision. The current study sought to minimize these types of errors.

A *pseudo-discrepancy* can occur when the same material is distributed differently by the compared standards in their respective organizational structures. For example, "estimation" is treated in the NAEP Mathematics Framework as a specific subtopic and also as an "estimation" objective in the content area of Number Properties and Operations, whereas the CCSS-M distribute "estimation" across multiple standards. As such, if the study methodology relied on a literal comparison of words, there would be a finding of discrepancy. The expert panels that participated in the current study were instructed to conduct a more deliberate evaluation of the topic "estimation" that transcended the organizational location of the topic in the text. It was expected that this type of evaluation would reduce the occurrence of pseudo-discrepancy errors.

Similarly, the same term might occur in both standards, leading to a finding of a match based on the literal occurrence of a word or topic. However, the meaning of the word, or the topic, in each context might be quite different, causing a *pseudo-match*. To decrease the occurrence of this type of interpretive error, the panels in this study were asked to evaluate and compare what was said about the topics for each standard and not just rely on the words used.

In addition to errors of pseudo-discrepancy and pseudo-match, there can be errors analogous to drawing inferences beyond the precision of the measurements being made. These are errors of *pseudo-precision*. When the meaning of one text is being compared with the meaning of another text, and the texts, although on the same broad topic, are not organized in parallel, care must be taken in how fine-grained the analysis is. Analyses that are too fine-grained could lead to results that are misleading. Therefore, panelists were asked to make broader judgments, at higher levels of analysis—for example, at the NAEP subtopic or CCSS-M cluster levels—where the differences in organizational structures are less likely to lead to pseudo-precision.

NAEP and the CCSS-M: Risks and Benefits

The large number of states and territories that have adopted the CCSS-M as their state standards has significantly reduced the variation in standards among the states. It is the hope that this will lead to a corresponding reduction in variation among states in curriculum: what instructional materials are used, what gets taught, what gets tested, and what gets learned. NAEP, by its mission, is independent of any particular curriculum. Given this curricular agnostic perspective, we asked the question: What changes, if any, should the National Assessment Governing Board (the Governing Board) consider for NAEP in response to the adoption of the CCSS-M across so many states?

Table 1 lists some possible findings from the comparison of NAEP and CCSS-M, and the risks and benefits associated with each. Each of these "if . . . then" propositions poses consequences for NAEP. As shown in the table, the seriousness of the consequences ranges from medium to high. This study is designed to provide data on the types of findings listed in the first three scenarios.

Table 1. Alignment of NAEP, CCSS-M, and Non-CCSS-M Content and the Consequences for NAEP

IF	THEN	Seriousness of Consequence for NAEP
1. If content is included in the CCSS-M at the grade level assessed by NAEP, but NAEP does not assess it …	Then growth in that content could go undetected by NAEP and NAEP will underestimate growth.	High
2. If content is included in NAEP, but not in the CCSS-M …	Then NAEP growth estimates could be diluted by inclusion of untaught content and NAEP will underestimate current growth; however, NAEP could continue to provide estimates of students' performance in areas of interest for long-term trends.	Medium
3. If there is a large degree of overlap between NAEP assessment objectives and the CCSS-M content standards, but there are states that adopt non-CCSS-M content …	Then growth in non-CCSS-M content will go undetected and NAEP will underestimate growth.	Medium
4. If NAEP item samples have grade level by content-strand interactions (e.g., items sample third-grade place value, fifth-grade graphing, and fourth-grade fractions) …	Then anchor items on the scale and perhaps standard setting may be off.	Medium
5. If NAEP item samples have content by complexity interactions different from the CCSS-M (e.g., higher complexity items with fractions, lower complexity items with operations) …	Then complexity will be confounded with content and the scale could be distorted.	Medium

Note: This list is not intended to be exhaustive.

Conducting Content Alignment Studies: A Review of the Literature

The CCSS-M have been adopted by an overwhelming majority of states; therefore, it is imperative that they be examined to determine whether there is alignment between the standards and the NAEP Mathematics Framework, given that NAEP results are used to make comparisons of student achievement across the states, U.S. territories, and the District of Columbia. Conducting an alignment study between a newly implemented set of standards and a previously used set of standards or assessments allows researchers to determine whether the newer set addresses the same or similar attributes (such as focus, coherence, or rigor) as the older set. The results of an alignment analysis comparing the NAEP Mathematics Framework and the CCSS-M also allow policymakers to make wise decisions about what changes, if any, should be made in the NAEP Mathematics Framework.

Content alignment refers to the degree to which content coverage is the same in two or more frameworks. According to the National Assessment Governing Board (n.d.), it is important to note that regardless of whether the focus of the alignment study is on a framework's attributes or content coverage, alignment refers more to the relationship between the two frameworks (or documents) and less to particular characteristics of either of the documents.

Different methodologies have been used in the various alignment studies that have been conducted over the past decade. Early approaches to the study of alignment were developed by Webb (2002, 2005), Porter (2002, 2006), and Achieve, Inc. (2002). All three approaches use panels made up of individuals with expertise in the content area under study. In each approach, panelists, individually or collectively, rate the degree of alignment using specific criteria. A consensus can be reached by the panel members or there may be interest in reporting the variability that exists among them. The three approaches differ, however, in the types of judgments made by the panelists and in the information that is produced in the alignment study. A detailed discussion of the three approaches and the design to guide implementation of content alignment studies for 12th-grade NAEP assessments in mathematics and reading (as well as other assessments that are used to provide indicators for reporting the preparedness of 12th graders on NAEP in these subjects) can be found at www.nagb.org/publications/design-document-final.pdf.

In fact, there are several different alignment study designs that can be employed: (a) standards to standards; (b) standards to assessment items; (c) assessment items to assessment items; (d) assessment items to assessment frameworks (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007; Everson, Kim, & Butvin, 2009); and (e) assessment frameworks to assessment frameworks. The current study employs a hybrid standards-to-assessment framework design.

Curricular Alignment and the CCSS-M

Interest in the relationships, and particularly the alignment, among standards, assessments, and U.S. students' performance on international as well as national assessments emerged in the late 1990s with the release of the original Third International Mathematics and Science Study (TIMSS) data (Schmidt, McKnight, Valverde, Houang, & Wiley, 1997). The results revealed a downward trend in the performance of U.S. students in Grades 4 through 12 relative to the performance of students in other countries. More than two decades later, the message has not changed. Results from international studies such as TIMSS and the Program for International Student Assessment (PISA), as well as national assessment results from NAEP, echo the mediocre performance of U.S. students, especially in mathematics.

Astute observers of these trends recognize that there are several factors related to low performance (Kilpatrick, Swafford, & Findell, 2001; Schmidt et al., 2001). Some of these factors are embodied in the nature of the curricula (Stancavage et al., 2008). These curricula include not only the written or intended curriculum, but the implemented curriculum (what and how it is taught), the learned curriculum (how and how much of it is learned), and ultimately, the assessed curriculum (how it is assessed). Researchers who study the alignment of intended and assessed curricula and the effects of that alignment on the learned curriculum often operationalize the intended curriculum as curricular content standards, the assessment curriculum as assessment frameworks, and the learned curriculum as student performance or achievement (Porter, 2002; Schmidt & Maier, 2009).

Prior to beginning an alignment study, it is common to identify the criteria that will be used to make judgments about alignment. Quite often, the criteria for excellence or important characteristics of that which is to be examined or compared are identified. In the release of the 1997 TIMSS results, the criterion for excellence that was used to make comparisons among countries consisted of the curriculum standards of all countries whose eighth-grade students performed at the top of the international distribution. These countries were referred to as the A+ countries, and three characteristics of their curriculum standards were identified as important: focus, coherence, and rigor (Schmidt, Wang, & McKnight, 2005). In the 1997 TIMSS study release, a measure of *focus* was defined as "the number of topics covered at each grade that was also aggregated over the first eight grades, by counting the total number of topic-by-grade combinations covered in elementary and middle school" (Schmidt & Houang, 2012, p. 235). Essentially, a set of standards possesses the characteristic of focus to the extent that it has a relatively small number of topics. In addition, Schmidt and Houang (2007) defined a topic-grade combination as coverage of a topic at a particular grade.

Schmidt et al. (2005) considered coherence as the most important characteristic of a set of curriculum standards. They defined *coherence* as a sequence of topics and performances, articulated over time, that is logical and reflects, where appropriate, the sequential and hierarchical nature of the disciplinary content from which the subject matter derives. Thus, coherence refers not only to the coverage of topics within the standards, but more importantly, to whether the sequence in which the topics are covered is consistent with the logical structure of the subject matter from which it is derived. Based on this definition, an international model of coherence, referred to as the A+ model, was derived by an examination and vetting of the coherence found in the national standards of the top-achieving TIMSS countries by a group of mathematicians. Schmidt and Houang (2007) also identified quantitative indicators for both focus and coherence, calculated measures for each of the countries in the A+ group, and related focus and coherence to student achievement. The results of that study suggested that focus is an integral part of the concept of coherence, and their joint influence is positively related to performance on the TIMSS mathematics test.

Schmidt and Houang (2012) also undertook a multicomponent, comprehensive study of the CCSS-M. First, the CCSS-M were compared with the A+ model for congruence. Next, the CCSS-M were compared with state standards to determine the level of congruence. Using data from the Teacher Education and Development Study in Mathematics and *Mathematics Teaching in the 21st Century*, state standards for 50 states were compared with the CCSS-M. The cognitive demand of the CCSS-M and the state standards by grade level was also evaluated using four levels: (1) knowledge—memorizing definitions; (2) performing routine procedures; (3) solving routine problems; and (4) mathematics reasoning, including nonroutine problem solving. Schmidt and Houang (2012) considered cognitive demand to be an indication of a topic's depth, related to the third characteristic of the A+ model—rigor. Last, the authors examined the relationship between the CCSS-M and student achievement, as measured by NAEP, through a simple linear regression. The regression analysis tested the hypothesis that states with standards more congruent to the CCSS-M had higher scores on NAEP in 2009.

A two-dimensional approach was used that consisted of a topic/content specification dimension as well as a performance expectation, or cognitive demand, dimension. To assess the congruence between topic/content and cognitive demand, a matrix was formed with topics in the rows and grades across the columns. Congruence was measured by a combination of focus and coherence. The model of congruence in the Schmidt and Houang (2012) study was the CCSS-M. There were five indicators of congruence that were combined to form one overall measure:

1. A dichotomous (0 or -1) indicator that assessed whether a topic was introduced at an earlier grade level than in the CCSS-M. For every topic for which this was the case, a negative one was added to the indicator; however, a zero was assigned when the topic was introduced on the same grade level as in the CCSS-M.
2. An indicator of *focus* that was calculated by adding a negative one each time a topic was covered at a grade level for which it was not intended in the CCSS-M. These occurrences were then summed over all topics.
3. An indicator of the number of times a topic was not covered at a grade level for which it was intended in the CCSS-M. Every time this occurred, a negative one was added to the topic indicator and summed over all topics.
4. An indicator of whether a topic was covered later than the CCSS-M intended (e.g., decimals were covered in Grade 5 when the CCSS-M had indicated that decimals should not have been covered after Grade 3). Each time this occurred, a negative one was added to the topic indicator and summed across all topics.

5. An indicator of whether a topic was covered across consecutive grades, but was covered in only certain grades in the CCSS-M (e.g., in Grades 5 through 8 versus Grades 5, 6, and 8). These occurrences were coded as in indicator 2 above.

According to Schmidt and Houang (2012), the five indicators were summed across all topics to produce a negative value, which indicated the degree of lack of congruence between the standards and the CCSS-M and, more specifically, the degree of deviation from the CCSS-M. To facilitate interpretation of the results, the overall scale for measuring congruence was converted from a negative scale to a positive one, ranging from 0 to 1,000—with 1,000 indicating perfect agreement with the set of standards that represented the model of congruence

The results showed that the CCSS-M are coherent and focused when compared with the A+ model, even though the CCSS-M contain three additional topics (and the topics in the CCSS-M are not ordered in the same way as in the A+ model). Only three topics in the A+ model were introduced at earlier grades than in the CCSS-M, but several topics were introduced earlier in the CCSS-M than in the A+ model. Overall, there were no significant differences between the CCSS-M and the A+ model (i.e., they are congruent), as the two had a degree of consistency of 85 percent.

The results also revealed that from a maximum of 1,000 points on the measure of congruence with the CCSS-M, states ranged in scores from 662 to 826, with a mean of 762 (SD = 33.5). The 50 states were placed into five categories ranging from *most like CCSS mathematics* to *least like CCSS mathematics* based on their congruence with the CCSS-M. The most congruent states were California, Florida, Georgia, Indiana, Alabama, Minnesota, Oklahoma, Michigan, Mississippi, and Washington; the least congruent states were Arizona, Nevada, Iowa, Kansas, Louisiana, New Jersey, Wisconsin, Rhode Island, and Kentucky.

With regard to the *focus* component of the congruence measure, the CCSS-M required slightly fewer topics than the state standards at Grades 1 through 5, but there was little difference between the CCSS-M and the state standards at Grades 6 through 8. Furthermore, when examining cognitive demand, only 3 percent of the state standards reached the highest level—level 4: "mathematics reasoning, including non-routine problem solving." By way of contrast, 61 percent of the state standards were at the lowest level—level 1: "knowledge—memorizing definitions."

The results of a simple linear regression, which included all 50 states, revealed a weak relationship between CCSS-M congruence (that is, congruence between the CCSS-M and state standards) and performance on state NAEP. The states were then divided into two groups: Group 1 consisted of states with standards that varied in their level of congruence with the CCSS-M and the NAEP scores; Group 2 had a high level of congruence with the CCSS-M, but lower NAEP scores. Correlations between the level of CCSS-M-state congruence and performance on state NAEP were then calculated for each group. These analyses revealed there was a positive relationship between congruence and NAEP scores in Group 1, but there was no significant relationship between congruence and NAEP scores in Group 2. After controlling for this group difference, the results showed that states with standards that are more congruent to the CCSS-M generally had higher NAEP mathematics scores.

The current study assesses the alignment of the NAEP Mathematics Framework and the CCSS-M. Unlike Schmidt and Houang (2012), who examined focus and coherence in the CCSS-M, the current study does not specifically examine the extent to which there is focus and coherence in the NAEP framework. Nevertheless, the results of the study could very well lead to the following question: What does the extent of alignment between the NAEP Mathematics Framework and the CCSS-M tell us about the focus and coherence of the NAEP Mathematics Framework, and what effect will that have on NAEP's role as a monitor of student performance in the context of the CCSS-M?

Methodology

In the absence of CCSS-M assessments (which were under development at the time of the study), this alignment study focuses primarily on the conceptual match between the subtopics and objectives in the NAEP Mathematics Framework and the CCSS-M content standards. The Governing Board oversees the development of the NAEP Mathematics Framework, which describes the specific knowledge and skills to be assessed at Grades 4, 8, and 12. The *2011 NAEP Mathematics Assessment Framework* was used in comparisons with the CCSS-M. A subsequent study that compares items from the CCSS-M assessments with the NAEP items will answer other important questions.

The NAEP Mathematics Framework is organized into five broad areas of mathematics content:

- **Number Properties and Operations (NPO),** including computation and understanding of number concepts
- **Measurement (M),** including use of instruments, application of processes, and concepts of area and volume
- **Geometry (G),** including spatial reasoning and applying geometric properties
- **Data Analysis, Statistics, and Probability (DASP),** including graphical displays and statistics
- **Algebra (A),** including representations and relationships

Each content area is divided into subtopics, and each subtopic consists of one or more objectives. These divisions are not intended to separate mathematics into discrete, nonoverlapping elements. Rather, they are intended to provide a helpful classification scheme that describes the universe of mathematical content assessed by NAEP.

The CCSS-M consist of two components: the Standards for Mathematical Content and the Standards for Mathematical Practice. The two components operate in concert to provide school mathematics experiences that, according to the authors, are "substantially more focused and coherent in order to improve mathematics achievement …" in the United States. The CCSS-M set grade-specific content standards for Grades K–8 and subject-specific standards for high school. The grade-level standards are organized into standards, clusters, and content domains. Each content domain consists of clusters of related standards. Standards define what students should understand and be able to do. (See Appendix A for a detailed discussion of how the NAEP Mathematics Framework and the CCSS-M are organized.)

For the current study, two mappings were conducted: (a) CCSS-M content standards to NAEP Mathematics Framework subtopics and objectives; and (b) NAEP Mathematics Framework subtopics and objectives to CCSS-M content standards.

Mappings

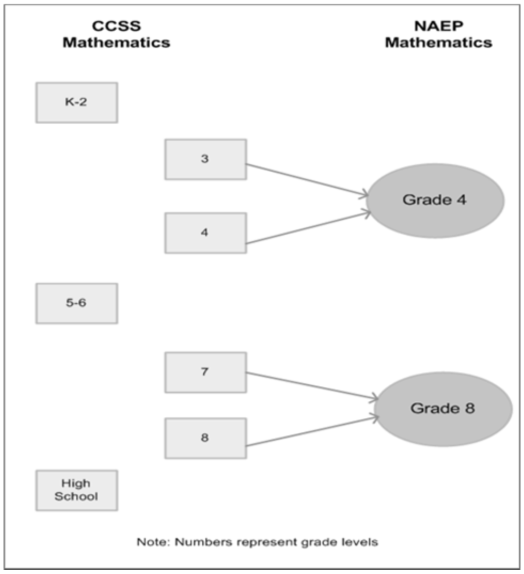
Mapping 1: CCSS-M Standards to NAEP Mathematics Framework (CCSS-M → NAEP)

The mapping from the CCSS-M to the NAEP Mathematics Framework subtopics and objectives was expected to provide answers to the following question, which became **Research Question 1**:

> Which CCSS-M clusters and standards in Grades 3 and 4 or Grades 7and 8 are not represented at all or are not explicitly addressed among the subtopics and objectives for Grade 4 or Grade 8, respectively, in the current NAEP Mathematics Framework? Where there is good representation, in what ways are the CCSS-M clusters/standards and NAEP subtopics/objectives different (i.e., in concept meaning or perspective, specificity of coverage, coverage by grade level, or cognitive demand or complexity)?

Although the CCSS-M span Grades K–8 and high school, Figure 1 shows the specific grade-level mappings referenced in Research Question 1—clusters and standards from Grades 3 and 4 in the CCSS-M to subtopics and objectives for Grade 4 in the NAEP Mathematics Framework, and clusters and standards from Grades 7 and 8 in the CCSS-M to subtopics and objectives for Grade 8 in the NAEP Mathematics Framework. For each mapping, we used the CCSS-M grade that is the same as the grade assessed by NAEP and the CCSS-M grade that is one grade below the grade assessed by NAEP. The absence of arrows means that there was no direct comparison of those grade-level clusters and standards with the subtopics and objectives of the grades assessed by NAEP; however, where important differences or similarities occurred, they were noted.

Figure 1. Mapping From the CCSS-M to the NAEP Mathematics Framework



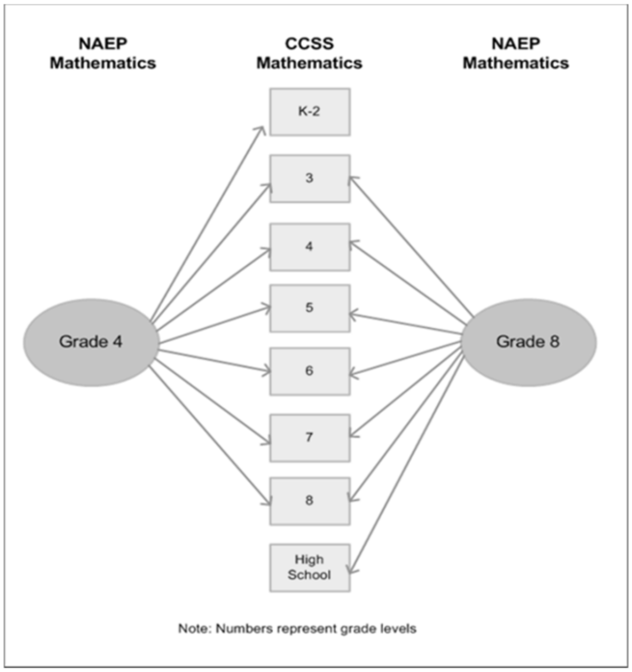
Mapping 2: NAEP Mathematics Framework to CCSS-M Standards (NAEP → CCSS-M)

Any specific Grade 4 or Grade 8 NAEP mathematics objective may suggest a broader breadth of content than any of the CCSS-M specific grade-level standards. Thus, the mapping from NAEP subtopics and objectives to the CCSS-M standards was expected to provide an answer to the following question, which became **Research Question 2**:

> Which NAEP subtopics/objectives for Grade 4 and Grade 8 are not addressed on grade level or have been deemphasized in the CCSS-M?

Figure 2 illustrates how the comparisons for Research Question 2 were operationalized. Each objective in the NAEP Mathematics Framework for Grade 4 and Grade 8 was matched to one or more standards in the CCSS-M. The standards in the CCSS-M that were matched to the objectives in the NAEP Mathematics Framework could be on the grade level of the grade assessed by NAEP or below or above the grade level. The arrows in Figure 2 that extend from Grade 4 and Grade 8 indicate that a "match" could occur across a wide band of clusters and standards in CCSS-M grades. The absence of an arrow from Grade 4 or Grade 8 to a particular CCSS-M grade indicates that a match is not likely to occur among objectives for the grade assessed by NAEP and the standards for that grade in the CCSS-M.

Figure 2. Mappings From the NAEP Mathematics Framework to the CCSS-M



Note: Numbers represent grade levels

Design of the Alignment Study

Specification of Criteria to Determine the Degree of Alignment Between the Two Frameworks

Two criteria were used to describe the degree of alignment between the CCSS-M and the NAEP Mathematics Framework subtopics and objectives: the extent of content coverage and the grade at which the content was covered. The extent of content coverage was rated using four descriptive levels:

- Covered with few differences
- Covered with differences related to specificity
- Covered with differences related to conceptual understanding
- Not covered.

The study also sought to determine the match between the K–8 grades in which the CCSS-M content is supposed to be taught and the grades at which matched objectives appear in the NAEP Mathematics Framework. A mismatch in content by grade could result in an underestimation of students' achievement. For example, if content appears in the NAEP Grade 4 assessment, but that content does not appear in the CCSS-M until later grades, then students who take the NAEP Grade 4 assessment would not have had an opportunity to learn the content. Similarly, if the content appears in the NAEP Grade 4 assessment, but the content is introduced in earlier grades at a level that is less mature than that assessed at Grade 4, then students may not be able to handle the cognitive demand or complexity of the content on the NAEP Grade 4 assessment.

In both cases, students may be underprepared to respond successfully to items or tasks in the NAEP Grade 4 assessment; hence, their mathematics achievement is likely to be underestimated by NAEP. On the other hand, if content that appears in the NAEP Grade 4 assessment is taught in earlier grades in ways that become increasingly more cognitively demanding, then students who take that assessment are better prepared to respond successfully to items or tasks on the NAEP Grade 4 assessment.

Panelists' Procedures for Conducting the Alignment Analysis

Use of Expert Panels: Fourteen experts were divided into four mathematics content panels—two panels each for Grade 4 and Grade 8. At each grade level, one panel addressed the research questions using the NAEP mathematics content areas of Number Properties and Operations and Algebra, while the other panel addressed the research questions using the NAEP mathematics content areas of Measurement; Geometry; and Data Analysis, Statistics, and Probability. Also, two panels examined the alignment of CCSS-M clusters and standards in each of the K–8 grades with the NAEP Grade 4 subtopics and objectives, and two panels examined the alignment of CCSS-M clusters and standards in each of Grades 3 through 8 as well as high school with the NAEP Grade 8 subtopics and objectives.

Composition of Panels: Each panel for Grades 4 and 8 consisted of three or four experts. Experts were drawn from the following four groups: elementary and secondary school teachers and/or school-based mathematics specialists; mathematics educators; mathematicians; and mathematics consultants. Panels were formed based on participants' school-level teaching experience, gender, race/ethnicity, and knowledge of the NAEP Mathematics Framework and the CCSS-M.

Panel Procedures: Panelists reviewed information prior to attending a panel meeting in person. At the panel meeting, panelists discussed their independent judgments about the answers to Research Questions 1 and 2. Then, as a panel, they were asked to reach a consensus about answers to the research questions and to write a panel summary.

To facilitate the panel's review and comparison of the NAEP subtopics and objectives and the CCSS-M clusters and standards and to assist the panelists in answering the research questions, two preliminary mappings—CCSS-M → NAEP and NAEP → CCSS-M—were conducted by Deborah Holtzman, one of the authors of this paper. The results, which were referred to as "Deb's Analysis," were recorded on spreadsheets and sent to the respective panelists.[2]

"Deb's Analysis" was done to reduce the voluminous amount of information about the alignment of CCSS-M clusters and standards with NAEP subtopics and objectives into a manageable quantity. It would not have been possible to ask the panelists to have done this work given the large number of hours these preliminary analyses required. For Grade 4 and Grade 8, the analysis consisted of examining each set of standards organized under a CCSS-M cluster and writing a statement about the extent to which each cluster was covered in a set of NAEP objectives organized by subtopic and grade. "Deb's Analysis" made the examination of the information more manageable, and it also provided a perspective, as a starting point, for panelists to express different levels of agreement with the judgments made about the alignment of the CCSS-M and NAEP Mathematics Framework.

For the CCSS-M → NAEP mapping, "Deb's Analysis" matched groups of standards within each cluster and content domain in the CCSS-M for Grades 3 and 4 with the appropriate objectives, subtopics, and content areas for Grade 4 in the NAEP Mathematics Framework. For example, the CCSS-M Grade 3 standards 8 and 9 in the content domain "Operations and Algebraic Thinking," cluster A "Solve problems involving the four operations and identify and explain patterns in arithmetic" (notated as 3.OA.A.8 and 3.OA.A.9), were matched with the NAEP Grade 4 content area "Number Properties and Operations," subtopic 3 "Number Operations," objective f "Solve application problems involving numbers and operations" (notated as 4NPO3f), and subtopic 5 "Properties of Number Operations," objective e "Apply basic properties of operations" (notated as 4NPO5e). In addition, 3.OA.A.9 was matched to a NAEP Grade 4 objective in the algebra content area, subtopic 1 (4A1a).

---

[2] Deborah Holtzman is a Ph.D.-level analyst with expertise in mathematics education at American Institutes for Research.

Similar comparisons were conducted between the CCSS-M for Grades 7 and 8 and the Grade 8 objectives, subtopics, and content areas in the NAEP Mathematics Framework.

For the NAEP → CCSS-M mapping, subtopics and objectives for Grades 4 and Grade 8 in the NAEP Mathematics Framework were compared with standards and clusters in the CCSS-M for Grades K–8. The goal of this mapping was to identify in what grades and to what degree content objectives in the NAEP framework were aligned with standards in the CCSS-M. Thus, for each objective in the NAEP framework for Grade 4 and Grade 8, "Deb's Analysis" identified the grade(s) in which a similarly stated standard was found in the CCSS-M. Furthermore, a judgment statement was recorded about the extent of the content alignment with the NAEP objectives for Grades 4 and 8.

Panelists were asked to review "Deb's Analysis" and indicate whether they agreed or disagreed with each of the mappings. The panelists were also asked to write comments at the standards level for the CCSS-M → NAEP mapping and at the objective level for the NAEP → CCSS-M mapping, in cases where they did not mark "Agree." The purpose of the comments was to note any perceived misinterpretations or additional information needed in "Deb's Analysis." Finally the panelists were asked to review their ratings of agreement and comments across all standards and objectives and to write summaries of their conclusions. The summaries were to be written at the cluster level for the alignment of the CCSS-M to NAEP and at the subtopic level for the alignment of NAEP to the CCSS-M.

The panelists completed these assignments prior to attending a two-day meeting in person. The results of their preliminary work were used to frame panel discussions and to create panel summaries for each CCSS-M cluster and NAEP subtopic comparison.

A leader for each panel was selected from among its members. The panel leader was charged with facilitating the panel's discussions and submitting the panel's cluster and subtopic summaries.

Analysis and Reporting of Findings

Panelists made two types of judgments—one at a more micro level and one at a more macro level—about the alignment of the NAEP Mathematics Framework and the CCSS-M. The micro-level judgments were related to their individual levels of agreement with the results from "Deb's Analysis." The macro-level judgments were related to their collective level of agreement, in the form of panel summaries, about each CCSS-M cluster and NAEP subtopic.

The findings of the study are represented both qualitatively and quantitatively. The *qualitative* findings are represented by identifying the specific NAEP content that is not covered well in the CCSS-M and the specific CCSS-M content that is not covered well in the NAEP framework, based on the panelists' judgments. Content that is covered well and matched by grade level in the two documents carries no major negative consequences for NAEP. Content that is not aligned well may result in negative consequences for NAEP—some of which were noted in Table 1. The *quantitative* aspects of the study are related, in part, to the "spread" of the content alignment across CCSS-M grades. This spread, or the number of grades in which NAEP objectives are addressed in the CCSS-M, speaks to the extent of coverage between the CCSS-M and NAEP frameworks. Both types of findings are captured in the Results and Discussion section below, separately by cluster in the CCSS-M (Tables 2 through 5) and by subtopic in the NAEP Mathematics Framework (Tables 6 and 7). The tables present the panel summaries and also use shading to denote differences in the extent of content coverage and the amount of spread across grades.

There was no attempt to represent findings in terms of correlation coefficients or other statistical representations of alignment.

Results and Discussion

This section presents the results and discussion of the two mappings—CCSS-M → NAEP and NAEP→ CCSS-M—in connection with the panelists' considerations of the answers to Research Questions 1 and 2.

The results of the analysis and subsequent discussion could potentially serve at least two purposes: (1) provide valuable information about the level of student preparedness in the CCSS-M for the mathematics knowledge and skills that the NAEP assessment is designed to measure in Grade 4 and Grade 8; and (2) make recommendations to NCES regarding broad issues related to the content comparison of NAEP and the CCSS-M, including the extent of alignment that is appropriate to support NAEP's continuing role as an independent monitor.

Research Question 1: *Which CCSS-M clusters and standards in Grades 3 and 4 or Grades 7 and 8 are not represented at all or are not explicitly addressed among the subtopics and objectives for Grade 4 or Grade 8, respectively, in the current NAEP Mathematics Framework? Where there is good representation, in what ways are the CCSS-M clusters/standards and NAEP subtopics/objectives different (i.e., in concept meaning or perspective, specificity of coverage, coverage by grade level, or cognitive demand or complexity)?*

Results for CCSS-M Grades 3 and 4 → NAEP Grade 4

To answer Research Question 1, four panels examined the CCSS-M → NAEP mapping. Two panels examined specifically the alignment of CCSS-M clusters and standards in Grades 3 and 4 with the NAEP Grade 4 subtopics and objectives, and two panels examined specifically the alignment of CCSS-M clusters and standards in Grades 7 and 8 with the NAEP Grade 8 subtopics and objectives. The rationale for targeting two adjacent grades in the CCSS-M was to determine the nature of the alignment of the CCSS-M clusters/standards with the NAEP Grade 4 and Grade 8 framework objectives *at or immediately below* Grade 4 and Grade 8, respectively.

All panelists had access to the results of "Deb's Analysis" as a starting point for making individual judgments about content coverage by grade level between the subtopics and objectives in the NAEP Mathematics Framework and the clusters and standards in the CCSS-M. When the panelists convened for a two-day meeting, the individual panelists' judgments were used to form, by consensus, panel summaries for each CCSS-M cluster and NAEP subtopic.

Table 2 presents the panel summaries that describe the alignment between the CCSS-M standards for Grade 3 and the NAEP subtopics and objectives for Grade 4. Each CCSS-M cluster for Grade 3 is listed in the left-hand column of the table. In addition, there is a visual representation of the nature of the content coverage (or alignment) between the CCSS-M and NAEP, as judged by the panelists, in the right-hand column. Table 3 is set up identically to Table 2, but compares the CCSS-M clusters for Grade 4 and the subtopics and objectives in the NAEP framework for Grade 4.

The panel summaries reveal that the content coverage between the CCSS-M and NAEP could be described in essentially four ways: (1) covered with few differences; (2) covered with differences related to specificity; (3) covered with differences related to conceptual understanding; and (4) not covered. For the purposes of this report, (3) and (4) are combined and illustrated together.

An example of *(1) covered with few differences* is found in Table 3 for CCSS-M Grade 4 content domain "Number and Operations: Fractions," cluster C "Understand decimal notation for fractions, and compare decimal fractions." (notated 4.NF.C) The panel summary for this cluster stated the following: "This cluster is closely aligned with NAEP objectives 4NPO1b, 4NPO1e, and 4NPO1i in subtopic Number Sense and 4NPO3a in subtopic Number Operations.*"* Furthermore, there were no statements from the panel about major differences between the CCSS-M standards and the NAEP objectives. This type of content coverage is denoted by a pattern.

Examples of alignment results that illustrate *(2) covered with differences related to specificity* are numerous. For example, in Table 2, the panel summary for the CCSS-M Grade 3, content domain "Measurement and Data," cluster A "Solve problems involving measurement and estimation" (3.MD.A), stated the following: "The CCSS-M are more detailed in their requirements and also more specific in connecting problem solving and measurement data." In another example in Table 3, the panel summary for the CCSS-M Grade 4 content domain "Operations and Algebraic Thinking," cluster A "Use the four operations with whole numbers to solve problems" (4.OA.A), indicated that: "Computational objectives for CCSS-M and NAEP are aligned in NAEP objectives 4NPO3e and 4NPO3f; however, the representation of multiplication as a comparative operation in CCSS-M is not included (or specified) in NAEP." This type of content coverage is denoted by dark gray.

The two types of alignment that could potentially have more negative consequences for NAEP in its role as a monitor—because they could result in NAEP underestimating student performance—are related to *(3) covered with differences related to conceptual understanding* and *(4) not covered*. An example of alignment (3) can be found in Table 2 for the CCSS-M Grade 3 content domain "Operations and Algebraic Thinking," cluster A "Represent and solve problems involving multiplication and division" (3.OA.A). Here the panel summary noted: "CCSS-M goes beyond the NAEP objectives, which concentrate primarily on procedural skill… It is unclear whether both sets of expectations hold the same conceptual understanding." Differences in cognitive demand could also be related to differences in conceptual understanding. For example, within the same content domain, the panel summary for cluster D, "Solve problems involving four operations, and identify and explain patterns in arithmetic" (3.OA.D), stated: "[S]tandards in this cluster require students to '*explain* patterns in arithmetic,' whereas in NAEP objective 4A1a, the expectation is to '*extend* numerical patterns'." These types of content coverage are denoted by light gray.

Table 2. Coverage of the CCSS-M Grade 3 Clusters in the NAEP Grade 4 Mathematics Framework[1,2]

CCSS-M Grade 3 Clusters	Panel Summaries on Alignment of CCSS-M Grade 3 With NAEP Grade 4	Coverage in the NAEP Grade 4 Mathematics Framework
3.OA: Operations and Algebraic Thinking		
Cluster A: Represent and solve problems involving multiplication and division.	Both the CCSS-M and the NAEP framework expect students to solve problems involving multiplication and division. The CCSS-M in this cluster are mapped to the NAEP Grade 4 subtopic Number Operations, objectives 4NPO3b, 4NPO3c, 4NPO3e, and 4NPO3f. Panelists note that the CCSS-M go beyond the NAEP Grade 4 objectives, which concentrate primarily on procedural skill. It is unclear whether both sets of expectations hold the same conceptual understanding.	░
Cluster B: Understand properties of multiplication and the relationship between multiplication and division.	Although conceptually aligned, the CCSS-M in this cluster clearly set the groundwork for algebraic expressions, which are not covered in the NAEP Grade 4 framework. Some content is covered in the NAEP subtopic Properties of Number and Operations, Grade 4 objective 4NPO5e.	▓
Cluster C: Multiply and divide within 100.	Topical coverage is aligned; however, the CCSS-M expectation includes both fluency and from memory whereas the NAEP Grade 4 objectives, 4NPO3b and 4NPO3c, include the use of a calculator.	▓
Cluster D: Solve problems involving the four operations, and identity and explain patterns in arithmetic.	The CCSS-M expect students to solve two-step word problems with equations. It is unclear whether the expectation of application problems found in NAEP Grade 4 objectives 4NPO3f or 4NPO5e includes two-step problems. Also, standards in this cluster require students to “explain patterns in arithmetic,” whereas in the NAEP Grade 4 objective 4A1a, the expectation is to “extend numerical patterns.”	░
3.NBT: Number and Operations in Base Ten		
Cluster A: Use place value understanding and properties of operations to perform multidigit arithmetic.	There is an explicit expectation that understanding of place value is used to round whole numbers, and fluency is used to add and subtract and to multiply one-digit numbers by multiples of 10. The explicit expectation of rounding is not included in the NAEP Grade 4 objectives. Rather, rounding is mentioned, parenthetically, in the NAEP Grade 4 objective 4NPO2b, which states: “Makes estimates appropriate to ... whole numbers ... by … selecting the appropriate method of estimation (e.g., rounding).” Additionally, the CCSS-M expect fluency, whereas the NAEP Grade 4 framework allows calculators and provides guidelines for what computations will be assessed with and without the use of calculators. Some content coverage of the standards in this CCSS-M cluster can also be found in objectives in three NAEP Grade 4 subtopics: Number Sense—4NPO1a; Number Operations—4NPO3a, b, and e; and Properties of Number and Operations—4NPO5e.	░

3.NF: Number and Operations: Fractions		
Cluster A: Develop understanding of fractions as numbers.	Conceptual understanding of fractions as numbers, especially using a number line, is an expectation in the CCSS-M; however, this expectation is absent in the NAEP Grade 4 framework. The framework suggests models as representations of fractions in the NAEP Grade 4 objective 4NPO1e. Both the CCSS-M standard 3.NF.A.3d and the NAEP Grade 4 objective 4NPO1i address "comparing fractions," but the CCSS-M make explicit the validity of comparisons in the context of the same whole. Furthermore, reasoning about the size of fractions in the CCSS-M expects a lot more than simply comparing numbers as indicated in the NAEP Grade 4 objectives.	░
3.MD: Measurement and Data		
Cluster A: Solve problems involving measurement and estimation.	The CCSS-M are more detailed in their requirements and also more specific in connecting problem solving and measurement data. The panelists thought that the NAEP Grade 4 objectives 4NPO3f, 4M1c, and 4M1e aligned well. The CCSS-M focus on time, volume, and weight only—not on temperature, as does the NAEP Grade 4 objective 4M1b, which specifically mentions temperature.	▓
Cluster B: Represent and interpret data.	The standards in this cluster on solving problems related to a data set do not appear to be as tightly focused as the NAEP Grade 4 objectives. Similarly, the standards' focus on measuring and plotting the measurements in a line plot does not seem to be fully captured by the NAEP Grade 4 objectives.	▓
Cluster C: Geometric measurement: understand concepts of area and relate area to multiplication and to addition.	The standards in this cluster—3.MD.C.5 through 3.MD.C.7d—make up a much more specific, prescriptive, and detailed treatment of student learning outcomes than the NAEP Grade 4 objective, 4M1g, which simply states "solve problems involving area of squares and rectangles." The CCSS-M describe the process of measuring area in much greater detail. The CCSS-M are also very specific about representing the distributive property using areas of rectangles. This treatment continues in the NAEP Grade 4 objectives, but is not nearly as specific.	▓
Cluster D: Geometric measurement: recognize perimeter.	Both the CCSS-M and NAEP Grade 4 framework address solving problems involving perimeter; however, the CCSS-M are more specific and focused than the NAEP Grade 4 objectives. For example, problems in the CCSS-M might involve rectangles with the same perimeter and different areas or with the same area and different perimeters. The relevant NAEP Grade 4 objective 4M1f simply states, "solve problems involving perimeter of plane figures."	▓
3.G: Geometry		
Cluster A: Reason with shapes and their attributes.	This cluster is another example of the CCSS-M being more targeted than what would be found in the NAEP Grade 4 framework, especially at the standard level. The NAEP Grade 4 framework and the CCSS-M also use slightly different language around definition, classification, categories, and so on. The standard 3.G.A.2 in this cluster is more about fractions than about geometry. Also, the NAEP Grade 4 objective 4NPO1e, in the subtopic Number Sense, seems a better fit for the CCSS-M standard 3.G.A.2 than any of the NAEP Grade 4 objectives for Geometry.	░

▩	Covered with few differences	▓	Covered with differences related to specificity	░	Covered with differences related to conceptual understanding

[1] *Notation for the CCSS-M:* Grade level, content domain, cluster, standard number within domain. For example, 3.OA.D.8 is read as Grade 3, Operations and Algebraic Thinking, Cluster D, Standard 8.

[2] *Notation for NAEP objectives:* Grade level, content area, subtopic, objective. For example, 4NPO1i is read as Grade 4, Number Properties and Operations, Subtopic 1, Objective i.

Table 3. Coverage of the CCSS-M Grade 4 Clusters in the NAEP Grade 4 Mathematics Framework[1,2]

CCSS-M Grade 4 Clusters	Panel Summaries on Alignment of CCSS-M Grade 4 With NAEP Grade 4	Coverage in the NAEP Grade 4 Mathematics Framework
4.OA: Operations and Algebraic Thinking		
Cluster A: Use the four operations with whole numbers to solve problems.	Computational standards in the CCSS-M are aligned with NAEP Grade 4 objectives 4NPO3e and 4NPO3f; however, the representation of multiplication as a comparative operation found in the CCSS-M is not among the NAEP objectives. The standard 4.OA.A.3 in this cluster also includes estimation strategies (e.g., rounding) to determine the reasonableness of an answer. A similar expectation can be found in the NAEP Grade 4 objectives 4NPO2b and 4NPO2c under the subtopic Estimation.	▓
Cluster B: Gain familiarity with factors and multiples.	In the CCSS-M, whole numbers in the range of 1 to 100 are classified as prime or composite. Although factor pairs for 1 to 100 are determined, the CCSS-M do not specify prime or composite factorizations. In fact, in the CCSS-M, there is no mention of prime factorization per se. In NAEP Grade 4 objective 4NPO5b, however, there is an expectation to "recognize, find, or use factors, multiples, or prime factorization."	▓
Cluster C: Generate and analyze patterns.	Similar topical coverage of patterns can be found across NAEP Grade 4 objectives 4A1a, 4A1b, 4A1c, and 4A1d. Although the generation of patterns using a rule is a common expectation in the CCSS-M and the NAEP Grade 4 framework, this CCSS-M cluster also expects students to be able to analyze patterns and explain attributes of the elements of the pattern. This "analysis of patterns" expectation is not found among the NAEP Grade 4 objectives.	▓
4.NBT: Number and Operations in Base Ten		
Cluster A: Generalize place value understanding for multidigit whole numbers.	The connection between place value and comparing and ordering whole numbers is not specifically made in the NAEP Grade 4 objectives. Rounding is a strategy explicitly called for in the CCSS-M, but is offered as an example of an estimation strategy in the NAEP Grade 4 framework. (See objective 4NPO2b under the NAEP Grade 4 subtopic Estimation.)	░
Cluster B: Use place value understanding and properties of operations to perform multidigit arithmetic.	Illustrations and explanations of computational results by using equations, rectangular arrays, and/or area models are included in the CCSS-M in this cluster, but are not included in the NAEP Grade 4 framework. Fluency in adding and subtracting multidigit whole numbers is an expectation unique to the CCSS-M.	░

4.NF: Number and Operations: Fractions		
Cluster A: Extend understanding of fractions equivalence and ordering.	The NAEP Grade 4 framework does not sufficiently address fractions as a quantity. The CCSS-M include the statement “Recognize that comparisons are valid only when the two fractions refer to the same whole.” This is an important concept on which the NAEP Grade 4 framework is silent. The NAEP framework does not include the symbols <, >, or =, nor does it connect estimation to comparing numbers. Topics related to comparing fractions can be found in the NAEP Grade 4 subtopics Number Sense (objective 4NPO1i) and Estimation (objective 4NPO2a), but there is no reference to using visual fraction models to compare fractions as in the CCSS-M.	░
Cluster B: Build fractions from unit fractions.	Both the CCSS-M in this cluster and NAEP Grade 4 objective 4NPO3a address addition and subtraction of fractions; however, the CCSS-M approach to building fractions with the use of unit fractions and operations on whole numbers is unique to the CCSS-M. Further, NAEP Grade 4 objectives do not include multiplication of fractions.	░
Cluster C: Understand decimal notation for fractions, and compare decimal fractions.	This cluster is closely aligned with NAEP Grade 4 objectives 4NPO1b, 4NPO1e, and 4NPO1i in the subtopic Number Sense and NAEP Grade 4 objective 4NPO3a in the subtopic Number Operations. These objectives in the NAEP framework cover “representing numbers using models” (as in the case of decimal fractions), comparing decimal fractions, and operations on fractions and decimals.	▩
4.MD: Measurement and Data		
Cluster A: Solve problems involving measurement and conversion of measurements.	Alignment is good; however, some differences are related to specificity. For example, the span of the NAEP Grade 4 subtopics and objectives that map to this CCSS-M cluster reflects the tendency of the CCSS-M to draw together topics from multiple NAEP Grade 4 objectives, including Number Properties and Operations—4NPO3f; Measurement—4M1b, 4M1f, and 4M1g; and Algebra—4A1e. Standard 4.MD.A.2 in this cluster covers solving problems involving simple fractions and decimals. Solving problems involving multiplication or division with fractions or decimals is not represented in any Grade 4 objective in the NAEP framework; rather, this expectation is covered in the NAEP Grade 8 objective 8NPO3f.	▓
Cluster B: Represent and interpret data.	Alignment is good, with exceptions worth noting. The CCSS-M include expectations that students will be given multiple opportunities to analyze and interpret data that they have collected or been given. The CCSS-M require that students know and use multiple ways of representing data and be able to communicate and justify their thinking. The CCSS-M place more emphasis on line plots than do the objectives in the NAEP Grade 4 framework. The NAEP Grade 4 objectives that map to this cluster are in Number Properties and Operations—4NPO3f; and Data Analysis, Statistics, and Probability—4DASP1a and 4DASPb.	▓
Cluster C: Geometric measurement: understand concepts of angle and measure angles.	The CCSS-M’s approach to measurement appears to be a better balance of building a conceptual basis for later procedural skills, whereas the approach in the NAEP Grade 4 framework seems to be more procedural. This CCSS-M cluster is mapped to the NAEP Grade 4 Geometry objective 4G1c, which states: “Identify or draw angles and other geometric figures in the plane.” Measuring angles and drawing angles of a specific measure are emphasized in the CCSS-M, but are less specific in the NAEP Grade 4 objective 4G1c.	░

4.G: Geometry *Cluster A:* Draw and identify lines and angles, and classify shapes by properties of their lines and angles.	Alignment between the CCSS-M and the NAEP Grade 4 objectives 4G1c and 4G1d is good for this CCSS-M cluster.	▩

▩ Covered with few differences

■ Covered with differences related to specificity

□ Covered with differences related to conceptual understanding

[1] *Notation for the CCSS-M:* Grade level, content domain, cluster, standard number within domain. For example, 3.OA.D.8 is read as Grade 3, Operations and Algebraic Thinking, Cluster D, Standard 8.

[2] *Notation for NAEP objectives:* Grade level, content area, subtopic, objective. For example, 4NPO1i is read as Grade 4, Number Properties and Operations, Subtopic 1, Objective i.

Discussion of the Extent of Alignment Between CCSS-M Grades 3 and 4 and NAEP Grade 4

The computational requirements in the CCSS-M at Grades 3 and 4 are matched by the requirements in the objectives in the NAEP framework at Grade 4. One exception to matching computational demands is that the CCSS-M include multiplication of fractions by whole numbers, but the Grade 4 NAEP objectives do not. The CCSS-M also emphasize representing quantitative relationships in a real-world problem by an expression or equation as well as in two-step problems, whereas Grade 4 NAEP objectives do not. An item-to-item comparison in subsequent assessment alignment studies will reveal if these differences are of concern when it comes to NAEP's ability to continue to assess states' educational progress and thereby provide valid information.

Some of the specific understandings in the CCSS-M number domains that are not included in the NAEP framework at Grade 4 are (1) understanding that place value in base 10 implies that each place is worth 10 times as much as the place to its right, (2) illustrating and explaining a multiplication calculation by using equations, (3) using rectangular arrays and/or area models in problem solving, (4) understanding fractions as numbers, and (5) generating fraction equivalence (except by "comparison").

The CCSS-M Grade 3–4 measurement domain is more detailed and specific than are the NAEP Grade 4 objectives. This situation could eventually lead to differences in emphases between Grade 4 measurement items used for NAEP and the CCSS-M. For example, although both the CCSS-M and the NAEP framework include area of rectangles at Grade 4, only the CCSS-M ask for understanding of the connection between area and multiplication and the additivity of areas. In addition, the CCSS-M specify fractional and decimal lengths for measurement problems, while the NAEP framework is not as specific.

A close examination should be undertaken by the Governing Board of each cluster in the CCSS-M for which there is coverage in the NAEP framework at Grade 4, but "with differences related to specificity" or "with differences related to conceptual understanding," or where there is "no coverage in the NAEP framework" at Grade 4.

Importantly, an item-to-item comparison in subsequent studies will reveal if these differences between the CCSS-M and the Grade 4 NAEP subtopics and objectives are associated with variances in the tests.

Results for CCSS-M Grades 7 and 8 → NAEP Grade 8

Table 4 provides panel summaries that describe the alignment between the CCSS-M standards for Grade 7 and the subtopics and objectives in the NAEP Mathematics Framework for Grade 8. Each CCSS-M cluster for Grade 7 is listed in the left-hand column of the table. There is also a visual representation of the nature of the content coverage (or alignment) between the CCSS-M and the NAEP Grade 8 framework, as judged by the panelists. Table 5 is set up identically to Table 4, but compares the CCSS-M clusters for Grade 8 and the NAEP subtopics and objectives at Grade 8.

The panel summaries for the CCSS-M for Grade 7 and Grade 8 reveal that the content coverage between the CCSS-M and the NAEP subtopics and objectives for Grade 8 framework could be described in the same way as the analyses reported for Grades 3 and 4 above: (1) covered with few differences, (2) covered with differences related to specificity, (3) covered with differences related to conceptual understanding, and (4) not covered. For the purposes of this report, (3) and (4) are combined and illustrated together.

An example of *(1) covered with few differences* is found in Table 4 for cluster A "Use properties of operations to generate equivalent expressions" in the CCSS-M Grade 7 content domain "Expressions and Equations" (7.EE.A). The panel summary for this cluster stated: "The CCSS-M specify rational coefficients, while the NAEP Grade 8 framework does not." This type of difference is not major. This was the only cluster in this mapping where coverage between the CCSS-M at Grade 7 and the NAEP subtopics and objectives at Grade 8 were judged to have few differences.

There are several examples of alignment *(2) covered with differences related to specificity*. For example, in Table 4, for cluster B, "Solve real-life and mathematical problems using numerical and algebraic expressions and equations" in CCSS-M Grade 7 content domain "Expressions and Equations" (7.EE.B), the panel summary stated: "…the CCSS-M standard 7.EE.B.3 in this cluster includes 'assess the reasonableness of answers using mental computation and estimation strategies,' which is not explicitly emphasized in NAEP." Another example is in Table 5 for cluster A, "Know that there are numbers that are not rational, and approximate them by rational numbers," in the CCSS-M Grade 8 content domain "Number System" (i.e., 8.NS), where the panel summary stated: "…the CCSS-M for Grade 8 address irrational numbers more explicitly than the NAEP Grade 8 framework. The NAEP Grade 8 framework addresses irrational numbers in two subtopics—Number Sense (objective 8NPO1e) and Estimation (objective 8NPO2a)—where "common irrational numbers such as e and π are applied in contexts."

An example of alignment (*3) covered with differences related to conceptual understanding* can be found in cluster A, "Analyze proportional relationships and use them to solve real-world and mathematical problems," in the CCSS-M for the Grade 7 content domain "Ratios and Proportional Relationships" (7.RP.A). Panelists noted: "Even though there is somewhat of a match between the NAEP Grade 8 objectives … and the CCSS-M cluster, the NAEP Grade 8 objectives do not require the depth of conceptual understanding called for in the CCSS-M." An example of a cluster in which panelists observed that there was "no coverage" can be found in cluster A, "Define, evaluate and compare functions," in the CCSS-M Grade 8 content domain "Functions" (8.F.A). Panelists noted: "…the CCSS-M standard 8.F.A.1 is not addressed in the NAEP Grade 8 framework."

Table 4. Coverage of the CCSS-M Grade 7 Clusters in the NAEP Grade 8 Mathematics Framework[1,2]

CCSS-M Grade 7 Clusters	Panel Summaries on Alignment of CCSS-M Grade 7 With NAEP Grade 8	Coverage in the NAEP Grade 8 Mathematics Framework
7.RP: Ratios and Proportional Relationships		
Cluster A: Analyze proportional relationships and use them to solve real-world and mathematical problems.	Even though there are similarities between the NAEP Grade 8 objectives 8M1i (i.e., solving problems involving ratios) and 8NPO4b, 8NPO4c, and 8NPO4d (i.e., using fractions to represent ratios and proportions) and the standards in this CCSS-M cluster, the NAEP Grade 8 objectives do not require the depth of conceptual understanding called for in the CCSS-M. Items in the NAEP Grade 8 assessment generated from these NAEP objectives could be solved by setting up proportions without understanding the underlying concepts related to proportionality.	░
7.NS: The Number System		
Cluster A: Apply and extend previous understandings of operations with fractions.	The computational aspect of CCSS-M standard 7.NS.A.1 in this cluster is addressed in the NAEP Grade 8 objective 8NPO3a (i.e., perform computations with rational numbers); however, there is no mention of number line representations of fractions in the NAEP Grade 8 objective, as there is in standard 7.NS.A.1. CCSS-M standard 7.NS.A.2d refers to terminating or repeating decimal forms. Division of rational numbers is inferred in NAEP Grade 8 objective 8NPO3a, but explicit knowledge of terminating or repeating decimals is not. Other standards in this cluster map onto NAEP Grade 8 objectives 8NPO3d and 8NPO3e.	░
7.EE: Expressions and Equations		
Cluster A: Use properties of operations to generate equivalent expressions.	Standard 7.EE.A.1 in this cluster is addressed in NAEP Grade 8 objectives across two content areas: Number Properties and Operations—8NPO5e; and Algebra—8A3c. The CCSS-M specify rational coefficients, while the NAEP Grade 8 objectives do not. This latter difference in expectation is not major.	▩
Cluster B: Solve real-life and mathematical problems using numerical and algebraic expressions and equations.	Standard 7.EE.B.3 in this cluster includes "assess the reasonableness of answers using mental computation and estimation strategies," which is not emphasized in the NAEP Grade 8 framework. Also, 7.EE.B.3 emphasizes "apply properties of operations to calculate with numbers in any form," whereas the NAEP Grade 8 objective only references the calculations. 7.EE.B.3 also includes performing operations with tools to solve numeric problems, not just linear algebraic expressions. In addition, 7.EE.B.4a includes "compare an algebraic solution to an arithmetic solution, identifying the sequence of the operations used in each approach," which is not included in the NAEP Grade 8 framework.	▓

7.G: Geometry		
Cluster A: Draw, construct, and describe geometrical figures and describe the relationships between them.	The objectives in the NAEP Grade 8 framework do not focus as much on work with triangles as do the CCSS-M. By focusing on triangles, the CCSS-M pave the way to formal high school work with triangle congruence criteria.	░
Cluster B: Solve real-life and mathematical problems involving angle measure, area, surface area, and volume.	Standard 7.G.B.4 in this cluster calls for an informal derivation of the relationship between the circumference and area of a circle, which does not appear in any of the objectives in the NAEP Grade 8 framework. The CCSS-M also have a greater focus on solving for unknown angle measures in preparation for standard 8.G.A.5 and the standards in the Geometry domain for high school. In the CCSS-M, there are more obvious opportunities for employing the Standards for Mathematical Practice.	■
7.SP: Statistics and Probability		
Cluster A: Use random sampling to draw inferences about a population.	Variability among sample means is not addressed in the NAEP Grade 8 objectives. The standards in this cluster focus on variation, generating a sample, and randomness as a tool for making samples representative. The standards are mapped to the NAEP Grade 8 objectives 8DASP3a and 8DASP3b under the subtopic Experiments and Samples.	■
Cluster B: Draw informal comparative inferences about two populations.	The standards in this cluster focus on making informal comparisons between two populations using measures of variability and central tendency. The NAEP Grade 8 objective 8DASP2d, which states: "Using appropriate statistical measures, compare ... two different populations…," infers the use of measures of variability and central tendency for making comparisons, but is not as specific as the standards in this cluster.	■
Cluster C: Investigate chance processes and develop, use, and evaluate probability models.	The CCSS-M are more specific about expectations and results from greater versus fewer numbers of trials. The CCSS-M elaborate more fully the idea of sample space. The relevant NAEP Grade 8 objectives under the subtopic Probability include 8DASP4a, b, c, d, e, f, g, and j.	■

▩ Covered with few differences

■ Covered with differences related to specificity

░ Covered with differences related to conceptual understanding

[1] *Notation for the CCSS-M:* Grade level, content domain, cluster, standard number within domain. For example, 3.OA.D.8 is read as Grade 3, Operations and Algebraic Thinking, Cluster D, Standard 8.

[2] *Notation for NAEP objectives:* Grade level, content area, subtopic, objective. For example, 4NPO1i is read as Grade 4, Number Properties and Operations, Subtopic 1, Objective i.

Table 5. Coverage of the CCSS-M Grade 8 Clusters in the Grade 8 NAEP Mathematics Framework[1,2]

CCSS-M Grade 8 Clusters	Panel Summaries on Alignment of CCSS-M Grade 8 With NAEP Grade 8	Coverage in the NAEP Grade 8 Mathematics Framework
8.NS: The Number System		
Cluster A: Know that there are numbers that are not rational, and approximate them by rational numbers.	The Grade 8 CCSS-M cover irrational numbers more broadly than the NAEP Grade 8 framework. The NAEP Grade 8 framework addresses rational numbers in objectives in two subtopics—Number Sense (objective 8NPO1e) and Estimation (objective 8NPO2a)—and focuses on “common irrational numbers,” such as e and π in applied contexts.	▓
8.EE: Expressions and Equations		
Cluster A: Work with radical and integer exponents.	The standards in this cluster address exponents more specifically and radicals/roots more conceptually than the objectives in the NAEP Grade 8 framework. For example, neither standard 8.EE.A.1 (“laws of integer exponents”) nor standard 8.EE.A.2 (“represent solutions to equations of the form $x^2= p$ and $x^3 = p$, where p is a positive rational number” and “know that $\sqrt{2}$ is irrational”) is covered in the NAEP Grade 8 framework. The CCSS-M expect students to perform operations with numbers expressed in scientific notation, including multiplicative comparisons. The NAEP Grade 8 objectives 8NPO1f, 8NPO2d, and 8A3c cover scientific notation, estimating square and cube roots, and performing basic operations on roots, respectively.	░
Cluster B: Understand the connections between proportional relationships, lines, and linear equations.	The intent of this cluster is to address the connections between proportional relationships, lines, and equations. This is not a focus in the NAEP Grade 8 framework. Standard 8.EE.B.5 in this cluster is mapped to the following NAEP Grade 8 objectives: 8NPO4c, 8A1f, 8A2a, 8A2b, and 8A4d. These NAEP objectives appear in the subtopics Ratios and Proportional Reasoning; Patterns, Relations, and Functions; Algebraic Representations; and Equations and Inequalities. Standard 8.EE.B.6 is not covered in the NAEP Grade 8 framework.	▓
Cluster C: Analyze and solve linear equations and pairs of simultaneous linear equations.	The standards in this cluster are not covered well in the NAEP Grade 8 framework. Standard 8.EE.C.8, “Analyze and solve simultaneous systems of linear equations,” is not covered at all in the NAEP Grade 8 framework. Standard 8.EE.C.7a, “Give examples of linear equations with different number of solutions,” also is not covered. In addition, the NAEP Grade 8 framework does not address linear equations with the distributive property, as called for in this cluster. What is addressed is found in the NAEP Grade 8 objective 8A4a, “Solve linear equations or inequalities,” and Grade 8 objective 8A4c, “Analyze situations or solve problems using linear equations and inequalities with rational coefficients symbolically or graphically.”	░

8F: Functions		
Cluster A: Define, evaluate, and compare functions.	Standard 8.F.A.1 is not addressed in the NAEP Grade 8 framework; however, components of the remaining standards, 8.F.A.2 and 8.F.A.3, are mapped to the following NAEP Grade 8 objectives: 8A1c, 8A1e, and 8A1f, which address patterns, relations, and functions; 8A2a, 8A2b, and 8A2f, which address algebraic representations; and 8A4d, which focuses on interpretations of relationships between symbolic linear expressions and their graphical representations.	░
Cluster B: Use functions to model relationships between quantities.	The standards in this cluster are mapped to the following NAEP Grade 8 objectives: 8A1c and 8A1e; 8A2a, 8A2b, and 8A2f; and 8A5a. These NAEP Grade 8 objectives are subsumed under the subtopics Patterns, Relations, and Functions; Algebraic Representations; and Mathematical Reasoning in Algebra, respectively.	▓
8G: Geometry		
Cluster A: Understand congruence and similarity using physical models, transparencies, or geometry software.	The NAEP Grade 8 framework does not have the same explicit focus on triangles and angles that appears in standard 8.G.A.5. In the NAEP Grade 8 framework, transformations are not explicitly connected to congruence and similarity. In the CCSS-M, transformations provide the undergirding for an understanding of these ideas. The CCSS-M also provide for the use of technology as a tool for work in this cluster. The properties of transformations are made explicit in the CCSS-M, but not in the NAEP Grade 8 framework. Standards in this cluster are mapped to the following NAEP Grade 8 objectives: 8G2c,8G2e, 8G2f, 8G3f, and 8G4d.	░
Cluster B: Understand and apply the Pythagorean Theorem.	The CCSS-M go further than the NAEP Grade 8 framework in expectations of fluency with the Pythagorean Theorem. Standard 8.G.B.6 specifies “explain a proof” and standard 8.G.B.7 covers work in two and three dimensions. Some of these concepts are mapped to the NAEP Grade 8 objective 8G3d.	▓
Cluster C: Solve real-world and mathematical problems involving volume of cylinders, cones, and spheres.	Standard 8.G.C.9 requires work with volume of a sphere, cone, or cylinder. This standard can be mapped to the NAEP Grade 8 Measurement objective 8M1h, which focuses on solving problems involving the volume or surface area of rectangular solids, cylinders, prisms, or composite shapes.	▓
8SP: Statistics and Probability		
Cluster A: Investigate patterns of association in bivariate data.	The CCSS-M and the NAEP Grade 8 framework include work with scatterplots, but the CCSS-M go beyond finding a line of best fit and interpreting slope to having students make scatterplots and interpret various patterns of distribution. The CCSS-M also cover modeling relationships between quantities using scatterplots. Bivariate categorical data are missing from the NAEP Grade 8 framework. The standards in this cluster are mapped to the NAEP Grade 8 objectives from Data Analysis, Statistics, and Probability (8DASP2e) and Algebra (8A1f).	▓

▨	Covered with few differences	▓	Covered with differences related to specificity	░	Covered with differences related to conceptual understanding

[1] *Notation for the CCSS-M:* Grade level, content domain, cluster, standard number within domain. For example, 3.OA.D.8 is read as Grade 3, Operations and Algebraic Thinking, Cluster D, Standard 8.

[2] *Notation for NAEP objectives:* Grade level, content area, subtopic, objective. For example, 4NPO1i is read as Grade 4, Number Properties and Operations, Subtopic 1, Objective i.

Discussion of the Extent of Alignment Between CCSS-M Grades 7 and 8 and NAEP Grade 8

There are some differences between the CCSS-M at Grades 7–8 and the NAEP framework at Grade 8 that are related to conceptual understanding; these may lead to differences in learning and in the development of the respective assessments. The emphasis in the CCSS-M's "ratio and proportionality" on unit rate (constant of proportionality) is not matched in the NAEP framework at Grade 8; however, the NAEP framework covers the CCSS-M topics in ratio and proportionality. The CCSS-M make explicit the use of number lines in specifying understanding of number systems, whereas the NAEP framework at Grade 8 does not.

Expressions and Equations (algebra) is one content domain in the CCSS-M for which students may be learning mathematics that goes untested and undetected by NAEP at Grade 8. This is perhaps the most dangerous risk to the NAEP mission, given the national priority on algebra for all. It is fundamental to NAEP's mission that its assessments be able to detect progress in this high-priority domain. By not testing what the CCSS-M recommend should be taught, NAEP risks underestimating progress. Increases in student enrollment in Algebra I in eighth grade have already exposed NAEP to this risk, even prior to the development of the CCSS-M.

Whereas Expressions and Equations in the CCSS-M begins the study of topics traditionally taught in Algebra I in the United States, the NAEP framework's treatment of expressions and equations at Grade 8 is more typical of prealgebra. The CCSS-M reflect the migration of Algebra I content to lower grades in the United States over the last two decades. At the time the NAEP Mathematics Framework was originally written, few American eighth graders took Algebra I. The number of eighth graders enrolled in Algebra I has increased substantially—from approximately 15 to 20 percent in the late 1980s and early 1990s to approximately 30 percent in 2009 (Stein, Kaufman, Sherman, & Hillen, 2011). Many of the same topics appear in prealgebra and in Algebra I, but with a real difference in depth, rigor, and technical demand. It appears that something like this difference exists between the NAEP Mathematics Framework and the CCSS-M in Expressions and Equations. As an example, the CCSS-M, but not the NAEP Mathematics Framework, require the use of properties of operations to generate equivalent expressions, laws of exponents, the correspondences between proportional relationships, lines and equations, and analyze and solve linear equations and pairs of simultaneous linear equations. (However, the CCSS-M do not complete the study of Algebra I topics in Grade 8, only going as far as systems of linear equations. Polynomials and quadratic formulas, for example, are in the CCSS-M for high school, not Grade 8.)

Geometry may be another area where the CCSS-M at Grades 7 and 8 go further than the NAEP Mathematics Framework and expose NAEP to underestimating progress. The CCSS-M are more explicit about the mathematical understandings associated with a given topic than is the NAEP framework in geometry at Grade 8. The topics are mostly aligned, but differ in their specificity. In the NAEP framework, for example, students apply the Pythagorean Theorem to solve problems, but understanding and proof are not explicit objectives, as they are in the CCSS-M. Although both the NAEP framework and the CCSS-M have a transformational approach to geometry, the properties of transformations are made explicit in the CCSS-M, but not in the NAEP framework (nor does the NAEP framework have the same explicit focus on triangles and angles that appears in the CCSS-M).

In statistics, the CCSS-M explicitly call for a comparison that involves the use of both a measure of central tendency and a measure of variability. The NAEP framework at Grade 8 does not explicitly call for the use of both measures; rather, it calls for the “use of appropriate statistical measures.” The CCSS-M also include bivariate categorical data, whereas the NAEP framework does not. Otherwise, the alignment is adequate.

In both Grade 4 and Grade 8, the NAEP Mathematics Framework’s approach to broader mathematical expertise is spotty compared with the CCSS-M’s approach. (For instance, the NAEP framework does not have anything comparable to the CCSS-M’s Standards for Mathematical Practice.) Finally, the NAEP framework has incorporated mathematical reasoning in several places, but lacks the explicitness of the CCSS-M.

Research Question 2: *Which NAEP subtopics/objectives for Grades 4 and Grade 8 are not addressed on grade level or have been deemphasized in the CCSS-M?*

Four panels examined the NAEP → CCSS-M mapping to answer Research Question 2. For this mapping, the panels were organized by grade levels assessed by NAEP and by the content areas in the NAEP Mathematics Framework: Grade 4 and Grade 8, Number Properties and Operations and Algebra; and Grade 4 and Grade 8, Measurement; Geometry; and Data Analysis, Statistics, and Probability.

Results for NAEP Grade 4 → CCSS-M

Table 6 presents a graphical representation of the alignment between the subtopics and objectives in the NAEP Mathematics Framework for Grade 4 and the CCSS-M for Grades 1–8. The graphical representation was produced by shading all grade levels where the CCSS-M were matched with objectives in the NAEP Mathematics Framework under a subtopic. For example, the subtopic Number Sense has six objectives. The CCSS-M standards that were matched with the six objectives in the NAEP Mathematics Framework for Grade 4, Number Sense, included the following: 2.NBT.A.1, 2.MD.B.6, 2.NBT.A.3, 2.G.A.2, 2.G.A.3, 3.NF.A.2, 4.NBT.A.2, and 5.NBT.A.3a. These standards represent an alignment spread across Grades 2–5. The different kinds of shading in Table 6 represent different levels of alignment or coverage.

Table 6 reveals that all but one subtopic in the NAEP Grade 4 framework under the content area Number Properties and Operations is covered to some extent in the CCSS-M during or prior to Grade 4. The only exception is the NAEP subtopic Ratios and Proportional Reasoning, which is initially introduced in the CCSS-M at Grade 5. Under the content area Algebra, three of the six subtopics are covered in the Grade 4 CCSS-M: patterns, relations, and functions; algebraic representations; and mathematical reasoning with algebra. The depth of coverage for two algebra subtopics—namely, Variables, Expressions, and Operations; and Equations and Inequalities—is minimal in the CCSS-M, with gaps at Grade 4 and Grade 5.

The two subtopics under the NAEP content area of Measurement are covered in Grade 2 through Grade 5 in the CCSS-M. All objectives in the NAEP framework in geometry at Grade 4 have some coverage in the Grade 4 CCSS-M. There is some concern, however, that there is a difference in specificity between the NAEP objectives in the subtopic Position, Direction, and Coordinate Geometry at Grade 4 and the CCSS-M. The NAEP subtopic Dimension and Shape is covered in the CCSS-M in Grade 2 through Grade 4. Furthermore, objectives in the subtopic Mathematical Reasoning in Geometry are inferred in the CCSS-M across Grade 2 through Grade 6, in part because mathematical reasoning is part of the Standards for Mathematical Practice and is therefore infused throughout the CCSS-M. Finally, there are quite a few gaps in the coverage of objectives in the NAEP content area of Data Analysis, Statistics, and Probability.

More details about the CCSS-M coverage for each Grade 4 NAEP subtopic are provided below and in Appendix B.

Table 6. Coverage of NAEP Grade 4 Mathematics Subtopics in the CCSS-M Grades 1–8

NAEP Subtopic	Where Taught in the CCSS-M?							
Number Properties and Operations (by Grade)								
	1	2	3	4	5	6	7	8
Number Sense		▨	▨	▨	▨			
Estimation		■		■		■		
Number Operations		▨	▨	▨	▨	▨		
Ratios and Proportional Reasoning					░	░		
Properties of Number and Operations		■	■	■				
Mathematical Reasoning Using Number		■	■	■	■			
Algebra (by Grade)								
	1	2	3	4	5	6	7	8
Patterns, Relations, and Functions			▨	░	░		▨	
Algebraic Representations				░	▨			
Variables, Expressions, and Operations			▨			▨		
Equations and Inequalities			▨					
Mathematical Reasoning in Algebra		■	■	■	■			

Measurement (by Grade)

	1	2	3	4	5	6	7	8
Measuring Physical Attributes		■	■	■	■			
Systems of Measurement		■		■	■			

Geometry (by Grade)

	1	2	3	4	5	6	7	8
Dimension and Shape		░	░	░				
Transformation and Shapes and Preservation of Properties				░				■
Relationships Between Geometric Figures	■	■		■				
Position, Direction, and Coordinate Geometry				■	■	▩		■
Mathematical Reasoning in Geometry		■	■	■	■	■		

Data Analysis, Statistics, and Probability (by Grade)

	1	2	3	4	5	6	7	8
Data Representation					▩	▩		
Characteristics of Data Set						░		
Probability							░	

▩ Covered with few differences

■ Covered with differences related to specificity

░ Covered with differences related to conceptual understanding

Number Properties and Operations (NPO): Grade 4

The six subtopics under the Number Properties and Operations (NPO) content area in the NAEP framework are number sense, estimation, number operations, ratios and proportional reasoning, properties of numbers and operations, and mathematical reasoning using numbers. The following descriptions identify the primary areas where there is not a match between the subtopics and objectives in the NAEP framework at Grade 4 and the CCSS-M.

For *Number Sense*, NAEP objective 4NPO1b, which refers to using a two-dimensional model for representing numbers, and objective 4NPO1d, which refers to writing or renaming whole numbers, have been deemphasized in the CCSS-M. Otherwise, the panelists agreed that there is good alignment between the objectives in the NAEP subtopic Number Sense and the CCSS-M.

For *Estimation*, several CCSS-M standards refer to estimation or cover estimation in the context of solving a word problem. The panelists noted that the NAEP subtopic Estimation is not covered so much in the CCSS-M content standards as in the CCSS-M's Standards for Mathematical Practice. Furthermore, the CCSS-M address estimation much more in the context of measurement than in the context of number properties and operations.

For *Number Operations*, references to the operations of addition, subtraction, multiplication, and division of whole numbers; fractions with like denominators; and decimals to the hundredths place are covered in the CCSS-M, especially in domains 2.NBT to 5.NBT, 2.OA to 4.OA, 3.NF, and 3.MD to 4.MD. The explicit reference to the use of the calculator as a method for dealing with multiplication of large whole numbers in the NAEP framework could not be found in the CCSS-M.

The *Ratios and Proportional Reasoning* subtopic contains one NAEP Grade 4 objective: use of "simple ratios to describe problem situations." Information in this subtopic does not appear in the CCSS-M until Grade 5 (in standard 5.NF.B.3) and Grade 6 (in standard 6.RP.A.1).

For *Properties of Numbers and Operations*, the NAEP Grade 4 objective "identifying odd and even numbers" (4NPO5a) is covered in the CCSS-M in standard 2.OA.C.3—two grades below the grade at which it is assessed by NAEP. Beyond Grade 2, even and odd numbers are not the subject of any standard in the CCSS-M. Otherwise, this subtopic receives good coverage in Grades 2 through 5 in the CCSS-M.

The only objective in the subtopic *Mathematical Reasoning Using Numbers* focuses on explaining or justifying "a mathematical concept or relationship." For example, one might be asked to explain why 15 is odd or why 7 minus 3 does not equal 3 minus 7. It is instructive to note that "mathematical reasoning" appears in other subtopics in the NAEP Mathematics Framework: "mathematical reasoning in algebra" and "mathematical reasoning in geometry." Expectations for explanation and justification are evident throughout the CCSS-M, in part because in the CCSS-M, mathematical reasoning is linked to the mathematical practices and not necessarily to any particular content standard. In different ways, the NAEP framework and the CCSS-M treat "reasoning" in a distributed way. Both provide evidence that "reasoning" is pervasive in mathematics.

Measurement: Grade 4

There are two subtopics under the NAEP content area of Measurement at Grade 4: measuring physical attributes and systems of measurement.

Measuring Physical Attributes is covered in several CCSS-M standards before Grade 4. The coverage starts as early as Grade 2 (e.g., 2.MD.A.1–4) and extends to Grade 5 (5.NF.B.4b), where students are specifically expected to solve area problems in which figures have fractional sides. Panelists noted that some of the less explicit attention to estimation in the CCSS-M standards might be compensated for by the emphasis on the mathematical practice "precision." It was noted by the panelists that in many measurement situations, an exact measurement is not called for; thus, "appropriate precision" can be read as an endorsement of estimation when it is needed. Panelists noted that measurement of temperature was completely absent from the CCSS-M.

For *Systems of Measurement*, the panelists observed that all but one of the NAEP Grade 4 objectives in this subtopic are covered in the CCSS-M's Standards for Mathematical Practice. These objectives focus on selecting or using an appropriate type and size of unit for the attribute being measured and determining situations in which highly accurate measurement is important. The one exception is the NAEP Grade 4 objective 4M2b, which focuses on solving problems involving conversions and is covered in the Grade 4 and Grade 5 CCSS-M standards 4.MD.A.1 and 5.MD.A.1.

Geometry: Grade 4

The NAEP content area of Geometry consists of five subtopics: dimension and shape; transformation of shapes and preservation of properties; relationships between geometric figures; position, direction, and coordinate geometry; and mathematical reasoning in geometry.

For *Dimension and Shape*, two of the NAEP Grade 4 objectives are covered in the CCSS-M standards much earlier than at Grade 4—specifically, in kindergarten and Grades 2 and 3 (K.G.A.1, K.G.A.3, 2.G.A.1, and 3.G.A.1). The panelists agreed that, in general, the content coverage for the treatment of solid figures in the CCSS-M is almost nonexistent after kindergarten.

Three of the four objectives under the subtopic *Transformation of Shapes and Preservation of Properties* are covered either in the Grade 4 or Grade 8 CCSS-M. Symmetrical figures, lines of symmetry, and attributes of area are covered in the CCSS-M in 4.G.A.2 and 4.G.A.3, and the identification of images that result from flips (reflections), slides (translations), and turns (rotations) is covered in the CCSS-M in 8.G.A.3 and 8.G.A.4. The NAEP Grade 4 objective 4G2e, which focuses on matching or drawing congruent figures in a given collection, is not explicitly covered in the CCSS-M.

The conceptual match between the NAEP Grade 4 subtopic *Relationships Between Geometric Figures* and the CCSS-M was mixed. The match at the objective level ranged from "covered with few differences" for the description and comparison of properties of simple and compound figures composed of triangles, squares, and rectangles to "covered with differences related to specificity" for the objective that focuses on recognizing two-dimensional faces or three-dimensional shapes. The deemphasis on two- and three-dimensional shapes in the CCSS-M also was mentioned in the discussion on the subtopic of Dimensions and Shape. An objective involving patterns, which is under the subtopic Patterns, Relations, and Functions in the NAEP Algebra content area, appears in the CCSS-M as the context of "patterns of geometric figures." The panelists noted that there is also somewhat of a match between the NAEP Grade 4 objective 4G3a involving geometric patterns and the CCSS-M geometry standard 5.G.A.2.

For the subtopic *Position, Direction, and Coordinate Geometry*, the panel judged that "the subtopic is covered for the most part in a nice progression." Parallelism and perpendicularity are covered in the Geometry domain at Grades 4 and 8 in the

CCSS-M, and the concept of representing geometric figures using rectangular coordinates is covered in standards 5.G.B.3 and 6.G.A.3.

Data Analysis, Statistics, and Probability: Grade 4

At Grade 4, the NAEP content area of Data Analysis, Statistics, and Probability includes three of the four subtopics that also appear at Grade 8 and Grade 12. The subtopics are data representation, characteristics of data sets, and probability. The subtopic excluded from Grade 4 is experiments and samples.

It is fair to say that the only subtopic in this NAEP Grade 4 content area that has adequate coverage in the CCSS-M is *Data Representation*. Even with that analysis, the panelists noted different emphases. For example, data representation in the NAEP Grade 4 framework can take the form of pictographs, bar graphs, circle graphs, line graphs, line plots, and tallies, whereas in the CCSS-M, line graphs are not addressed nor is there any mention of tallies. Tables are mentioned in the Grade 4 CCSS-M, but in the context of a very specific mandated activity (e.g., record measurement equivalents in a two-column table).

The subtopics *Characteristics of Data Sets* and *Probability* could not be matched to any of the CCSS-M standards in Grades 3 through 5. All of the objectives in these two subtopics are introduced in the CCSS-M at Grade 6 or 7, where they appear as standards 6.SP.A.2, 6.SP.B.5c, and 7.SP.C.7.

Algebra: Grade 4

The NAEP content area of Algebra consists of five subtopics: patterns, relations, and functions; algebraic representations; variables, expressions, and operations; equations and inequalities; and mathematical reasoning in algebra.

According to the panelists, the *Patterns, Relations, and Functions* subtopic of the NAEP Grade 4 framework exhibits more dissonance with the CCSS-M than any of the other subtopics in the Algebra content area. The panelists suggested that the concept of "pattern" in the CCSS-M conveys something slightly different from anything found in the NAEP framework at Grade 4. For example, the NAEP Grade 4 "pattern" objectives ask one to (a) recognize, describe, or extend a pattern, or (b) given a pattern or sequence, construct a rule that can generate the terms of the pattern or sequence. The CCSS-M standards 3.OA.D.9, 4.OA.C.5, and 5.OA.B.3 emphasize generating a pattern from a rule and analyzing and explaining patterns.

For *Algebraic Representations*, the emphasis in NAEP on translating between the different forms of representations (symbolic, numerical, verbal, or pictorial) of whole number relations is not explicitly referenced in the CCSS-M. In the CCSS-M, the emphasis is on using different types of representation; hence, translation is implied rather than explicit. The NAEP Grade 4 objective on graphing or interpreting points with whole numbers or letters on a grid is covered in the Grade 5 CCSS-M standards 5.G.A.1 and 5.G.A.2.

The remaining three subtopics—*Variables, Expressions, and Operations; Equations and Inequalities;* and *Mathematical Reasoning in Algebra*—are all covered in the CCSS-M. It was noted by the panelists that two of these subtopics (variables, expressions, and operations; and equations and inequalities) are covered in the CCSS-M as part of solving word problems in Grade 3 and Grade 4 and that the third subtopic (mathematical reasoning in algebra) is mentioned as part of understanding the operations and computations in base 10.

Discussion of the Extent of Alignment Between NAEP Grade 4 and the CCSS-M

At Grade 4, most of the content in the NAEP framework is also included in the CCSS-M. For example, alignment between the NAEP framework and the CCSS-M was quite good for the content domain Number Properties and Operations, with only one subtopic misaligned by grade level—ratio and proportional reasoning. The objectives of ratio and proportional reasoning are introduced at a later grade level in the CCSS-M than in NAEP. In the Algebra content area, the match is good, with two exceptions: (a) the treatment of patterns has a different perspective in the CCSS-M than in NAEP; and (b) the CCSS-M emphasize generating patterns from rules while the NAEP framework emphasizes inferring the next step in a pattern or inferring a rule from a pattern. Whether these differences in perspective will lead to different kinds of test items can only be determined in a future comparative item-to-item study. Even if the items differ in some systematic way, it remains an empirical question how this difference will affect performance. The Measurement and Geometry content areas in the NAEP Grade 4 framework and the CCSS-M do not show major differences.

The clearest difference between the NAEP Grade 4 framework and the CCSS-M is in Data Analysis, Statistics, and Probability. The NAEP framework has substantially more emphasis on data and probability by Grade 4 than do the CCSS-M. It is worth noting, however, that this difference disappears by Grade 8. The CCSS-M concentrate data and probability in fewer and later grades (particularly in Grade 7) than does the NAEP framework. This may lead to a scenario in which students taught under the CCSS-M but tested by NAEP will encounter data and probability constructs they have not been taught, a circumstance which could depress overall NAEP scores. It would be possible, and worthwhile, to study the correlation between the CCSS-M implementation and performance on the Data Analysis, Statistics, and Probability subscale of NAEP over time.

Results for NAEP Grade 8 → CCSS-M

Table 7 illustrates the alignment of Grade 8 NAEP subtopics and objectives with the CCSS-M for Grade 1 to Grade 8. Overall, the NAEP Grade 8 objectives in the content areas Number Properties and Operations, Algebra, and Geometry have very good coverage in the CCSS-M in Grade 6 through Grade 8. Gaps in coverage in the CCSS-M for NAEP Grade 8 objectives appear in the content areas Measurement and Data Analysis, Statistics, and Probability.

The NAEP Grade 8 subtopics appear to have fewer gaps (or greater coherence across grade bands) than the NAEP Grade 4 subtopics. NAEP Grade 8 objectives in the subtopic Number Operations are mapped to standards in the CCSS-M across seven continuous grades—from Grade 2 to Grade 8. Four of the six NAEP Grade 8 subtopics under the content area Number Properties and Operations have continuous coverage across at least six grades. Similarly, the subtopics in the content area Geometry—including Dimension and Shape, Relationships Between Geometric Figures, and Mathematical Reasoning in Geometry—are covered to various degrees across six continuous grades.

More details about the CCSS-M coverage for each Grade 8 NAEP subtopic are provided below and in Appendix C.

Table 7. Coverage of Grade 8 NAEP Mathematics Subtopics in the CCSS-M Grades 1–8

NAEP Subtopic	Where Taught in the CCSS-M?							
	Number Properties and Operations (by Grade)							
	1	2	3	4	5	6	7	8
Number Sense			▩	▩	▩	▩	▩	▩
Estimation			░	░	░	░	░	░
Number Operations		▩	▩	▩	▩	▩	■	▩
Ratios and Proportional Reasoning					▩	▩	▩	▩
Properties of Number and Operations				▩	▩	▩	▩	
Mathematical Reasoning			▩	▩	▩	▩	▩	▩
Algebra (by Grade)								
	1	2	3	4	5	6	7	8
Patterns, Relations, and Functions				▩	▩			▩
Algebraic Representations						▩	▩	▩
Variables, Expressions, and Operations						▩	▩	▩
Equations, and Inequalities						▩	▩	▩
Mathematical Reasoning in Algebra						▩	▩	▩
Measurement (by Grade)								
	1	2	3	4	5	6	7	8
Measuring Physical Attributes		▩	▩	▩	▩	▩	▩	
Systems of Measurement				░				
Measurement in Triangles							▩	▩

Geometry (by Grade)

	1	2	3	4	5	6	7	8
Dimension and Shape			■	■	▒	▒	▒	▒
Transformation and Shapes and Preservation of Properties				▦		■	■	▦
Relationships Between Geometric Figures			▦	▦	▦	▦	▦	▦
Position, Direction, and Coordinate Geometry				▦		▦	▦	▦
Mathematical Reasoning in Geometry			▒	▒	▒	▒	▒	▒

Data Analysis, Statistics, and Probability (by Grade)

	1	2	3	4	5	6	7	8
Data Representation						▒	▒	▒
Characteristics of Data Set						■	■	■
Experiments and Samples							▒	
Probability							▦	

▦ Covered with few differences

■ Covered with differences related to specificity

▒ Covered with differences related to conceptual understanding

Number Properties and Operations: Grade 8

Most of the NAEP Grade 8 objectives under the subtopic *Number Sense* are covered in the CCSS-M prior to Grade 8. For example, NAEP objective 8NPO1a, in which place value is used to model and describe integers and decimals, is not mentioned in the CCSS-M beyond Grade 5. NAEP Grade 8 objectives 8NPO1b, 8NPO1g, and 8NPO1h—which include modeling rational numbers, modeling and applying absolute value, and comparing rational numbers using various representations (e.g., fractions, decimals, percentages or integers)—are covered in the CCSS-M in Grade 6 and Grade 7. (The specific CCSS-M standards that are matched to this subtopic are 6.RP.3b and 7.ND.2d.) Expressing or interpreting numbers using scientific notation from real-life contexts (8NPO1f) is the only NAEP Grade 8 objective in the Number Sense subtopic that appears to be introduced for the first time in eighth grade in the CCSS-M.

The NAEP Grade 8 objectives in the *Estimation* subtopic focus primarily on the accuracy and appropriateness of estimation in a particular context or situation. For example, the NAEP Grade 8 objective 8NPO2d, which covers estimation of square roots or cube roots of numbers less than 1,000, is very similar to the CCSS-M Grade 8 standard 8.NS.A.2, which focuses on the use of rational approximations of irrational numbers for comparing the size of irrational numbers. There is one important difference between the expectation in the NAEP Grade 8 objective and the expectation in CCSS-M standard 8.NS.A.2: namely, 8.NS.A.2 involves a two-step process (first, the estimation of the irrational number by a rational approximation;

second, a comparison of the rational approximations). Furthermore, standard 8.NS.A.2 does not identify an upper limit (e.g., 1,000) for selecting examples nor does it explicitly mention the use of calculators or computers to verify results, as is the case for NAEP Grade 8 objective 8NPO2c. There are CCSS-M practice standards that refer to estimation; however, there are no CCSS-M content standards at Grade 8 that are specifically about estimation.

The NAEP Grade 8 subtopic *Number Operations* covers performing operations on rational numbers, interpreting the results of number operations, and solving application problems involving rational numbers and operations. The objectives under this subtopic allow for the use of exact answers or estimates "as appropriate" for problem solving, whereas CCSS-M standard 7.EE.B.3 calls for assessing the "reasonableness of answers using mental computation and estimation strategies."

The NAEP Grade 8 subtopic *Ratios and Proportional Reasoning* includes the use of fractions to represent and express ratios and proportions in problem situations, particularly solving problems involving percent increase and decrease, interest rates, and part/whole relationships. All of these objectives are covered in the CCSS-M with standards 6.RP.A and 7.RP.A. The NAEP Grade 8 objective 8NPO4c, "using proportional reasoning to model and solve problems," is also covered in the CCSS-M.

The NAEP objectives in the subtopic *Properties of Numbers and Operations* are mapped to standards that are introduced and taught prior to Grade 8. Prime and composite numbers are covered in the CCSS-M by standard 4.OA.B.4. Greatest common factors and least common multiples are mentioned in the CCSS-M standard 6.NS.B.4, and the application of basic properties of operations is covered in the CCSS-M at Grade 6 and Grade 7. Operations with odd and even numbers and rules of divisibility, however, are not specifically mentioned in the CCSS-M.

The NAEP subtopic *Mathematical Reasoning Using Number,* which includes "explaining operations with two or more fractions," is represented in a standard in the CCSS-M for Grade 5 that involves multiplication of fractions as well as in a standard for Grade 6 that involves division of a fraction by a fraction. The panelists also noted that even though an objective in this subtopic calls for explanations and justifications of mathematical concepts or relationships, "justifications" are seldom asked for in the CCSS-M for Grade 6 through Grade 8. They are, however, mentioned in the CCSS-M Standards for Mathematical Practice, which apply to all grades.

Measurement: Grade 8

The subtopic *Measuring Physical Attributes* has six objectives, all of which are covered in the CCSS-M, but at various grade levels from Grade 2 to Grade 7. Three NAEP Grade 8 objectives—8M1b, which focuses on comparing objects with respect to some measurement attribute; 8M1c, which asks individuals to estimate the size of an object with respect to a measurement attribute; and 8M1e, which requires individuals to use an appropriate measurement instrument or create a given unit of measure—are mapped to standards in grades much lower than Grade 8 (e.g., Grade 2 and Grade 3) in the CCSS-M. The remaining three objectives under this subtopic—8M1f, 8M1h, and 8M1i—all involve solving problems related to perimeter, area, volume, rates, and population density. These latter three objectives are covered in Grade 5 to Grade 7; however, concepts of density, including population density, do not appear in the CCSS-M until high school.

Two of the NAEP Grade 8 objectives in S*ystems of Measurement*—focus on estimation and determining the appropriate size of a unit of measurement—are not matched with any of the CCSS-M standards. Instead, both NAEP Grade 8 objectives (and the closely parallel and Grade 4 objectives) are more aligned with several of the CCSS-M Standards for Mathematical Practice, particularly SMP5 and SMP6. (See Mathematical Practice Standards in Appendix A.)

The NAEP Grade 8 objectives under the subtopic *Measurement in Triangles* focus on solving problems involving indirect measurement. These objectives are covered in the CCSS-M by 7.G.A.1, 7.G.B.6, 8.G.A.4, and 8.EE.B.6.

Geometry: Grade 8

Under the subtopic *Dimension and Shape,* NAEP objective 8G1a, which refers to drawing or describing a path of shortest length between points to solve problems in context, was judged by the panelists to be "not covered" in the CCSS-M. However, upon close examination of the CCSS-M Grade 8 standards, it appears that 8.G.B.8, which refers to applying the Pythagorean Theorem to find the distance between two points in a coordinate system, could be a conceptual match for NAEP objective 8G1a. All other objectives under this subtopic are covered in the CCSS-M for Grade 6 and Grade 7, with the exception of objective 8G1b, which asks individuals to identify a geometric object given a written description of its properties. This latter objective is covered in the CCSS-M at Grades 3, 4, and 5 (standards 3.G.A.1, 4.G.A.2, and 5.G.B.3).

The NAEP Grade 8 objectives in the subtopic *Transformation of Shapes and Preservation of Properties* are covered for the most part in CCSS-M standards 8.G.A.2, 8.G.A.3, and 8.G.A.4; however, the foundational understandings of combining, subdividing, and changing shapes of plane figures and solids are in the CCSS-M for Grade 6 and Grade 7 (standards 6.G.A.1, 7.G.A.3, and 7G.B.4). In addition, lower levels of cognitive demand, which ask individuals to "identify" or "recognize" lines of symmetry in plane figures, appear in the CCSS-M for Grade 4 (standard 4.G.A.3).

For the NAEP Grade 8 subtopic *Relationships Between Geometric Figures*, the panelists noted that there is a strong match between the NAEP objectives and the CCSS-M for Grade 3 through Grade 8. The CCSS-M that were matched with the NAEP Grade 8 objectives in this subtopic included standards 3.G.A.1, 4.G.A.1, 4.G.A.2, 5.G.B.3, 5.G.B.4, 6.G.A, 7.G.A, and 8.G.A–C.

The NAEP Grade 8 objectives in the subtopic *Position, Direction, and Coordinate Geometry* cover the grade span from Grade 4 to high school with a gap at Grade 5 in the CCSS-M. NAEP objective 8G4a (which focuses on describing relative positions of points and lines using geometric ideas of midpoint, parallelism, and perpendicularity) is first introduced in CCSS-M standards 4.G.A1, 4.G.A.2, and 4.G.A.3. Furthermore, for standards 8.G.A.1–5, students use congruence, similarity, or geometric software to meet NAEP objective 8G4a.

The NAEP objective under the subtopic *Mathematical Reasoning in Geometry* is not specifically covered in the CCSS-M, but is conceptually aligned with the CCSS-M Standards for Mathematical Practice.

Data Analysis, Statistics, and Probability: Grade 8

The objectives under the NAEP Grade 8 subtopic *Data Representation* that focus on (a) reading, interpreting, interpolating or extrapolating from data; and (b) graphing and solving problems related to data (8DAS1a, 8DASP1b) are covered in the CCSS-M in standards 6.SP.A.2, 7.SP.A.1, 8.SP.A.3, and 8.SP.A.4. The remaining NAEP objectives (8DASPc, 8DASPd, and 8DASPe), which focus on (a) solving problems by estimating; (b) determining whether information in a graph is represented effectively and appropriately; and (c) comparing/contrasting the effectiveness of different representations of the same data, are reflected in the CCSS-M Standards for Mathematical Practice. The specific standards that apply are SMP1—make sense of problems and persevere in solving them; SMP2—reason abstractly and quantitatively; SMP3—construct viable arguments; SMP5—use appropriate tools strategically; and SMP6—attend to precision. Circle graphs, which appear in NAEP Grade 8 objective 8DASP1b, are deemphasized in the CCSS-M.

In the NAEP Grade 8 subtopic *Characteristics of Data Sets*, the mean and median are covered in CCSS-M standards 6.SP.A.3 and 7.SP.B.4 as "measures of center;" however, there is no specific reference to mode in the CCSS-M. Also, the CCSS-M seem to place greater emphasis on understanding and interpreting the measures of center and spread than on calculating them. The NAEP Grade 8 objective 8DASP2c, on outliers, is covered by two CCSS-M standards (6.SP.B.5c and 8.SP.A.1); however, these standards do not specifically address the *effect* of outliers on measures of central tendency and spread as does the NAEP objective.

The NAEP Grade 8 subtopic *Experiments and Samples* is covered somewhat in the CCSS-M, primarily in Grade 7. The NAEP Grade 8 objectives focus broadly on issues related to sampling design, whereas the CCSS-M focus only on the need for a sample to be random. The NAEP objective 8DASP3d, "evaluate the design of an experiment," is covered in the CCSS-M high school statistics and probability content domain.

The NAEP Grade 8 subtopic *Probability* is covered in the CCSS-M at Grade 7 in standard 7.SP (all clusters). The panelists noted that while there is a strong match between the NAEP framework and the CCSS-M for this subtopic, the NAEP framework goes further than the CCSS-M in including a focus on independent and dependent events. The CCSS-M address the probability of independent and dependent events in high school statistics.

Algebra: Grade 8

Within the NAEP Grade 8 subtopic *Patterns, Relations, and Functions*, objectives related to numerical or geometric patterns and sequences are covered in the CCSS-M for the elementary grades in 4.OA.C.5 and 5.NBT.A.2, but are not found in any of the standards for the middle grades (Grade 6 through Grade 8). However, objectives related to linear functions—that is, how to calculate their slopes and intercepts and how they contrast with nonlinear functions—are covered in the CCSS-M by Grade 8 (in standards 8.SP.A.1 and 8.SP.A.2).

The NAEP Grade 8 subtopic *Algebraic Representations* has objectives that are covered throughout the CCSS-M for Grade 6 through Grade 8. For example, NAEP objective 8A2c (graphing or interpreting points on a rectangular coordinate system) and objective 8A2d (solving problems involving coordinate pairs) are covered in standard 6.NS.C.8. Objective 8A2f (identifying or representing functional relationships in meaningful contexts) is covered in standards 8.EE.B.5 and 8.F.B.5. Analyzing or interpreting linear relationships—objective 8A2b—is covered in standard 8.F.A.3.

The NAEP Grade 8 subtopic *Variables, Expressions, and Operations* has two objectives. NAEP objective 8A3b, which deals with writing algebraic expressions, equations, or inequalities, is covered in CCSS-M standards 6.EE.B.6, 6.EE.B.7, 6.EE.B.8, and 7.EE.A.2. NAEP objective 8A3c, which focuses on performing basic operations and using appropriate tools on linear expressions, is addressed broadly in the CCSS-M content domain Expressions and Equation at Grades 5–8. Objective 8A3c also is covered by SMP5 in the CCSS-M Standards for Mathematical Practice (see Appendix A).

Solving equations and inequalities and interpreting the meaning of the equal sign are covered in the NAEP Grade 8 subtopic *Equations and Inequalities.* There also is a focus on demonstrating how to use and evaluate common formulas. These areas are covered in the CCSS-M for Grade 6 through Grade 8, primarily in the content domain Expressions and Equation.

The NAEP Grade 8 subtopic *Mathematical Reasoning in Algebra* asks that individuals make, validate, and justify conclusions and generalizations about linear relationships. This topic is covered in the CCSS-M content domain Expressions and Equation at Grades 6 and 8 and in the Standards for Mathematical Practice.

Discussion of the Extent of Alignment Between NAEP Grade 8 and the CCSS-M

Every content area in the NAEP Grade 8 framework has been covered in the CCSS-M by Grade 8 and, in most cases, is initially presented at an earlier grade. Under ideal conditions, it is not likely that students taking the NAEP Grade 8 assessment would encounter topics that they have not been taught. Thus, the risk of underestimating growth by diluting scores with untaught material is small for the NAEP Grade 8 assessment.

There are some differences in specificity and conceptual understandings between the NAEP Mathematics Framework and the CCSS-M, and these differences might matter when assessing students. In Number Properties and Operations, the NAEP framework treats "estimation" as a content area whereas the CCSS-M distribute estimation among other content domains and the Standards for Mathematical Practice. Some topics in measurement are covered in much lower grades in the CCSS-M than in the NAEP framework. This could lead to "less mature" versus "more mature" differences in the conceptualization of these topics. To be certain of this will require item-to-item comparisons in subsequent studies. In Data Analysis, Statistics, and Probability, experimental design and conditional probability are not taught until high school in the CCSS-M, but get some attention in the NAEP Grade 8 framework.

Conclusions and Recommendations

When Should NAEP Change the Yardstick?

When a set of common standards on which common assessments would be based, and which nearly all states would adopt, became a reality in 2010 with the introduction of the CCSS, it became necessary for NAEP to attend to shifting definitions and emphases of subject matter competence and to determine how these might affect claims about progress or lack thereof on a national, state, or district level (National Center for Education Statistics, 2012).

Historically, the NAEP frameworks have aspired to represent the union of all the various state curricula while reaching beyond these curricula to lead as well as reflect. As a result, NAEP often has pushed on the leading edge of what the nation's children know and should able to do. The introduction of the CCSS-M provides both new opportunities and challenges for NAEP. As the nation moves toward widespread implementation of instruction and assessment based on the CCSS-M, NAEP must balance the goals of comparability over time (i.e., maintaining trend) with keeping itself relevant.

NAEP in the Context of the CCSS-M

This study found the preponderance of content in the CCSS-M also is found in the NAEP Mathematics Framework, but with some differences. The differences are potentially important and should receive attention in the normal revision of the framework and the assessments. Four types of discrepancies were observed. Compared with the NAEP framework, the CCSS-M have

1. More rigorous content in eighth-grade algebra and geometry
2. More extensive and systematic treatment of mathematical expertise (found in the Standards for Mathematical Practice)
3. A more conceptual perspective on many mathematical topics, explicitly stating the mathematics to be understood rather than the type of problem to be solved
4. Some content taught at higher grades than is assessed in the fourth-grade NAEP assessment. For example, the study of proportional relationships is concentrated in Grades 6 and 7, and data sets and probability are taught in Grades 6 and 7, respectively.

These are important differences and these areas should be considered a priority in the normal revision of the NAEP Mathematics Framework.

The study also found that the CCSS-M include the preponderance of content included in the NAEP framework by the grade level assessed, with several important exceptions noted in the results reported above. Subsequently, where content is assessed by NAEP, but not included in the CCSS-M, analyses should be conducted to estimate the effect that dropping this content from the curriculum that align with the CCSS-M might have on overall NAEP scores. This should be done to avoid misinterpreting this effect as a general decline in mathematics achievement, when it may be due to a specific decline in a subdomain that has been intentionally deemphasized in the CCSS-M.

Recommendations and Next Steps

Based on the results of our research, we offer the following recommendations with respect to the NAEP Mathematics Framework and the NAEP mathematics assessments in the context of the CCSS-M:

- NAEP should continue to maintain its role as an independent monitor of the academic achievement of the nation's students.
- NAEP should not aim to be a replica of the assessments that are based on the CCSS-M, but should make use of advances in item development technology associated with the CCSS-M assessments, particularly those related to assessing mathematical practices—an area that has not been a strong point for NAEP, especially when designing items of high complexity.
- NAEP should review its mathematics framework to ensure that objectives remain current and reflect the coursetaking patterns of the nation's students (e.g., algebra I enrollment in eighth grade versus ninth grade and the placement of content assessed on the fourth-grade NAEP, such as proportionality and probability, in higher grades in the CCSS-M curriculum).
- NAEP should continue to lead improvements in item design and should pay particular attention to avoiding items biased toward a characterization of mathematics as merely a domain of problems organized as topics. The items should also assess conceptual understanding of the mathematics, explanations of solutions, reasoning and content, and manipulation of expressions or equations for a purpose.
- NAEP should consider improving its strategy for assessing mathematical expertise, perhaps expanding and adding a broader set of objectives to the assessment frameworks that cut across content areas and focus on what in the CCSS-M are called "mathematical practices." A move in this direction can already be seen for "mathematical reasoning." In 2005, mathematical reasoning appeared only in the Geometry content area in NAEP; however, by 2013, mathematical reasoning appeared as a subtopic in Number Properties and Operations, Algebra, and Geometry in Grades 4 and 8. NAEP has extensive experience in assessing skills in reading and writing and should draw on this expertise to do something similar in mathematics.
- NAEP should continue to serve as a leader, especially in the areas of scoring, interpreting, and reporting assessment data and information from different sources (e.g., providing linkages among district, state, national, and international assessments).
- When the CCSS-M items are available, NAEP should carry out a study comparing how well NAEP items reflect the CCSS-M standards and how well the CCSS-M items fit into the NAEP Mathematics Framework.

A major trend to which NAEP must respond if it is to remain relevant in the future is outlined in the report titled *NAEP: Looking Ahead—Leading Assessment Into the*

Future (National Center for Education Statistics, 2012). That trend is NAEP's capacity to assess a broader set of learning outcomes. NAEP needs to remain responsive in a changing and dynamic curriculum and assessment milieu. Whether the issues are related to high-stakes versus low-stakes, status versus growth, or assessment *of* learning versus assessment *for* learning, NAEP's role must be clear and unambiguous. If change is coming to NAEP, and particularly the NAEP frameworks, it must be deliberate and not reactionary, thoughtful and not superfluous. NAEP has undergone notable changes to meet expanded new demands in the past. NAEP also can meet new demands successfully—not only now, but also in the future in the context of the Common Core State Standards.

References

Achieve, Inc. (2002). *Benchmarking and alignment of standards and testing*. Retrieved from http://www.achieve.org/files/TR566.pdf

Common Core State Standards Initiative. (2010). *Common Core State Standards for Mathematics*. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

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

Everson, H. T., Kim, Y. Y., & Butvin, H. (2009). *An empirical look at the 2007 and 2009 NAEP Reading Assessment Frameworks: A content alignment study*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academies Press.

National Assessment Governing Board. (n.d.). *Design of content alignment studies in mathematics and reading for 12th grade NAEP and other assessments to be used in preparedness research studies*. Retrieved from http://www.nagb.org/publications/design-document-final.pdf

National Assessment Governing Board. (2010). *Mathematics framework for the 2011 National Assessment of Educational Progress*. Washington, DC: Author.

National Center for Education Statistics. (2012). *NAEP: Looking ahead—Leading assessment into the future*. Retrieved from http://www.nces.ed.gov/nationsreportcard/pdf/Future_of_NAEP_Panel_White_Paper.pdf

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*(7), 3–14.

Porter, A. (2006). Measuring alignment. *National Council on Measurement in Education Newsletter*, *14,* 4. Retrieved from http://www.ncme.org/default/assets/File/pdf/newsletter/vol14_4_Dec2006.pdf

Schmidt, W. H., & Houang, R. T. (2007). Lack of focus in the mathematics curriculum: Symptom or cause? In T. Loveless (Ed.), *Lessons learned: What international assessments tell us about math achievement* (pp. 65–84). Washington, DC: Brookings Press.

Schmidt, W. H., & Houang, R. T. (2012). Curricular coherence and the Common Core State Standards for Mathematics. *Educational Researcher, 41*, 294–308.

Schmidt, W. H., & Maier, A. (2009). Opportunity to learn. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 541–559). New York: Rutledge for American Educational Research Association.

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H. A., Wiley, D. E., Cogan, L. S., & Wolfe, R. G. (2001). *Why schools matter: A cross-national comparison of curriculum and learning.* San Francisco: Jossey-Bass.

Schmidt, W. H., McKnight, C., Valverde, G. A., Houang, R. T., & Wiley, D. E. (1997*). Many visions, many aims, Volume 1: A cross-national investigation of curricular intentions in school mathematics.* Dordrecht, Netherlands: Kluwer.

Schmidt, W. H., Wang, H. C., & McKnight, C. C. (2005). Curriculum coherence: An examination of U.S. mathematics and science content standards from an international perspective. *Journal of Curriculum Studies*, *35*, 528–529.

Stancavage, F., Shepard, L., McLaughlin, D., Holtzman, D., Blankenship, C., & Zhang, Y. (2009). *Sensitivity of NAEP to the effects of reform-based teaching and learning in middle school mathematics.* A publication of the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

Stein, M. K., Kaufman, J. H., Sherman, M., & Hillen, A. F. (2011). Algebra: A challenge at the crossroads of policy and practice. *Review of Educational Research, 81*(4), 453–492.

Webb, N. L. (2002). *An analysis of the alignment between mathematics standards and assessments for three states.* Paper presented at the American Educational Research Association Annual Meeting, New Orleans, LA.

Webb, N. L. (2005). *Alignment analysis of mathematics standards and assessments.* Unpublished report.

Appendix A. Features of the NAEP Mathematics Framework and the CCSS-M

NAEP Mathematics Framework: An Assessment Framework

The National Assessment Governing Board oversees the development of the NAEP Mathematics Framework—a framework that describes the specific knowledge and skills to be assessed in each NAEP content area and grade level. The various stakeholders to whom NAEP results are made available are the same stakeholders who provide input on the framework: content experts, school personnel, teachers, parents, policymakers, and others.

The mathematics knowledge and skills that are assessed in NAEP must necessarily take into account the constraints of a large-scale assessment, with its limitations on time, space, and resources. In practical terms, this means that the frameworks are developed with the understanding that some concepts, skills, and activities in mathematics as taught are not suitable to be assessed by NAEP even though they may be very important components of a school curriculum.

The Grade 4 and Grade 8 objectives in the *2011 Mathematics Framework* were used as the basis for making comparisons with the CCSS-M in the current study and have served as the basis for the NAEP assessment at these grade levels since 2005. These are the same Grade 4 and Grade 8 objectives that are in the *2013 Mathematics Framework*. Therefore, the results of the analyses about the alignment of the 2011 mathematics framework with the CCSS-M are applicable to the 2013 mathematics framework as well.

The NAEP Mathematics Framework is organized into five broad areas of mathematics content:

- **Number Properties and Operations (NPO),** including computation and understanding of number concepts
- **Measurement (M),** including use of instruments, application of processes, and concepts of area and volume
- **Geometry (G),** including spatial reasoning and applying geometric properties
- **Data Analysis, Statistics, and Probability (DASP)**, including graphical displays and statistics
- **Algebra (A),** including representations and relationships

Each content area is divided into subtopics, and each subtopic consists of one or more objectives. These divisions are not intended to separate mathematics into discrete, nonoverlapping elements. Rather, they are intended to provide a helpful classification scheme that describes the universe of mathematical content assessed by NAEP.

Number Properties and Operations measures students' understanding of ways to represent, calculate, and estimate with numbers. It consists of the following subtopics: number sense, estimation, number operations, ratios and proportional reasoning, properties of number operations, and mathematical reasoning using numbers. At Grade 4, objectives cover number properties and operations and focus on computation with or understanding of whole numbers and common fractions and decimals. At Grade 8, students are expected to compute with rational and common irrational numbers as well as solve problems using proportional reasoning and apply properties of select number systems.

Measurement assesses students' knowledge of units of measurement for such attributes as capacity, length, area, volume, time, angles, and rates. It consists of the following subtopics: measuring physical attributes, systems of measurement, and measurement in triangles. At Grade 4, objectives focus on customary units, such as inch, quart, pound, and hour, and common metric units, such as centimeter, liter, and gram. Length as a geometric attribute also is addressed. At Grade 8, students are expected to know how to use rates and square units for measuring area and surface area, cubic units for measuring volume, and degrees for measuring angles.

Geometry measures students' knowledge and understanding of shapes in two and three dimensions and relationships between shapes, such as symmetry and transformations. It consists of the following subtopics: dimension and shape; transformation of shapes and preservation of properties; relationships between geometric figures; position, direction, and coordinate geometry; and mathematical reasoning in geometry. At Grade 4, objectives focus on simple figures such as cubes and spheres. At Grade 8, the focus is on properties of plane figures, especially parallel and perpendicular lines, angle relationships in polygons, cross-sections of solids, and the Pythagorean Theorem.

Data Analysis, Statistics, and Probability consists of the following subtopics: data representation, characteristics of data sets, experiments and samples, and probability. At Grade 4, objectives focus on how data are collected and organized, how to read and interpret various representations of data, and basic concepts of probability. At Grade 8, the student is expected to know how to organize and summarize data in various formats, such as tables, charts, and graphs; analyze statistical claims; and solve problems involving probability.

Algebra measures students' understanding of patterns, using variables, algebraic representation, and functions. At Grade 4, objectives focus on students' understanding of algebraic representation, patterns, and rules; graphing points on a line or a grid; and using symbols to represent unknown quantities. At Grade 8, the focus is on understanding patterns and functions; algebraic expressions, equations, and inequalities; and algebraic representations, including graphs.

Levels of Complexity in the Framework

In addition to the content dimension of the objectives of the NAEP framework, there is a complexity dimension that classifies items into three levels of complexity: (1) low, (2) moderate, and (3) high.

The objectives that generate *low-complexity* items usually are statements of recall and recognition of previously learned concepts and principles. The following statements are typical of the demands of objectives that might lead to low-complexity items:

- Recall or recognize a fact, term, or property
- Recognize an example of a concept
- Compute a sum, difference, product, or quotient
- Evaluate an expression in an equation
- Solve a one-step problem
- Draw or measure a simple geometric figure

The objectives that generate *moderate-complexity* items involve more flexibility of thinking as well as informal methods of reasoning and problem solving. These objectives bring together skills and knowledge from various content areas. The following statements are typical of what may lead to moderate-complexity items:

- Solve a word problem using multiple steps
- Provide justification for steps in a solution process
- Extend a pattern
- Retrieve information from a graph, table, or figure and use it to solve a problem

High-complexity items are generated from statements that require more abstract thinking, planning, analysis, and creative thought. The following are examples of statements of objectives that may generate high-complexity items:

- Perform a procedure with multiple decision points
- Generate a pattern
- Formulate an original problem, given a scenario
- Describe, compare, and contrast solution methods
- Analyze the assumptions of a mathematical model
- Solve a novel problem

The final form of the assessment depends on the assessment blueprint or test specifications. These define how the content and the levels of complexity of the items are to be distributed.

CCSS-M: A Curriculum Framework

The CCSS-M framework consists of two components: the Standards for Mathematical Content and the Standards for Mathematical Practice. The two components operate in concert to provide school mathematics experiences that, according to the authors, are "substantially more focused and coherent in order to improve mathematics achievement …" in the United States. The CCSS-M set grade-specific content standards for Grades K–8 and subject-specific standards for high school. The grade-level standards are organized into clusters and content domains. The standards define what students should understand and be able to do, clusters are groups of related standards, and content domains are larger groups of related clusters. The following is an example showing the organization of one cluster in the Geometry content domain for Grade 4.

Geometry: 4.G ***(CCSS-M—Content Domain)***

A. Draw and identify lines and angles, and classify shapes by properties of their lines and angles*. (Cluster)*

4.G.A.1. Draw points, lines, line segments, rays, angles (right, acute, obtuse), and perpendicular and parallel lines. Identify these in two-dimensional figures. *(Standard)*

4.G.A.2. Classify two-dimensional figures based on the presence or absence of parallel or perpendicular lines, or the presence or absence of angles of a specified size. Recognize right triangles as a category, and identify right triangles. *(Standard)*

4.G.A.3. Recognize a line of symmetry for a two-dimensional figure as a line across the figure such that the figure can be folded along the line into matching parts. Identify line-symmetric figures and draw lines of symmetry. *(Standard)*

The organization of the CCSS-M hierarchy is very similar to the organization of the NAEP Mathematics Framework. For NAEP Grades 4, 8, and 12, groups of objectives form subtopics, and groups of subtopics are subsumed under content areas. For example,

Geometry: Grade 4 ***(NAEP—Content Area)***

Dimension and shape *(Subtopic)*

a. Explore properties of paths between points. *(Objective)*
b. Identify or describe (informally) real-world objects using simple plane figures (e.g., triangles, rectangles, squares, and circles) and simple solid figures (e.g., cubes, spheres, and cylinders.) *(Objective)*
c. Identify or draw angles and other geometric figures in the plane. *(Objective)*

Items for the NAEP assessments are constructed from the statements of the objectives. Items for the CCSS-M assessments will be constructed from statements at the content standards level in concert with the appropriate Standards for Mathematical Practice.

The Standards for Mathematical Practice describe different types of practices and habits of mind that mathematics educators at all levels should seek to develop in their students. These practices and mindsets are not new to the mathematics education community. They are based on important processes and proficiencies embedded in the core work of the National Council of Teachers of Mathematics (NCTM) and the National Research Council (NRC), respectively. The "processes" are based on the NCTM's process standards of problem solving, reasoning and proof, communication, representation, and connections. The "proficiencies," or habits of mind, are based on the mathematical proficiencies described in the National Research Council's report, *Adding It Up* (Kilpatrick, Swafford, & Findell, 2001): adaptive reasoning, strategic competence, conceptual understanding, procedural fluency, and productive disposition. Unlike the content standards for mathematics, which are grade-specific, the eight Standards for Mathematical Practice are consistent at and apply to all grades (kindergarten through high school). Because they play an important role in decisions about the level of the alignment of the NAEP objectives with the CCSS-M content standards and in the design of assessment items based on the CCSS-M, a brief description of each of the Standards for Mathematical Practice (SMP) is provided below.

1. **Make sense of problems and persevere in solving them (SMP1).** Mathematically proficient students check their answers to problems using different methods, and they continually ask themselves, "Does this make sense?"
2. **Reason abstractly and quantitatively (SMP2).** Mathematically proficient students bring two complementary abilities to bear on problems involving quantitative relationships: the ability to decontextualize and the ability to contextualize. The ability to decontextualize allows them to abstract a given situation, represent it symbolically, and manipulate the representing symbols as if they have a life of their own. On the other hand, the ability to contextualize allows students to add meaning to the symbols involved. Quantitative reasoning involves creating a coherent representation of a problem, knowing and flexibly using different and appropriate properties of operations, and computing them accurately.
3. **Construct viable arguments and critique the reasoning of others (SMP3).** Mathematically proficient students make conjectures and build a logical progression of statements to explore the truth of their conjectures. They justify their conclusions, communicate them to others, and respond to the arguments of others. If there is a flaw in an argument, they can explain what it is. They reason inductively about data, making plausible arguments that take into account the context from which the data arose.
4. **Model with mathematics (SMP4).** Mathematically proficient students can apply the mathematics they know to solve problems arising in everyday life, society, and the workplace. In the early grades, this may involve writing an addition equation to describe a situation. In the middle grades, a student might apply proportional reasoning to plan a school event or analyze a problem in the community. In high school, a student might use geometry to solve design problems or use a function to describe how one quantity of interest depends on another. They interpret their mathematical results in the context of the situation and reflect on whether the results make sense, possibly improving the model if it has not served its purpose.
5. **Use appropriate tools strategically (SMP5).** Mathematically proficient students consider the available tools when solving a mathematical problem. These tools might include pencil and paper, concrete models, a ruler, a protractor, a calculator, a spreadsheet, a computer algebra system, a statistical package, or dynamic geometry software. Proficient students are sufficiently familiar with tools appropriate for their grade or course to make sound decisions

about when each of these tools might be helpful, recognizing both the insight to be gained and their limitations.

6. **Attend to precision (SMP6).** Mathematically proficient students try to communicate precisely to one another. They are careful about specifying units of measure and labeling axes to clarify the correspondence with quantities in a problem. They calculate accurately and efficiently, expressing numerical answers with a degree of precision appropriate for the problem context.
7. **Look for and make sure of structure (SMP7).** Mathematically proficient students look closely to discern a pattern or structure. In the early grades, students might notice that 3 + 7 is the same as 7 + 3, or they may sort a collection of shapes according to how many sides the shapes have. Later, they may recognize that 7 × 8 equals 7 × (5 + 3), which is the same as 7 × 5 + 7 × 3.
8. **Look for and express regularity in repeated reasoning (SMP8).** Mathematically proficient students notice if calculations are repeated and look for both general methods and mathematically sound shortcuts. They continually evaluate the reasonableness of their intermediate results.

The authors of the CCSS-M suggest that designers of curricula, professional development, instruction, and assessments should attend to the need to connect mathematical practices to mathematical content. Expectations in content standards that begin with the word "understand" are often good opportunities to connect to the practices of the content. For example, CCSS-M standard 4.NF.B.3b (a Grade 4 standard in the content domain of Number and Operations: Fractions) states that students will:

> "Understand a fraction a/b with a > 1 as the sum of fractions 1/b and decompose a fraction into a sum of fractions with the same denominator in more than one way, recording each decomposition by an equation. Justify decompositions, e.g., by using a visual fractional model."

According to this standard, students are expected to extend previous understandings about how fractions are built, composed, and decomposed from unit fractions. In addition, students are expected to use the meaning of fractions and multiplication to multiply a fraction by a whole number (e.g., 3/8 = 3 × 1/8 = (1/8 + 1/8 + 1/8). Evident in this content standard and the accompanying example are at least three of the Standards for Mathematical Practice: SMP2, SMP4, and SMP7.

Appendix B. Coverage of Grade 4 NAEP Mathematics Objectives in the CCSS-M

Table B-1. Coverage of Grade 4 NAEP Mathematics Objectives in the CCSS-M[1,2]

NAEP content area: Number properties and operations

NAEP subtopic: (1) Number sense			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
4NPO1a	(a) Identify place value and actual value of digits in whole numbers.	2.NBT.A.1, 4.NBT.A.1	
4NPO1b	(b) Represent numbers using models such as base 10 representations, number lines, and two-dimensional models.	2.MD.B.6	Exception: Two-dimensional models for representing numbers are not covered.
4NPO1c	(c) Compose or decompose whole quantities by place value (e.g., write whole numbers in expanded notation using place value: 342 = 300 + 40 + 2).	2.NBT.A.1, 4.NBT.A.2	
4NPO1d	(d) Write or rename whole numbers (e.g., 10 = 5 + 5, 12 – 2, or 2 × 5).	2.NBT.A.3	
4NPO1e	(e) Connect model, number word, or number using various models and representations for whole numbers, fractions, and decimals.	2.NBT.A.3, 2.MD.B.6, 2.G.A.2, 2.G.A.3, 3.NF.A.2, 4.NBT.A.2, 5.NBT.A.3a	
4NPO1i	(i) Order or compare whole numbers, decimals, or fractions.	2.NBT.A.4, 4.NBT.A.2, 5.NBT.A.3b	
NAEP subtopic: (2) Estimation			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
4NPO2a	(a) Use benchmarks (well-known numbers used as meaningful points for comparison) for whole numbers, decimals, or fractions in contexts (e.g., ½ and .5 may be used as benchmarks for fractions and decimals between 0 and 1.00).		
4NPO2b	(b) Make estimates appropriate to a given situation with whole numbers, fractions, or decimals by • Knowing when to estimate, • Selecting the appropriate type of estimate, including overestimate, underestimate, and range of estimate, or • Selecting the appropriate method of estimation (e.g., rounding).	2.MD.A.3, 4.NBT.A.3	

4NPO2c	(c) Verify solutions or determine the reasonableness of results in meaningful contexts.	4.OA.A.3, 6.EE.B.5	Also covered in the Standards for Mathematical Practice.

NAEP subtopic: (3) Number operations

NAEP objective ID	NAEP objective	Where taught in the CCSS-M?	Comments
4NPO3a	(a) Add and subtract: • Whole numbers, or • Fractions with like denominators, or • Decimals through hundredths.	2.OA.A.1, 2.OA.B.2, 2.NBT.B.6, 2.NBT.B.7, 3.NBT.A.2, 4.NF.B.3c, 5.NBT.B.7	
4NPO3b	(b) Multiply whole numbers: • No larger than two-digit by two-digit with paper-and-pencil computation, or • Larger numbers with use of Calculator	3.NBT.A.3, 5.NBT.B.5	Use of calculators is not mentioned in the CCSS-M.
4NPO3c	(c) Divide whole numbers: • Up to three digits by one digit with paper-and-pencil computation, or • Up to five digits by two digits with use of calculator.	3.OA.C.7, 4.NBT.B.6, 5.NBT.B.6	
4NPO3d	(d) Describe the effect of operations on size (whole numbers).	3.NF.A.2a–2b, 3.NF.A.3, 5.NBT.A.2	The match of CCSS-M standards with this objective is more indirect than direct.
4NPO3e	(e) Interpret whole-number operations and the relationships between them.	3.OA.A.1, 3.OA.A.2, 3.OA.B.6	
4NPO3f	(f) Solve application problems involving numbers and operations.	1.OA.D.8, 2.OA.A.1, 2.MD.C.8, 3.OA.A.3, 3.OA.D.8, 3.MD.C.7b, 3.MD.D8, 4.OA.A.2, 4.OA.A.3, 4.NF.B.4c, 4.MD.A.2, 4.MD.C.7, 5.NF.A.2, 5.NF.B.3, 5.NF.B.6, 5.NF.B.7c	

NAEP subtopic: (4) Ratios and proportional reasoning

NAEP objective ID	NAEP objective	Where taught in the CCSS-M?	Comments
4NPO4a	(a) Use simple ratios to describe problem situations.	5.NF.B.3, 6.RP.A.1	

NAEP subtopic: (5) Properties of numbers and operations

NAEP objective ID	NAEP objective	Where taught in the CCSS-M?	Comments
4NPO5a	(a) Identify odd and even numbers.	2.OA.C.3	
4NPO5b	(b) Identify factors of whole numbers.	3.OA.A.4, 4.OA.B.4	
4NPO5e	(e) Apply basic properties of operations.	3.OA.A.4, 4.OA.B.4	

NAEP subtopic: (6) Mathematical reasoning using numbers

NAEP objective ID	NAEP objective	Where taught in the CCSS-M?	Comments
4NPO6a	(a) Explain or justify a mathematical concept or relationship (e.g., explain why 15 is an odd number or why 7–3 is not the same as 3–7).	2.NBT.B.9, 3.NFA.3, 4.NF.A.1, 5.NBT.A.2, 5.NF.B.5b	Also covered in the Standards for Mathematical Practice.

NAEP content area: Algebra

NAEP subtopic: (1) Patterns, relations, and functions

NAEP objective ID	NAEP objective	Where taught in the CCSS-M?	Comments
4A1a	(a) Recognize, describe, or extend numerical patterns.	3.OA.D.9	
4A1b	(b) Given a pattern or sequence, construct or explain a rule that can generate the terms of the pattern or sequence.	4.OA.C.5, 5.OA.B.3, 5.NBT.A.2	
4A1c	(c) Given a description, extend or find a missing term in a pattern or sequence.	5.OA.B.3	
4A1d	(d) Create a different representation of a pattern or sequence given a verbal description.		Not found in the CCSS-M
4A1e	(e) Recognize or describe a relationship in which quantities change proportionally.	7.RP.A.2a	

NAEP subtopic: (2) Algebraic representations

NAEP objective ID	NAEP objective	**Where taught in the CCSS-M?**	**Comments**
4A2a	(a) Translate between the different forms of representations (symbolic, numerical, verbal, or pictorial) of whole-number relationships (such as from a written description to an equation or from a function table to a written description).	8.F.A.2	The content in this standard may be more than what is expected at fourth grade.
4A2c	(c) Graph or interpret points with whole-number or letter coordinates on grids or in the first quadrant of the coordinate plane.	5.G.A.1, 5.G.A.2	

NAEP subtopic: (3) Variables, expressions, and operations

NAEP objective ID	NAEP objective	Where taught in the CCSS-M?	Comments
4A3a	(a) Use letters and symbols to represent an unknown quantity in a simple mathematical expression.	3.OA.A.3, 6.EE.A.2a, 6.EE.C.9	
4A3b	(b) Express simple mathematical relationships using number sentences.	6.EE.A.2b	

NAEP subtopic: (4) Equations and inequalities			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
4A4a	(a) Find the value of the unknown in a whole-number sentence.	3.OA.A.4	
NAEP subtopic: (5) Mathematical reasoning in algebra			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
4A5a	(a) Verify a conclusion using algebraic properties.	Taught throughout the CCSS-M content standards.	Also covered in the Standards for Mathematical Practice.

NAEP content area: Measurement

NAEP subtopic: (1) Measuring physical attributes			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
4M1a	(a) Identify the attribute that is appropriate to measure in a given situation.		Not found in the CCSS-M.
4M1b	(b) Compare objects with respect to a given attribute, such as length, area, volume, time, or temperature.	2.MD.A.2, 2.MD.A.4	
4M1c	(c) Estimate the size of an object with respect to a given measurement attribute (e.g., length, perimeter, or area using a grid).	2.MD.A.3, 3.MD.A.2	
4M1e	(e) Select or use appropriate measurement instruments, such as a ruler, meter stick, clock, thermometer, or other scaled instruments.	2.MD.A.1	
4M1f	(f) Solve problems involving perimeter of plane figures.	3.MD.D.8, 4.MD.A.3	
4M1g	(g) Solve problems involving area of squares and rectangles.	3.MD.C.5, 3.MD.C.6, 3.MD.C.7, 4.MD.A.3, 5.NF.B.4b	
NAEP subtopic: (2) Systems of measurement			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
4M2a	(a) Select or use an appropriate type of unit for the attribute being measured, such as length, time, or temperature.	2.MD.A.1	
4M2b	(b) Solve problems involving conversions within the same measurement system, such as conversions involving inches and feet or hours and minutes.	4.MD.A.1, 5.MD.A.1	
4M2d	(d) Determine appropriate size of unit of measurement in problem situation involving such attributes as length, time, capacity, or weight.	4.MD.A.1	
4M2e	(e) Determine situations in which a highly accurate measurement is important.		May be covered in the Standards for Mathematical Practice.

NAEP content area: Geometry

NAEP subtopic: (1) Dimension and shape

NAEP objective ID	NAEP objective	Where taught in the CCSS-M?	Comments
4G1a	(a) Explore properties of paths between points.	4.G.A.1	
4G1b	(b) Identify or describe (informally) real-world objects using simple plane figures (e.g., triangles, rectangles, squares, and circles) and simple solid figures (e.g., cubes, spheres, and cylinders).	K.G.A.1, 2.G.A.1,	
4G1c	(c) Identify or draw angles and other geometric figures in the plane.	4.G.A.1	
4G1f	(f) Describe attributes of two- and three-dimensional shapes.	K.G.A.3, K.G.B.4, 3.G.A.1	

NAEP subtopic: (2) Transformation of shapes and preservation of properties

NAEP objective ID	NAEP objective	Where taught in the CCSS-M?	Comments
4G2a	(a) Identify whether a figure is symmetrical or draw lines of symmetry.	4.G.A.3	
4G2c	(c) Identify the images resulting from flips (reflections), slides (translations), or turns (rotations).	8.G.A.3, 8.G.A.4	
4G2d	(d) Recognize which attributes (such as shape and area) change or do not change when plane figures are cut up or rearranged.	4.G.A.3 (introduction)	See High School Geometry, Congruence.A.2, A.3, and B.6
4G2e	(e) Match or draw congruent figures in a given collection.	8.G.A.2	See High School Geometry, Congruence.B.7 and B.8

NAEP subtopic: (3) Relationships between geometric figures

NAEP objective ID	NAEP objective	Where taught in the CCSS-M?	Comments
4G3a	(a) Analyze or describe patterns of geometric figures by increasing number of sides, changing size or orientation (e.g., polygons with more and more sides).	3.G.A.1, 5.G.B.3	
4G3b	(b) Assemble simple plane shapes to construct a given shape.	1.G.A.2	
4G3c	(c) Recognize two-dimensional faces of three-dimensional shapes.	7.G.A.2	
4G3f	(f) Describe and compare properties of simple and compound figures composed of triangles, squares, and rectangles.	1.G.A.2, 2.G.A.1, 4.G.A.2	

NAEP subtopic: (4) Position, direction, and coordinate geometry

NAEP objective ID	NAEP objective	Where taught in the CCSS-M?	Comments
4G4a	(a) Describe relative positions of points and lines using the geometric ideas of parallelism or perpendicularity.	4.G.A.1, 4.G.A.2, 5.G.A.1, 8.G.A.1c	
4G4d	(d) Construct geometric figures with vertices at points on a coordinate grid.	6.G.A.3	

NAEP subtopic: (5) Mathematical reasoning in geometry			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
4G5a	(a) Distinguish which objects in a collection satisfy a given geometric definition and explain choices.	5.G.B.3, 5.G.B.4	

NAEP content area: Data Analysis, Statistics, and Probability

NAEP subtopic: (1) Data representation			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
	Pictographs, bar graphs, circle graphs, line graphs, line plots, tables, and tallies.		
4DASP1a	(a) Read or interpret a single set of data.	6.SP.A.2	
4DASP1b	(b) For a given set of data, complete a graph (limits of time make it difficult to construct graphs completely).	5.MD.B.2, 6.SP.B.4	
4DASP1c	(c) Solve problems by estimating and computing within a single set of data.	6.SP.B.5	
NAEP subtopic: (2) Characteristics of data sets			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
4DASP2b	(b) Given a set of data or a graph, describe the distribution of data using median, range, or mode.	6.SP.A.2, 6.SP.B.5c	
4DASP2d	(d) Compare two sets of related data.	7.SP.B.3, 7.SP.B.4	
NAEP subtopic: (3) Probability			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
4DASP4a	(a) Use informal probabilistic thinking to describe chance events (i.e., likely and unlikely, certain and impossible).	7.SP.C.5	
4DASP4b	(b) Determine a simple probability from a context that includes a picture.	7.SP.C.6, 7.SP.C.7	
4DASP4e	(e) List all possible outcomes of a given situation or event.	7.SP.C.7	
4DASP4g	(g) Represent the probability of a given outcome using a picture or other graphic.	7.SP.C.6	

[1] *Notation for CCSS-M standards*: Grade level, content domain, cluster, standard number within domain. For example, 3.OA.D.8 is read as Grade 3, Operations and Algebraic Thinking, Cluster D, Standard 8.

[2] *Notation for NAEP objectives:* Grade level, content area, subtopic, objective. For example, 4NPO1i is read as Grade 4, Number Properties and Operations, Subtopic 1, Objective i.

Appendix C. Coverage of Grade 8 NAEP Mathematics Objectives in the CCSS-M

Table C-1. Coverage of Grade 8 NAEP Mathematics Objectives in the CCSS-M[1,2]

NAEP content area: Number properties and operations

NAEP subtopic: (1) Number Sense			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8NPO1a	(a) Use place value to model and describe integers and decimals.	5.NBT.A.1-3, 5.NBT.B.6-7	Place value of decimals is covered in Grade 5, but not mentioned beyond Grade 5. Negative integers are introduced in Grade 6.
8NPO1b	(b) Model or describe rational numbers or numerical relationships using number lines and diagrams.	3.NF.A.2, 4.NF.B.4, 5.NF.B.6, 6NS.A.1, 7.RP.A	Modeling is covered in the Standards for Mathematical Practice and the high school mathematics standards.
8NPO1d	(d) Write or rename rational numbers.	3.NF.A.3, 4.NF.A, 4.NF.B, 5.NBT.A.3, 6.NS.C,7.RP.A, 7.NS.A, 8.NS.A	
8NPO1e	(e) Recognize, translate, or apply multiple representations of rational numbers (fractions, decimals, and percents) in meaningful contexts.	6.NS.C, 7.RP, 7.NS.A, 8.NS.A	
8NPO1f	(f) Express or interpret numbers using scientific notation from real-life contexts.	8.EE.A.3, 8.EE.A.4	
8NPO1g	(g) Find or model absolute value or apply to problem situations.	6.NS.C.7, 7.NS.A.1c	
8NPO1h	(h) Order or compare rational numbers (fractions, decimals, percents, or integers) using various models and representations (e.g., number line).	3.NF., 4.NF.A., 4.NF.B, 4.NF.C 5.NBT.A.3b, 5.NF.B.5a, 6.NS.C	
8NPO1i	(i) Order or compare rational numbers including very large and small integers, and decimals and fractions close to zero.		This objective is very similar to 8NPO1h; however, the CCSS-M do not address "very large and small integers" or "decimals and fractions close to zero."

NAEP subtopic: (2) Estimation			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8NPO2a	(a) Establish or apply benchmarks for rational numbers and common irrational numbers (e.g., π) in contexts.	8.EE.A.2	

8NPO2b	(b) Make estimates appropriate to a given situation by: • Identifying when estimation is appropriate, • Determining the level of accuracy needed, • Selecting the appropriate method of estimation, or • Analyzing the effect of an estimation method on the accuracy of results.		Covered in the Standards for Mathematical Practice.
8NPO2c	(c) Verify solutions or determine the reasonableness of results in a variety of situations, including calculator and computer results.	4.OA.A.3, 6.EE.B.5;	Also covered in the Standards for Mathematical Practice. Use of calculator or computer is not specifically addressed in the CCSS-M.
8NPO2d	(d) Estimate square or cube roots of numbers less than 1,000 between two whole numbers.	8.NS.A.2	
NAEP subtopic: (3) Number Operations			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8NPO3a	(a) Perform computations with rational numbers.	3.OA, 3.NBT, 3.NF, 4.OA, 4.NBT, 4.NF, 5.NBT.B.5, 5.NBT.B.6, 5.NBT.B.7, 5.NF, 6.RP.A, 6.NS.B.2, 6.NS.B.3, 7.RP.A, 7.NS.A, 7.EE.A, 8.EE.A.4	
8NPO3d	(d) Describe the effect of multiplying and dividing by numbers, including the effect of multiplying or dividing a rational number by: • Zero, or • A number less than zero, or • A number between zero and one, • One, or • A number greater than one.	3.OA.C.7, 3.NBT.A.3, 4.NF.A, 4.NF.B, 5.NF.B.3, 5.NF.B.4, 5.NF.B.5, 5.NF.B.7 6.NS.A.1, 7.NS.A.2b	
8NPO3e	(e) Interpret rational number operations and the relationships between them.	7.NS.A	
8NPO3f	(f) Solve application problems involving rational numbers and operations using exact answers or estimates as appropriate.	7.RP.A, 7.NS.A, 7.EE.B.3	
NAEP subtopic: (4) Ratios and proportional reasoning			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8NPO4a	(a) Use ratios to describe problem situations.	6.RP.A, 7.RP.A	

8NPO4b	(b) Use fractions to represent and express ratios and proportions.	6.RP.A, 7.RP.A	
8NPO4c	(c) Use proportional reasoning to model and solve problems (including rates and scaling).	7.RP.A.1, 7.RP.A.2, 7.RP.A.3, 8.EE.B.5	
8NPO4d	(d) Solve problems involving percentages (including percent increase and decrease, interest rates, tax, discount, tips, or part/whole relationships).	6.RP.A, 7.RP.A.3, 7.EE.A	
NAEP subtopic: (5) Properties of number and operations			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8NPO5a	(a) Describe odd and even integers and how they behave under different operations.	Inferred in 4.OA.C.5	
8NPO5b	(b) Recognize, find, or use factors, multiples, or prime factorization.	4.OA.A.1, 4.OA.B.4, 4.NF.B.4, 6.NS.B.4	Factors and multiples are covered, but prime factorization is not.
8NPO5c	(c) Recognize or use prime and composite numbers to solve problems.	4.OA.B.4	No reference to using prime numbers to solve problems
8NPO5d	(d) Use divisibility or remainders in problem settings.	4.OA.A.3	Remainders are mentioned; however, divisibility is not specifically covered in the CCSS-M.
8NPO5e	(e) Apply basic properties of operations.	5.NF.B.4, 5.NF.B.7, 6.NS.A, 6.NS.B, 7.NS.A	
NAEP subtopic: (6) Mathematical reasoning using numbers			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8NPO6a	(a) Explain or justify a mathematical concept or relationship (e.g., explain why 17 is prime).		Covered in the Standards for Mathematical Practice.
8NPO6b	(b) Provide a mathematical argument to explain operations with two or more fractions.	5.NF.A, 5.NF.B, 6.NS.A.1,	Also covered in the Standards for Mathematical Practice.

NAEP content area: Algebra

NAEP subtopic: (1) Patterns, Relations, and Functions			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8A1a	(a) Recognize, describe, or extend numerical and geometric patterns using tables, graphs, words, or symbols.	4.OA.C.5, 8.SP.A.4	
8A1b	(b) Generalize a pattern appearing in a numerical sequence, table, or graph using words or symbols.	4.OA.C.5, 8.SP.A.4	

8A1c	(c) Analyze or create patterns, sequences, or linear functions given a rule.	8.F.B.4	
8A1e	(e) Identify functions as linear or nonlinear or contrast distinguishing properties of functions from tables, graphs, or equations.	8.F.A.2	
8A1f	(f) Interpret the meaning of slope or intercepts in linear functions.	8.EE.B.5, 8.F.A.3	Applied in Modeling in high school.
NAEP subtopic: (2) Algebraic representations			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8A2a	(a) Translate between different representations of linear expressions using symbols, graphs, tables, diagrams, or written descriptions.	8.F.A.2	
8A2b	(b) Analyze or interpret linear relationships expressed in symbols, graphs, tables, diagrams, or written descriptions.	8.F.A.3	
8A2c	(c) Graph or interpret points represented by ordered pairs of numbers on a rectangular coordinate system.	6.NS.C.6b, 6.NS.C.6c, 7.RP.A.2a	
8A2d	(d) Solve problems involving coordinate pairs on the rectangular coordinate system.	6.NS.C.8	Further developed in High School Geometry: HSG.B.7
8A2f	(f) Identify or represent functional relationships in meaningful contexts, including proportional, linear, and common nonlinear (e.g., compound interest, bacterial growth) in tables, graphs, words, or symbols.	8.EE.B.5, 8.F.B.5	Further developed in High School Functions and Modeling.
NAEP subtopic: (3) Variables, expressions, and operations			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8A3b	(b) Write algebraic expressions, equations, or inequalities to represent a situation.	6.EE.A.2, 6.EE.B.6-8, 7.EE.A.2	
8A3c	(c) Perform basic operations, using appropriate tools, on linear algebraic expressions (including grouping and order of multiple operations involving basic operations, exponents, roots, simplifying, and expanding).	6.EE.A, 7.EE.A, 8.EE.A.2-4	
NAEP subtopic: (4) Equations and inequalities			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8A4a	(a) Solve linear equations or inequalities (e.g., $ax + b = c$ or $ax + b = cx + d$ or $ax + b > c$).	6.EE.B, 7.EE.B, 8.EE.C	
8A4b	(b) Interpret "=" as an equivalence between two expressions and use this interpretation to solve problems.	1.OA.D.7, 6.EE.B, 7.EE.B	Notation of equivalence introduced in Grade 1.

8A4c	(c) Analyze situations or solve problems using linear equations and inequalities with rational coefficients symbolically or graphically (e.g., ax + b = c or ax + b = cx + d).	6.EE.B.7, 6.EE.B.8, 6.EEB.9, 6.G.A, 7.EE.B.4, 8.EE.C.7	
8A4d	(d) Interpret relationships between symbolic linear expressions and graphs of lines by identifying and computing slope and intercepts (e.g., know in y = ax + b, that a is the rate of change and b is the vertical intercept of the graph).	8.EE.B	Also covered in the Standards for Mathematical Practice.
8A4e	(e) Use and evaluate common formulas (e.g., relationship between a circle's circumference and diameter [C = πd], distance, and time under constant speed).	5.MD.C.5b, 6.EE.C.9, 6.G.A.2, 7.G.B.4	

NAEP subtopic: (5) Mathematical reasoning in algebra

NAEP objective ID	NAEP objective	Where taught in the CCSS-M?	Comments
8A5a	(a) Make, validate, and justify conclusions and generalizations about linear relationships.	6.EE.B.5, 8.EE.B	Also covered in the Standards for Mathematical Practice.

NAEP content area: Measurement

NAEP subtopic: (1) Measuring physical attributes

NAEP objective ID	NAEP objective	Where taught in the CCSS-M?	Comments
8M1b	(b) Compare objects with respect to length, area, volume, angle measurement, weight, or mass.	2.MD.A.4	
8M1c	(c) Estimate the size of an object with respect to a given measurement attribute (e.g., area).	2.MD.A.2, 3.MD.A.2	
8M1e	(e) Select or use appropriate measurement instrument to determine or create a given length, area, column, angle, weight, or mass.	2.MD.A.1, 3.MD.C.5, 3.MD.C.6	
8M1f	(f) Solve mathematical or real-world problems involving perimeter or area of plane figures, such as triangles, rectangles, circles, or composite figures.	3.MD.D.8, 4.MD.A.3, 6.G.A.1, 7.G.B.4	
8M1h	(h) Solve problems involving volume or surface area of rectangular solids, cylinders, prisms, or composite shapes.	5.MD.C.3, 5MD.C.4, 5.MD.C.5, 6.G.A.2, 7.G.B.6,	
8M1i	(i) Solve problems involving rates such as speed or population density.	6.RP.A.2, 6.RP.A.3b, 7.RP.A.2b	Concepts of density, including population density, do not appear in the CCSS-M until high school.

NAEP subtopic: (2) Systems of measurement			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8M2a	(a) Select or use an appropriate type of unit for the attribute being measured, such as length, area, angle, time, or volume.	4.MD.A.1	
8M2b	(b) Solve problems involving conversions within the same measurement system, such as conversions involving square inches and square feet.	4.MD.A.1, 6.RP.A.3d,	
8M2c	(c) Estimate the measure of an object in one system given the measure of that object in another system and the approximate conversion factor. For example: • Distance conversion: 1 kilometer is approximately 5/8 of a mile. • Money conversion: U.S. dollars to Canadian dollars. • Temperature conversion: Fahrenheit to Celsius.	4.MD.A.1, 6.RP.A.3d	Covered in the Standards for Mathematical Practice.
8M2d	(d) Determine appropriate size of unit of measurement in problem situation involving such attributes as length, area, or volume.		Covered in the Standards for Mathematical Practice.
8M2e	(e) Determine appropriate accuracy of measurement in problem situations (e.g., the accuracy of each of several lengths needed to obtain a specified accuracy of a total length) and find the measure to that degree of accuracy.		Covered in the Standards for Mathematical Practice.
NAEP subtopic: (3) Measurement in triangles			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8M3a	(a) Solve problems involving indirect measurement, such as finding the height of a building by comparing its shadow with the height and shadow of a known object.	7.G.A.1, 7.G.B.6, 8.G.A.4	

NAEP content area: Geometry

NAEP subtopic: (1) Dimension and shape			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8G1a	(a) Draw or describe a path of shortest length between points to solve problems in context.	Context application in 8.G.B.8, Pythagorean Theorem.	
8G1b	(b) Identify a geometric object given a written description of its properties.	3.G.A.1, 4.GA.2, 5.G.B.3	

8G1c	(c) Identify, define, or describe geometric shapes in the plane and in three-dimensional space given a visual representation.	6.G.A.4	
8G1d	(d) Draw or sketch from a written description polygons, circles, or semicircles.	6.G.A.3, 7.G.A.1, 7.G.A.2	
8G1e	(e) Represent or describe a three-dimensional situation in a two-dimensional drawing from different views.	7.G.A.3	
8G1f	(f) Demonstrate an understanding about the two- and three-dimensional shapes in our world through identifying, drawing, modeling, building, or taking apart.	6.G.A.3, 7.G.A.1, 7.G.A.2	
NAEP subtopic: (2) Transformation of shapes and preservation of properties			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8G2a	(a) Identify lines of symmetry in plane figures or recognize and classify types of symmetries of plane figures.	4.G.A.3	
8G2c	(c) Recognize or informally describe the effect of a transformation on two-dimensional geometric shapes (reflections across lines of symmetry, rotations, translations, magnifications, and contractions).	8.G.A.3	
8G2d	(d) Predict results of combining, subdividing, and changing shapes of plane figures and solids (e.g., paper folding, tiling, cutting up, and rearranging pieces).	6.G.A.1, 7.G.A.3, 7.G.B.4, 7.G.B.6	The foundational understandings are addressed in these standards. “Predicting,” per se, with respect to this objective, is not specifically evident in the CCSS-M.
8G2e	(e) Justify relationships of congruence and similarity, and apply these relationships using scaling and proportional reasoning.	8.G.A.2, 8.G.A.4	
8G2f	(f) For similar figures, identify and use the relationships of conservation of angle and proportionality of side length and perimeter.	8.G.A.4	
NAEP subtopic: (3) Relationships between geometric figures			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8G3b	(b) Apply geometric properties and relationships in solving simple problems in two and three dimensions.	6.G.A, 7.G.A	
8G3c	(c) Represent problem situations with simple geometric models to solve mathematical or real-world problems.	6.G.A, 7.G.A, 8.G.A	

8G3d	(d) Use the Pythagorean Theorem to solve problems.	8.G.B.7, 8.G.B.8	
8G3f	(f) Describe or analyze simple properties of, or relationships between, triangles, quadrilaterals, and other polygonal plane figures.	3.G.A.1, 5.G.B.3, 5.G.B.4, 6.G.A, 7.G.A, 8.G.A.2-4	
8G3g	(g) Describe or analyze properties and relationships of parallel or intersecting lines.	4.G.A.1, 4.G.A.2, 8.G.A.1c	
NAEP subtopic: (4) Position, direction, and coordinate geometry			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8G4a	(a) Describe relative positions of points and lines using the geometric ideas of midpoint, points on common line through a common point, parallelism, or perpendicularity.	High School Geometry	
8G4b	(b) Describe the intersection of two or more geometric figures in the plane (e.g., intersection of a circle and a line).	High School Geometry	
8G4c	(c) Visualize or describe the cross-section of a solid.	7.G.A.3	
8G4d	(d) Represent geometric figures using rectangular coordinates on a plane.	6.G.A.3, 8.G.A.3	
NAEP subtopic: (5) Mathematical reasoning in geometry			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8G5a	(a) Make and test a geometric conjecture about regular polygons.		Covered in the Standards for Mathematical Practice.

NAEP content area: Data Analysis, Statistics, and Probability

NAEP subtopic: (1) Data representation			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
	Histograms, line graphs, scatterplots, box plots, bar graphs, circle graphs, stem and leaf plots, frequency distributions, and tables.		
8DASP1a	(a) Read or interpret data, including interpolating or extrapolating from data.	6.SP.A.2, 7.SP.A.1, 7.SP.B.4	No mention of interpolating or extrapolating from data in the CCSS-M.
8DASP1b	(b) For a given set of data, complete a graph and then solve a problem using the data in the graph (histograms, line graphs, scatterplots, circle graphs, and bar graphs).	6.SP.B.4, 8.SP.A.1-3,	Solving problems from data in graphs is addressed in the elementary grades in Measurement and Data.

8DASP1c	(c) Solve problems by estimating and computing with data from a single set or across sets of data.	7.SP.A.2, 7.SP.B.3-4	Also covered in the Standards for Mathematical Practice.
8DASP1d	(d) Given a graph or a set of data, determine whether information is represented effectively and appropriately (histograms, line graphs, scatterplots, circle graphs, and bar graphs).		Covered in the Standards for Mathematical Practice.
8DASP1e	(e) Compare and contrast the effectiveness of different representations of the same data.	7.SP.B.3	Also covered in the Standards for Mathematical Practice.
NAEP subtopic: (2) Characteristics of data			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8DASP2a	(a) Calculate, use, or interpret mean, median, mode, or range.	6.SP.A.3, 6.SP.B.5c, 7.SP.A.2, 7.SP.B.3, 7.SP.B.4	The CCSS-M use the term “measure of center” and refer only to the mean and median.
8DASP2b	(b) Describe how mean, median, mode, range, or interquartile ranges relate to distribution shape.	6.SP.B.5d, 7.SP.B.4	
8DASP2c	(c) Identify outliers and determine their effect on mean, median, mode, or range.	6.SP.B.5c, 8.SP.A.1	
8DASP2d	(d) Using appropriate statistical measures, compare two or more data sets describing the same characteristic for two different populations or subsets of the same population.	7.SP.B.4	
8DASP2e	(e) Visually choose the line that best fits given a scatterplot and informally explain the meaning of the line. Use the line to make predictions.	8.SP.A.2	
NAEP subtopic: (3) Experiments and samples			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8DASP3a	(a) Given a sample, identify possible sources of bias in sampling.	7.SP.A.2	Bias, per se, is not mentioned in the CCSS-M.
8DASP3b	(b) Distinguish between a random and nonrandom sample.	7.SP.A.1	Coverage of nonrandom samples is inferred.
8DASP3d	(d) Evaluate the design of an experiment.		Covered in High School Statistics and Probability
NAEP subtopic: (4) Probability			
NAEP objective ID	**NAEP objective**	**Where taught in the CCSS-M?**	**Comments**
8DASP4a	(a) Analyze a situation that involves probability of an independent event.	7.SP.C.5	

8DASP4b	(b) Determine the theoretical probability of simple and compound events in familiar contexts.	7.SP.C.6	
8DASP4c	(c) Estimate the probability of simple and compound events through experimentation or simulation.	7.SP.C.8b, 7.SP.C.8c	
8DASP4d	(d) Use theoretical probability to evaluate or predict experimental outcomes.	7.SP.C.7	
8DASP4e	(e) Determine the sample space for a given situation.	7.SP.A.2	
8DASP4f	(f) Use a sample space to determine the probability of possible outcomes for an event.	7.SP.C.8a, 7.SP.C.6	
8DASP4g	(g) Represent the probability of a given outcome using fractions, decimals, and percent.	7.SP.C.5	Representation of probability using fractions is explicit in Grade 7 in the CCSS-M; representation using decimals and percent is implicit.
8DASP4h	(h) Determine the probability of independent and dependent events. (Dependent events should be limited to a small sample size.)		Covered in High School Statistics and Probability.
8DASP4j	(j) Interpret probabilities within a given context.	7.SP.C.6	

[1] *Notation for the CCSS-M:* Grade level, content domain, cluster, standard number within domain. For example, 3.OA.D.8 is read as Grade 3, Operations and Algebraic Thinking, Cluster D, Standard 8.

[2] *Notation for NAEP objectives:* Grade level, content area, subtopic, objective. For example, 4NPO1i is read as Grade 4, Number Properties and Operations, Subtopic 1, Objective i.

Appendix D. NAEP and CCSS-M Alignment Study Panel Assignments—July 2012

Elementary (Grade 4): Number Properties and Operations; Algebra

Sandra Alberti, Panel Leader
Sharon Gaines
Roger Howe
Tad Wantanabe

Elementary (Grade 4): Measurement; Geometry; Data Analysis, Statistics, and Probability

William Bush, Panel Leader
Brittany Gaines
Andy Isaacs
Norman Mattox

Secondary (Middle Grade 8): Number Properties and Operations; Algebra

Elaine Abbas
Diane Briars, Panel Leader
Jason McNeil

Secondary (Middle Grade 8): Measurement; Geometry; Data Analysis, Statistics, and Probability

Pamela Beck
Brad Findell
Carole Philip, Panel Leader

A Study of NAEP Reading and Writing Frameworks and Assessments in Relation to the Common Core State Standards in English Language Arts

Karen K. Wixson
The University of North Carolina at Greensboro

Sheila W. Valencia
University of Washington, Seattle

Sandra Murphy
University of California, Davis

Gary W. Phillips
American Institutes for Research

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

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

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

Executive Summary

Since its first assessment in 1969, the National Assessment of Educational Progress (NAEP) has made a unique contribution to our understanding of American education. It is the only national source of information on the educational achievement of U.S. students, and it is the only vehicle by which states can compare the progress of their students against a common standard. Assessment results reported by NAEP complement the states' own reports of progress under No Child Left Behind (NCLB) and track the status of achievement gaps for traditionally disadvantaged student groups.

NAEP is carried out under the guidance of the National Assessment Governing Board (Governing Board) and the National Center for Education Statistics (NCES). Throughout the course of its history, NAEP has frequently sought to improve by studying its own processes, instruments, and procedures. In keeping with this tradition, in fall 2011, NCES asked the NAEP Validity Studies (NVS) Panel, which operates under contract to NCES, to undertake two inter-related studies, one in reading/writing and one in mathematics, to examine the content of the current NAEP frameworks and item pools at Grades 4, 8, and 12 in relation to the Common Core State Standards (CCSS). The primary question under investigation is whether NAEP can continue to serve as an independent monitor of student achievement and state assessments following the implementation of the CCSS.

This report addresses the relations between the NAEP reading and writing frameworks and the CCSS in English language arts (CCSS-ELA), and the relations between the NAEP reading and writing items and the CCSS-ELA. It does not address the relations between NAEP reading and writing items and items developed by the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (Smarter Balanced) to assess the CCSS-ELA because those items were not available at the time of this study.

The report concludes with recommendations to NCES regarding broader issues on the alignment between NAEP reading and writing and CCSS-ELA, including the extent of alignment that is appropriate to support NAEP's role as an independent monitor of student achievement.

Purpose and Methods

To address the broad charge to the NVS Panel to evaluate NAEP as a potential monitor of CCSS-ELA achievement, two expert panels were convened—one for reading and one for writing. Listening and speaking were not included in the analysis because there are no NAEP assessments in these areas.

The study directors were NVS Panel members Karen Wixson and Gary Phillips, and the subject area directors were Sheila Valencia (reading) and Sandra Murphy (writing). Additional content experts with extensive knowledge and experience with NAEP and/or CCSS-ELA were invited to participate in either the reading or writing analyses.

The following comparative analyses were designed by the study directors and carried out by the expert panels, separately for reading and writing:

NAEP Frameworks to CCSS-ELA Documents

The purpose of these analyses was to determine the similarities and differences between the conceptualization and content of the NAEP reading and writing frameworks and the CCSS-ELA documents. All CCSS-ELA documents and NAEP reading and writing framework documents were analyzed using a structured qualitative protocol. This method was used to accommodate the basic differences in the purposes of CCSS-ELA and the NAEP frameworks. The CCSS-ELA documents represent a detailed framework and exemplars for what should be taught and what students should know and be able to do in K–12 in English language arts and in literacy in history/social studies, science, and technical subjects. By contrast, the NAEP documents are assessment frameworks and do not expressly seek to influence curricular decisions. These differences in purpose translate into different aspects/elements included in each. With these basic differences in mind, the analyses enumerate the similarities and differences the panelists believe are important to consider in light of the charge to advise NVS regarding the potential of NAEP to serve as an independent monitor of CCSS-ELA.

NAEP Reading Passages/Writing Prompts, Scoring Guides, and Anchor Papers to CCSS-ELA Documents

The purpose of these analyses was to study the alignment between the NAEP reading passages and writing prompts, scoring guides, and anchor papers and the CCSS-ELA general guidelines for the types of reading and writing students should do. Reading analysis focused on three aspects of text as defined by both qualitative and quantitative criteria described in CCSS-ELA documents: (1) range of text types, (2) quality of text, and (3) text complexity. Writing analysis focused on three elements of the NAEP writing assessment in relation to the CCSS-ELA standards and sample papers: (1) NAEP scoring guides (criteria for valued dimensions of writing), (2) NAEP anchor papers (illustrations of performance levels), and (3) NAEP prompts (qualities, range of purposes, audiences).

NAEP Reading Items/Writing Prompts to CCSS-ELA Anchor/Grade-Level Standards

The purpose of the final analyses was a detailed examination of the NAEP reading items and writing prompts at Grades 4, 8, and 12 in relation to the specific anchor CCSS-ELA standards. These analyses were designed to evaluate more precisely the alignment between NAEP items and the standards and to determine whether there are CCSS-ELA standards that are not addressed by NAEP items/prompts. In total, the Reading Panel analyzed 146 reading items across Grades 4, 8, and 12, and the Writing Panel analyzed 80 prompts, 8 scoring guides, and 36 anchor papers.

Overall Conclusions of the Reading and Writing Panels

The Reading and Writing Panel members recognize the different purposes of NAEP and CCSS-ELA and feel strongly that NAEP should retain its independence from any particular curriculum and serve as a general assessment of reading and writing performance. Overall, the panels are cautiously optimistic that, with attention to the specific issues identified in this report and a systematic program of special studies to inform future assessments, NAEP could continue to serve as an independent monitor of student achievement in an era of CCSS. In the area of reading assessment, NAEP should consider revisions related to reading and knowledge building in the disciplines, text selection (including digital texts) and complexity, integration of reading and writing, and assessment of academic vocabulary. In the area of writing, NAEP should consider revisions related to writing in response to text and research, integrating writing into discipline-specific assessments, expanding the use of technology, and providing more extended time for writing to accommodate different types of writing tasks and conditions.

The panels also judge that NAEP could serve as an intellectual tool to promote the design and use of quality assessments apart from CCSS. With attention to the recommendations in this report, NAEP could be in an excellent position to lead the way for forward-looking reading and writing assessment. Indeed, the panels encourage NAEP to consider the future and changes in literacy demands as they conceptualize literacy assessment. NAEP's ability to sample a wide variety of student performance on a range of texts and tasks through its matrix sampling design is consistent with the range of literacy performances expected by CCSS-ELA and places it in an excellent position to engage in the kind of special studies needed, both to assess these complex standards and to serve as an external point of comparison useful to future revisions of the CCSS-ELA.

Because of the timing of the study, the panels could not determine the degree of alignment between NAEP and new assessments under development by Smarter Balanced and PARCC. This is an important consideration because the ability of NAEP to serve as an independent monitor may be judged by a comparison of student achievement on NAEP with achievement on the new assessments; alternatively, it may be judged by the degree of alignment between NAEP assessments and the framing concepts in the CCSS-ELA documents rather than simply the new assessments. Furthermore, at this point in time, the potential impact of CCSS documents and specific standards on curriculum and assessment is unknown, most especially the integration of reading and writing, technology, and knowledge building in the disciplines. The CCSS documents integrate writing and reading across the disciplines, call for extended writing tasks that involve reading and research, and convey the expectation that students will use technology "strategically and capably." The extent to which these elements will be operationalized in the new assessments and/or in classroom instruction is not clear, but the panels believe these issues are integral to the next iterations of literacy assessment and to students' success in their careers and college. Consequently, there will need to be additional studies to evaluate the fit of new CCSS assessment items to CCSS standards and to compare CCSS assessment items to NAEP items. In cases in which NAEP and new CCSS assessment do not align, it will be important to look at the areas of nonalignment found in the studies reported here as a possible explanation for the nonalignment. Furthermore, it will be important to define the specific contribution NAEP should make and the role it should play. These issues will need to be addressed as new assessments are implemented and evaluated and as curricula and instruction change to reflect successful implementation of CCSS-ELA.

The panel advises that the CCSS-ELA reading and writing anchor standards, which are research based and consistent across grade levels, are most consistent with the NAEP reading and writing frameworks in contrast to the CCSS-ELA grade-level standards. Furthermore, the panel suggests that NAEP interpret the anchor standards broadly and conceptually rather than specifically and procedurally. Because some of the anchor standards include multiple parts or specifics that could confound or constrain test development, the panel encourages NAEP to bring a "generous" reading to the anchor standards.

Specific Conclusions and Recommendations for NAEP Reading

1. Panel members find that many aspects of the current NAEP reading assessment are consistent with conceptualizations of the reading process found in the research and in CCSS-ELA documents:
    - Cognitive focus aligned with research
    - Broad range of text types
    - High quality and appropriate length of texts used in assessment
    - Attention to literary and informational comprehension
    - Use of text pairs
    - Attention to reader-text interactions in item development
    - Inclusion of writing in response to reading
    - Parsimony and elegance in crafting questions to align with specific texts
    - Thoughtful, meaningful items—well sequenced and crafted

    Panelists also recognize the different purposes of NAEP and CCSS-ELA and feel strongly that NAEP should retain its independence from any particular curriculum and serve as a general assessment of reading comprehension. In addition, NAEP's ability to sample a wide variety of student performance on a range of texts and tasks through its matrix sampling is consistent with the range of reading performances expected by CCSS-ELA and should be preserved.

    The panel believes that NAEP could build upon these strengths as they consider several recommendations and issues to enhance its relevance to the CCSS-ELA and reflect emerging areas of reading assessment. These recommendations follow.

2. CCSS-ELA has made clear the expectation to increase the "rigor" and "complexity" of texts students read at each grade level as well as progressively across grade levels. In contrast, the NAEP approach is to use texts that are judged to be within the currently recognized range of difficulty for the targeted grade. Nevertheless, the panel finds that the NAEP reading selections at Grades 4 and 8 generally fall within (or above) the quantitative ranges called for in the CCSS-ELA, while the Grade 12 NAEP passages are consistently less difficult than called for by CCSS-ELA quantitative indexes. The panel suggests that NAEP consider passages that include more complexity at the upper grade levels in terms of perspective taking, bias, competing accounts, trustworthiness of the sources, craft, conceptual issues, etc., that might allow for assessing deeper, closer reading. The panel cautions, however, that text difficulty should not be judged solely on quantitative measures—a position supported by both CCSS-ELA and NAEP. The complex issue of text difficulty, including differences between assessment and instructional-level texts, the interplay of text and reading

items/tasks, and assessments that reliably measure across the ability range should be explicitly addressed as NAEP prepares for future assessments.

3. The panel finds that the NAEP framework for constructing items to align with cognitive targets is compatible with the CCSS-ELA anchor standards and should continue to be used for item development.
4. Panel members caution NAEP to be cognizant of the lack of research base, inconsistencies, and specificity of the "learning progressions" embodied by the K–12 grade-level standards in CCSS-ELA.
5. NAEP items align most often with CCSS-ELA Anchor Standards R1–5. Anchor Standards 6–9 are least well represented in the assessments. The panel suggests that NAEP examine how it might place additional focus on assessing point of view, bias, perspectives, and such (Standard R6) and explore strategies (including the use of special studies) for assessing standards related to building knowledge (Anchor Standards R6–9).
6. Many of the NAEP short-constructed and extended-constructed response reading items are aligned with both CCSS-ELA reading and writing anchor standards. Given the emphasis on writing in response to text in the CCSS-ELA writing standards, the panelists suggest that NAEP investigate the possibility of double scoring these items for both reading and writing.
7. An important area of difference between CCSS-ELA and NAEP is the manner in which disciplinary reading is addressed. The *conceptual framing* for CCSS-ELA positions disciplinary reading for the purposes of building new knowledge in the specific discipline. In contrast, the NAEP Reading Framework subsumes disciplinary texts under "informational texts," sampled from varied content areas. Although these differences exist in the *framing sections* of CCSS-ELA and NAEP documents, the panel finds them to be far less evident when comparing of NAEP items and CCSS-ELA anchor standards or grade-level standards. As a result, the panel was uncertain about the degree to which specific disciplinary reading outcomes would be operationalized when the CCSS-ELA standards are implemented.

   The panel suggests that NAEP adopt a more systematic treatment of discipline-specific texts in the text selection process. However, at the same time, it is unclear what the focus should be for assessing these texts—general understanding or disciplinary knowledge building, especially given the difficulties of attending to issues of prior knowledge and topic familiarity in an assessment like NAEP. Overall, the issue of disciplinary text—the purpose, outcomes, and text selection—needs to be addressed and clarified in future NAEP frameworks and assessments.
8. There is a general sense that NAEP's practice of restricting text selection to material written for general audiences may have had the overall effect of constraining the texts that appear on NAEP more than intended. The panelists suggest that NAEP would be more consistent with the CCSS-ELA if it were to consider inclusion of more dense text and texts that are representative of textbook or workplace reading—these are typically less explicit and controlled than texts currently used in NAEP. At the same time, NAEP needs to accommodate a wide range of reading abilities, including students performing at and below the *Basic* achievement level, especially at fourth grade.

9. The CCSS-ELA documents include attention to classic literature, well-known documents, and popular texts. Attention to these sorts of texts may be appropriate in an instructional setting; however, issues of familiarity (prior knowledge) and length are likely to make these types of texts inappropriate for inclusion in NAEP. NAEP might want to clarify for CCSS-ELA consumers how and why texts used for assessment must necessarily differ in some respects from those used in school and the workplace.

10. NAEP should consider using digital text and information displayed in graphs and charts. These text types are called for in CCSS-ELA, and panelists generally feel that a current (and forward looking) assessment of 21st century literacy should include online reading and research. They suggest that NAEP consult existing research regarding the similarities and differences between "traditional" and Internet/online reading to inform future assessment development. Some panelists also feel that NAEP should reconsider the role and nature of more procedural/ functional texts both in the real-world and academic contexts as well as more 12th-grade passages that align with the types of texts typically assigned in college.

11. There are differences in how NAEP and CCSS-ELA address vocabulary. NAEP focuses on a particular type of vocabulary and format for assessment purposes—word meaning in the context of a given passage; CCSS-ELA takes a much broader perspective on vocabulary as an essential element of ELA with a definite emphasis on discipline-specific and academic vocabulary. The panel recommends that NAEP consider both the reading anchor standards and the language anchor standards as it evaluates its existing approach and possible new approaches to vocabulary assessment.

12. The CCSS-ELA include K–5 standards for foundational skills, while NAEP assessments target comprehension beginning at Grade 4. The panelists caution that fourth-grade assessments developed specifically to measure CCSS-ELA may include items testing foundational skills as well as literature/informational standards. Because foundational skills are not part of NAEP, comparisons of fourth-grade performance across different assessments may need to take this into account.

Specific Conclusions and Recommendations for NAEP Writing

1. Panel members find much to commend in the current NAEP writing assessment, reflecting as it does a conceptualization of writing found in both research and in the CCSS-ELA documents. Both NAEP and CCSS-ELA present writing as a social, communicative activity; emphasize the importance of audience, purpose, and task; and treat rhetorical flexibility as an important component of skilled performance. NAEP and CCSS-ELA are aligned in other important ways as well. They address similar broad domains of writing and identify and discuss essentially the same valued characteristics of effective writing—development of ideas, organization, and language facility and conventions.

   In light of these strengths, the panel concludes that NAEP should continue to serve as an independent monitor of student achievement in writing in an era of CCSS. The panel also concludes that NAEP should build upon these strengths as it considers ways to reflect emphases in writing curricula in current practice, research, and the CCSS that are not well addressed by the current assessment. These issues and recommendations follow.

2. The CCSS-ELA clearly emphasizes integration of the language arts, while NAEP does not. In particular, CCSS-ELA emphasizes writing about reading and writing from sources (writing based on research). NAEP assessment tasks rely primarily on background knowledge and personal experience. Panelists recommend that NAEP consider including writing in response to print and/or nonprint texts and writing based on research (writing from sources) in future assessments.
3. The CCSS-ELA is explicit in acknowledging that the teaching of writing is a shared responsibility across disciplines and that writing activities within the disciplines are integrated with content learning. While the NAEP Writing Framework acknowledges the situated nature of writing and its importance in all disciplines, it does not address the special skills, strategies, or domain-specific vocabulary associated with writing in the disciplines. Panelists recommend that NAEP consider including writing tasks, especially those that are structured around deep knowledge of subject matter, in NAEP's discipline-specific assessments, either as part of the regular NAEP assessment or as a probe study. Furthermore, NAEP should consider tracking domain-specific vocabulary along with general vocabulary.
4. At present, NAEP limits the role that technology plays in assessment to students' use of a computer for composing and editing with a limited set of commonly available tools. CCSS-ELA, on the other hand, conveys a portrait of college- and career-ready students who "use technology and digital media strategically and capably." Panelists recommend that NAEP consider expanding the use of technology in writing, either as part of the regular NAEP assessment or as a probe study. They also note, however, that if students are to have a wider range of technology-enabled options in the regular NAEP assessment, they would need to have more time to compose as well as to understand the options presented in whatever platform is used in the assessment.
5. NAEP assesses on-demand writing in an abbreviated time frame, while CCSS-ELA emphasizes writing under a variety of conditions and conveys specific expectations for students' use of writing processes such as planning, revising, editing, and rewriting. Panelists recommend that NAEP consider investigating ways to allow different amounts of time for different kinds of tasks. Providing more extended time frames could encourage revising and/or accommodate some of the more complex reading/writing tasks found in the CCSS-ELA. Panelists also suggest that NAEP consider conducting special studies of extended tasks as they are being used in schools.

In Closing

The Reading and Writing Panels appreciate the opportunity to analyze NAEP in light of the CCSS-ELA and the literacy demands of the 21st century. The hope is that the detailed analyses and recommendations contained in the full report will provide the NVS with both information and perspectives that will help it move forward.

CONTENTS

Introduction

Since its first assessment in 1969, the National Assessment of Educational Progress (NAEP) has made a unique contribution to our understanding of American education. It is the only national source of information on the educational achievement of U.S. students, and it is the only vehicle by which states can compare the progress of their students against a common standard. Assessment results reported by NAEP complement the states' own reports of progress under No Child Left Behind (NCLB) and track the status of achievement gaps for traditionally disadvantaged student groups.

NAEP is carried out under the guidance of the National Assessment Governing Board (Governing Board) and the National Center for Education Statistics (NCES). Throughout the course of its history, NAEP has frequently sought to improve by studying its own processes, instruments, and procedures. In keeping with this tradition, in fall 2011, NCES asked the NAEP Validity Studies (NVS) Panel, which operates under contract to NCES, to undertake two interrelated studies, one in reading/writing and one in mathematics, to examine the content of the current NAEP frameworks and item pools at Grades 4, 8, and 12 in relation to the Common Core State Standards (CCSS). The primary question under investigation is whether NAEP can continue to serve as an independent monitor of student achievement and state assessments following the implementation of the CCSS.

This report addresses the relations between the NAEP reading and writing frameworks and the CCSS in English language arts (CCSS-ELA), and the relations between the NAEP reading and writing items and the CCSS-ELA. It does not address the relations between NAEP reading and writing items and items developed by the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (Smarter Balanced) to assess the CCSS-ELA because those items were not available at the time of this study.

The report concludes with recommendations to NCES regarding broader issues on the alignment between NAEP reading and writing and CCSS-ELA, including the extent of alignment that is appropriate to support NAEP's role as an independent monitor of student achievement.

NAEP Frameworks and Common Core State Standards

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

To fulfill this mandate, the Governing Board, working through its contractors, produces an assessment framework for each subject area. These frameworks are replaced or updated periodically, balancing the need to stay current with the field against an interest in measuring trends over time.

The framework documents are intended to portray the NAEP assessments to a broad audience of educators and the general public as well as to inform the test developers. The frameworks explicate the structure of the knowledge domain to be assessed, describe the broad outlines of the assessment, define the achievement levels that will be used to report the assessment, and present a set of sample questions.

Reading Framework

The Reading Framework employed in this study has been operational since 2009. It is the second Reading Framework approved by the Governing Board and replaces the framework that was used in NAEP from 1992 to 2007. As noted above, the framework is intended for a broad audience. A more detailed technical document, the *Reading Assessment and Item Specifications for the National Assessment of Educational Progress*, provides information to guide passage selection, item development, and other aspects of test development. Both the framework and the specifications documents are available to the public at http://www.nagb.org/publications/frameworks.html.

Through the framework, the Governing Board has defined several parameters for the reading assessment. First, the assessment will measure reading comprehension in English. On the assessment, students will be asked to read passages written in English and to answer questions about what they have read. Second, because this is an assessment of reading comprehension and not listening comprehension, the assessment does not allow passages to be read aloud to students as a test accommodation. Third, under Governing Board policy, the framework "shall not endorse or advocate a particular pedagogical approach, but shall focus on important, measurable indicators of student achievement…" (National Assessment Governing Board, 2010a, p. iii). Although broad implications for instruction may be inferred from the assessment, NAEP does not specify how reading should be taught, nor does it prescribe a particular curricular approach to teaching reading.

The NAEP Reading Framework results from the work of many individuals and organizations involved in reading and reading education, including researchers, policymakers, educators, and other members of the public. Their work was guided by scientifically based literacy research that conceptualizes reading as a dynamic cognitive process as reflected in the following definition of reading:

"Reading is an active and complex process that involves:

- Understanding written text.
- Developing and interpreting meaning.
- Using meaning as appropriate to type of text, purpose, and situation" (National Assessment Governing Board, 2010a, p. iv).

This definition applies to the assessment of reading achievement on NAEP and is not intended to be an inclusive definition of reading or to describe a reading curriculum.

Writing Framework

The Writing Framework employed in the study became operational in 2011 for Grades 8 and 12. (Grade 4 was assessed on a pilot basis only, using the new framework, in 2011.) The framework describes, for a general audience, how the assessment should measure students' writing at Grades 4, 8, and 12. Both the framework and the more technical specifications document are available to the public at http://www.nagb.org/publications/frameworks.html. This is the second Writing Framework approved by the Governing Board; it replaces the framework that was used in the NAEP from 1998 to 2007.

Given expanding contexts for writing in the 21st century, the NAEP Writing Framework is designed to support the assessment of writing as a purposeful act of thinking and expression used to accomplish many different goals. Although NAEP cannot assess all contexts for student writing, the framework defines an assessment that offers opportunities to understand students' ability—in an "on-demand" writing situation—to make effective choices for their writing in relation to a specified purpose and audience. In this respect, the assessment reflects writing situations common to both academic and workplace settings, in which writers are often expected to respond to on-demand writing tasks.

In addition, the assessment is designed to provide important information about the impact of new technologies on writing in K–12 education—including the impact of word processing software—and about the extent to which students at Grade 12 are prepared to meet postsecondary expectations.

For the assessment, students at all three grades complete two 30-minute on-demand writing tasks. Students have the flexibility to make rhetorical choices that help shape the development and organization of ideas and the language of their responses. Using age- and grade-appropriate writing tasks, the assessment evaluates writers' ability to achieve three purposes common to writing in school and in the workplace: to persuade; to explain; and to convey experience, real or imagined.

The scoring guides for each of these three purposes focus on three broad features of writing (development of ideas, organization of ideas, and language facility/conventions) and describe six levels of performance. Anchor papers (selected pieces of student work) illustrate expectations for performance at each of the six levels. Taken together, the tasks, scoring guides, and anchor papers define the assessment.

The NAEP Writing Framework results from the work of a diverse array of individuals and organizations involved in writing and writing education, including researchers, policymakers, educators, and other members of the public. Their work was guided by scientifically based research that conceptualizes writing as a relationship or negotiation between the writer and reader to satisfy the aims of both parties. As a result, the Writing Framework focuses on writing for communicative purposes and on the relationship of the writer to his or her intended audience, as reflected in the following definition of writing:

"Writing is a complex, multifaceted and purposeful act of communication that is accomplished in a variety of environments, under various constraints of time, and with a variety of language resources and technological tools" (National Assessment Governing Board, 2010b, p. 3).

Common Core State Standards

The Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects (CCSS-ELA), like most content standards, are designed to provide a consistent, clear understanding of what students are expected to be taught and, thus, to learn. They are designed to be robust and relevant to the real world, reflecting the knowledge and skills that young people need for success in college and careers. The concept of college and career readiness is a driving force behind the CCSS-ELA. College and career readiness (CCR) standards for the end of 12th grade were developed first. They then served as the basis for the development of the K–12 grade-level standards, which are intended to be learning progressions that lead to achievement of the CCR.

The development of the CCSS was led by the states, not a federal agency, under the auspices of the National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO). As a state-led initiative, the CCSS are designed to improve on current state standards by creating fewer, clearer, and higher level standards. The CCSS-ELA are also reported to be internationally benchmarked to help ensure that all students are prepared to succeed in a global economy and society.

It is also worth noting what the CCSS-ELA do not define. First, the CCSS-ELA are not intended to define all that can or should be taught; they are not intended to be a curriculum. Rather, they are intended to provide specification of the goals that should be achieved through curriculum. Second, the CCSS-ELA do not define how teachers should teach. Third, they do not define the nature of advanced work beyond the CCSS or the interventions needed for students well below grade level. Finally, they do not define the full range of supports for English language learners and students with special needs.

The CCSS-ELA are the culmination of an extended, broad-based effort to create the next generation of K–12 standards in order to help ensure that all students are college and career ready in literacy no later than the end of high school. The CCSS-ELA consists of several documents. The main body of the CCSS-ELA includes introductory material and the standards themselves. The standards are presented separately for each area of the language arts—reading, writing, speaking and listening, and language. Within each of these areas, there are two types of standards. First, there are the CCSS college and career readiness anchor standards. These standards are the same for all grades, K–12. Second, there are grade-level standards, which "unpack" the CCSS anchor standards at each grade level. A unique feature of the standards in Grades 6–12 is the addition of CCSS anchor and grade-level standards in reading and writing in the subject areas—history/social studies, science, and technical subjects.

In addition to the introductory materials and standards, the CCSS-ELA documents include three appendixes. Appendix A elaborates on text complexity, foundational reading skills, and a skills progression for language development. Appendix B provides sample reading texts and performance tasks, and Appendix C provides samples of quality writing at each grade level. These appendixes are integral to understanding and implementing the standards.

The CCSS-ELA documents build on the foundation laid by states in their decades-long work on crafting high-quality education standards. The introductory material states that the standards also draw on the most important international models as well as research and input from numerous sources, including state departments of education; scholars; assessment developers; professional organizations; educators from kindergarten through college; and parents, students, and other members of the public. In their design and content, refined through successive drafts and numerous rounds of feedback, the standards represent an effort to synthesize the best elements of standards-related work to date and represent an advance over that previous work.

The CCSS-ELA standards provide an integrated view of English language arts. There is integration of all of the areas of the language arts (reading, writing, listening/speaking, and language) across Grades K–12 and integration between two areas of the language arts (reading and writing) across the subject areas of history/social studies, science, and technical subjects at Grades 6–12. It is important to note that the 6–12 reading and writing standards in history/social studies, science, and technical subjects are not meant to replace content standards in those areas but rather to supplement them. States may incorporate these reading and writing standards into their standards for those subjects or adopt them separately as content area literacy standards.

In addition to the integrated and disciplinary focus of the CCSS-ELA Grade 6–12 standards, the Grade 12 standards are intended to define the English language arts skills and understandings required for college and career readiness. As a natural outgrowth of meeting this intent, the standards also lay out a vision of what it means to be a literate person in the 21st century. Therefore, the skills and understandings that students are expected to demonstrate are intended to have wide applicability outside the classroom or workplace.

- "Students who meet the Standards readily undertake the close, attentive reading that is at the heart of understanding and enjoying complex works of literature.
- They habitually perform the critical reading necessary to pick carefully through the staggering amount of information available today in print and digitally.
- They actively seek the wide, deep, and thoughtful engagement with high-quality literary and informational texts that builds knowledge, enlarges experience, and broadens worldviews.
- They reflexively demonstrate the cogent reasoning and use of evidence that is essential to both private deliberation and responsible citizenship in a democratic republic" (National Governors Association & Council of Chief State School Officers, 2010, p. 3).

In short, students who meet the standards are expected to develop the skills in reading, writing, listening/speaking, and language that are the foundation for any creative and purposeful use of language.

Purpose and Methods

To address the broad charge to the NVS Panel to evaluate NAEP as a potential monitor of CCSS-ELA achievement, two expert panels were convened—one for reading and one for writing. Listening and speaking were not included in the analysis because there are no NAEP assessments in these areas.

The study directors were NVS Panel members Dr. Wixson and Dr. Phillips, and the subject area directors were Dr. Valencia (reading) and Dr. Murphy (writing). Additional content experts with extensive knowledge and experience with NAEP and/or CCSS-ELA were invited to participate in either the reading or writing analyses. All agreed. The names of these content experts are listed in the appendix.

The following comparative analyses were designed by the study directors and carried out by the expert panels, separately for reading and writing:

1. **NAEP Frameworks to CCSS-ELA Documents**—to analyze the similarities and differences between the conceptualization and content of the NAEP reading and writing frameworks and the CCSS-ELA documents
2. **NAEP Reading Passages/Writing Prompts, Scoring Guides, and Anchor Papers to CCSS-ELA Documents**—to analyze the alignment between the NAEP reading passages and writing prompts, scoring guides, and anchor papers and the CCSS-ELA general guidelines for the types of reading and writing students should do
3. **NAEP Items/Writing Prompts to CCSS-ELA Anchor/Grade-Level Standards**—to analyze the alignment of the actual NAEP reading items and writing prompts at Grades 4, 8, and 12 with specific anchor CCSS-ELA standards

Activity 1. NAEP Frameworks to CCSS-ELA Documents

This activity was a qualitative analysis of the similarities and differences between the NAEP reading and writing frameworks and the CCSS-ELA documents to determine how the domains are conceived, defined, organized, and parsed. All CCSS-ELA documents (including the CCSS-ELA Appendixes A, B, and C) and NAEP reading and writing framework documents were used for this analysis. The analyses were conducted by five expert panel members for each subject area, including study director Dr. Wixson and either Dr. Valencia (reading) or Dr. Murphy (writing).

The choice of a qualitative, descriptive set of procedures for making the comparisons, as opposed to a traditional alignment methodology, was primarily driven by the nature of the NAEP reading and writing frameworks. The methods used in traditional alignment studies would require that the NAEP frameworks be parsed into standards/objectives that do not reflect the basic intent of these documents.

After considering several different approaches to this comparative analysis, the study directors agreed to ask expert panel members to respond individually to the five questions listed below and then hold several conference calls to deliberate and come to a consensus. In conducting Activity 1 of this study, the panelists were cognizant of the basic differences that exist in the purposes of CCSS-ELA and the NAEP frameworks. As crafted, CCSS-ELA documents represent a detailed framework, with exemplars, for what should be taught and what students should be able to do in K–12 in English language arts and in literacy in history/social studies, science, and technical subjects. By contrast, the NAEP documents are assessment frameworks and do not expressly seek to influence curricular decisions. These differences in purpose translate into different aspects/elements being included in each. With these basic differences in mind, the analyses enumerated the similarities and differences that the panelists believed are important to consider in light of the charge to advise NCES regarding the potential of NAEP to serve as an independent monitor of student achievement under CCSS-ELA. The questions driving this analysis were:

1. What similarities and differences are important to consider in the conceptualization of reading or writing (depending on your group) as reflected in the NAEP framework and the CCSS-ELA documents?
2. Starting with the NAEP framework, what aspects/elements of NAEP reading or writing (depending on your group) are addressed in the overview of the CCSS-ELA, the appendixes, and the standards for Grades 4, 8, and 12? Where are the NAEP elements addressed in the CCSS-ELA documents? What, if anything, is in the NAEP framework that is not in CCSS-ELA overview and other documents?
3. Starting with CCSS-ELA documents, including the overview, the grade-level standards for Grades 4, 8, and 12, and the appendixes, what aspects of reading or writing are not addressed in the NAEP framework?
4. What elements identified as present in CCSS-ELA standards and associated documents, but not in the NAEP framework, do you consider important for the purposes of assessment? Where, or in what ways, might they be addressed in a NAEP assessment?
5. What additional issues, beyond those identified above, do you think are important to address as NAEP considers alignment with CCSS-ELA? Please help us understand the issues and why they are important to a national assessment.

Once the panel members had been contacted and had agreed to participate in this activity, separate conference calls were held with the Reading and Writing Panels to go over the task and address panelists' questions. The panelists then submitted individual written responses to the five questions. The study directors prepared a draft summary of the comparisons for review and discussion by the panelists in subsequent conference calls. Information from the individual panelists' analyses and the conference calls was synthesized and then shared with panelists for their review and comment.

Activity 2. NAEP Reading Passages/Writing Prompts, Scoring Guides, and Anchor Papers to CCSS-ELA Documents

Once Activity 1 was concluded, Activities 2 and 3 were conducted concurrently. A total of nine reading experts and nine writing experts, including the study directors (Dr. Wixson, Dr. Valencia/Dr. Murphy) participated in Activity 2 and Activity 3 (see the appendix). The Reading Panel met on September 11–12, 2012, and the Writing Panel met on October 26–27, 2012. Several observers from NCES and AIR also attended these meetings.

Activity 2 focused on the relations between aspects of the NAEP assessments (specifically, the reading passages, and the writing prompts with their associated scoring guides and anchor papers) and the CCSS-ELA documents. The study directors developed the methods used for comparing specific dimensions of the NAEP assessments to the CCSS-ELA documents. Each panel member conducted individual analyses with an emphasis on one of the three grade levels—4, 8, or 12 (approximately three to four panelists per grade level)—prior to the face-to-face meetings. The following describes the processes specific to each subject area panel.

Analysis of NAEP Reading Passages

CCSS-ELA documents place a major emphasis on describing the types of texts students should read. Therefore, prior to the face-to-face meeting, panelists evaluated each of the reading passages in the pool of NAEP passages at their assigned grade level and a selected sample of passages from the other grade levels. Across grade levels, a total of 28 reading blocks (20 containing a single reading selection and 8 containing two reading selections) from the 2009 and 2011 NAEP reading assessments were used for this analysis. All blocks were analyzed by three to six panelists to establish a consensus.

The analysis focused on three aspects of text as defined by both the qualitative and quantitative criteria described in CCSS-ELA documents: (1) range of text types, (2) quality of text, and (3) text complexity. All panelists provided a written analysis of each NAEP passage they were assigned in response to the following questions:

- How does this passage fit within the range of types of texts called for by the CCSS-ELA at the designated grade level?
- How does this passage fit with the dimensions of passage quality (i.e., levels of meaning or purpose, structure, language conventions and clarity, knowledge demands) called for by the CCSS-ELA at this grade level ?
- How are the passage qualities similar to/different from the passage qualities called for in CCSS-ELA?
- How does the complexity of the passage fit with both the qualitative and quantitative criteria called for by CCSS-ELA at the designated grade level?

In addition, panelists wrote summary reports for the passages they evaluated in response to the following question: “How well does the pool of NAEP passages at your target grade level reflect what is called for in CCSS-ELA in terms of range, quality, and complexity? Explain your reasoning and indicate what, if any, changes NAEP should consider making in its passage selections.” These written analyses were reviewed and assembled by the study directors. At the face-to-face meetings, panelists met in grade-level subgroups to develop a consensus analysis for each grade level that was then shared and discussed with the entire panel.

Analysis of NAEP Writing Prompts, Scoring Guides, and Anchor Papers

CCSS-ELA documents prioritize adapting writing to purpose and audience as well as dimensions of writing such as clarity, coherence, development, organization, and use of language and conventions. Evidence of attention to these elements can be found in particular artifacts associated with the assessment: the prompts, focused holistic scoring guides (one for each purpose), and sets of anchor papers (each set illustrating performance levels 1 through 6) that, taken together, define the assessment. Therefore, prior to the face-to-face meeting, panelists were asked to conduct individual analyses of these artifacts. Panelists read all of the prompts, the three scoring guides, and two sets of anchor papers (one for each of two prompts) at one assigned grade level (4, 8, or 12). Each anchor set contained six papers, one for each score level, 1–6. In addition, to give panelists some background for discussions with the panel as a whole and a sense of the progression of expectations across the three grade levels, panelists were asked to read two prompts along with their corresponding scoring guides and anchor sets at the other two grade levels. At the face-to-face meetings, panelists worked in grade-level groups to establish a consensus. All prompts, scoring guides, and anchor papers were read and discussed by three to six panelists. Each panelist completed three individual summary sheets, one for the scoring guide analysis, one for the prompt analysis, and one for the anchor set analysis. A total of 80 prompts, 8 scoring guides, and 6 sets of anchor papers were used for these analyses.

Scoring Guide Analysis. For the scoring guide analysis, panelists wrote responses to three questions about the extent to which the NAEP scoring guides for their assigned grade levels were consistent with the emphasis in the CCSS-ELA standards and accompanying documents on (1) particular types/purposes for writing; (2) particular dimensions of writing (development, organization, language facility and conventions); and (3) adapting writing to purpose, audience, and task. Panelists were asked: "Explain your reasoning and discuss the implications, if any, for the design of the NAEP scoring guides."

Anchor Paper Analysis. Appendix C of the CCSS-ELA documents includes sample papers that portray the level of quality that students would be expected to achieve in order to meet (or exceed) grade-level expectations. In the NAEP assessment, scores of 4 are characterized as "sufficient," scores of 5, "skillful," and scores of 6, "excellent." For the anchor paper analysis, individual panelists provided written responses to the question, "How do the NAEP writing samples at score level 4 and above from your assigned anchor sets compare with the writing samples at this grade level in Appendix C of the CCSS-ELA? Explain your reasoning and discuss the implications, if any, for the design of the NAEP writing assessment."

Prompt Analysis. For the prompt analysis, panelists wrote responses to three questions about how well the pool of NAEP prompts for their assigned grade level fit with the information in the CCSS-ELA standards and accompanying documents with regard to particular text types and purposes; range of tasks, purposes, and audiences; and the emphasis on adapting writing to task, audience, and purpose. For all three questions, panelists were asked: "Explain your reasoning and indicate what, if any changes NAEP should consider making in its prompts."

The panelists' written analyses were reviewed and assembled by the study directors. At the face-to-face meetings, panelists met in grade-level subgroups to develop consensus analyses for writing prompts, scoring guides, and anchor papers that were then shared and discussed with the entire panel.

Activity 3. NAEP Reading Items/Writing Prompts to CCSS-ELA Anchor/Grade-Level Standards

Panelists examined reading items and writing prompts at Grades 4, 8, and 12 and identified the anchor standard(s) and grade-level standard(s) with which each item/prompt was most closely aligned. These analyses were designed to evaluate more precisely the alignment of NAEP items and prompts to the standards and to determine whether there are CCSS-ELA standards that are not addressed by NAEP items/prompts. For this analysis, actual NAEP items as well as scoring guides were used. Because NAEP reading items often require readers to draw on multiple sources of information, interpret text, and use a variety of skills and strategies, and because writing prompts sometimes elicit more than one type of writing, reading items and writing prompts sometimes aligned with multiple CCSS-ELA standards. Therefore, based on their expert judgment, panelists rated each item as *strongly aligned, moderately aligned,* or *weakly aligned* with specific standards. This provided an opportunity for panelists to go beyond a simple matching to indicate alignment; it permitted them to evaluate the strength of alignment across multiple standards.

During each of the face-to-face meetings (reading and writing), panelists first worked as an entire group to complete the task using one set of reading items or one writing prompt for each of the three grade levels. The goal here was to clarify and revise the task as needed and to reach agreement on panelists' alignment judgments across different types of assessment tasks. After working through these initial sets of items/prompts, panelists completed additional sets of items/prompts individually for their assigned grade levels. Individual ratings were then compared in grade-level groups. Grade-level groups created consensus ratings, which were shared and discussed with the entire group in an effort to examine trends and unique attributes across the grade levels. In total, the Reading Panel analyzed 146 reading items (including scoring guides for constructed response questions) across Grades 4, 8, and 12, and the Writing Panel analyzed 80 prompts, 8 scoring guides, and 36 anchor papers.

Both the Reading and Writing Panels found this task very challenging, largely because of highly variable levels of specificity found in the grade-level standards. As a result of difficulties in matching NAEP items/prompts to grade-level standards, both panels decided to use only the anchor standards for this analysis. The panels judged the anchor standards to best represent the content and intent of the CCSS-ELA. Although this was challenging, too, both panels felt this analysis resulted in a fair description of which standards are/are not covered by NAEP items and prompts. We further discuss this decision to use anchor standards rather than grade-level standards in the Results section for Activity 3 that follows.

Results

This section includes the results for reading, followed by those for writing. Within each subject area, summaries of the findings from each of the separate analyses are presented first, followed by overall conclusions and recommendations for that subject area.

The paper concludes with an overall set of conclusions and recommendations that span the two subject areas.

Reading Findings

Summary of Comparison Between NAEP Reading Framework and CCSS-ELA Documents (Activity 1)

The following describes similarities and differences between the NAEP Reading Framework and the CCSS-ELA in the areas of definition/conceptualization, cognitive processes, text types, text difficulty, vocabulary, and foundational skills. The focus is on similarities and differences with implications for NAEP's role as an independent monitor, after acknowledging that there are important differences in the *purposes* of these documents that are reflected in differences in the scope and specificity of the documents.

Definition/Conceptualization. Both NAEP and CCSS-ELA consider reading to be a complex, interactive process that is influenced by the reader, text, and context of reading. NAEP does so explicitly with its definition of reading, and CCSS-ELA does so implicitly as it describes a "vision" of what it means to be a literate person in the 21st century.

Despite the basic similarities in the conceptualization of reading in the CCSS-ELA and the NAEP Reading Framework, the Reading Panel identified some differences that could have implications for NAEP. One difference arises from the extent to which a focus on disciplinary reading is integral to the conceptual framing of English language arts in the standards documents. The CCSS-ELA documents have dedicated standards for reading in history/social studies, science, and technical subjects that are not matched within the NAEP framework. For CCSS-ELA, disciplinary reading is related to knowledge in two ways: (1) reading in the discipline serves as a way for readers to build new knowledge from text related to specific subject matter, and (2) background knowledge in the discipline or specific subject matter is necessary for deep comprehension. Disciplinary knowledge, therefore, is both the outcome of deep reading in a specific content area and a requirement to enable deep reading. In contrast, the NAEP Reading Framework subsumes disciplinary texts under "informational texts" and samples from "varied content areas." NAEP's approach to disciplinary reading is more one of assessing general comprehension, aligned with the cognitive targets, rather than specific knowledge building. This is not surprising given an assessment context that is not tied to curriculum and in which differential levels of prior knowledge and familiarity could confound the interpretation of students' performance.

A second issue that emerged as a result of this analysis focused on the CCSS-ELA grade-level standards. The panel raised concerns about the validity and specificity of the grade-level standards that might influence Activity 3 (comparing NAEP items to CCSS-ELA anchor/grade-level standards). They recommended more in-depth attention to this issue in the design and implementation of Activity 3 (see below).

Cognitive Processes. The NAEP Reading Framework explicitly defines "the mental processes or kinds of thinking that underlie reading comprehension" (National Assessment Governing Board, 2010a, p. 39) in terms of three cognitive targets—locate/recall, integrate/interpret, critique/evaluate. The CCSS-ELA anchor standards make no mention of the cognitive processes that readers might engage in when reading to achieve particular standards, although it could be argued that they might be inferred from the wording of the anchor or grade-level standards. In contrast to NAEP's emphasis on cognitive processes, the CCSS-ELA documents focus on the outputs/behaviors (i.e., what students should know and be able to do) to demonstrate their performance of the standards.

An important difference between NAEP and CCSS-ELA from the standpoint of assessment is the matter of what "develops" across grade levels. Specifically, within NAEP, the same cognitive targets are specified across grades, and the level of text complexity varies. In contrast, the complexity of both the texts *and* the grade-level standards (outputs/behaviors) is designed to increase across grades within the CCSS-ELA. If NAEP is aligned at the level of the anchor standards, rather than the grade-level standards, this is not an issue because those standards remain the same across grades.

Text Types. Both NAEP and CCSS-ELA identify two general types of text—literary and informational—and both assert that proficient readers must be able to demonstrate reading processes across a range of text types/subtypes, with an increasing presence of informational texts as students move up the grade levels. However, NAEP provides a much more elaborate system for specifying both genre and text elements than does the CCSS-ELA. Although it is likely that the more detailed NAEP specifications would fulfill the general text type categories identified in CCSS-ELA, the exemplar texts provided in the CCSS-ELA Appendix B and the list of text types of texts recommended in the main body of the CCSS-ELA standards document include additional text types (e.g., classic and traditional texts) that are not typically included in NAEP. Another area specifically noted in the CCSS-ELA that is not addressed in the NAEP framework is students' ability to read digital text.

Text Difficulty. Both NAEP and CCSS-ELA attend to a range of factors that influence "comprehensibility" of text or "text complexity." NAEP attends to text difficulty primarily through a set of qualitative factors (National Assessment Governing Board, 2010a, pp. 29–32) applied by "expert judgment," as well as the use of story/concept maps and "at least two research-based readability formulas." CCSS-ELA addresses text difficulty through guidelines for measuring text complexity provided in Appendix A, in which the importance of both quantitative and qualitative factors is acknowledged.

An important element of the CCSS-ELA documents with regard to text difficulty is the intention to increase the “rigor” and “complexity” of texts students read at each grade level as well as progressively across grade levels. In contrast, the NAEP approach is to use texts that are judged to be within the currently recognized range of difficulty for the targeted grade. This issue of text difficulty and what counts as grade-level text must be carefully analyzed as NAEP explores its role as a monitor of CCSS-ELA.

The wording for some grade-level standards in CCSS-ELA includes explicit references to supports for lower performing students—for example, “with prompting and support” (National Assessment Governing Board, 2010a, p. 11) or “with scaffolding as needed at the high end of the [text complexity] range” (National Assessment Governing Board, 2010a, p. 37). This wording makes sense from a developmental perspective and from an instructional perspective. However, in an assessment, where students’ reading abilities vis-à-vis the CCSS-ELA standards are being tested, the level of prompting or support is irrelevant because students must function independently. Consequently, the issue of the difficulty level of the texts selected for NAEP comes back into play. Panelists discussed this issue especially given that NAEP has been concerned about the reliability and validity of data for low-performing students. The panelists were also aware that some of the new CCSS-ELA assessments might integrate adaptive testing strategies that could provide students with texts of varying difficulty levels.

Vocabulary. There is a noticeable distinction between the NAEP and CCSS-ELA in the treatment of vocabulary. NAEP focuses on a particular type of vocabulary for assessment purposes—word meaning in the context of a given passage—while CCSS-ELA takes a much broader perspective on vocabulary as an essential element of ELA and also places a definite emphasis on discipline-specific and academic vocabulary.

Foundational Skills. An obvious difference between CCSS-ELA and the NAEP framework is attention to foundational skills for K–5 in the CCSS-ELA. Although it is not common practice to assess foundational skills above Grade 3 in large-scale assessments, it is possible that newer assessments of the CCSS-ELA may include foundational skills. If that happens, NAEP will need to revisit issues of alignment with CCSS-ELA for its fourth-grade assessment.

Summary of Comparison Between NAEP Reading Passages and Descriptions of Texts and Exemplars in CCSS-ELA Documents (Activity 2)

The following section provides a summary of the panelists’ evaluation of the pool of 28 NAEP passages at Grades 4, 8, and 12 in relation to the CCSS-ELA descriptions of the range, quality, and complexity of texts that students are expected to encounter in instruction at different grade levels. Some NAEP passages are administered only at one grade level (4, 8, or 12), and others are administered at two grades (4 and 8 or 8 and 12).

Range. Range is defined in CCSS-ELA documents as the types of texts students should encounter within literature and informational reading (e.g., stories, poems, myths, and disciplinary texts in history and science). There was general agreement among the panelists that at Grades 4 and 8 the pool of NAEP passages reviewed was fairly representative of the kinds of texts called for in the CCSS-ELA. At Grade 12, some differences were noted, such as the inclusion of documents in the CCSS-ELA that were more focused on academic content than are found in NAEP.

Although there was general agreement that there is reasonably good alignment between NAEP passages and the text types called for in CCSS-ELA, panelists were concerned that there was limited variability among the pool of NAEP passages representing each text type. It was also observed that while canonical texts are emphasized in the CCSS-ELA, they are not as present in the NAEP passages, although some do exist at Grades 8 and 12.

Panelists further noted that there are several types of texts included in CCSS-ELA that were not included in the NAEP item pool for 2009–2011 or called for in the Reading Framework. At Grade 4, there was no representation of drama, forms (documents) or information displayed in charts and graphs, or digital texts. At Grade 8, there was no representation of drama, and there were no examples of document reading, although some of the passages did include charts, tables, and other graphic elements. Furthermore, there was no Web-based or media-like information represented in NAEP, although these types of texts are called for in the CCSS-ELA documents. At Grade 12, it was noted that the NAEP passages had no instances of drama or of digital or online texts. Although documents were present in NAEP at Grade 12, there were questions about the relevance of the selected documents for "college and career readiness." It was also noted that NAEP seemed to be missing the kinds of texts college freshmen and sophomores are expected to read, including philosophical treatises, texts from times and contexts greatly dissimilar to our own, research reports, and, above all, textbooks.

At the same time, NAEP includes some types of passages not referenced in CCSS-ELA. Specifically, NAEP passages draw from a broader range of text types that readers interact with in everyday life, such as popular magazines and newspapers. The CCSS-ELA exemplars do not include as broad an array of reading material found in various contexts of life, including career, college, and citizenship.

Panelists also considered the issue of how well NAEP passages address the CCSS-ELA emphasis on subject-matter reading at Grades 8 and 12. This seems to be an area of difference. For example, at Grade 8 it was noted that the science texts in NAEP did not include scientific explanations but relied heavily on passages from sources like *Highlights*, with little attention paid to the actual science, but more to the social/political/health implications. However, the panelists also noted that, compared to the attention reading in the disciplines receives in CCSS-ELA documents, the actual treatment of disciplinary texts in CCSS-ELA standards appears to be quite generic and does not explicitly address the manner in which texts should be read and evaluated differently from one discipline to another. Based on this observation, the panelists concluded that, even though NAEP does not specifically privilege reading in the disciplines, NAEP's treatment of informational/disciplinary texts might not be all that different from the CCSS-ELA.

Quality. "Quality" of texts as described in CCSS-ELA relates to "literary merit and value," "rich content," and "cultural and historical significance." Quality is also defined in CCSS-ELA through lists of "quality" texts and through excerpts in CCSS-ELA Appendix B from "exemplar" texts. NAEP seeks texts that are characterized by "high quality literary and informational material," and "evidencing the characteristics of good writing, coherence and appropriateness for each grade level." The NAEP text selection criteria intended to lead to the use of quality texts are numerous and detailed. NAEP also provides citations to documents that further define the facets of text quality.

In general, the panel judged that the "quality" of the NAEP texts is similar to that of the CCSS-ELA exemplars. The literary texts in NAEP are comparable to the literary exemplars in CCSS-ELA Appendix B, although the CCSS-ELA exemplars include multiple excerpts from canonical texts at all grade levels, and NAEP has none at Grade 4 and few at Grades 8 and 12. Similarly, the quality of the informational texts in NAEP is comparable to the informational exemplars in CCSS-ELA Appendix B.

Complexity. Appendix A of the CCSS-ELA provides a description of how to evaluate text *complexity* using three broad dimensions: quantitative measures, qualitative criteria, and reader and task factors. Quantitative dimensions focus on various readability formulas, and Appendix A includes a chart showing the computer-generated numeric Lexile levels appropriate for different grade bands. Qualitative dimensions are described in terms of levels of meaning, structure, language conventionality and clarity, and knowledge demands. CCSS-ELA documents suggest that reader and task factors be determined locally with reference to variables such as student motivation and knowledge, as well as the purpose and complexity of the reading task. Because NAEP and CCSS-ELA deal with reader and task factors differently, the panelists attended only to quantitative and qualitative dimensions of complexity called for in CCSS-ELA in their analysis of NAEP reading passages.

The panel found that NAEP fourth- and eighth-grade passages are appropriately complex according to CCSS-ELA *quantitative* criteria. Using the quantitative criteria in CCSS-ELA Appendix A, the overwhelming majority of the Grade 4 passages fall in the fourth- to fifth-grade complexity band, and several Grade 4 passages could be placed in the sixth- to eighth-grade band. Similarly, the quantitative measures of eighth-grade NAEP passages are solidly within the revised quantitative Lexile guidelines in CCSS-ELA. As might be expected, the cross-grade NAEP passages designated for inclusion in both the Grade 4 and Grade 8 assessments are generally below the intended eighth-grade range, but this seems appropriate given NAEP's purposes for cross-grade administration.

At Grade 12, however, the NAEP passages are consistently less difficult than the CCSS-ELA quantitative criteria called for at Grade 12; the cross-grade passages designed to be administered to both Grades 8 and 12 fall within the quantitative guidelines for Grade 8. The difference in 12th-grade passage difficulty between the two frameworks begs the question as to whether CCSS-ELA texts are too challenging, NAEP passages are too easy, or whether other factors account for the discrepancy. It may be that more complex or challenging texts can be used when instructional support is provided, but that text difficulty may need to be reconsidered within CCSS-ELA when associated assessments are developed. Furthermore, text difficulty needs to be considered alongside the demands of specific assessment items about the text in order to determine *comprehension* difficulty. As described in the CCSS-ELA appendixes, some texts that appear easy using quantitative measures can be quite difficult to understand at a deep level, and, conversely, some texts that appear to be difficult can be easy to understand when more surface-level comprehension is expected. One panelist, with many years of experience as a college-level ELA expert, expressed the view that many of the 12th-grade exemplars from the CCSS-ELA are inappropriately difficult for 12th grade and would challenge many college students even near the end of their undergraduate programs.

Although the NAEP passages appear to be largely within the quantitative guidelines provided by CCSS-ELA, there are some qualitative differences in complexity that are apparent across all grade levels when the NAEP passages are compared to the CCSS-ELA exemplars. In general, NAEP appears to employ literature that does not include many complex literary devices, whereas CCSS-ELA exemplars tend to include more texts with this characteristic. When NAEP literary passages do contain some metaphorical language and literary devices, they do not seem to be as complex as CCSS-ELA calls for, and related comprehension items do not seem to require sophisticated interpretation. Turning to informational texts, panelists found that the NAEP informational passages have relatively simple levels of meaning and require less in terms of conceptual understanding. In general, the language of the NAEP passages is syntactically and semantically less complex and includes less technical vocabulary than CCSS-ELA exemplars.

NAEP passages have reader-friendly structures and a conversational style, which often includes an engaging introduction. The narratives often follow simple story grammar; the nonfiction texts are typically chronological or problem/solution. As with many authentic texts, visuals (e.g., photos, charts, graphs, etc.) are sometimes ornamental and sometimes functional in delivering information. In addition, the level of prior knowledge needed to read NAEP passages is generally low, and references to other texts are generally not present. Although it might be helpful to know "a little bit" about the topic, topical knowledge does not seem essential to the comprehension of important ideas.

Finally, the panel noted several cautions for NAEP as it considers issues of text complexity in light of the CCSS-ELA recommendations. First, the CCSS-ELA includes reference to students reading independently as well as with scaffolding and support. The fact that assessments do not provide reading support has implications for how difficult assessment texts should be at various grade levels. Second, data do not yet exist to determine whether an assessment that is aligned with the CCSS-ELA recommendations for complexity would be able to provide estimates of achievement across the proficiency span. Third, the panel noted that NAEP should consider the text-task-reader interaction as it evaluates complexity and not rely solely on quantitative alignment with CCSS-ELA; for individual students, particular NAEP items (or CCSS-ELA tasks) can require complexity of thinking that may or may not be indicated by an analysis of text complexity alone.

Summary of Alignment Between NAEP Reading Items and CCSS-ELA Anchor and Grade-Level Standards (Activity 3)

Anchor Standards and Grade-Level Standards. As indicated previously, the panelists raised concerns about the validity and consistency of grade-level standards following Activity 1. Nevertheless, they tried to use grade-level standards to examine a sample set of items from each grade level. After the grade-level standard exercise and considerable discussion, the panel unanimously agreed that aligning NAEP items with grade-level standards was so problematic that it did not make sense to continue with this part of the analysis. Two issues are relevant here.

First, there were multiple instances in which the grade-level standards associated with a particular anchor standard did not appear to form learning progressions that clearly build across grade levels or are more developmentally complex at the higher grade levels. Moreover, panelists could not identify research that supported the placement of specific knowledge/skills at specific grade levels or the developmental progression of a specific anchor standard across the grades.

For example, the grade-level standards developed for Anchor Standard R1 emphasize different skills across Grades 3–5, and there is no clear sequence of complexity or difficulty across the grades.

Anchor Standard R1—Read closely to determine what the text says explicitly and to make logical inferences from it; cite specific textual evidence when writing or speaking to support conclusions drawn from the text.

> Grade 3—Ask and answer questions to demonstrate understanding of a text, referring explicitly to the text as the basis for the answers.
>
> Grade 4—Refer to details and examples in a text when explaining what the text says explicitly and when drawing inferences from the text.
>
> Grade 5—Quote accurately from a text when explaining what the text says explicitly and when drawing inference from the text.

In other cases, such as Anchor Standard R9 (analyze how two or more texts address similar themes or topics in order to build knowledge or to compare approaches authors take), the associated grade-level standards remain identical across several grades (Grades 6, 7, and 8).

Second, sometimes the grade-level standards include so much specificity (which is also not consistent across grade levels) that it was difficult, if not impossible, to reliably identify a standard that closely aligned with each NAEP item. For example, the grade-level standards for reading Anchor Standard 3 focus on identifying and describing characters, settings, and major events in stories at kindergarten and Grades 1, 4, and 5; however, the standards for Grades 2 and 3 focus only on characters.

Similarly, some of the grade-level standards associated with Anchor Standards 4 and 9 identify a particular genre or specific types of texts only at specific grades:

> Anchor Standard R4 (Grade 4)—Determine the meaning of words and phrases as they are used in a text, including those that allude to significant characters found in mythology (e.g., Herculean).
>
> Anchor Standard R9 (Grades 9–10)—Analyze seminal U.S. documents of historical and literary significance (e.g., Washington's Farewell Address, the Gettysburg Address, Roosevelt's Four Freedoms speech, King's "Letter from Birmingham Jail"), including how they address related themes and concepts.

As a result of efforts to try to align NAEP reading items with grade-level standards, the panelists determined that it would be most appropriate to examine reading items in relation to the anchor standards for reading (National Governors Association & Council of Chief State School Officers, p. 10) that apply to all grades, K–12. Furthermore, the panel determined that it was necessary to interpret the anchor standards broadly and conceptually rather than specifically and procedurally. As several of the examples above demonstrate, even at the anchor standards level, the standards often include multiple parts or specifics that would be difficult to find in a single NAEP reading item. For example, Anchor Standard R2 states, "Determine central ideas or themes of a text and analyze their development; summarize the key supporting details." Often, a NAEP reading item addresses either the first or second part of this standard but not both.

Item Alignment. Across the pool of items at all three grade levels, the majority of items were identified through consensus as "*strongly aligned*" to one of the first five anchor standards for reading. Although there was some variability across grade levels, the overall percentage of items that was determined to be strongly aligned with each of the first five standards is listed below:

Key Ideas and Details

> R1—Read closely to determine what the text says explicitly and to make logical inferences from it; cite specific textual evidence when writing or speaking to support conclusions drawn from the text. (36 percent of NAEP items strongly aligned)
>
> R2—Determine central ideas or themes of a text and analyze their development; summarize the key supporting details and ideas. (13 percent of NAEP items strongly aligned)
>
> R3—Analyze how and why individuals, events, and ideas develop and interact over the course of a text. (8 percent of NAEP items strongly aligned)

Craft and Structure

> R4—Interpret words and phrases as they are used in a text, including determining technical, connotative, and figurative meanings, and analyze how specific word choices shape meaning or tone. (19 percent of NAEP items strongly aligned)

> R5—Analyze the structure of texts, including how specific sentences, paragraphs, and larger portions of the text (e.g., a section, chapter, scene, or stanza) relate to each other and the whole. (10 percent of NAEP items strongly aligned)

In addition, the majority of reading items (75 percent) was judged to be related to more than one of these five anchor standards; these were double or triple coded to indicate they were also *moderately* or *weakly* aligned with multiple standards. Considering the nature of the NAEP reading assessment, the alignment with these five reading anchor standards seems appropriate.

The reading anchor standards that are least, or not at all, aligned with the NAEP reading assessment fall under the category of integration of knowledge and ideas and specifically address using and evaluating multimedia texts (Anchor Standard R7), evaluating arguments and claims (Anchor Standard R8), and using multiple texts to build knowledge (Anchor Standard R9). The panel suggested that NAEP might consider new strategies for addressing some aspects of these standards but was mindful of the challenges that would be introduced in the NAEP context by the role of prior knowledge in these standards, especially in relation to disciplinary reading.

The panel also found that a small number of reading items could be aligned with one or more of the language and writing anchor standards. Specifically, vocabulary items that are integrated into the main reading NAEP are often aligned with:

> L4—Determine or clarify the meaning of unknown and multiple-meaning words and phrases by using context clues, analyzing meaningful word parts, and consulting general and specialized reference materials, as appropriate.

The panel also noted instances in which short-constructed response and extended-constructed response items in the NAEP reading assessment are aligned with both writing and reading standards. Writing Anchor Standards W1 and W9 are most likely to be assessed as part of NAEP reading and to offer the possibility of double scoring (for reading and writing).

> W1—Write arguments to support claims in an analysis of substantive topics or texts, using valid reasoning and relevant and sufficient evidence.
>
> W9—Draw evidence from literary or informational texts to support analysis, reflection, and research.

Overall Reading Conclusions and Recommendations

1. Panel members find that many aspects of the current NAEP reading assessment are consistent with conceptualizations of the reading process found in the research and in CCSS-ELA documents:
    - Cognitive focus aligned with research
    - Broad range of text types
    - High quality and appropriate length of texts used in assessment
    - Attention to literary and informational comprehension

- Use of text pairs
- Attention to reader-text interactions in item development
- Inclusion of writing in response to reading
- Parsimony and elegance in crafting questions to align with specific texts
- Thoughtful, meaningful items—well sequenced and crafted

As a result, the panel is cautiously optimistic that, with attention to the specific issues identified in this report and a systematic program of special studies to inform future assessments, NAEP could continue to serve as an independent monitor of student achievement in an era of CCSS.

Panelists also recognize the different purposes of NAEP and CCSS-ELA and feel strongly that NAEP should retain its independence from any particular curriculum and serve as a general assessment of reading comprehension. In addition, NAEP's ability to sample a wide variety of student performance on a range of texts and tasks through its matrix sampling is consistent with the range of reading performances expected by CCSS-ELA and should be preserved.

The panel believes that NAEP could build upon these strengths as they consider several recommendations and issues to enhance its relevance to the CCSS-ELA and reflect emerging areas of reading assessment. These recommendations follow

2. CCSS-ELA has made clear the expectation to increase the "rigor" and "complexity" of texts students read at each grade level as well as progressively across grade levels. In contrast, the NAEP approach is to use texts that are judged to be within the currently recognized range of difficulty for the targeted grade. Nevertheless, the panel finds that the NAEP reading selections at Grades 4 and 8 generally fall within (or above) the quantitative ranges called for in the CCSS-ELA, while the Grade 12 NAEP passages are consistently less difficult than called for by CCSS-ELA quantitative indexes. The panel suggests that NAEP consider passages that include more complexity at the upper grade levels in terms of perspective taking, bias, competing accounts, trustworthiness of the sources, craft, conceptual issues, etc., that might allow for assessing deeper, closer reading. The panel cautions, however, that text difficulty should not be judged solely on quantitative measures—a position supported by both CCSS-ELA and NAEP.

   Three issues should be considered in regard to text complexity: (1) differences in the level of complexity that students can handle in texts used for *instruction* versus texts used for *assessment*, (2) NAEP's historical difficulty obtaining valid data for low-performing students, and (3) the interplay of reading items/task and text in determining reading comprehension difficulty. NAEP should explicitly consider each of these three issues as it deals with text complexity in future assessments.

3. The panel finds that the NAEP framework for constructing items to align with cognitive targets is compatible with the CCSS-ELA anchor standards and should continue to be used for item development. There is not a one-to-one alignment of cognitive targets to anchor standards because CCSS-ELA standards describe what students should be able to do rather than articulate the mental processes or thinking that underlie these competencies. In general, however, the locate and

recall items align with reading Anchor Standard 1 and the integrate/interpret and critique/evaluate items fall across all of the other anchor standards (2–9).

4. Panel members caution NAEP to be cognizant of the lack of research base, inconsistencies, and specificity of the "learning progressions" embodied by the K–12 grade-level standards in CCSS-ELA. The panel advises NAEP to use the reading anchor standards, which are research based and consistent across grade levels, to determine alignment, rather than the grade-level standards. Furthermore, the panel suggests that NAEP interpret the anchor standards broadly and conceptually rather than specifically and procedurally. Because some of the anchor standards include multiple parts or specifics that could confound or constrain test development (and instruction), we encourage NAEP to bring a "generous" reading to the anchor standards as they consider issues of alignment.
5. NAEP items align most often with CCSS-ELA Anchor Standards 1–5. Anchor Standards 6–9 are less well represented. The panel suggests that NAEP examine how it might place additional focus on assessing point of view, bias, perspectives, and such (Anchor Standard 6), which may require selecting different types of texts as well as crafting new types of items. In addition, the panel suggests that NAEP explore possible strategies and limitations for expanding coverage of Anchor Standards 7–9 (which represent integrating of knowledge and ideas), even though these standards may be difficult to assess in NAEP because they require students to draw on prior knowledge and build new knowledge using text.
6. Many of the NAEP short-constructed and extended-constructed response reading items are aligned with both CCSS-ELA reading and writing anchor standards. Given the emphasis on writing in response to text in the CCSS-ELA writing standards, the panelists suggest that NAEP investigate the possibility of double scoring these items for both reading and writing.
7. An important area of difference between CCSS-ELA and NAEP is the manner in which disciplinary reading is addressed. The *conceptual framing* for CCSS-ELA positions disciplinary reading for the purposes of building new knowledge in the specific discipline. In contrast, the NAEP Reading Framework subsumes disciplinary texts under "informational texts," sampled from varied content areas. The treatment of these texts in NAEP assumes little prior knowledge and relies on general comprehension questions rather than more subject-matter specific comprehension. Although these differences exist in the *framing sections* of CCSS-ELA and NAEP documents, the panel finds them to be far less evident when comparing NAEP items and CCSS-ELA anchor standards or grade-level standards. As a result, the panel was uncertain about the degree to which specific disciplinary reading outcomes would be operationalized when the CCSS-ELA standards are implemented.

   The panel suggests that NAEP adopt a more systematic treatment of discipline-specific texts in the text selection process. However, at the same time, it is unclear what the focus should be for assessing these texts—general understanding or disciplinary knowledge building, especially given the difficulties of attending to issues of prior knowledge and topic familiarity in an assessment like NAEP. One suggestion might be to use cross-text blocks to assess knowledge building across disciplinary texts (minimizing prior knowledge) and to use other informational texts to assess more general comprehension. Overall, the issue of disciplinary text—the purpose, outcomes, and text selection—needs to be addressed and clarified in future NAEP frameworks and assessments.

8. There is a general sense that NAEP's practice of restricting text selection to material written for general audiences may have had the overall effect of constraining the texts that appear on NAEP more than intended. The panelists suggest that NAEP would be more consistent with the CCSS-ELA if it were to consider inclusion of more dense text and texts that are representative of textbook or workplace reading—these are typically less explicit and controlled than texts currently used in NAEP. At the same time, NAEP needs to accommodate a wide range of reading abilities, including students performing at and below the *Basic* achievement level, especially at fourth grade.
9. The CCSS-ELA documents include attention to classic literature, well-known documents, and popular texts. Attention to these sorts of texts may be appropriate in an instructional setting, however, issues of familiarity (prior knowledge) and length are likely to make these types of texts inappropriate for inclusion in NAEP. NAEP might want to clarify for CCSS-ELA consumers how and why texts used for assessment must necessarily differ in some respects from those used in school and the workplace.
10. NAEP should consider using digital text and information displayed in graphs and charts. These text types are called for in CCSS-ELA, and panelists generally feel that a current (and forward looking) assessment of 21st century literacy should include online reading and research. They suggest that NAEP consult existing research regarding the similarities and differences between "traditional" and Internet/online reading to inform future assessment development. Some panelists also feel that NAEP should reconsider the role and nature of more procedural/functional texts both in the real-world and academic contexts as well as more 12th-grade passages that align with the types of texts typically assigned in college.
11. There are differences in how NAEP and CCSS-ELA address vocabulary. NAEP focuses on a particular type of vocabulary and format for assessment purposes—word meaning in the context of a given passage; CCSS-ELA takes a much broader perspective on vocabulary as an essential element of ELA with a definite emphasis on discipline-specific and academic vocabulary. The panel recommends that NAEP consider both the reading anchor standards and the language anchor standards as it evaluates its existing approach and possible new approaches to vocabulary assessment.
12. The CCSS-ELA include K–5 standards for foundational skills, while NAEP assessments target comprehension beginning at Grade 4. The panelists caution that fourth-grade assessments developed specifically to measure CCSS-ELA may include items testing foundational skills as well as literature/informational standards. Because foundational skills are not part of NAEP, comparisons of fourth-grade performance across different assessments may need to take this into account.

Writing Findings

Summary of the Comparison Between NAEP Writing Framework and CCSS-ELA Documents (Activity 1)

The following describes similarities and differences between the NAEP Writing Framework and the CCSS-ELA in the areas of definition/conceptualization, domains of writing, dimensions of writing, incorporation of technology, writing processes, and range of writing. The focus is on similarities and differences with implications for NAEP's role as an independent monitor, after acknowledging that there are important differences in the purposes of these documents.

Definition/Conceptualization. Both NAEP and CCSS-ELA emphasize the situated, social nature of writing. NAEP, for example, defines writing as "…a complex, multifaceted and purposeful act of communication…" (National Assessment Governing Board, 2010b, p. 3) and explains that "Writing is a social act—not only do writers always write for a purpose, but they usually write to communicate ideas to others" (National Assessment Governing Board, 2010b, p. 4). In keeping with this view of writing, both documents emphasize the importance of audience, purpose, and task in writing, and both documents treat rhetorical flexibility as an important component of skilled performance.

An important difference in conceptualization is that while the CCSS-ELA standards are integrated in multiple ways, the treatment of ELA in NAEP is not integrated. Reading and writing are treated in separate frameworks in NAEP, and there is little integration across the modes in NAEP assessments with the exception of the use of some "constructed response" writing in the NAEP assessment of reading. In contrast, integration of the modes is a "key design" consideration in the CCSS-ELA. CCSS-ELA integrates reading, writing, speaking, and listening, and the individual standards reflect this integration. For example, as articulated in Anchor Standard W9, students are expected to "Draw evidence from literary or informational texts to support analysis, reflection, and research." Because the standards are integrated, most of the sample writing tasks in the CCSS-ELA are integrated as well, requiring students to read (or view or listen) and then write in response to a text or set of texts. NAEP does not assess these sorts of integrated tasks. Although brief reading passages may accompany some writing prompts in the NAEP assessment of writing, they serve primarily as stimuli for writing rather than as material for analysis or as sources of information. CCSS-ELA, in contrast, emphasizes writing about reading and writing from sources.

CCSS-ELA also integrates writing across the disciplines. Although the NAEP framework deals with very broad domains of writing, it does not address the special skills and strategies of writing in the disciplines. While the NAEP framework is confined to writing in ELA, CCSS-ELA spans writing in the content areas of history/social studies, science, and technical subjects as well. In the writing standards for literacy in history/social studies, science, and technical subjects for Grades 6–12, students are expected to write about discipline-specific content, be aware of the norms and conventions of each discipline, and acquire and use discipline-specific vocabulary.

Domains of Writing. Both NAEP and CCSS-ELA describe similar, broad domains of writing, although they describe them in different terms. NAEP defines the domains in terms of three broad purposes for writing: *to persuade, to explain,* and *to convey experience*. CCSS-ELA describes them as types of writing: *arguments, informative/explanatory texts,* and *narratives*. Both NAEP and CCSS-ELA acknowledge that the identified domains subsume a wide range of products, genres, and forms. Both also acknowledge that the borders of the domains are porous; that is, that writers create texts that blend types using strategies such as embedding narrative elements within a largely expository structure or employing narrative structures for informational, explanatory, or persuasive purposes. Finally, both NAEP and CCSS-ELA identify similar domains of writing when describing how relative emphasis should change across the grade levels. Both recommend increased emphasis in the upper grades on writing *to explain* (*informational/explanatory writing* in CCSS-ELA terms) and *to persuade* (*argument* in CCSS-ELA terms).

Dimensions of Writing. Both NAEP and CCSS-ELA identify and discuss essentially the same valued dimensions of effective writing: development, organization, language facility, and conventions. These dimensions are articulated in the NAEP Writing Framework as criteria for evaluating responses and are threaded throughout the CCSS-ELA documents, in the anchor standards for writing and language, as well as the annotated samples of student writing in CCSS-ELA Appendix C.

Incorporation of Technology. Both NAEP and CCSS-ELA address the integral role that technology now plays in writing. However, in the NAEP framework, the role played by technology is currently limited to students' use of a computer "to compose and construct their responses using word processing software…with the option to use commonly available tools" (National Assessment Governing Board, 2010b, p. 7). CCSS-ELA conveys a more expansive and comprehensive view of the role played by technology and digital tools—one that cuts across reading, writing, speaking, listening, and includes its use in research along with the expectation that students will "use technology and digital media strategically and capably" (National Governors Association & Council of Chief State School Officers, 2010, p. 7).

Writing Processes. NAEP and CCSS-ELA both acknowledge the role that writing processes play in the improvement of writing. However, while NAEP provides computer tools for drafting, revising, and editing, there are constraints on NAEP procedures that privilege first-draft writing and make time for significant planning and revision unlikely. CCSS-ELA, on the other hand, treats the management of writing processes, including collaboration with others, as an important component of writing ability that develops over time (see Anchor Standards W5 and W6). Performance expectations for what students are expected to be able to do in regard to writing processes are further elaborated in the CCSS in the K–12 grade-level standards. By Grades 11–12, students are expected to be able to "Develop and strengthen writing as needed by planning, revising, editing, rewriting, or trying a new approach, focusing on addressing what is most significant for a specific purpose and audience…" (National Governors Association & Council of Chief State School Officers, 2010, p. 46).

Range of Writing. NAEP assessments collect on-demand writing samples, and students have only 30 minutes to complete each writing sample. In contrast, the CCSS-ELA explicitly calls for students to write in both short and extended time frames (Anchor Standard W10). Extended time frames are more appropriate for the kinds of complex, integrated reading/writing tasks that CCSS-ELA emphasizes, and extended time frames can also accommodate more attention to writing processes such as planning, revising, and editing.

Summary of Comparison Between NAEP Writing Scoring Guides, Anchor Papers, and Prompts and CCSS-ELA Documents (Activity 2)

The following summarizes the results of analyses of the NAEP scoring guides, anchor papers, and prompts for writing in relation to the CCSS-ELA. A total of 80 prompts, 8 scoring guides, and 6 sets of anchor papers from the 2011 assessment (Grades 8 and 12) and pilot test (Grade 4) were used for this analysis.

Scoring Guides. NAEP provides focused holistic scoring guides for each of the three writing purpose assessed by NAEP. Panelists observed that these three types of scoring guides aligned well with expectations for the text types described in the CCSS-ELA anchor standards for writing. Although the labels are sometimes different, the features emphasized in the three dimensions of the NAEP scoring guides correspond very closely to those identified in CCSS-ELA as characterizing particular text types. The NAEP scoring guides for *persuade,* for example, evaluate text on the same features that CCSS highlights as required for a well-constructed *argument* (clear position, logical reasoning, strong evidence). Similarly, the *explain* scoring guides emphasize clarity and accuracy of *explanation*; and the *convey* scoring guides mirror the emphasis in CCSS *narrative* on effective, well-chosen details to convey experiences. Furthermore, the scoring guide analysis revealed an emphasis on audience and purpose that aligns well with CCSS-ELA Anchor Standard W4: “Produce clear and coherent writing in which the development, organization, and style are appropriate to task, purpose and audience.” Audience is explicit in all three types of guides in reference to both development of ideas and language facility: “Voice and tone are well controlled, showing an awareness of purpose and audience.”

However, the panelists also observed that: (1) CCSS-ELA specifies narrative structures, while the NAEP scoring guides leave the *To Convey* organization open; (2) CCSS-ELA requires the development of discipline-specific stances under explanation, while the NAEP scoring guides for *To Explain* appear less rigorous because they do not; and (3) CCSS-ELA specifies more sophisticated techniques of argument at the upper grades (such as counterclaims and careful evaluation of evidence) than are apparent in the NAEP scoring guides for *To Persuade*. While the NAEP scoring guides reflect dimensions of writing valued in the CCSS-ELA, and while they emphasize audience and purpose, they do not align well to the integrated academic, disciplinary, and evidence rich stances and tasks that CCSS-ELA emphasizes, particularly in the upper grades (11–12).

Anchor Papers. Panel members observed that NAEP anchor papers—all of which were produced “on demand” under timed and supervised testing conditions—are considered "first-draft" writing by NAEP. CCSS-ELA sample grade-level papers, on the other hand, were not produced consistently in uniform "testing/assessment" environments. Some of the CCSS-ELA sample papers were produced in extended time frames and benefited from feedback from teachers and peers. Other papers were produced under testing conditions that may have been different from those of NAEP. This made it somewhat difficult to compare the CCSS-ELA samples directly with the NAEP anchor papers. While there are some individual papers in the CCSS-ELA samples that are similar in quality to NAEP anchors, there are others that are widely divergent, particularly the CCSS-ELA samples at the upper grades that were produced in extended time frames. This finding suggests a lack of alignment between NAEP and part of the CCSS-ELA standard for range, W10: "Write routinely over extended time frames (time for research, reflection, and revision) and shorter time frames (a single sitting or a day or two)…."

Prompts. Panelists observed that the pool of writing prompts for the three purposes assessed by NAEP are broadly representative of the text types and purposes described in the CCSS-ELA anchor standards. In addition, the prompt coding revealed that the pool of prompts incorporates a wide variety of audiences (ranging from familiar to more distant), a range of publication types (websites, newspapers, online forums, books), a variety of genres and forms (letters, essays, reviews, reports, speeches), and a variety of topics and tasks. This finding suggests a relatively close alignment between NAEP and part of Anchor Standard W10: "Write…for a range of tasks, purposes, and audiences."

However, the panel also observed, and the coding of the prompts confirmed, that the pool of NAEP prompts relies primarily on personal experience or general background knowledge. The pool of prompts does not include the more extended kinds of tasks that would require "short as well as more sustained research projects" (Anchor Standard W7) or tasks that would require students to "integrate information" gathered "from multiple print and digital sources" (Anchor Standard W8). As pointed out in earlier sections, the range of the NAEP pool of tasks is limited by the constraints of the testing situation (30 minutes).

Summary of Alignment Between NAEP Writing Prompts and CCSS-ELA Anchor and Grade-Level Standards (Activity 3)

After some discussion, and in light of the concerns about the validity and consistency of grade-level standards raised by the Reading Panel, the Writing Panel decided that trying to locate NAEP prompts in relation to the grade-level standards would not be a useful activity. Instead, they decided to analyze the prompts in relation to the CCSS-ELA anchor standards and to gather information about the knowledge demands and range of audiences associated with the NAEP prompts reported previously.

As noted above, because NAEP reading items often require readers to draw on multiple sources of information, interpret text, and use a variety of skills and strategies, and because writing prompts sometimes appear to elicit more than one type of writing, reading items and writing prompts sometimes aligned with multiple CCSS-ELA standards. Therefore, based on their expert judgment, panelists rated each item/prompt as *strongly aligned, moderately aligned,* or *weakly aligned* with specific standards. This provided an opportunity for panelists to go beyond a simple matching to indicate degree of alignment; it permitted them to evaluate the strength of alignment across multiple standards.

Prompt Alignment. Across the pool of prompts coded at all three grade levels, all of the prompts were identified through consensus as *strongly aligned* to at least one of the first three anchor standards for writing. The overall percentage of prompts coded as strongly aligned with each of the first three standards is listed below:

> W1—Write arguments to support claims in an analysis of substantive topics or texts, using valid reasoning and relevant and sufficient evidence. (32 percent of NAEP prompts strongly aligned)
>
> W2—Write informative/explanatory texts to examine and convey complex ideas and information clearly and accurately through the effective selection, organization, and analysis of content. (35 percent of NAEP prompts strongly aligned)
>
> W3—Write narratives to develop real or imagined experiences or events using effective technique, well-chosen details, and well-structured event sequences. (33 percent of NAEP prompts strongly aligned)

Three of the prompts (4 percent) were coded as strongly aligned to more than one of these first three CCSS-ELA anchor standards and 23 of the prompts (29 percent) were coded as strongly aligned to one and weakly aligned to another. Panelists' comments indicated that prompts were double coded when they were viewed as being likely to elicit more than one type of writing. Some *To Convey* prompts, for example, appeared as likely to elicit some combination of description and explanation as to elicit narrative, particularly when the prompt asked students to convey what something was like. Some *To Persuade* prompts appeared as likely to elicit explanation as persuasion.

All of the prompts (100 percent) were coded as moderately aligned with another five of the CCSS-ELA anchor standards: writing Anchor Standards W4 and W5 and language Anchor Standards L1, L2, and L3. During whole-group discussion, these five writing and language standards were grouped by consensus into what the panel called a "bundle" and recorded as moderately aligned because the standards applied to all types of writing, more or less equally.

> W4—Produce clear and coherent writing in which the development, organization, and style are appropriate to task, purpose, and audience.
>
> W5—Develop and strengthen writing as needed by planning, revising, editing, rewriting, or trying a new approach.
>
> L1—Demonstrate command of the conventions of standard English grammar and usage when writing or speaking.
>
> L2—Demonstrate command of the conventions of standard English capitalization, punctuation, and spelling when writing.

> L3—Apply knowledge of language to understand how language functions in different contexts, to make effective choices for meaning or style, and to comprehend more fully when reading or listening.

Finally, a few of the prompts were also coded as weakly aligned to Anchor Standard L5 and aspects of Anchor Standard L6 related to vocabulary use.

> L5—Demonstrate understanding of figurative language, word relationships, and nuances in word meanings.
>
> L6—Acquire and use accurately a range of general academic and domain-specific words and phrases sufficient for reading, writing, speaking, and listening at the college and career readiness level; demonstrate independence in gathering vocabulary knowledge when considering a word or phrase important to comprehension or expression.

More specifically, in these cases, the prompts appear especially likely to elicit particular kinds of language specified in the standards, such as figurative language (Anchor Standard L5) or general academic and domain-specific words and phrases (Anchor Standard L6).

Several of the writing anchor standards are not aligned with the NAEP prompts because they refer to competencies not addressed by the NAEP writing assessment:

> W6—Use technology, including the Internet, to produce and publish writing and to interact and collaborate with others.
>
> W7—Conduct short as well as more sustained research projects based on focused questions, demonstrating understanding of the subject under investigation.
>
> W8—Gather relevant information from multiple print and digital sources, assess the credibility and accuracy of each source, and integrate the information while avoiding plagiarism.
>
> W9—Draw evidence from literary or informational texts to support analysis, reflection, and research.
>
> W10—Write routinely over extended time frames (time for research, reflection, and revision) and shorter time frames (a single sitting or a day or two) for a range of tasks, purposes, and audiences.

As noted above, in general, most of the panelists did not find trying to locate NAEP prompts in relation to the grade-level standards to be a useful activity. However, the Grade 12 group did attempt to code some of them, and the attempt informed the later deliberations of the panel. The Grade 12 group observed that, when judged against the grade-level standards, some of the NAEP 12th-grade prompts, in particular the *To Explain* and *To Persuade* prompts, appear more appropriate for lower grade levels (i.e., Grades 6, 7, and 8) than for Grade 12. They also observed that some of the prompts could be considered "on grade" only if the limitations of the test situation itself were taken into account. For example, to fulfill the expectations of the grade-level standard for argument at Grades 11–12, students would have to

"Introduce precise, knowledgeable claim(s), establish the significance of the claim(s), distinguish the claim(s) from alternate or opposing claim(s), and create an organization that logically sequences claims(s), counterclaims, reasons, and evidence." Students would also have to "Develop claim(s) and counterclaims fairly and thoroughly, supplying the most relevant evidence for each while pointing out the strengths and limitations of both in a manner that anticipates the audience's knowledge level, concerns, values, and possible biases." The panelists questioned whether it would be possible for students to fulfill these expectations in the 30 minutes allotted for writing to a prompt with access only to remembered evidence.

Overall Writing Conclusions and Recommendations

1. Panel members find much to commend in the current NAEP writing assessment, reflecting, as it does, a conceptualization of writing found in both research and in the CCSS-ELA documents. Both NAEP and CCSS-ELA present writing as a social, communicative activity; emphasize the importance of audience, purpose, and task; and treat rhetorical flexibility as an important component of skilled performance. NAEP and CCSS-ELA are aligned in other important ways as well. They address similar broad domains of writing, and identify and discuss essentially the same valued characteristics of effective writing: development of ideas, organization, and language facility and conventions. The NAEP scoring guides emphasize adapting writing to purpose, task, and audience (CCSS-ELA Anchor Standard W4), and the features highlighted in the three separate NAEP guides for *To Persuade*, *To Explain*, and *To Convey* are generally parallel to the features emphasized in the three broad types of writing described in CCSS-ELA writing standards 1, 2, and 3 (argument, informational/explanatory and narrative). The NAEP pool of prompts is also generally aligned with the CCSS-ELA "text types and purposes" described in the first three CCSS-ELA writing anchor standards. As noted above, panelists also observed that the pool of prompts contains a broad range of audiences and forms, an aspect of range described in CCSS-ELA Anchor Standard W10. The panel concludes that NAEP should build upon these features as they consider ways to enhance NAEP's alignment with CCSS-ELA, including measuring aspects of CCSS-aligned curricula not well addressed by the current assessment.

   The standards-to-framework and standards-to-assessment analyses also reveal several gaps in alignment between NAEP and CCSS-ELA. The panel concludes that NAEP should consider several recommendations to enhance its alignment with CCSS-ELA. These recommendations follow.

2. The CCSS-ELA clearly emphasizes integration of the language arts, while NAEP does not. In particular, CCSS-ELA emphasizes writing about reading and writing from sources (writing based on research). These emphases are threaded throughout the standards and featured prominently in Anchor Standard W9: "Draw evidence from literary or informational texts to support analysis, reflection, and research." Many of the example tasks and standards in the CCSS-ELA documents involve writing (or speaking) about what has been read. Tasks that require writing about reading and/or writing based on research are currently not included in the NAEP assessment. Instead, NAEP tasks rely primarily on background knowledge and personal experience. Panelists recommend that

NAEP consider including writing in response to print and/or nonprint texts and writing based on research (writing from sources), either by including such items in the assessment itself or by conducting a systematic collection of samples of such tasks that students have done in school or in curriculum embedded assessments to compare with students' performances on other sorts of tasks.

3. The CCSS-ELA is explicit in acknowledging that the teaching of writing is a shared responsibility across disciplines, assuming a single teacher of all subjects through Grade 5, and separate subjects (with separate writing standards) from Grade 6 on. In the CCSS-ELA, writing activities within the disciplines are integrated with content learning. Furthermore, the CCSS-ELA language standards, which apply to writing as well as reading, speaking, and listening, distinguish between general, academic, and domain-specific vocabulary (e.g., technical vocabulary within the disciplines). While the NAEP Writing Framework acknowledges the situated nature of writing and its importance in all disciplines, and while the NAEP writing assessment deals with purposeful writing skills and general and academic vocabulary, it does not address the special skills, strategies, or domain-specific vocabulary associated with writing in the disciplines. Writing from substantive disciplinary content is an important literacy skill not presently addressed in NAEP. Panelists recommend that NAEP consider including writing tasks, especially those that are structured around deep knowledge of subject matter, in NAEP's discipline-specific assessments, either as part of the regular NAEP assessment or as a probe study. Furthermore, NAEP should consider tracking domain specific vocabulary along with general vocabulary.
4. At present, NAEP limits the role that technology plays in assessment to students' use of a computer "to compose and construct their responses using word processing software…with the option to use commonly available tools." CCSS-ELA, on the other hand, conveys a portrait of college- and career-ready students who "use technology and digital media strategically and capably…" who "are familiar with the strengths and limitations of various technological tools and mediums" and who "can select and use those best suited to their communication goals." Panelists recommend that NAEP consider expanding the use of technology in writing, either as part of the regular NAEP assessment or as a probe study. They also note that if students are to have a wider range of technology-enabled options in the regular NAEP assessment, they would need to have more time to compose as well as to understand the options presented in whatever platform is used in the assessment.
5. At present, NAEP allows students 30 minutes to respond to a prompt. While NAEP thus assesses on-demand writing in an abbreviated time frame, CCSS-ELA emphasize writing under a variety of conditions and convey specific expectations for students' use of writing processes such as planning, revising, editing, and rewriting. While the NAEP Writing Framework acknowledges the roles played by writing processes in the improvement of writing, actually allowing time for significant revising and editing in the NAEP regular assessments would mean extending the current time frames. Similarly, tasks that require substantial reading before writing would require more time than currently allowed. Panelists recommend that NAEP consider investigating ways to allow different amounts of time for different kinds of

tasks. Providing more extended time frames could encourage revising and/or accommodate some of the more complex reading/writing tasks found in the CCSS-ELA. Panelists also suggest that NAEP consider conducting special studies of extended tasks as they are being used in schools.

Summary Conclusions by the Reading and Writing Panels

The Reading and Writing Panel members recognize the different purposes of NAEP and CCSS-ELA and feel strongly that NAEP should retain its independence from any particular curriculum and serve as a general assessment of reading and writing performance. Overall, the panels are cautiously optimistic that, with attention to the specific issues identified in this report and a systematic program of special studies to inform future assessments, NAEP could continue to serve as an independent monitor of student achievement in an era of CCSS. In the area of reading assessment, NAEP should consider revisions related to reading and knowledge building in the disciplines, text selection (including digital texts) and complexity, integration of reading and writing, and assessment of academic vocabulary. In the area of writing, NAEP should consider revisions related to writing in response to text and research, integrating writing into discipline-specific assessments, expanding the use of technology, and providing more extended time for writing to accommodate different types of writing tasks and conditions.

The panels also judge that NAEP could serve as an intellectual tool to promote the design and use of quality assessments apart from CCSS. With attention to the recommendations in this report, NAEP could be in an excellent position to lead the way for forward-looking reading and writing assessment. Indeed, the panels encourage NAEP to consider the future and changes in literacy demands as they conceptualize literacy assessment. NAEP's ability to sample a wide variety of student performance on a range of texts and tasks through its matrix sampling design is consistent with the range of literacy performances expected by CCSS-ELA and places it in an excellent position to engage in the kind of special studies needed, both to assess these complex standards and to serve as an external point of comparison useful to future revisions of the CCSS-ELA.

Because of the timing of the study, the panels could not determine the degree of alignment between NAEP and new assessments under development by Smarter Balanced and PARCC. This is an important consideration because the ability of NAEP to serve as an independent monitor may be judged by a comparison of student achievement on NAEP with achievement on the new assessments; alternatively, it may be judged by the degree of alignment between NAEP assessments and the framing concepts in the CCSS-ELA documents rather than simply the new assessments. Furthermore, at this point in time, the potential impact of CCSS documents and specific standards on curriculum and assessment is unknown, most especially the integration of reading and writing, technology, and knowledge building in the disciplines. The CCSS documents integrate writing and reading across the disciplines, call for extended writing tasks that involve reading and research, and convey the expectation that students will use technology "strategically and capably." The extent to which these elements will be operationalized in the new assessments and/or in classroom instruction is not clear but, the panels believe these issues are integral to the next iterations of literacy assessment and to students' success in their careers and college. Consequently, there will need to be additional studies to evaluate the fit of new CCSS assessment items to CCSS standards and to compare CCSS assessment items to NAEP items. In cases in which NAEP and new CCSS assessment do not align, it will be important to look at the areas of nonalignment found in the studies reported here as a possible explanation for the nonalignment. Furthermore, it will be important to define the specific contribution NAEP should make and the role it should play. These issues will need to be addressed as new assessments are implemented and evaluated and as curriculum and instruction change to reflect successful implementation of CCSS-ELA.

The Reading and Writing Panels appreciate the opportunity to analyze NAEP in light of the CCSS-ELA and the literacy demands of the 21st century. Several of our findings may provide the basis for immediate changes, and others may provide the impetus for special studies that could inform future NAEP assessments and issues of alignment with CCSS-ELA. We hope that the detailed analyses and recommendations will provide the NVS Panel with both information and perspectives that will help it move forward.

References

National Assessment Governing Board (September, 2010a). *Reading Framework for the 2011 National Assessment of Educational Progress*. Retrieved from http://www.nagb.org/content/nagb/assets/documents/publications/frameworks/reading-2011-framework.pdf

National Assessment Governing Board (September, 2010b). *Writing Framework for the 2011 National Assessment of Educational Progress*. Retrieved from http://www.nagb.org/content/nagb/assets/documents/publications/frameworks/writing-2011.pdf

National Governors Association & Council of Chief State School Officers (2010). *Common Core State Standards for English language arts and history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

Appendix A. Reading and Writing Panelists

Reading Panelists

Peter Afflerbach*
University of Maryland

Patricia Alexander*
University of Maryland

Judith Langer
University at Albany, State University of New York

Carol Lee
Northwestern University

P. David Pearson*
University of California, Berkeley

Cynthia Shanahan
University of Illinois at Chicago

Terry Underwood
California State University, Sacramento

Sheila Valencia*, **
University of Washington

Karen Wixson*, **
The University of North Carolina at Greensboro

Writing Panelists

Arthur Applebee*
University at Albany, State University of New York

Charles Bazerman
University of California, Santa Barbara

Beverly Chin*
University of Montana

Elyse Eidman-Aadahl*
National Writing Project

Sally Hampton*
Pearson Education

Sandra Murphy*, **
University of California, Davis

Peggy O'Neill
Loyola University Maryland

Dorothy Strickland
Rutgers University

Carl Whithaus
University of California, Davis

Note: All panelists participated in Activities 2 and 3.
* Also participated in Activity 1.
** Study Leads

The Relevance of Learning Progressions for NAEP

Lorrie Shepard
University of Colorado Boulder

Phil Daro
Strategic Education Research Partnership (SERP) Institute

Fran B. Stancavage
American Institutes for Research

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

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

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

Executive Summary

Learning progressions are one of the most important assessment design ideas to be introduced in the past decade. In the United States, several committees of the National Research Council (NRC) have argued for the use of learning progressions as a means to foster both deeper mastery of subject-matter content and higher level reasoning abilities. Consideration of learning progressions is especially important in the context of the new Common Core State Standards (CCSS) and Next Generation Science Standards (NGSS) that attend specifically to the sequencing of topics and skills across grades to ensure attainment of college and career expectations by the end of high school.

In this paper we address the question: Should more formally developed learning progressions be considered for the future design of the National Assessment of Educational Progress (NAEP)? After a brief overview of the research on learning progressions, we describe the idealized model whereby shared, instructionally grounded learning progressions—once developed—could be used to link classroom-level assessments with large-scale assessments such as NAEP. At the same time, we also consider potential problems. In particular, learning progressions—which require agreed-upon instructional sequences—could be problematic in the context of a national assessment program intended to be curriculum neutral (i.e., not favoring one state's or district's curriculum over another). Finally, we use a sample of NAEP and Balanced Assessment in Mathematics (BAM; Mathematics Assessment Resource Service, 2002, 2003) items to explore the possibility of constructing "quasi learning progressions" that could be used to illuminate the substantive meaning of the NAEP achievement results.

Can Formal Learning Progressions Be Incorporated in NAEP?

Multiple research traditions have contributed to our current understanding of learning progressions. What all of these approaches have in common is the shared understanding that learning progressions are an advancement beyond traditional curricular scope and sequence schema because they are based on research investigating and documenting how learning typically unfolds in a particular area of study. They also have either been empirically tested and revised or designed with this intent. *Thus, empirical verification and a recursive process of development are defining characteristics of learning progressions.* Importantly, these also are the features of learning progressions that ensure the close connections between assessment and instruction. Furthermore, it is because of these built-in and validated instructional supports that learning progressions hold such promise for the deepening of student learning.

The most significant impediment to implementing learning progressions for any large-scale assessment program is the fledgling state of research on learning progressions. Detailed, carefully wrought, and recursively tested progressions are rare, although the few that do exist demonstrate what is possible. A second impediment, in the case of NAEP, is the close linkage required for learning progressions between assessment tasks and instructional activities. The instructional grounding of learning progressions is a defining characteristic and core strength, but it also is a constraint if NAEP as a national assessment is required to be curriculum neutral. NAEP is intended to be an independent monitor of educational achievement in the United States over time and is used to report trends for states and important groups within the population. To enable fair comparisons, the national assessment should not favor one particular curriculum over another.[3]

If curriculum-linked learning progressions cannot be the primary or central building blocks for NAEP, the assessment must nonetheless be designed in such a way as to monitor the success of deeper curricular reforms where they occur. To continue to be an independent monitor and even a check on other assessments, NAEP must have a strategic vision that attends to both breadth and depth in representing subject-matter expertise.

In a recent white paper on the future of NAEP (National Center for Education Statistics, 2012), an expert panel recommended that NAEP domain specifications be broadened so as to enable linkages with multiple other assessments, as well as to assess advanced skills that may not be well distributed across the population. *Under such a design, the NAEP framework and reporting domain need not be the same as this comprehensive item pool, which might be thought of as a "super-assessment" domain or blueprint.* By beginning with special studies, as have been used in the past, to determine whether more advanced performance can be documented in those settings where reform curricula have been successfully implemented, assessment tasks tied to learning progressions in mathematics, science, or literacy could be embedded within the NAEP super-assessment framework. Both performance outcomes and the psychometric functioning of the assessment tasks could be compared for students with and without instructional opportunities tied directly to learning progressions curricula.

An Illustration of Quasi Learning Progressions for NAEP

The demand for curricular neutrality appears to render the use of learning progressions infeasible as a central means for developing NAEP, given the appeal of learning progressions as a way to illuminate the substantive meaning of achievement results. However, we considered the possibility of constructing "quasi learning progressions" to use as a NAEP reporting device.

Using both NAEP and BAM items, we constructed four hypothetical learning progressions representing subtopics in two of NAEP's content areas: Data Analysis and Probability, and Algebra. As a whole, BAM items are designed to tap higher levels of reasoning and application; therefore, they might be more like the kinds of assessment tasks developed to assess the CCSS.[4]

In constructing the quasi learning progressions, a critical conceptual decision was to order items by the typical instructional sequencing of topics, *not* by cognitive

---

[3] Many believe that adoption of the new CCSS now ensures much greater agreement among states as to how students move through topics, and thus creates the needed shared curriculum. However, a large gap remains between the general character of CCSS sequences and the specificity of actual learning progressions, which are much more dependent on specific curricular decisions.

[4] The inclusion of BAM items was possible because of an earlier study (Stancavage et al., 2009) in which NAEP and BAM items were scaled together.

complexity or perceived difficulty. The ordering process was conducted by coauthor Daro, using his knowledge of mathematics and research on mathematics learning, and reviewed by the other authors to confirm that items within each level were similar to each other in terms of the instructional topic addressed, and distinguishable from the next higher and next lower levels. We then plotted the relationship between *judged* levels of increasing proficiency on the intended construct and *empirical* evidence of item ordering for each of the four progressions, and evaluated the level of correlation between the two measures. Correlations were moderate, ranging from 0.41 to 0.60.

Based on this exercise, we conclude that such an approach is infeasible and likely to be misleading until there is more widespread implementation of the new standards and thereby greater congruence between hoped-for and empirical ordering of items. Although we can see ways to improve the meaningfulness of quasi learning progressions by eliminating misfitting items, in most cases these are not items that one would want to remove lightly. To anchor the scale with only the well-behaved items essentially moves more challenging items to a later place on the progression. These kinds of decisions can only be made after doing the kind of work that is required for the development of learning progressions (i.e., logical and expert-developed sequences must be tested in instructional contexts where students have had the opportunity to learn with the support of curricula specifically developed in conjunction with the intended progression).

CONTENTS

Introduction

Learning progressions are one of the most important assessment design ideas to be introduced in the past decade. The importance of their use in other countries, such as Australia and the Netherlands, reflects their fundamental characteristic, which is a much closer linkage between assessment and instruction than is true for typical large-scale assessment programs. In the United States, several committees of the National Research Council (NRC) have argued for the use of learning progressions as a means to foster both deeper mastery of subject-matter content and higher level reasoning abilities. Consideration of learning progressions is especially important in the context of the new Common Core State Standards (CCSS) and Next Generation Science Standards (NGSS) that attend specifically to the sequencing of topics and skills across grades to ensure attainment of college and career expectations by the end of high school.

Given the centrality of the CCSS and NGSS for current educational reforms, and the emphasis in these documents on the sequential deepening of content mastery and skill development over time, the question arises: Should more formally developed learning progressions be considered for the future design of the National Assessment of Educational Progress (NAEP)? In this paper, we provide a brief overview of the research on learning progressions and explain the combination of expert knowledge and empirical fieldwork needed to develop and test instructionally grounded learning progressions. We describe the idealized model whereby shared, instructionally grounded learning progressions—once developed—could be used to link classroom-level assessments with large-scale assessments such as NAEP. At the same time, we also consider potential problems. In particular, learning progressions—which require agreed-upon instructional sequences—could be problematic in the context of a national assessment program intended to be curriculum neutral (i.e., not favoring one state's or district's curriculum over another).

Due to the potential appeal of learning progressions as a way to illuminate the substantive meaning of achievement results, in this report we consider the possibility of constructing "quasi learning progressions" as a reporting device. We call them quasi progressions because they are developed after the fact, rather than being jointly constructed and field tested as a continuum of instructional and assessment tasks. Using data from a previous NAEP Validity Studies Panel investigation and an approach similar to the anchoring methodology used earlier in NAEP's history, we construct three quasi learning progressions for eighth-grade mathematics. This exercise illustrates the potential benefits of using sequenced exemplar items to give meaning to the numerical score scale. At the same time, misfitting items illustrate the difficulty of meeting both the logical and empirical requirements of learning progressions in multidimensional assessment domains.

Definition

Learning progressions are known by various terms: progress maps, progress variables, developmental continua, progressions of developing competence, profile strands, learning trajectories, and learning lines. According to Masters and Forster (1996, p. 4), "A progress map describes the knowledge, skills and understandings of a learning area in the sequence in which they typically develop and provides examples of the kinds of performances and student work typically observed at particular levels of attainment." Similarly, in *Taking Science to School: Learning and Teaching Science in Grades K–8*, learning progressions were defined as "descriptions of the successively more sophisticated ways of thinking about a topic that can follow one another as children learn about and investigate a topic over a broad span of time" (Duschl, Schweingruber, & Shouse, 2007). Although order is an implied characteristic of learning progressions, making it possible to quantify increases in proficiency, learning progressions are distinguished from other score scales by their attention to substantive markers of increasing proficiency. They are "criterion-referenced," in Glaser's (1963) original sense of the term, meaning that they are grounded in actual criterion performance and illustrate explicitly how performance has to improve to move higher on the score scale.

In part because of their sudden popularity and also because of their emergence in very different research literatures, the idea of learning progressions cannot be reduced to a single agreed-upon and precise definition. Early work in Australia using the term "progress maps" was informed by Rasch model scaling and therefore attended more to psychometric requirements (Masters, Adams, & Wilson, 1990; West Australian Ministry of Education, 1991). Other early work, also in Australia and the United States, focused on emergent literacy and was similar to parallel work in the United States examining early childhood mathematics learning. These latter efforts focused on instructional tasks that could be ordered on a continuum that also served assessment purposes (Baroody, 1984; Fuson, 1992). Some learning progressions are quite broad and general, depicting the mastery of a content domain over several grade levels. Other learning progressions are very detailed and focus on increasing mastery within a single unit of instruction. In the earliest grades, progressions may be affected by biological development, although the rate at which children proceed can clearly be influenced by instructional supports. Most learning progressions do not, however, imply some underlying latent trait. Rather, they reflect curricular and instructional choices within which may lie some "natural" orderings of difficulty. For example, multiplication may be easier than subtraction, depending on how they are taught, but two-digit subtraction will nearly always be easier than three-digit subtraction.

Why the Appeal? Learning Progressions in the Context of the Common Core

Unlike standards documents from the early 1990s that emphasized what students should "know and be able to do" at a given grade level, the CCSS are oriented toward cumulative growth in knowledge and skills across grade levels. The English language arts grade-level standards, for example, "define end-of-year expectations and a cumulative progression" leading to college and career readiness (National Governors Association Center for Best Practices and Council of Chief State School Officers, 2010a, p. 4). The specific reading standards establish "a grade-by-grade 'staircase' of increasing text complexity that rises from beginning reading to the college and career readiness level" (p. 8). Similarly, authors of the mathematics standards attended both to the hierarchical logic of disciplinary structures and to research on "how students' mathematical knowledge, skill, and understanding develop over time" (National Governors Association Center for Best Practices and Council of Chief State School Officers, 2010b, p. 4), with the intention of empirically verifying these sequences even more rigorously in the future.

Some of the popular rhetoric surrounding the CCSS makes it appear as if the sequential nature of the standards arose primarily from an exercise in backwards planning intended to ensure arrival at the endpoint of college and career readiness. Unfortunately, using "college and career readiness" as a short-hand summary for learning goals sometimes obscures the important underlying reform principle that links sequencing of learning goals with the need for greater rigor and depth of understanding. Most policymakers today are familiar with findings from more than a decade ago that attributed the poor performance of U.S. students on international comparisons to our "mile-wide and inch-deep curricula" (Schmidt, McKnight, & Raizen, 1997). In subsequent investigations, Schmidt and colleagues identified the features of "curriculum coherence" that distinguished the curricula of top-performing nations from the unfocused and repetitive curricula fostered by U.S. state and district standards documents. Surprisingly, for those who assume that academic excellence requires covering *more* topics, curriculum documents from high-performing countries included fewer topics per grade than is typical of U.S. standards because in high-performing countries topics were introduced, studied in greater depth, and then intentionally removed from the curriculum. In contrast, topics "linger" in U.S. curricula once they are introduced.

Fewer topics in the A+-rated countries naturally implied more focus. More importantly, however, the sequencing of topics in high-performing countries also appeared to be more carefully orchestrated to build on concepts from one grade to the next. Schmidt, Wang, and McNight (2005) concluded that standards meet a criterion of coherence "if they specify topics, including the depth at which the topic is to be studied as well as the sequencing of the topics, both within each grade and across the grades, in a way that is consistent with the structure of the underlying discipline" (p. 554). A basic goal of the CCSS is not only to design the standards to reflect the structure of the discipline or skill dimension, but also to make this structure visible to students as part of their understanding and mastery of the subject matter.

A word of caution is required, however, before assuming that the CCSS meet a technical definition of formal learning progressions. The same must be said of the NGSS despite their focus on core ideas that are "teachable and learnable over multiple grades at increasing levels of depth and sophistication." As we explain in the next section, elaborating within the broader standards frameworks to establish formal learning progressions will require a much more detailed codevelopment of instructional and assessment materials based on both expert judgment and empirical verification. Authors of the CCSS are aware that local variability and limitations in the research base make it impossible to say with certainty that topic A should always come before topic B. In describing the CCSS in mathematics, they note the following:

> …*grade placements for specific topics have been made on the basis of state and international comparisons and the collective experience and collective professional judgment of educators, researchers and mathematicians. One promise of common state standards is that over time they will allow research on learning progressions to inform and improve the design of standards to a much greater extent than is possible today.* (Common Core State Standards Initiative, 2012)

Thus, it might be useful to think of the grade-to-grade continua underlying the CCSS and NGSS as "learning sequences" and reserve the term learning progressions for more carefully developed progressions that meet the technical definition. Or, at a minimum, given the popular and pervasive use of "learning progressions" talk, it should be acknowledged that Common Core progressions are hypothetical and preliminary and are expected to be refined by further research and development.

Instructional Benefits and Requirements for the Development of Learning Progressions

The sudden policy interest in learning progressions as a reform strategy has led to some confusion about terminology and, more fundamentally, about the defining characteristics of learning progressions and what they can promise to do. This is due largely to the rapid merging and comingling of multiple research traditions. For example, mathematics education and science education have distinct research literatures, respectively, on learning trajectories and learning progressions. Some approaches to learning progressions have a decidedly measurement or assessment focus, meaning that the goal of research projects in this tradition is to produce a specific measurement instrument. Other approaches come from contemporary improvements in learning research—focusing on children's thinking and the need to design instructional tasks that directly build on students' intuitive understandings and prior experiences, but without attempting to score or quantify the level of student attainment. Assessment may be nearly invisible in the latter case. Some progressions are quite general and cover broad age spans as is intended for the CCSS. Examples provided by Masters and Forster (1996) are from national curricula for England and Wales, Australia, Hong Kong, and Canadian provinces. Other progressions, such as the "Sinking and Floating" example developed at the Stanford Education Assessment Laboratory (Ayala et al., 2008), mark progress over a single unit of study.

What all of these approaches have in common is the shared understanding that learning progressions are an advancement beyond traditional curricular scope and sequence schema because they are based on research investigating and documenting how learning typically unfolds in a particular area of study. They also have either been empirically tested and revised or designed with this intent. *Thus, empirical verification and a recursive process of development are defining characteristics of learning progressions.* Importantly, these are also the features of learning progressions that ensure the close connections between assessment and instruction. Furthermore, it is because of these built-in and validated instructional supports that learning progressions hold such promise for the deepening of student learning.

In a recent report summarizing research on learning progressions in science, Corcoran, Mosher, and Rogat (2009) identified five essential components of learning progressions as shown in Table 1.

Table 1. Essential Components of Learning Progressions

1. Learning targets or clear end points that are defined by societal aspirations and analysis of the central concepts and themes in a discipline;
2. Progress variables that identify the critical dimensions of understanding and skill that are being developed over time;
3. Levels of achievement or stages of progress that define significant intermediate steps in conceptual/skill development that most children might be expected to pass through on the path to attaining the desired proficiency;
4. Learning performances which are the operational definitions of what children's understanding and skills would look like at each of these stages of progress, and which provide the specifications for the development of assessments and activities which would locate where students are in their progress; and,
5. Assessments that measure student understanding of the key concepts or practices and can track their developmental progress over time.

Source: Corcoran et al., 2009, p. 15.

This distillation makes a useful distinction between the encompassing term, learning progressions, and the more detailed specification of skills required for "progress variables" as noted in step 2. In calling out these steps, the authors drew from the grand conceptual steps (steps 1 and 3) laid out in *Taking Science to School*, and the more detailed steps followed by Smith, Wiser, Anderson, and Krajcik (2006), to create progress variables, learning performances, and assessments of key concepts and practices in their construction of a learning progression for matter and the atomic-molecular theory. To be complete, we note that the conceptual steps described in *Taking Science to School* begin with a prior step that "anchors" learning progressions at one end "by what is known about the concepts and reasoning of students entering school" (Corcoran et al., 2009, p. 219).

As part of their synthesis project, Corcoran et al. (2009) and their panel of experts identified further the following possible benefits of learning progressions, which again emphasized the coconstruction of instructional materials and assessment tasks.

- They should provide a more understandable basis for setting standards, with tighter and clearer ties to the instruction that would enable students to meet them;
- They would provide reference points for assessments that report in terms of levels of progress (and problems) and signal to teachers where their students are, when they need intervention, and what kinds of intervention or ongoing support they need;
- They would inform the design of curricula that are efficiently aligned with what students need to progress;
- They would provide a more stable conception of the goals and required sequences of instruction as a basis for designing both pre- and in-service teacher education.
- The empirical evidence on the relationship between students' instructional experiences and the resources made available to them, and the rates at which they move along the progressions, gathered during their development and

> ongoing validation, can form the basis for a fairer set of expectations for what students and teachers should be able to accomplish, and thus a fairer basis for designing accountability systems and requirements. (pp. 9–10)

An example from mathematics serves to highlight the grounding of learning progressions in children's thinking and their subsequent linking to instructional interventions. Drawing on their own work for more than a decade and that of others, Clements and Sarama (2009) described early childhood mathematics progressions (trajectories) for counting, early arithmetic, spatial thinking, geometric shapes, and geometric measurement (e.g., length, area). Their work is instructive because it illustrates both the research process needed to develop learning progressions and the subsequent use of progressions to support and thereby accelerate and deepen student learning.

The learning progression for counting from Clements and Sarama (2009) is presented in Table 2. They note that this progression comprises three subtrajectories: verbal counting (knowing the number names), object counting, and counting strategies. These three subtrajectories build from one to the next but also become increasingly interrelated. "To count a set of objects, children must not only know verbal counting but must also learn (a) to coordinate verbal counting with objects by pointing to or moving the objects and (b) that the last counting word names the cardinality of ('how many objects in') the set" (p. 21). To establish the steps or levels in the progression, the researchers synthesized clinical interview and observational findings from dozens of prior studies. They developed descriptive labels and recognizable counting behavioral markers for each step. Then, importantly, Clements and Sarama developed instructional tasks for each level that would foster the kind of thinking required at that level. For example, as most parents know, touching each object while counting helps move "reciters" to the next level, "corresponders." To move from counting to understanding the "how many" question (the cardinality principle), children are first asked, "how many do I have?" after one object is added or removed. Next, they are asked "how many?" with surprise additions or subtractions of two or three.

Table 2. Learning Trajectory for Counting

Pre-Counter *Verbal* No verbal counting. Names some number words with no sequence. **Chanter** *Verbal* Chants "sing-song" or sometime indistinguishable number words.
Reciter *Verbal* Verbally counts with separate words, not necessarily in the correct order above "five."
Reciter (10) *Verbal* Verbally counts to ten, with *some* correspondence with objects, but may either continue an overly rigid correspondence or exhibit performance errors (e.g., skipping, double-counting). **Corresponder** Keeps one-to-one correspondence between counting words and objects (one word for each object), at least for small groups of objects laid in a line. May answer a "how many?" question by re-counting the objects, or violate 1-1 or word order to make the last number word be the desired or predicted word.
Counter (Small Numbers) Accurately counts objects in a line to 5 and answers the "how many" question with the last number counted. When objects are visible, and especially with small numbers, begins to understand cardinality. **Counter (10)** Counts arrangements of objects to 10. May be able to write numerals to represent 1–10. Accurately counts a line of 9 blocks and says there are nine. Verbal counting to 20 is developing.
Producer (Small Numbers) Counts out objects to 5. Recognizes that counting is relevant to situations in which a certain number must be placed. Produces a group of 4 objects.
Counter and Producer (10+) Counts and counts out objects accurately to 10, then beyond (to about 30). Has explicit understanding of cardinality (how numbers tell how many). Keeps track of objects that have and have not been counted, even in different arrangements. Writes or draws to represent 1 to 10 (then 20, then 30).
Counter Backward from 10 *Verbal and Object*
Counter from N (N + 1, N – 1) *Verbal and Object* Counts verbally and with objects from numbers other than 1 (but does not yet keep track of the *number* of counts).
Skip Counter by 10s to 100 *Verbal and Object* Skip counts by tens up to 100 or beyond with understanding; e.g., "sees" groups of 10 within a quantity and counts those groups by 10 (this relates to multiplication and algebraic thinking).
Counter to 100 *Verbal* Counts to 100. Makes decade transitions (e.g., from 29 to 30) starting at any number.
Counter On Using Patterns *Strategy* Keeps track of a few counting acts, but only by using numerical patterns.
Skip Counter *Verbal and Object* Counts by fives and twos with understanding.
Counter of Imagined Items *Strategy* Counts mental images of hidden objects.
Counter On Keeping Track *Strategy* Keeps track of counting acts numerically, first with objects, then by "counting counts." Counts up 1 to 4 *more* from a given number.
Counter of Quantitative Units/Place Value Understands the base-ten numeration system and place-value concepts, including ideas of counting in units and multiples of hundreds, tens, and ones. When counting groups of 10, can decompose into 10 ones if that is useful.
Counter to 200 *Verbal and Object* Counts accurately to 200 and beyond, recognizing the patterns of ones, tens, and hundreds.
Number Conserver Consistently conserves number (i.e., believes number has been unchanged) even in face of perceptual distractions such as spreading out objects of a collection.
Counter Forward and Back *Strategy* Counts "counting words" (single sequence or skip counts) in either direction. Recognizes that decades sequence mirrors single-digit sequence.

Source: Clements & Sarama, 2009, pp. 30-41.

Clements and Sarama (2007a) used this extensive program of research to develop the Building Blocks curriculum and computer software to support learning in both early numeracy and geometry. The impact on student learning of carefully designed interventions tailored to specific levels of learning progressions was documented in a comparative study conducted in preschool programs serving low-income families (Clements & Sarama, 2007b). Within state-funded preschool and Head Start school sites, classrooms were assigned to treatment or control groups. Control classrooms continued to receive the existing preschool curriculum. Participants were assessed at the beginning and end of the school year using individual interview protocols designed to cover the same topics as the curriculum but without mirroring the instructional activities. The statistical and practical significance of the effects was dramatic. For the Number and Geometry outcome measures, the effect-size differences between the treatment and control groups at the time of the post assessment were .85 and 1.47, respectively. Similar effects were also obtained for differential gains from pre- to post-assessment for the treatment group compared with the control group. The fact that instructional supports targeted to each level of the progressions were so effective provides additional evidence as to the validity of the progressions.

Clements and Sarama (2009) describe their progressions as *developmental progressions,* meaning that they represent natural sequences that are affected by biology. They use the example of infants and children first learning to crawl, then walk, then run, skip, and jump. Although biological readiness may also affect the *order* of skill development in mathematics and other early learning, Clements and Sarama (2009) emphasize that development may be fast or slow depending on learning opportunities. Many decades ago psychologists believed that development proceeded at a fixed pace and could not be hurried. On the contrary, contemporary learning research has demonstrated that learning affects and interacts with development—hence the interest in instructional moves specifically targeted to developmental stages. Virtually all researchers studying learning progressions recognize that development is strongly affected by learning opportunities and specific instructional contexts. As noted by Masters and Forster (1996), a learning progression is "NOT a description of 'natural' sequences of development only. A progress map is the result both of 'natural' sequences of student development and common conventions for the content and delivery of curricula, and may be elucidated by systematic research into student learning" (p. 11).

In addition to guiding instructional interventions, other potential benefits of learning progressions are more directly applicable to large-scale assessment applications. However, these benefits also derive from the connectedness of learning progressions to particular instructional practices. The NRC report, *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001), outlined key requirements for reforming assessment systems if they are to capitalize on recent findings from cognitive science research and measurement theory. Of their three requirements for assessment systems—comprehensiveness, coherence, and continuity—the latter two can best be met by the use of learning progressions. *Comprehensiveness* refers to the completeness with which various learning goals are represented by the assessment system. *Coherence* addresses the relationship among assessments at different levels of the system. In the past, large-scale assessments have been misaligned with classroom tasks and learning goals or, when they were made coherent, it was by creating classroom work and assessments that imitated external tests. If classroom formative assessments and large-scale assessments were designed around shared learning progressions instead, the resulting system would be conceptually coherent even if classroom materials would need to be developed at a much finer grain size. Last, *Knowing What Students Know* (Pellegrino, et al., 2001) recommended that ideal assessment systems be designed to be *continuous* as follows:

> *Assessments should measure student progress over time, akin more to a videotape record than to the snapshots provided by the current system of on-demand tests. To provide such pictures of progress, multiple sets of observations over time must be linked conceptually so that change can be observed and interpreted. Models of student progression in learning [emphasis added] should underlie the assessment system, and tests should be designed to provide information that maps back to the progression. With such a system, we would move from "one-shot" testing situations and cross-sectional approaches for defining student performance toward an approach that focused on the processes of learning and an individual's progress through that process. (pp. 256–257)*

Imagine a coherent and continuous system whereby classroom instructional activities and formative assessment tasks are developed, in tandem, as part of a learning progression. Then when it comes time to build the large-scale assessment, representative tasks are developed to measure progress along that same learning progression. Forster and Masters (2004) described just such a system developed by the Australian Council for Educational Research (ACER). They confess that they did not set out initially to build both classroom-level and linked national assessments, but having done so, they make a strong case for the resulting synergies and coherence. Their national survey assessment was built subsequent to the development of classroom-level curriculum and assessment materials but was closely tied to them, using the same underlying progressions.

ACER first created a Developmental Assessment Resource for Teachers (DART) "to assist teachers in assessing students' knowledge, skills, and understandings in English (language arts) at the elementary (Australian 'primary') level" (p. 52). Although the emphasis was on helping teachers to assess students' classroom work by providing assessment tasks, scoring guides, and samples of student work, the nature of the project also helped teachers develop a deep and shared understanding of the new national English curriculum framework that had been released that same year. Assessment materials were designed around common themes, videotapes were provided to set the theme, and teachers were encouraged to develop their own materials consistent with the theme. Later, the National School English Literacy Survey (NSELS) was developed based on the DART model and was able to use the same mix of classroom-based, teacher-scored authentic literacy tasks. In addition, because of a shared curriculum, the national survey could use tasks that called on the same themes as the classroom-level assessments. For example, a Year 3 poem on the NSELS about mosquitoes related to a film that children had watched as part of the DART myths and legends theme. To ensure reliability and comparability, external assessors joined teachers in scoring, but the national survey tasks were still highly congruent with typical classroom practices. According to Forster and Masters (2004), progress maps for each of the skill areas (reading, writing, spelling, and speaking) provided the “conceptual backbone” that made possible this kind of coherence between their classroom-level and accountability assessments.

Challenges to Implementing Learning Progressions With NAEP

The most significant impediment to implementing learning progressions for any large-scale assessment program is the fledgling state of research on learning progressions. Clements and Sarama's (2009) detailed, carefully wrought, and recursively tested early mathematics progressions are rare. They are an existence proof demonstrating what is possible, but similarly created progressions do not exist across grades and subject matters. Several progressions have been constructed in the sciences for matter and atomic molecular theory (Smith et al., 2006), evolution (Catley, Lehrer, & Reiser, 2004), complex reasoning about biodiversity (Songer, Kelcey, & Gotwals, 2009), force and motion (Alonzo & Steedle, 2009), genetics (Duncan, Rogat, & Yarden, 2009), and carbon cycling in socioecological systems (Mohan, Chen, & Anderson, 2009). Note that, as with all progressions, these are acknowledged to be working hypotheses or draft progressions. They are research based in that prior evidence and experience supports the reasoning that went into authoring the progressions. They have also been field tested, in many cases undergoing multiple iterations and revisions. However, although these development projects reflect the integration of big ideas and practices that are called for in the NGSS, they still have not worked out how multiple progressions of this type would be brought together in a coherent curriculum. The learning sequences embedded in the CCSS are even less well developed. They are research based in the sense that they use available research evidence about which concepts appear easier than others and, once mastered, facilitate subsequent learning. Expert judgment has been used to fill in the gaps. But the CCSS have not been empirically tested as to the rate at which progress is likely to occur and with what affordances, nor is there research knowledge yet about concurrent pursuit of these standards and the extent to which concurrence might foster (or impede) joint progress.

A second impediment, in the case of NAEP, is the close linkage required for learning progressions between assessment tasks and instructional activities. The instructional grounding of learning progressions is a defining characteristic and core strength, but it is also a constraint if NAEP as a national assessment is required to be curriculum neutral. NAEP is intended to be an independent monitor of educational achievement in the United States over time and is used to report trends for states and important groups within the population. To enable fair comparisons, the national assessment should not favor one particular curriculum over another. Therefore, it could not base its frameworks on specific curriculum-based progressions. In the past, we have argued that the national assessment should be comprehensive, reflecting the union of multiple curricular approaches (National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment, 1992), and, indeed, although not as broad as the sum of all possible state frameworks, NAEP has been found to have greater reach in terms of cognitive complexity than many state assessments (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007). Now, in the context of the CCSS, continuing to envision NAEP as the union of multiple curricula could contribute to a milewide, inch-deep problem if

NAEP does not explicitly attend to the depth-over-breadth conception of advanced performance.

Many believe that adoption of the new CCSS now ensures much greater agreement among states as to how students move through topics, and thus creates the needed shared curriculum. However, a large gap remains between the general character of CCSS sequences and the specificity of actual learning progressions, which are much more dependent on specific curricular decisions. The gap between general frameworks and specific curricula is particularly great if the intent of both is to aim for deeper understanding rather than superficial coverage. The ability to ask for deeper understanding, for example, in comparing character development in two different works of fiction requires that the test maker know what novels students have read. The demands of "going deeper" are especially great if we take seriously the relatively old finding from cognitive science research that thinking skills cannot be developed independent of content. When applied specifically to the NGSS and research on learning progressions in the sciences, this means that topics must be integrated with scientific practices; there are many ways of doing this that would still be consistent with the NGSS. Citing the *Taking Science to School* definition of learning progressions, Songer et al. (2009) argue that "successively more sophisticated ways of thinking about a topic…recognizes the inherent presence and interconnection of content knowledge with inquiry reasoning" (p. 611). In their development of a learning progression for complex reasoning about biodiversity, Songer et al. (2009) paired a biodiversity continuum (from "plants and animals differ" to "taxonomic diversity and abundance") with an inquiry reasoning progression based on evidence-based explanations. Had they picked "planning and carrying out investigations" or "analyzing and interpreting data"—other scientific practices that also require complex reasoning—the assessment and curricular tasks at the higher end of the progression would have looked quite different.

NAEP's History With Related Item-Anchoring Methodologies

When item response theory (IRT) was first introduced in the field of measurement, and later adopted as a NAEP's primary analytic model, one of its most desirable features was its ability to locate examinees and items on the same score continuum—thus making it possible to offer criterion-referenced interpretations of examinee's scores. Unfortunately, as researchers quickly realized, making statements about what examinees at a given score level "can do" depended greatly on the orderliness of the items being scaled, the criterion used to locate items on the scale, and the degree of relationship between unique items and more general descriptions of competencies. When they have not been specially designed to reflect sequential mastery, items do not march up the score continuum in tidy increments. The notion of a Guttman (1950) scale, whereby examinees can be located so that they fail all of the items above them on the scale and answer perfectly all of the items below them, simply does not occur in the world of achievement testing.

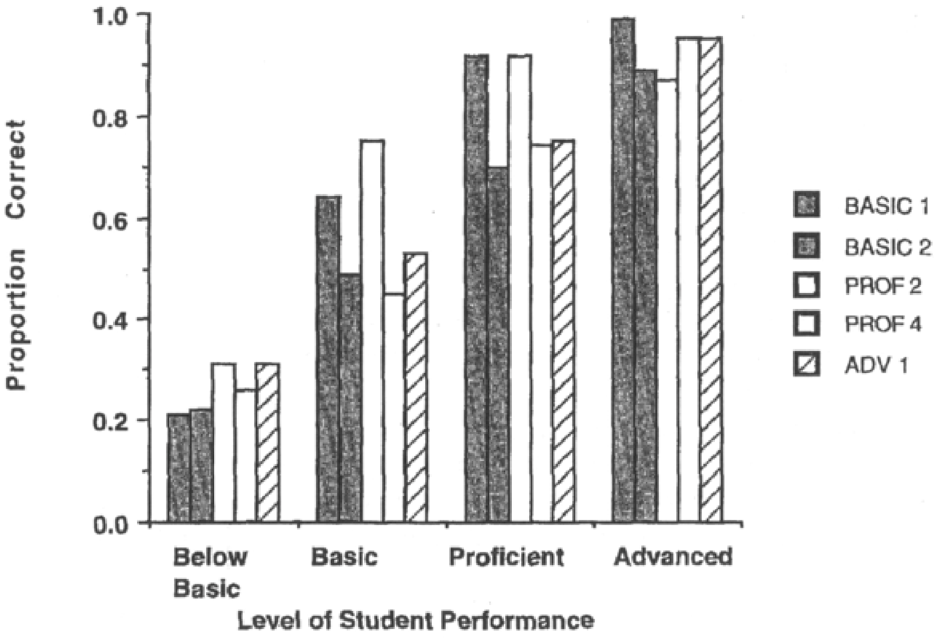
As described by Beaton and Allen (1992), item-anchoring methods were developed to identify the types of items that characterized performance at anchor points on the NAEP scale (150, 200, 250, 300, 350). The steps involved in creating anchor descriptions are as follows:

1. Form groups of examinees in close proximity to each anchor point.
2. For each item at each anchor point, calculate the proportion correct for the proximal group.
3. For each anchor point, determine which items could be answered correctly by a substantial majority of students at that level.
4. For succeeding anchor points, determine which items could be answered correctly by a substantial majority of examinees at that level but not by most of the students at the level of the next lower anchor point.
5. Given the sets of items identified at each anchor point, develop generalizations to describe the performance level characterized by these items.

In one of the earliest critiques of item-derived anchoring and criterion-referenced interpretations, Forsyth (1991) argued that in complex domains, such as NAEP mathematics and science assessments, learning could not possibly be expected to proceed uniformly for all examinees due to the different combinations of content, context, and cognitive processes. "Test developers face the enormous problems created by the interaction of an examinee's past experiences and the content of the item" (p. 5). Forsyth provided numerous examples of misinterpretations that were likely to occur because of the multidimensional nature of NAEP's composite scales. Most famously, Shanker (1990) assumed that only 6 percent of 17-year-olds could solve multistep math problems because such an item was used to anchor the 350-scale point, and only 6 percent of 17-year-olds scored above 350.

Linn (1998) further described the variations in item difficulties that could occur, not because of the level of proficiency associated with the skill or construct the item was intending to measure, but because of the particular question asked, the wording of distracters, and scoring rubrics in the case of open-ended questions. As an example, Linn noted the pattern shown in Figure 1 from Burstein et al. (1995/1996). When exemplar items were selected to illustrate the verbal descriptions of the 1992 mathematics achievement levels, the figure shows that, in some cases, a majority of students at a particular level could not answer an exemplar item selected for that level. The converse was also sometimes true, as when 77 percent of Basic-level students could answer one of the Proficient-level exemplars correctly and 79 percent of Proficient-level students could answer the Advanced-level exemplar correctly. As Linn notes, these obvious types of errors were eliminated in subsequent NAEP reports by applying statistical criteria in addition to logically matching items to verbal descriptions.

Figure 1. Proportion Correct by Achievement Level for Grade 4 Exemplar Items Selected to Illustrate Proficient and Advanced Exemplars That are Statistically Similar to Basic Exemplars



Source: Linn, 1998. Reprinted by permission of Taylor & Francis (http://www.tandfonline.com).

More recently, Schulz, Lee, and Mullen (2005) summarized difficulties with prior attempts to use individual items to make criterion-referenced descriptions of achievement and then proposed an alternative method using substantively identified testlets or domains of NAEP items that could be instructionally ordered. Using this method with eighth-grade NAEP mathematics data from 2000, they were able to show that performance on these expert- and teacher-identified domain-testlets was consistent with their expected instructional sequencing. Although these ordered domains do not have the detail of closely developed, curriculum-specific learning progressions, they do comport well with the broader grade-to-grade "progressions" envisioned for the CCSS, and therefore might well be a reasonable methodology to use with NAEP to help with scale interpretations.

We did not attempt to implement the Schulz et al. (2005) methodology for this paper because of cost constraints and because investment in such a study would make more sense sometime after the instructional sequencing based on the CCSS could reasonably be expected to be implemented. Nonetheless, for future reference, it may be useful here to elaborate on key features of the Schulz et al. methodology as distinct from item-anchoring methods.

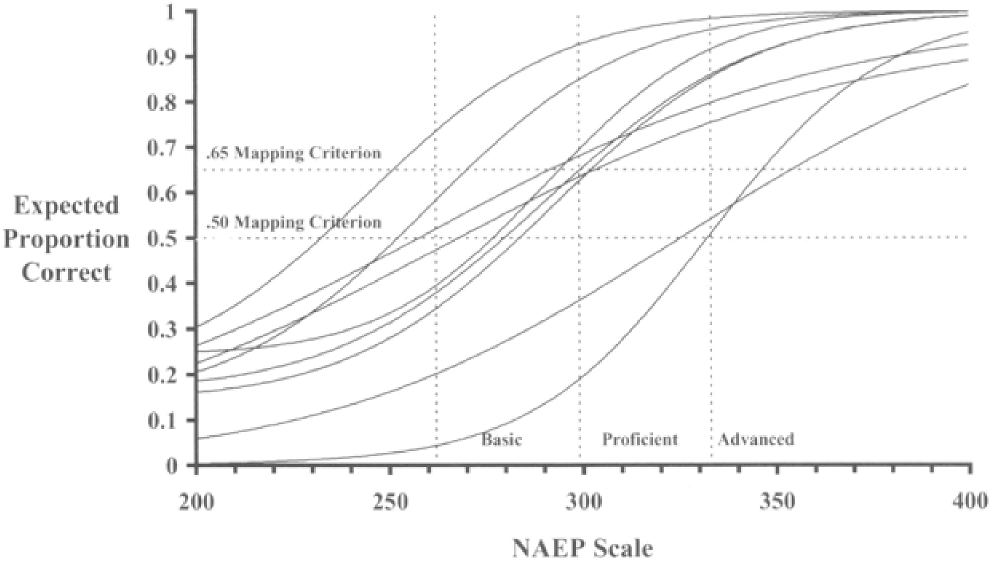
Schulz et al. (2005) created multiple domains within each NAEP content strand through a multistep approach. To begin, curriculum experts worked independently and then together to classify items into domain categories; a panel of teachers also classified items into domains. Final classifications were then determined by a domain classification team that used both sets of substantive classifications, in addition to item-difficulty parameters and teachers' ratings of instructional timing—both with respect to introduction and mastery of item content. Within both Geometry and Data Analysis, three teacher-ordered domains were preserved in the final analysis. However, for the Number Sense, Measurement, and Algebra content strands, greater numbers of teacher domains were collapsed when adjacent categories were overlapping too much in timing and difficulty. Figure 2 from Schulz et al. shows the extent to which individual items "misbehaved" within a single, seemingly homogeneous domain. In contrast, Figure 3 from Schulz et al. shows the more orderly progression of three final Number Sense domains (N1, N2, and N3), constituted as follows from finer grained teacher domains:

N1 Basic Computation with Positive Whole Numbers
  Addition and Subtraction of Integers in Context; Rounding and Place Value
  Models for Numbers and Operations

N2 Multiplication and Division
  Decimals

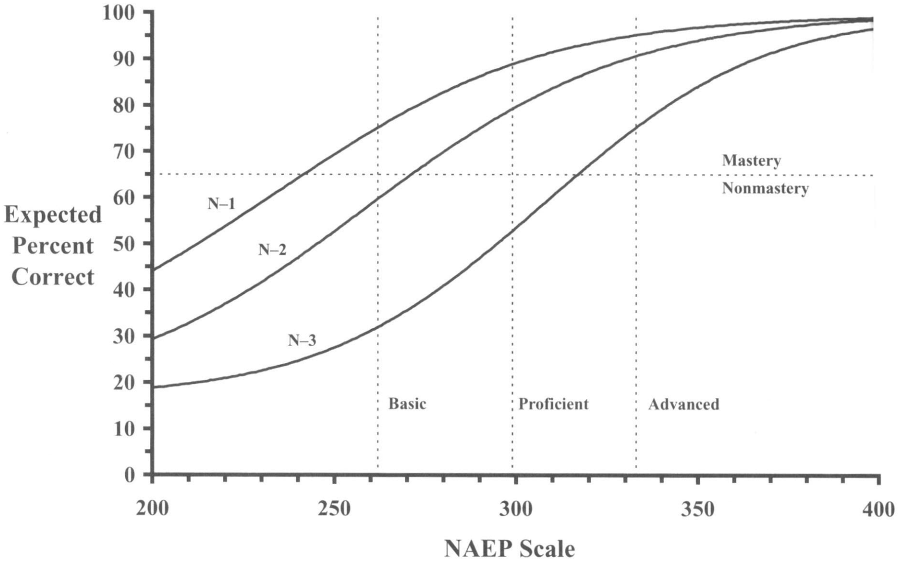
N3 Fractions and Ratios
  Rates and Percents
  Number Properties
  Scientific Notation and Exponents

Figure 2. Item Characteristic Curves in Domain D-2: Uses Graphs and Charts



Source: Schulz et al., 2005. Reprinted with permission from John Wiley and Sons.

Figure 3. Domain Characteristic Curves for Number Sense



Source: Schulz et al., 2005. Reprinted with permission from John Wiley and Sons.

An Illustration of Quasi Learning Progressions for NAEP

Although the demand for curricular neutrality appears to render the use of learning progressions infeasible as a central means for developing NAEP, given the appeal of learning progressions as a way to illuminate the substantive meaning of achievement results, we considered the possibility of constructing "quasi learning progressions" to use as a NAEP reporting device. To do this, we drew on NAEP's anchoring methodology as the psychometric techniques used to locate learning progression tasks and items on a score scale are essentially the same as the anchoring methods used historically by NAEP.

In their guiding document on the construction of progress maps, Masters and Forster (1996) distinguished between "top-down" and "bottom-up" methods for developing learning progressions. Top-down methods involve logically laying out a sequence based on expert judgments about typical pathways for knowledge and skills development. The CCSS and NGSS are examples of top-down methods, except that expert judgments may be strongly grounded in prior experience teaching or studying segments of the progressions. Bottom-up approaches begin and end with empirically gathered evidence and, in this sense, they are essentially norm-referenced approaches. In fact, Masters and Forster (1996) cited NAEP's 1990 Civics Report Card (ETS, 1990) with its item-anchoring method as an example of a bottom-up progress map.

For illustrative purposes, we proposed to construct three hypothetical learning progressions for Graphing, Statistics, and Equations representing two of NAEP's content areas: Data Analysis and Probability, and Algebra. Each of these specific objectives had sufficient numbers of items to make the exercise feasible. We elected to use items and item parameters from NAEP's 2005 eighth- grade mathematics assessment because of our prior work on this particular assessment (Daro et al., 2007; Stancavage et al., 2009) and because most items from the 2005 assessment have subsequently been released.[5] Therefore, it is possible to display various NAEP items illustrating features of the quasi progressions without violating the security of the items. In addition, in the Stancavage et al. study, the Balanced Assessment in Mathematics (BAM; Mathematics Assessment Resource Service, 2002, 2003) was also administered to approximately 2000 examinees and was concurrently scaled and equated to the NAEP scale. As a whole, BAM items were designed to tap higher levels of reasoning and application; therefore, they might be more like the kinds of assessment tasks developed to assess the CCSS.

Using his knowledge of mathematics and research on mathematics learning, study co-author Daro began the development of learning progressions by reviewing all of the items (NAEP and BAM) measuring each of the objectives. Items were ordered on a continuum to represent increasing mastery of the content objective. Items were *not* ordered by perceived difficulty. In particular, items that tapped multiple skills or relied on less familiar formats might be expected to be more difficult for students,

---

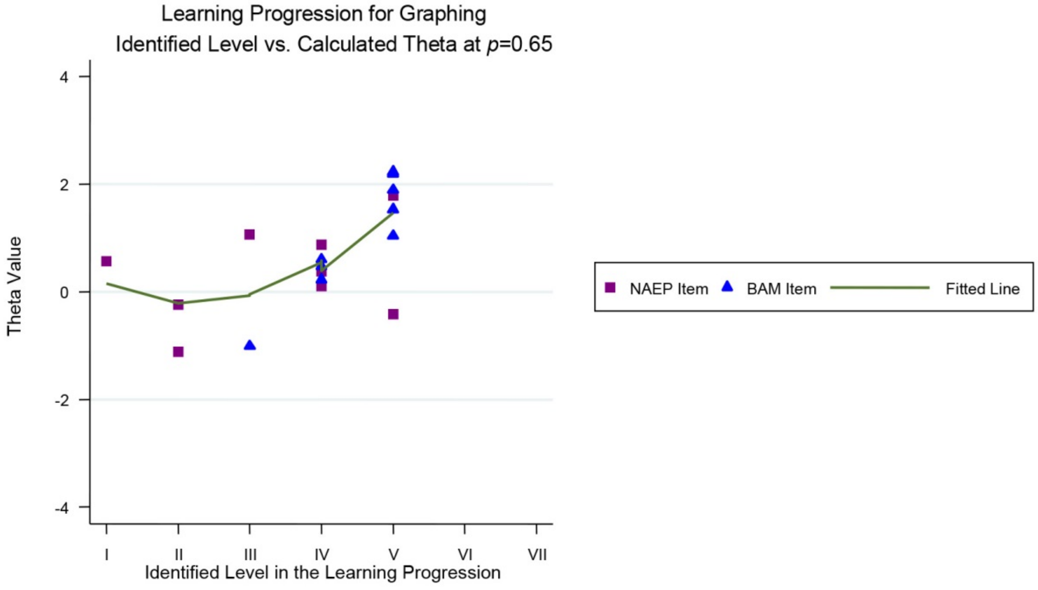
[5] The one item block from the 2005 assessment that has not been released was excluded from our exercise.

but such items were not placed at the higher end of the continuum if they called only for lower level mastery on the objective being rated. The ordering process was conducted following common-sense rules for essay grading and qualitative coding (i.e., only as many distinctions were made as could be reasonably described). Thus, Equations (branch 2) was said to have eight levels but Graphing had only six. Once ordered, items at each level were reviewed by the other authors to confirm that they were similar to each other in terms of the instructional topic addressed, and distinguishable from the next-higher and next-lower levels. Any differences were resolved by discussion among the authors and by using the item descriptors provided by NAEP and BAM, respectively. The items measuring Equations were sufficiently diverse that ultimately two different progressions were created (with some shared items), one calling for the procedural manipulations of equations (branch 1) and the other requiring that students develop equations to represent problem solutions (branch 2).

A critical conceptual decision made by Daro, in consultation with the other study authors, was to order items by the typical instructional sequencing of topics, *not* by cognitive complexity. For example, in statistics, measures of central tendency are usually taught before measures of variability. Very different progressions would have been produced had the ordering dimension been cognitive complexity, but postponing more complex reasoning about subject matter would be antithetical to the intentions of both the CCSS and learning progressions research, which aim to foster greater depth of thinking and reasoning *within* content objectives. For a given topic, of course, instruction usually proceeds from the simplest rendition of a core concept to medium complex and then highly complex understandings and applications of that concept. For two topics, usually taught in the order of A and then B, a highly simplistic ordering might expect to teach and ensure student mastery of all three levels of A before starting with the easiest version of B. In our experience, however, topics are not neatly finished before the next one begins and, in many cases, medium- and high-complexity understandings of any given topic require drawing connections and integrating knowledge and skills from multiple topics. Therefore, for the most part, we kept all items within a given instructional objective at the same level, regardless of whether they were of low, medium, or high complexity. Only when a more advanced application of a topic would typically be taught at a later time was it given a progression level of its own. For example, we created an Equations category called "Inversions" where students were asked to work backwards in applying a rule to a problem situation. Other experts might have argued that these items were just more advanced applications of an earlier level called "Using a rule without formally presenting the equation." We have tried to be as transparent as possible regarding the classification of items so that others may judge how much our findings could change if fundamentally different judgments were made about instructional sequencing. Two NAEP items were eliminated from the Graphing progression because they both involved number line representations that have been controversial with mathematicians. Some BAM items were eliminated or combined with companion items if IRT parameters could not be independently estimated due to the relatively small per-item sample size in the Stancavage et al. (2009) study.

Scatterplots were constructed to provide the simplest portrayal of the relationship between *judged* levels of increasing proficiency on the intended construct and *empirical* evidence of item ordering for each of the four progressions. The *x* axis represents the logically identified levels in the learning progression. The *y* axis represents the empirical value of the items; this empirical value is the value on the IRT score scale (theta value) corresponding to the probability of a correct response of .65 (RP 65). The theta score scale, defining the *y* axis, has a mean of 0 and a standard deviation of 1. Thus an item located at theta = 1 is a relatively difficult item because examinees would need to have a total test score of 1 standard deviation above the mean before they would have a 65 percent chance of getting this item correct. For NAEP items, this scale is the same as the appropriate NAEP subscale for eighth-grade mathematics (e.g., Algebra or Data Analysis and Probability). We have retained the theta metric rather than attempting to convert to a NAEP-like score scale to discourage overinterpretation of individual item locations, especially for BAM items that were calibrated to the NAEP scale using a sample that was not nationally representative. Figures 4–7 are the scatterplots for Graphing, Statistics, Equations branch 1, and Equations branch 2, respectively. Correlations were also computed for each item set overall and separately for NAEP and BAM items.

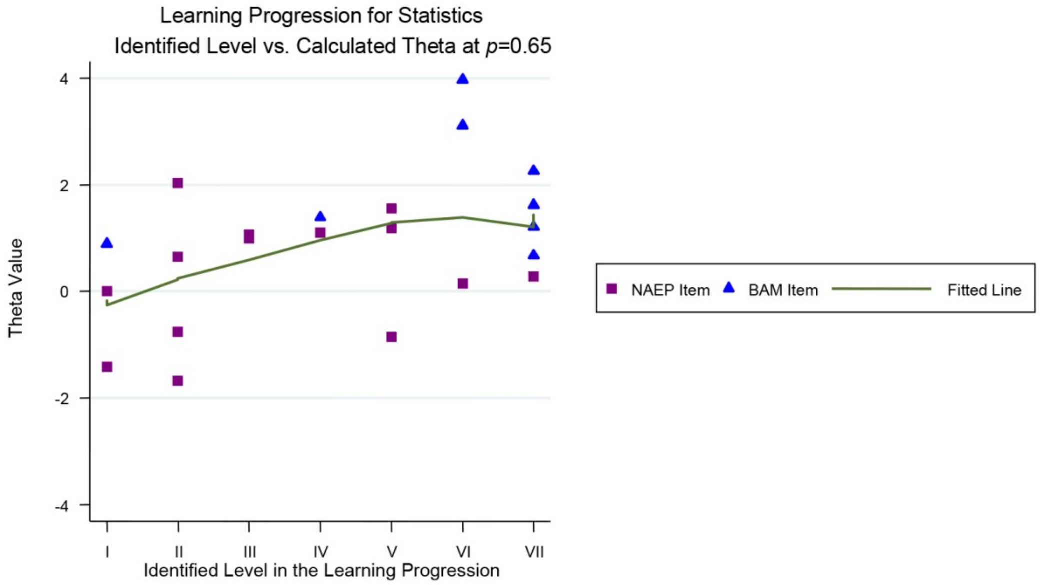
Figure 4. Scatterplot for Graphing (Theta at RP 65)



Correlations: Overall, 0.60 and BAM 0.93 both statistically significant at 0.05:
NAEP 0.34 n.s.; 18 items
Source: NAEP Items 2005 8th Grade Mathematics Assessment
Balanced Assesment in Mathematices (BAM)

Note: Theta at RP 65=value on the IRT score scale (theta value) corresponding to the probability of a correct response of .65.

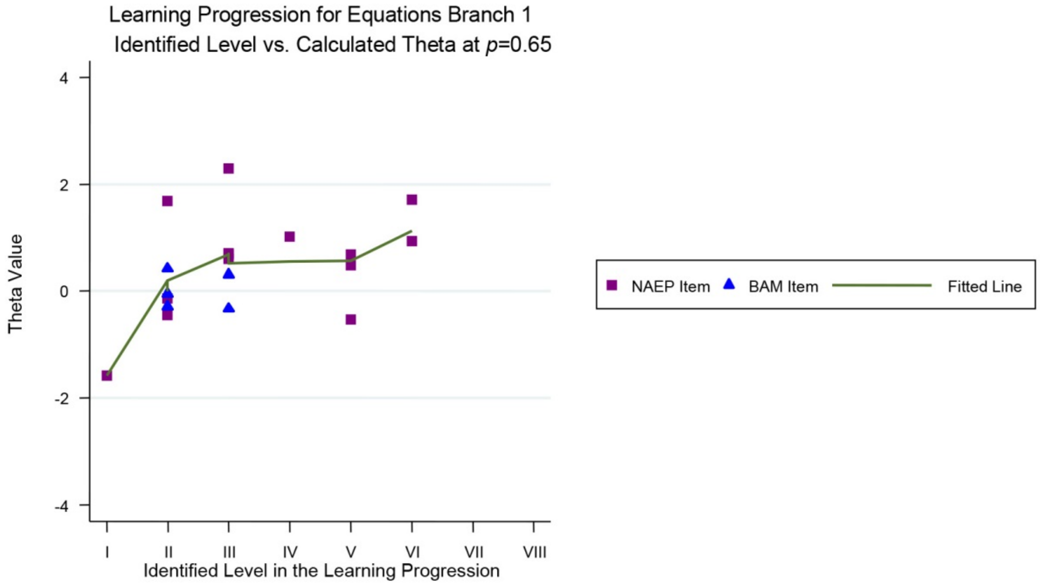
Figure 5. Scatterplot for Statistics (Theta at RP 65)



Correlations: Overall, 0.46 and statistically significant at 0.05:
NAEP 0.26 and BAM 0.25 both n.s.; 22 items
Source: NAEP Items 2005 8th Grade Mathematics Assessment
Balanced Assesment in Mathematices (BAM)

Note: Theta at RP 65=value on the IRT score scale (theta value) corresponding to the probability of a correct response of .65.

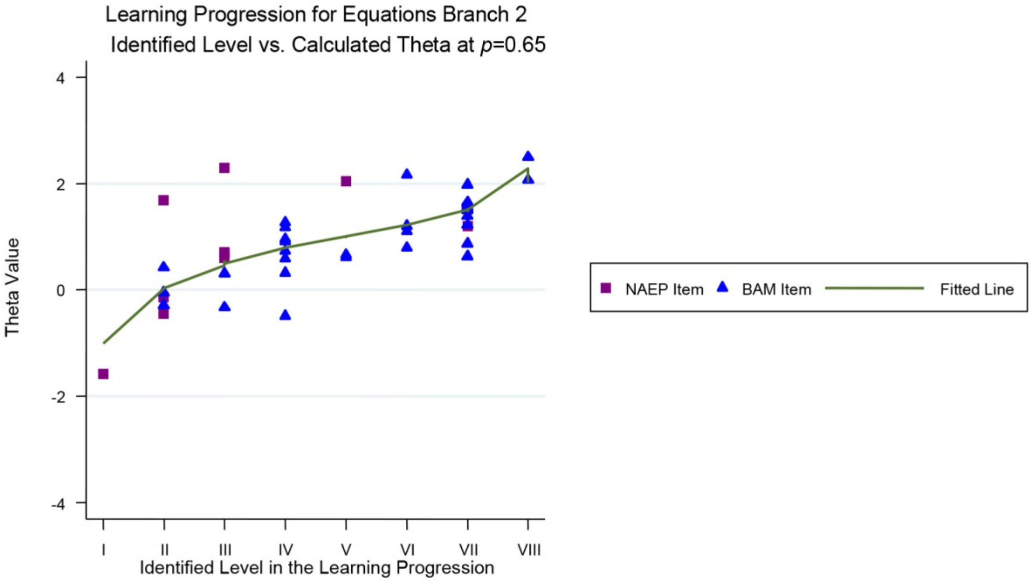
Figure 6. Scatterplot for Equations Branch 1 (Theta at RP 65)



Correlations: Overall, 0.41: BAM -0.06; NAEP 0.36 none are statistically significant; 19 items
Source: NAEP Items 2005 8th Grade Mathematics Assessment
Balanced Assesment in Mathematices (BAM)

Note: Theta at RP 65=value on the IRT score scale (theta value) corresponding to the probability of a correct response of .65.

Figure 7. Scatterplot for Equations Branch 2 (Theta at RP 65)



Correlations: Overall, 0.67: BAM 0.78 both statistically significant at 0.05:
NAEP 0.55 n.s.; 41 items
Source: NAEP Items 2005 8th Grade Mathematics Assessment
Balanced Assesment in Mathematices (BAM)

Note: Theta at RP 65=value on the IRT score scale (theta value) corresponding to the probability of a correct response of .65.

Using the combined NAEP/BAM data sets, the correlation between judged proficiency level and empirical theta was highest (r = .67) for the Equations branch 2 progression, followed by a correlation of .60 for the Graphing progression. The correlations between judged proficiency level and empirical difficulty were somewhat lower for the Statistics and Equations branch 1 progressions, at .46 and .41, respectively. However, even these more moderate correlations suggest that there is indeed a logical and somewhat shared ordering to instructional topics and corresponding student mastery. In general, the combined NAEP and BAM item sets exhibited stronger correlations than either set on its own. In the case of Statistics, combining the item sets improved the degree of relationship from .26 and .25 for the separate item sets to .46 overall. There were very few BAM items assigned to Equations branch 1, but they helped to increase the degree of relationship slightly, from .36 for NAEP items alone to .41 overall. In the case of Graphing and Equations branch 2, however, the logical ordering correlated better with empirical difficulty using BAM items alone rather than in combination with NAEP items.

The vertical spread in these plots illustrates the difficulty in developing assessment items that are so unidimensional that only a single construct determines the level of difficulty. Note also that this vertical spread or range of difficulty within nominally homogeneous groupings of items at each level is nearly identical to the range of difficulty found by Schulz et al. (2005) within domains ordered by instructional timing as illustrated in Figure 2. Several important ideas should be called out to help in interpreting items that are much easier or harder than expected given their location in the logical progression. First, these discrepancies could be caused by *construct-irrelevant variance*, which refers to features of an item that make it hard or easy but have nothing to do with the intended mathematical skill. Typical examples are when excessive verbal demands make an item too difficult for students who actually understand the mathematics, or when item distractors make an item too easy by increasing the possibility of picking the right answer without reasoning through the mathematics. More often, items will be more difficult than expected for the progression level because the mathematical demands are *multidimensional* (i.e., calling for reasoning and connections involving the intended progression construct along with other related mathematical constructs). The interconnecting of graphing skills with mastery of equations is one example. Multidimensionality of assessment items is closely related to our earlier discussion regarding the *degrees of cognitive complexity* within a given progression level. Had we sorted items within a topic category by complexity and moved the more challenging questions later in the progression, we would have reduced the vertical spread and increased the degree of fit between logical and empirical ordering because substantive multidimensionality is often the cause of increased difficulty. In our presentation of each progression, we draw attention to these more challenging and "misfitting" items, and encourage the reader to consider whether they are misplaced. Again, our argument is that to move such items higher in the progression would mean that the intention of the instructional sequence is to postpone reasoning and depth of understanding.

The issue of multidimensionality is also closely related to the issue of *curriculum-specificity*. Although orderings are usually widely shared within very narrow skill domains (e.g., adding fractions with like denominators always comes before unlike denominators), combining domains is usually an arbitrary decision made uniquely by each separate curriculum. For example, relating formulae and graphs comes much later in some curricula than others. We should also acknowledge that the apparent misfit in the scatterplots could be due to conceptual *inaccuracies* in our assignment of items to levels.

In our discussion of each progression, we refer to these types of explanations for within-level variations in item difficulty. Note that, for instructional purposes, within-level variation (from easiest to most challenging) could describe the sequencing of reasoning and deepening of understanding within a given unit of instruction, whereas the left-to-right sequencing of levels could describe the longer term ordering of concepts to be mastered over the course of many years of study. These two different orderings, within and across levels, are necessitated by the framing of this exercise in terms of the CCSS and the effort to represent mastery over broad reach of content. By contrast, Sztajn, Confrey, Wilson, and Edgington (2012), citing research on task analysis and discourse practices, argue that learning trajectories can guide teachers in responding to student thinking even within a single lesson focused on a specific task, but always with attention to the long-term goals of "fostering higher levels of sophistication over time" (p. 150). Although in many cases, we can make sense of the vertical spread instructionally, this heterogeneity illustrates the problem of using learning progressions to anchor the NAEP scale. The natural tendency would be to use the middle items that best fit the progression to anchor and describe the score scale, but for examinees scoring at any given score level this would ignore the complex items at that level that they cannot do as well as the easier items at higher levels that they can do.

Graphing Learning Progression

Data for the Graphing learning progression are presented in Table 3, while the text of the items is shown in Appendix Figure A1. Level I is represented by only one item, which asks students to follow directions to extend a pattern on a grid. The graphic knowledge involved is extremely simple, but the item has a higher than expected theta value (0.57), most likely because of the verbal demands of the item. Two items were classified as Level II items. Both involve locating a point on a grid and are relatively easy, with thetas of -1.12 and -0.24, respectively, although one can see the instructional progression from finding an intersection of number and letter dimensions on a map to formal coordinates. Level III items represent a slightly higher increment over Level II in that students must now determine an answer by locating the correct point on a curve that satisfies the problems' conditions. The first of these, A Swimming Race-item 2, which involves finding how long the winner took to swim the 50-meter race, is very easy (theta = -1.01). (Note that the full set of BAM items is shown in the figure, even though questions 3, 4, and 5, are discussed later in the progression.) By contrast, the second Level III item is quite challenging (theta = 1.06), presumably because eighth-grade students have not had experience estimating the value of a point on a curve that does not pass through a whole-number location on the grid. This could be thought of as an example of multidimensionality and/or curriculum specificity in that students would typically not be exposed to this type of question until much later in the curriculum, in the context of functions. However, the point estimation idea could be taught independent of functions, and this curricular decision would affect the fit of this item with Level III of the progression.

Items in Level IV Graphing all represent a greater knowledge of linear relationships and use of the coordinate system compared with Levels II and III. Theta locations range from 0.10 to 0.87. Level V items are a more significant step up, for the first time clearly linking Algebra and Graphing by asking students to relate linear formulas and graphs. With the exception of the first item in the level, all Level V items are quite difficult, requiring that students be 1.5 to 2 standard deviations above the mean before they have a 65 percent chance of getting the item correct. The first item is easier due to the instructions that tell students how to find the answer: "Graph the five points that represent the savings on the grid below and connect the points with a dotted line." Our observation that Level IV represents a small conceptual increment over prior levels, whereas Level V is a more significant step is consistent with the ordinal nature of the levels. No claim is made that these judgments represent an equal interval scale.

Table 3. Judged Levels and RP 65 Theta Locations for the Graphing Learning Progression

Item Identifier	Level	Theta	Level Description
XH000442	I	0.57	Follow directions to draw a line graph
VB335166	II	-1.12	Locate a point on a grid
VB434925	II	-0.24	
A Swimming Race-Item 2	III	-1.01	In a grid, locate a point on a curve
YJ000078	III	1.06	
VB429681	IV	0.1	Using lines to describe trends, find points
Vacations-Item 1	IV	0.23	
AP000711	IV	0.38	
Dollars-Item 1	IV	0.47	
Dollars-Item 2	IV	0.60	
VB434830	IV	0.87	
YJ000089	V	-0.42	Relate linear formula to graph
Dollars-Item 3	V	1.04	
Party-Item 5	V	1.53	
VB434934	V	1.79	
A Swimming Race-Item 4	V	1.89	
A Swimming Race-Item 5	V	2.19	
A Swimming Race-Item 3	V	2.23	

Note: RP 65 = value on the IRT score scale (theta value) corresponding to the probability of a correct response of .65

Statistics Learning Progression

Items for a possible Statistics learning progression are presented in Appendix Figure A2 with corresponding data shown in Table 4. Level I items require simple reading of information from graphical displays. The first item is correspondingly very easy (theta = -1.42). The next item is similar in terms of the mathematics elicited, but is much more difficult because of the demand characteristics of the item's format and language. Boxes of Candy item 2 (theta = 0.89) is an example of difficulty possibly due to curriculum specificity. It is conceptually simple for adults but could be difficult for eighth graders who might not yet have been taught about reading this type of information from bivariate plots. Level II items represent a step up from Level I items, asking students to produce a graph or describe relationships by extracting multiple pieces of information from graphs. Items in Level II vary tremendously in difficulty, from theta = -1.68 to 2.03, illustrating how much the particular demand characteristics of items affect the conclusion: "Yes, this student can interpret information from graphs."

Items addressing measures of central tendency comprise Level III. These items are relatively difficult, ranging from theta = 0.99 to 1.07. Level IV items tap more advanced understandings of central tendency. All three items are difficult, but the third item, which asked students to explain their reasoning for picking the median over the mean to represent the typical number of customers at Malcolm's Bike Shop over a five-day period, was almost impossibly difficult (theta = 7.62). Level V content returns to graphical interpretation and includes items that clearly would have been taught later than Level II graphical interpretation content, but note that how much later varies from one curriculum to the next. Level VI items test students' knowledge of sampling and variation, with theta values ranging from 0.15 to 3.97. Level VII items assess students' ability to interpret scatterplots and their use of sampling strategies to estimate large numbers. Theta values ranged from 0.28 to 2.26.

Table 4. Judged Levels and RP 65 Theta Locations for the Statistics Learning Progression

Item Identifier	Level	Theta	Level Description
VB335159	I	-1.42	Read from a graphical representation
HW000854	I	0.01	
Boxes of Candy-Item 2	I	0.89	
IY002250	II	-1.68	Interpret from a graphical representation
OM000557	II	-0.76	
YJ000102	II	0.65	
YJ000093	II	2.03	
VB335157	III	0.99	Measures of central tendency
VB434825	III	1.07	
IY002422	IV	1.1	Advanced measures of central tendency
Ages-Item 3	IV	1.39	
HL002246	IV	7.62	
VB417888	V	-0.86	Advanced graphical interpretation
VB434849	V	1.18	
YJ000060	V	1.56	
AP000506	VI	0.15	Indicators of variance
Best Guess-Item 2	VI	3.11	
Best Guess-Item 1	VI	3.97	
VB417891	VII	0.28	Measures of correlation and Estimation
Bacteria-Item 1	VII	0.67	
Boxes of Candy-Item 4	VII	1.21	
Bacteria-Item 2	VII	1.62	
Bacteria-Item 3	VII	2.26	

Note: RP 65 = value on the IRT score scale (theta value) corresponding to the probability of a correct response of .65

Equations Learning Progression Branch 1

The items measuring Equations were sufficiently diverse that ultimately two different progressions were created, one calling for procedural manipulations of equations (branch 1) and the other requiring that students develop equations to represent problem solutions (branch 2). Branch 1 appears in Table 5 and Appendix Figure A3, while branch 2 appears in Table 6 and Appendix Figure A4.

The first three levels of these two progressions are the same, but they separate into two distinct branches at Level IV. Level 1 is represented by a single item. It is an elementary-level, prealgebra item that asks students to figure out the missing value in a simple number sentence. Consistent with its judged level in the progression, the item also has a very easy theta value of -1.58. Level II asks students to evaluate an expression for a specific value or to complete a pattern by simple recursion. For example, in the first Apartment Numbers problem, students can complete the pattern by counting. In Boxes of Chocolates, the pictures help them see whether to "add two each time" or "add three each time." More advanced find-the-rule or develop-a-formula problems occur in later levels of the Equations learning progression branch 2. Theta values for Level II range from -0.45 to 0.42. This range excludes the last item in Level II, which we judged to be unusually difficult (theta = 1.69), due to construct irrelevant variance associated with format and linguistic demands. Items in Level III ask students to find and use an algebraic formula. They do not have to develop a formal equation, only recognize appropriate expressions. Theta values range from -0.33 to 0.71, with the exception of the final item, which has a theta value of 2.30. This last item is a bit odd as a test of algebra understanding and might better be used as a classroom activity to introduce the concept of slope.

Level IV has only one item and might therefore be combined with the next higher level, although we can imagine other similar items that test students' understandings of basic algebraic principles—in this case an understanding of the distributive property. This item is clearly more difficult than preceding levels (theta = 1.02), but is also more difficult than items in the subsequent level. Level V items ask students to manipulate equations, solving for *x*, or to identify equivalent expressions. Theta values range from -0.53 to 0.69. The last level in branch 1 asks students to use a formula to solve a problem. Problems of this type are more typically introduced as students begin formally working with functions. Correspondingly, the items are more difficult for students, with theta values of 0.93 and 1.71.

Table 5. Judged Levels and RP 65 Theta Locations for the Equations Learning Progression Branch 1

Item Identifier	Level	Theta	Level Description
HL000844*	I	-1.58	Supply the missing number
VB417883*	II	-0.45	Evaluate an expression for a specific value Determine an expression to model a scenario
Tilling Squares-Item 1*	II	-0.29	
VB434929*	II	-0.14	
Apartment Numbers-Item 1	II	-0.05	
Boxes of Chocolates-Item 1*	II	0.42	
EL001490*	II	1.69	
Emma's Models-Item1 *	III	-0.33	Determine equations Linear relationship between two quantities
Party-Item 1*	III	0.31	
VB335172*	III	0.60	
VB434848*	III	0.68	
VB335163*	III	0.71	
XH000443*	III	2.30	
VB335154	IV	1.02	Identify an equivalent algebraic expression
YJ000107	V	-0.53	Represent a quantitative relationship with an equation Solve for an algebraic equation
VB335169	V	0.48	
AP000710	V	0.69	
VB434852	VI	0.93	Functions
HW000857	VI	1.71	

Note: RP 65 = value on the IRT score scale (theta value) corresponding to the probability of a correct response of .65

*Same as Branch 2

Equations Learning Progression Branch 2

The two Equations progressions share the first three levels. All eight levels of branch 2 are shown in Table 6. Here we describe the unique levels of the second branch, beginning with Level IV. Although earlier levels required students to recognize and extend a number pattern, Level IV items require development of rules (rather than selecting a rule) and/or more significant extensions. The easiest item in this level—with a theta value of -0.49—asks for an extension of the pattern to the top apartment in the 10th house. The most difficult item (theta = 1.27) is also an extension of a pattern, but adds the challenge of understanding the geometry of the situation in order to calculate the number of white tiles that must be added each time. Items in Level V are quite similar to those in Level IV except that students must also explain their reasoning (i.e., they must give a verbal description of the pattern or rule). Items at Level VI also are similar to Level IV problems except that students are asked to invert their understanding of the rule—a slightly more complex task and one that would typically come after instruction focused on generating a rule and explain one's thinking about a pattern or rule. Note that none of these imply that instruction on one level is finished before moving on to the next, but we have tried to represent the sequencing of how these levels are typically introduced and perhaps how they might eventually be mastered. Items in Level VII go further and ask students to develop a formal expression for their conceptual rule. Although a few items at Level VII are easier than Level IV, as a set they are substantially more difficult, illustrating the important conceptual step required to move from pattern describing to formal algebraic representation. The two items in Level VIII ask students to conceptualize and relate two rules to find the problem solution. This last type of problem would be used to introduce and motivate the need for solving systems of equations.

Table 6. Judged Levels and RP 65 Theta Locations for the Equations Learning Progression Branch 2

Item Identifier	Level	Theta	Level Description
HL000844*	I	-1.58	Supply the missing number
VB417883*	II	-0.45	Evaluate an expression for a specific value Determine an expression to model a scenario
Tilling Squares-Item 1*	II	-0.29	
VB434929*	II	-0.14	
Apartment Numbers-Item 1	II	-0.05	
Boxes of Chocolates-Item 1*	II	0.42	
EL001490*	II	1.69	
Emma’s Models-Item 1*	III	-0.33	Determine equations Linear relationship between two quantities
Party-Item 1*	III	0.31	
VB335172*	III	0.60	
VB434848*	III	0.68	
VB335163*	III	0.71	
XH000443*	III	2.30	
Apartment Numbers-Item 2	IV	-0.49	Use a rule without formally presenting the equation
Cups-Item 5	IV	0.32	
Fish Ponds-Item 2	IV	0.59	
Party-Item 2	IV	0.73	
Design a Garden-Item 3	IV	0.91	
Cups-Item 3	IV	0.96	
Cups-Item 2	IV	1.17	
Tiling Squares-Item 2	IV	1.27	
Fish Ponds-Item 3	V	0.61	Explain reasoning
Vacations-item 3	V	0.66	
VB434859	V	2.05	
Fish Ponds-Item 4	VI	0.79	Inversions
Apartment Numbers-Item 3	VI	1.10	
Party-Item 4	VI	1.20	
Design a Garden-Item 4	VI	2.16	
Emma’s Models-Item 4	VII	0.63	Develop a formal expression
Fish Ponds-Item 5	VII	0.87	
EL001486	VII	1.20	
Fish Ponds-Item 6	VII	1.22	
Apartment Numbers-Item 4	VII	1.39	

Item Identifier	Level	Theta	Level Description
Tiling Squares-Item 4	VII	1.50	
Apartment Numbers-Item 5	VII	1.52	
Tiling Squares-Item 3	VII	1.56	
Tiling Squares-Item 5	VII	1.62	
Cups-Item 6	VII	1.65	
Party-Item 3	VII	1.98	
Cups-Item 7	VIII	2.07	System of two equations
Picking Apples-Item 3	VIII	2.50	

Note: RP 65 = value on the IRT score scale (theta value) corresponding to the probability of a correct response of .65

*Same as Branch 1

Conclusions

Learning progressions are a highly popular innovation in assessment and instructional design. The core principles of learning progressions have strong theoretical and research grounding, although specific, practical instantiations are rare, at least in U.S. contexts. Given the salience of hypothesized learning progressions in the design of the CCSS and NGSS, it is important to consider the relevance of formally developed learning progressions for the future design of NAEP.

The CCSS and NGSS are narrative documents, similar to past standards documents, and, as such, are likely to influence the crafting of the next NAEP frameworks in a variety of ways. In this paper we considered the relevance of more *formally developed learning progressions* for NAEP, which would involve more detailed development of instructional activities and corresponding assessment tasks tied to the frameworks. Because NAEP must be sufficiently robust to assess progress on the standards across multiple curricula (unlike assessments in countries with a single, national curriculum), it is highly unlikely that formal learning progressions could be the main building blocks of a newly design NAEP. Furthermore, even if the intention were to create Grade 4 and Grade 8 cross-sections for NAEP that are consistent with CCSS sequences, it is important to recognize that more formal progressions at the needed level of specificity do not yet exist, and developing and field testing progressions is a much more extensive and costly procedure than assessment design alone.

If curriculum-linked learning progressions cannot be the primary or central building blocks for NAEP, the assessment must nonetheless be designed in such a way as to monitor the success of deeper curricular reforms where they occur. To continue to be an independent monitor and even a check on other assessments, NAEP must have a strategic vision that attends to both breadth and depth in representing subject-matter expertise. In a recent white paper on the future of NAEP (National Center for Education Statistics, 2012), an expert panel recommended that NAEP domain specifications be broadened so as to enable linkages with multiple other assessments, including long-term trend versions of NAEP, international assessments, and state consortium assessments. *Under such a design, the NAEP framework and reporting domain need not be the same as this comprehensive item pool, which might be thought of as a "super-assessment" domain or blueprint.* Until now, a NAEP framework has always been used as the complete blueprint for the intended assessment. Items were developed to represent the framework, and performance was reported in terms of the intended framework. In contrast, the 2012 panel recommended a dynamic approach to constituting the content domain of NAEP administrations so as to address explicitly how changing definitions of subject-matter domains affect immediate outcomes and reports of progress over time. More specifically, the NAEP reporting framework as historically conceived would be situated within a larger, super-assessment domain. Like a series of Venn diagrams, other assessment domains would also be located within the super assessment, with carefully designed shared and unique item sets. By spiraling these various assessments together in a single NAEP administration, the means for linking and equating studies would be built in rather than requiring separate linking studies.

The panel also cautioned that NAEP may not be able to administer its most ambitious and innovative assessment tasks to random samples of students because a lack of opportunity to learn could make the assessment too difficult for the majority of students. Instead, the panel recommended that NAEP first conduct special studies, as have been undertaken in the past, to determine whether more advanced performance can be documented in those settings where reform curricula have been successfully implemented. Thus, assessment tasks tied to learning progressions in mathematics, science, or literacy could be embedded within the NAEP super-assessment framework, and both performance outcomes and the psychometric functioning of the assessment tasks could be compared for students with and without instructional opportunities tied directly to learning progressions curricula.

In this study, we used familiar anchoring methodology to construct four quasi learning progressions from existing NAEP items in combination with BAM items. This exercise allowed us to consider the feasibility of building example learning progressions into the NAEP item pool to enable their use as a reporting strategy. Based on this exercise, we conclude that such an approach is infeasible and likely to be misleading until there is more widespread implementation of new standards and thereby greater congruence between hoped-for and empirical ordering of items. Although we can see ways to improve the meaningfulness of quasi learning progressions by eliminating misfitting items, in most cases these are not items that one would want to remove lightly. In the case of items found to be unpredictably difficult because of construct irrelevant variance, removing the items would have an overall positive effect on assessment quality. However, this particular reason for misfitting items occurred relatively rarely. The more difficult problem has to do with items that did not fit the intended progression because of cognitive challenges often caused by multidimensionality and/or curriculum specificity that might not be as misfitting if students had more direct experience with this type of item. Such items should not be eliminated from the assessment because they represent the very ambitions of the new standards documents. To anchor the scale with only the well-behaved items essentially moves more challenging items to a later place on the progression. These kinds of decisions can only be made after doing the kind of work that is required for the development of learning progressions (i.e., logical and expert-developed sequences must be tested in instructional contexts where students have had the opportunity to learn with the support of curricula specifically developed in conjunction with the intended progression).

References

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*, 389–421.

Ayala, C. C., Shavelson, R. J., Ruiz-Primo, M. A., Brandon, P., Yin, Y., Furtak, E. M., et al. (2008). From formal embedded assessments to reflective lessons: The development of formative assessment suites. *Applied Measurement in Education*, *21*(4), 315–334.

Baroody, A. J. (1984). Children's difficulties in subtraction: Some causes and questions. *Journal for Research in Mathematics Education*, *15*(3), 203–213.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*(2), 191–204.

Burstein, L., Koretz, D., Linn, R., Sugrue, B., Novak, J., Baker, E. L., et al. (1995/1996). Describing performance standards: Validity of the 1992 National Assessment of Educational Progress achievement level descriptors as characterizations of mathematics performance. *Educational Assessment*, *3*(1), 9–51.

Catley, K., Lehrer, R., & Reiser, B. (2004, May 6–7). *Tracing a prospective learning progression for developing understanding of evolution.* Commissioned paper for the National Research Council Committee on Test Design for K–12 Science Achievement Workshop, Washington, DC.

Clements, D., & Sarama, J. (2007a). *Building Blocks—SRA Real Math, Grade PreK*, Columbus, OH: SRA/McGraw-Hill.

Clements, D., & Sarama, J. (2007b). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education*, *38*(2), 136–163.

Clements, D., & Sarama, J. (2009). Learning and teaching early math: The learning trajectories approach. New York: Routledge.

Common Core State Standards Initiative. (2012). *Mathematics: Introduction: How to read the grade level standards.* Retrieved from http://www.corestandards.org/Math/Content/introduction/how-to-read-the-grade-level-standards

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform.* New York: Columbia University, Teachers College: Center on Continuous Instructional Improvement, Consortium for Policy Research in Education.

Daro, P., Stancavage, F., Ortega, M., DeStefano, L., & Linn, R. (2007, September). *Validity study of the NAEP mathematics assessment: Grades 4 and 8*. (Conducted by the NAEP Validity Studies [NVS] Panel). Palo Alto, CA: American Institutes for Research.

Duncan, R. G., Rogat, A., & Yarden, A. (2009). A learning progression for deepening students' understanding of modern genetics across the 5th–12th grades. *Journal of Research on Science Teaching*, *46*(6), 655–674.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K–8*. Washington, DC: National Academies Press.

Educational Testing Service. (1990). *The Civics Report Card*. Princeton, NJ: Author.

Forster, M., & Masters, G. (2004). Bridging the conceptual gap between classroom assessment and system accountability. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education*. Chicago: University of Chicago Press.

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice*, *10*(3), 3–9, 16.

Fuson, K. C. (1992). Research on whole number addition and subtraction. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 243–275). New York: Macmillan.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*, 519–521.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffler (Ed.) et al., *Studies in social psychology in World War II, Vol. IV. Measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.

Linn, R. L. (1998). Validating inferences from National Assessment of Educational Progress achievement-level reporting. *Applied Measurement in Education*, *11*(1), 23–47.

Masters, G. N., Adams, R. A., & Wilson, M. (1990). Charting of student progress. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies. Supplementary volume 2* (pp. 628–634). Oxford, UK: Pergamon Press.

Masters, G., & Forster, M. (1996). *Progress Maps: Assessment resource kit*. Camberwell, Victoria, AU: Australian Council for Educational Research.

Mathematics Assessment Resource Service (MARS). (2002, 2003). *Balanced Assessment in Mathematics*. Nottingham, UK: Mathematics Assessment Resource Service, Shell Centre for Mathematical Education.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research on Science Teaching*, *46*(6), 675–698.

National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment. (1992). *Assessing student achievement in the states*. Stanford, CA: National Academy of Education.

National Center for Education Statistics. (2012). *NAEP: Looking ahead—Leading assessment into the future*. Washington, DC: Author. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/Future_of_NAEP_Panel_White_Paper.pdf

National Governors Association Center for Best Practices and Council of Chief State School Officers. (2010a). Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects. Washington, DC: Author.

National Governors Association Center for Best Practices and Council of Chief State School Officers. (2010b). *Common Core State Standards for mathematics*. Washington, DC: Author.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Dordrecht, Netherlands: Kluwer.

Schmidt, W. H., Wang, H. C., & McKnight, C. C. (2005). Curriculum coherence: An examination of U.S. mathematics and science content standards from an international perspective. *Journal of Curriculum Studies*, *37*(5), 525–559.

Schulz, E. M., Lee, W. C., & Mullen, K. (2005). A domain-level approach to describing growth in achievement. *Journal of Educational Measurement*, *42*(1), 1–26.

Shanker, A. (1990). A proposal for using incentives to restructure our public schools. *Phi Delta Kappa*, *71*, 345–357.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. Focus Article. *Measurement: Interdisciplinary Research and Perspectives, 14*, 1–98.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research on Science Teaching*, *46*(6), 610–631.

Stancavage, V. B., Shepard, L., McLaughlin, D., Holtzman, D., Blankenship, C., & Zhang, Y. (2009). *Sensitivity of NAEP to the effects of reform-based teaching and learning in middle school mathematics*. A publication of the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

Sztajn, P., Confrey, J., Wilson, P. H., & Edgington, C. (2012). Learning trajectory based instruction: Toward a theory of teaching. *Educational Researcher*, *41*(5), 147–156.

West Australian Ministry of Education. (1991). *First steps spelling developmental continuum*. Perth, AU: Author

Appendix A. Items in Learning Progressions

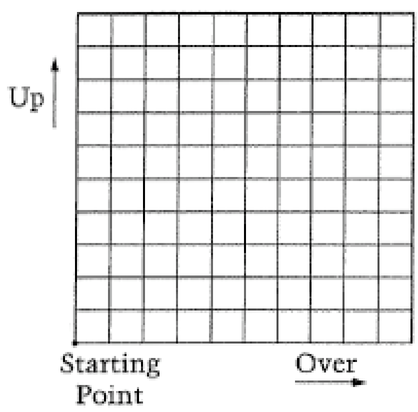
Figure A-1. Graphing Learning Progression

Level I

10. From the starting point on the grid below, a beetle moved in the following way. It moved 1 block up and then 2 blocks over, and then continued to repeat this pattern. Draw lines to show the path the beetle took to reach the right side of the grid.

Level I
Theta 0.57

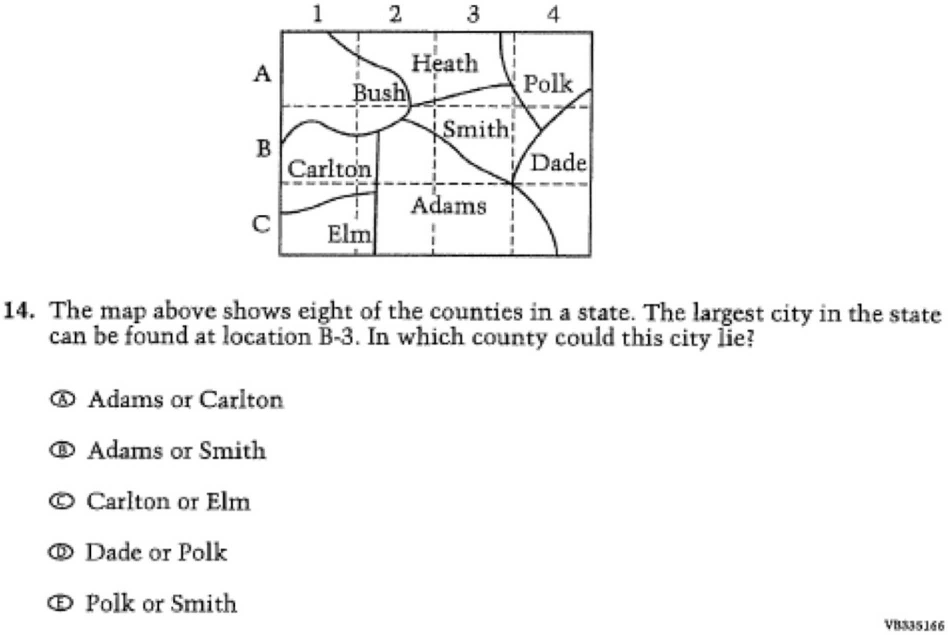
XH000442



Item XH000442.

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

Level II



14. The map above shows eight of the counties in a state. The largest city in the state can be found at location B-3. In which county could this city lie?

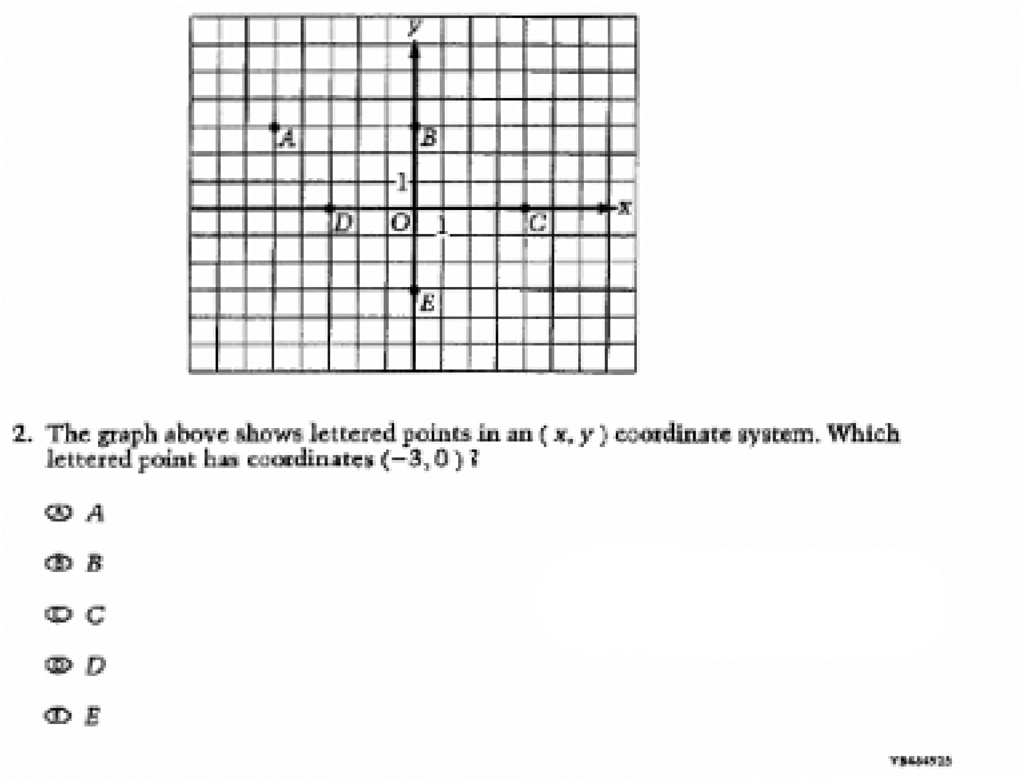
- Ⓐ Adams or Carlton
- Ⓑ Adams or Smith
- Ⓒ Carlton or Elm
- Ⓓ Dade or Polk
- Ⓔ Polk or Smith

VB335166

Level II
Theta -1.12

Item VB335166.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z2M12.



2. The graph above shows lettered points in an (x, y) coordinate system. Which lettered point has coordinates (-3, 0) ?

- Ⓐ A
- Ⓑ B
- Ⓒ C
- Ⓓ D
- Ⓔ E

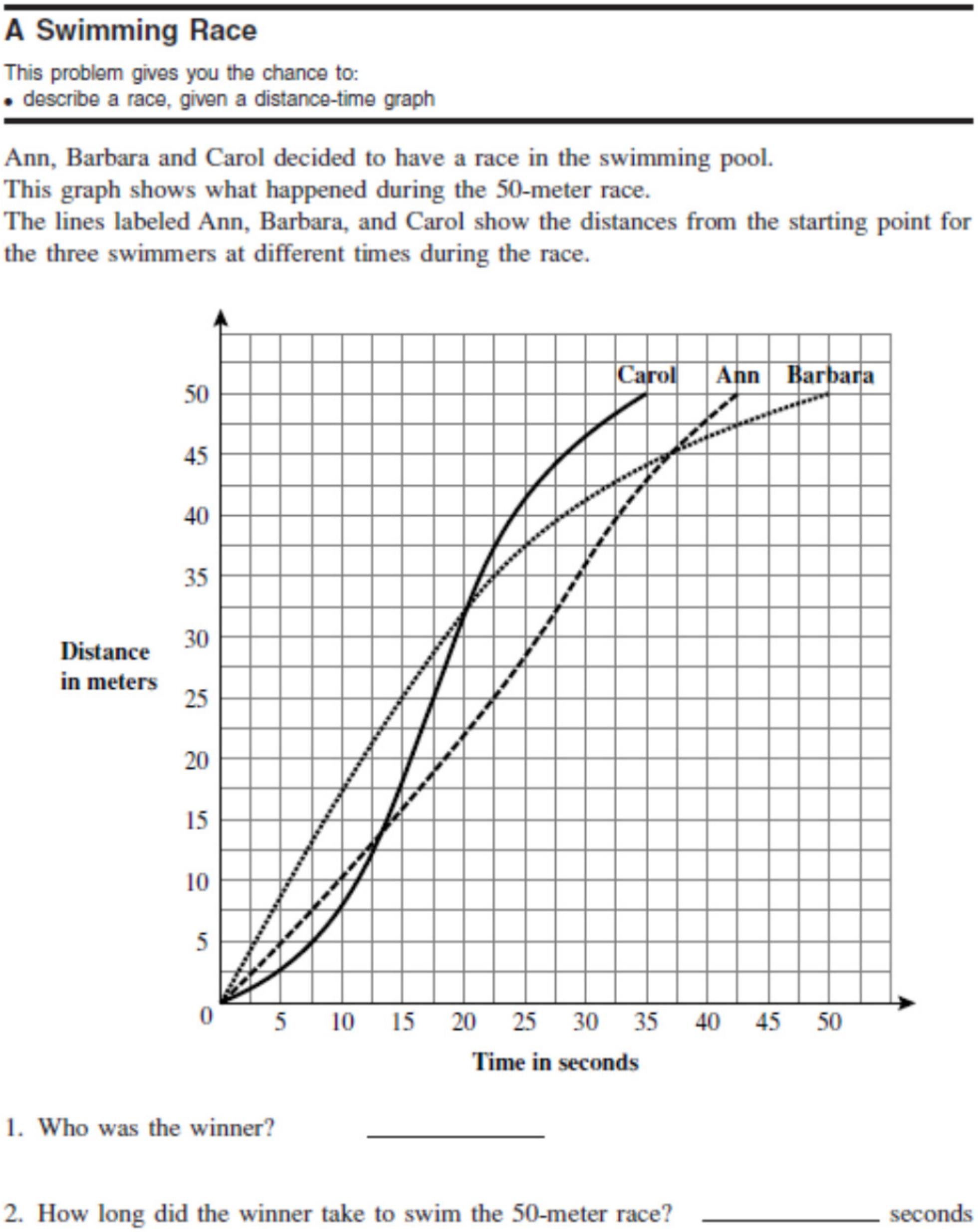
VB434925

Level II
Theta -0.24

Item VB434925.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z2M11.

Level III



A Swimming Race

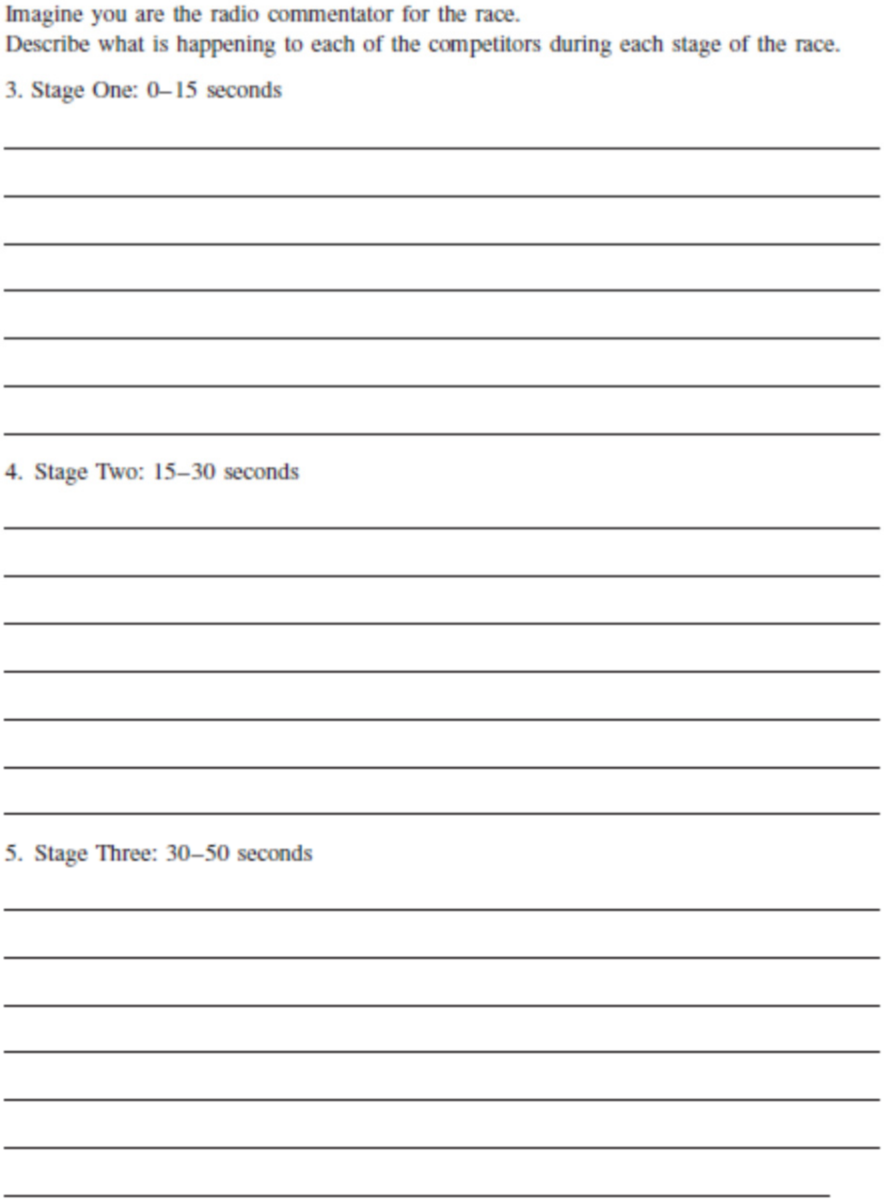
This problem gives you the chance to:
- describe a race, given a distance-time graph

Ann, Barbara and Carol decided to have a race in the swimming pool.
This graph shows what happened during the 50-meter race.
The lines labeled Ann, Barbara, and Carol show the distances from the starting point for the three swimmers at different times during the race.

1. Who was the winner? ____________

2. How long did the winner take to swim the 50-meter race? ____________ seconds

Level III
Theta -1.01



Imagine you are the radio commentator for the race.
Describe what is happening to each of the competitors during each stage of the race.

3. Stage One: 0–15 seconds

Level V
Theta 2.23

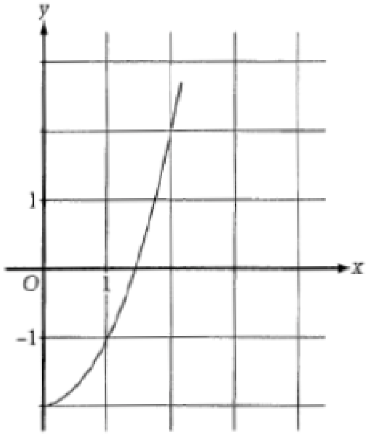
4. Stage Two: 15–30 seconds

Level V
Theta 1.89

5. Stage Three: 30–50 seconds

Level V
Theta 2.19

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.



11. On the curve above, what is the <u>best</u> estimate of the value of x when $y = 0$?

Ⓐ −2.0

Ⓑ 1.1

Ⓒ 1.4

Ⓓ 1.7

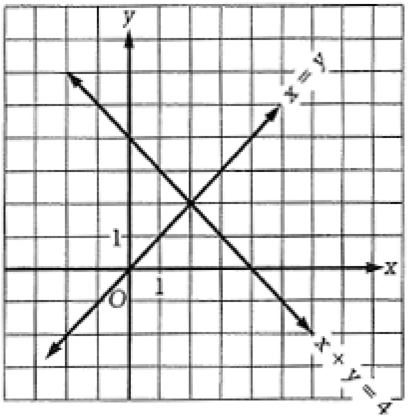
Ⓔ 1.9

Level III
Theta 1.06

Item YJ000078.

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

Level IV



VB429681

7. Which point is the solution to both equations shown on the graph above?

Ⓐ (0, 0)

Ⓑ (0, 4)

Ⓒ (1, 1)

Ⓓ (2, 2)

Ⓔ (4, 0)

Level IV
Theta 0.1

Item VB429681.

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

Vacations

This problem gives you the chance to:
- analyze relationships using graphs and algebra

Here is some information about how some students are paying for their summer vacations.

Carla: Her mom gave her $100 in January and Carla has saved $25 every month since, starting in February.

Arnie: Arnie put $150 in his piggy bank in January.

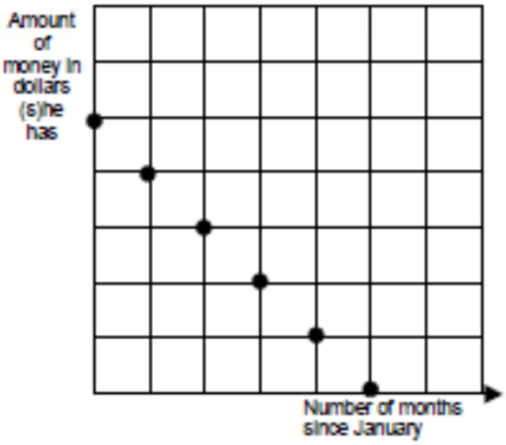
Sue: Sue booked her vacation in January. She had $250 in her piggy bank. Starting in February, she is paying $50 each month to the travel company.

Ben: Starting in February, Ben saves $30 every month.

Here are some graphs illustrating these situations.

Level IV
Theta 0.23

1. Match each person with a graph and explain how you decided.

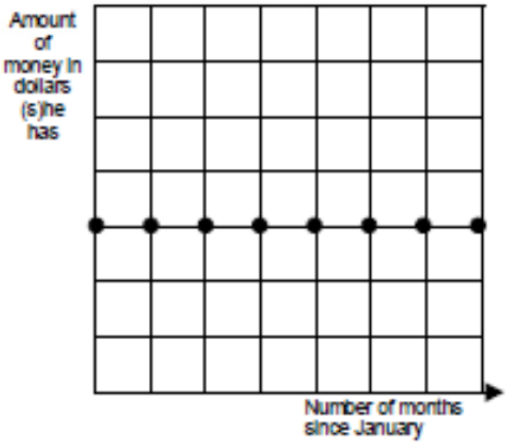


Name: ____________________

Reason: ____________________________

____

____

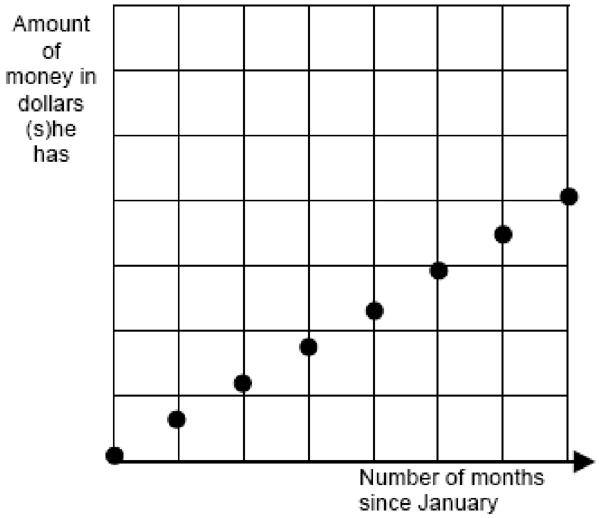


Name: ____________________

Reason: ____________________________

____

____

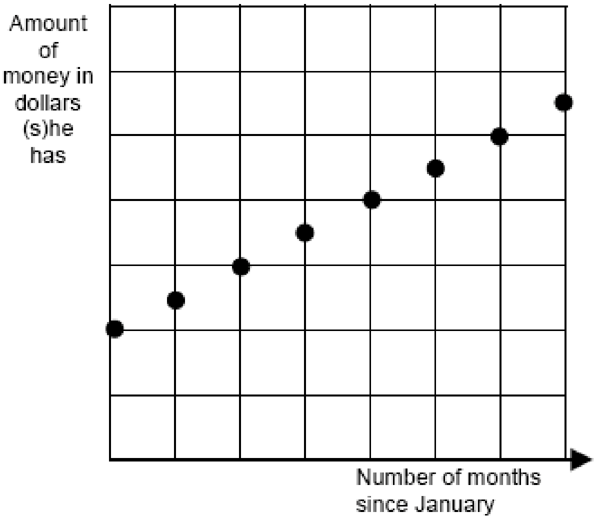


Name: ____________________

Reason: ____________________________

____

____



Name: ____________________

Reason: ____________________________

____

____

2. In these equations, \$$A$ is the amount of money and n is the number of months since January.

   $A = 250 - 50n$
   $A = 30n$
   $A = 150$

   a. Find the person for each of these equations.
   b. Write a formula for the fourth person.

   Carla ______________________________

   Arnie ______________________________

   Sue ________________________________

   Ben ________________________________

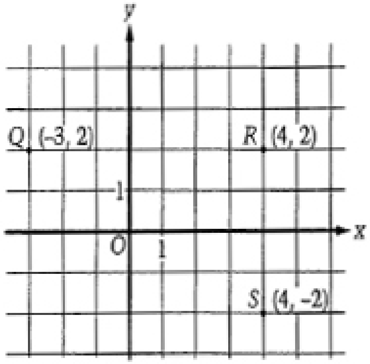
3. Write a possible description for this formula: $A = 50n + 150$

   ____

   ____

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Note: Only Item 1 from Vacations pertains to this progression. The remaining items 2 and 3 do not occur in this progression.



13. If the points *Q, R,* and *S* shown above are three of the vertices of rectangle *QRST,* which of the following are the coordinates of *T* (not shown) ?

Level IV
Theta 0.38

AP000711

Ⓐ (4, − 3)

Ⓑ (3, − 2)

Ⓒ (−3, 4)

Ⓓ (− 3 , − 2)

Ⓔ (−2, − 3)

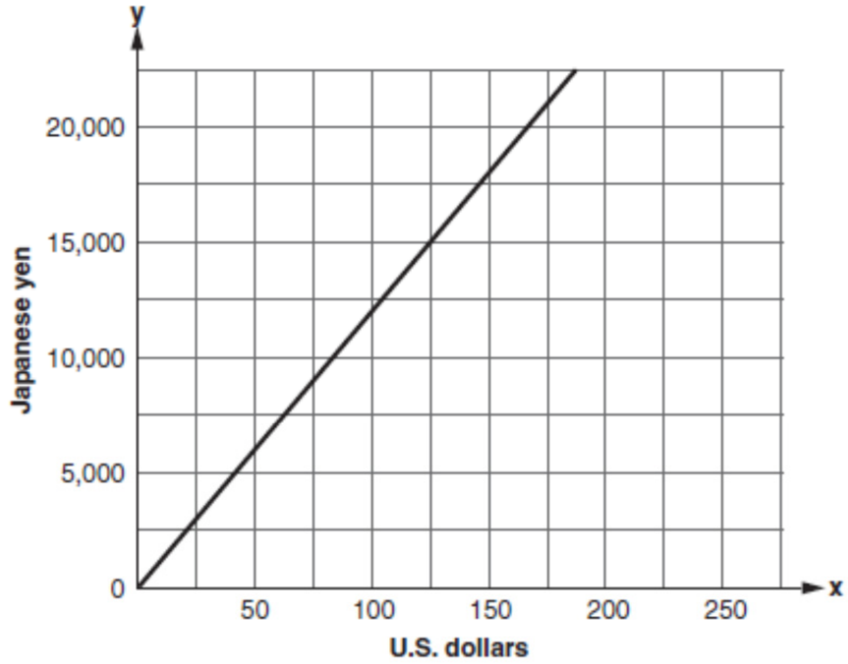
Item AP000711.

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

Dollars

This problem gives you the chance to:
- use a graph to convert currency

This graph can be used to convert between U.S. dollars and Japanese yen.



1. Use the graph to estimate how many Japanese yen you would get for 100 U.S. dollars.

On the graph, show how you found your answer. ________________

Level IV
Theta 0.47

2. Use the graph to find out how many U.S. dollars you would get for 20,000 Japanese yen.

Show how you found your answer. ________________

Level IV
Theta 0.60

3. Use the graph to estimate the number of Japanese yen you would get
for 1,000 U.S. dollars.
Explain how you figured it out. ________________

Level V
Theta 1.04

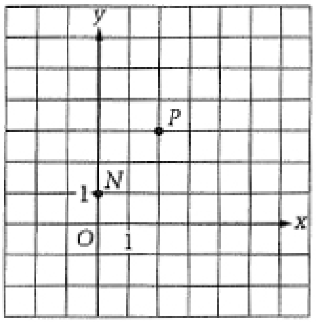
____

____

____

____

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.



VB434830

14. For the figure above, which of the following points would be on the line that passes through points N and P?

Ⓐ $(-2, 0)$

Ⓑ $(0, 0)$

Ⓒ $(1, 1)$

Ⓓ $(4, 5)$

Ⓔ $(5, 4)$

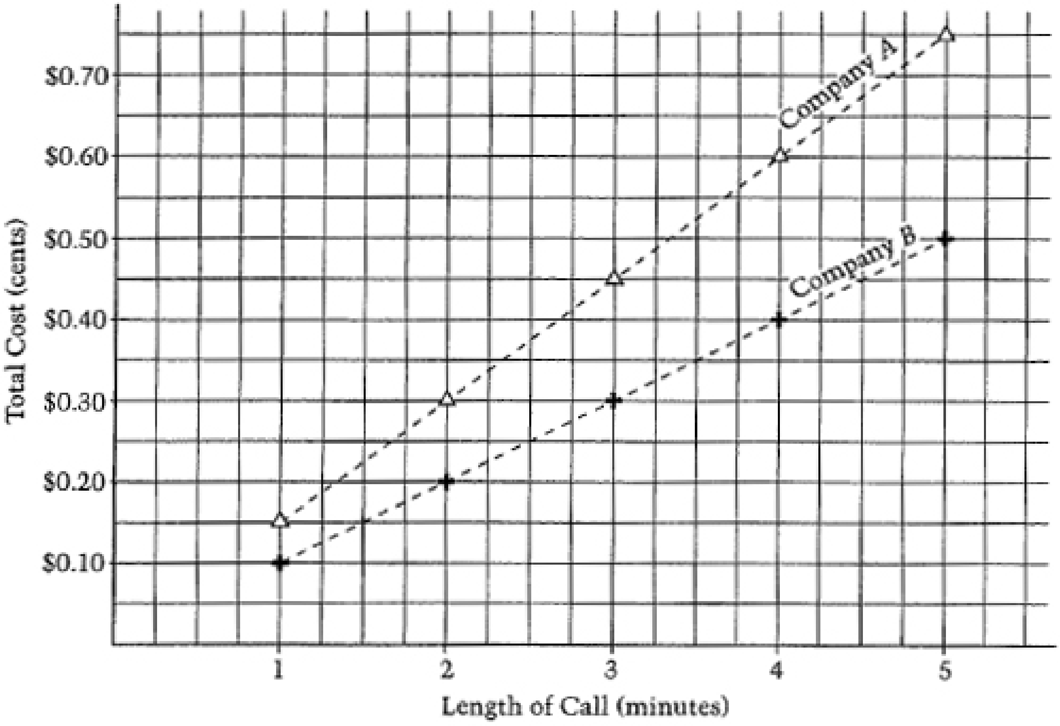
Level IV
Theta 0.87

Item VB434830.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M10.

Level V

20. The graph below shows the cost that two long-distance telephone companies each charge for calls of various lengths (in minutes).



a. What is the cost of a 4-minute call using Company *B* ?

____

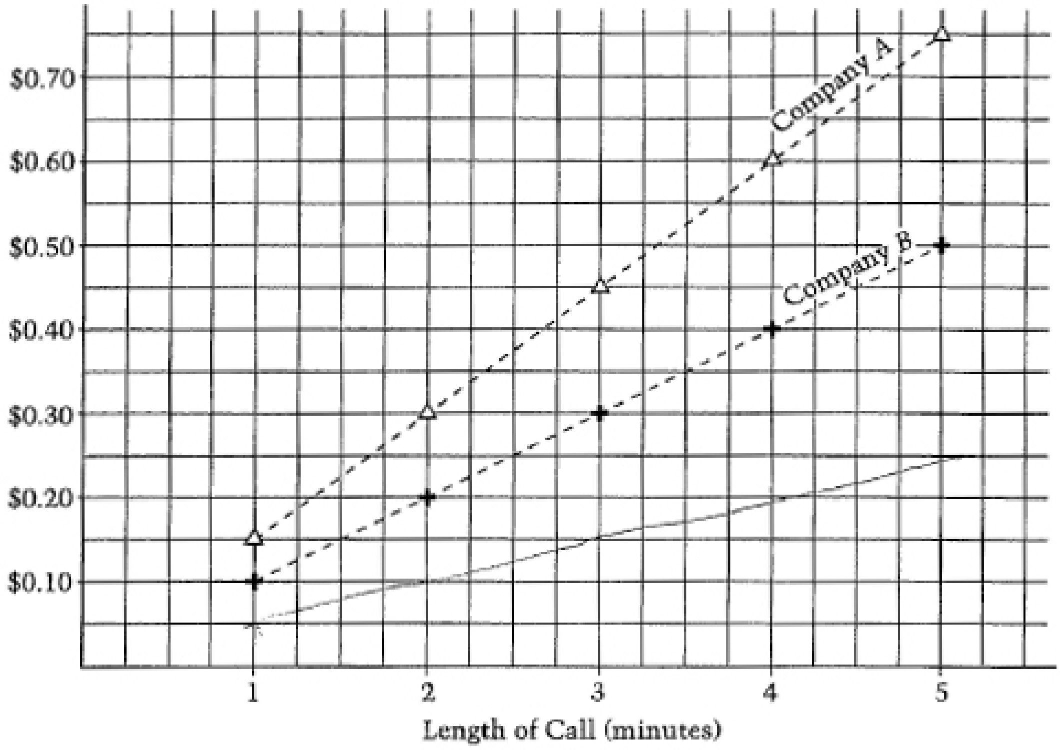
b. What is the cost per minute for a call using Company *B* ?

____

c. Determine the amounts of money saved (in cents) by using Company *B* instead of Company *A* when calls of 1, 2, 3, 4, and 5 minutes are made. Then graph the five points that represent the savings on the grid below and connect the points with a dotted line.

Level V
Theta -0.42

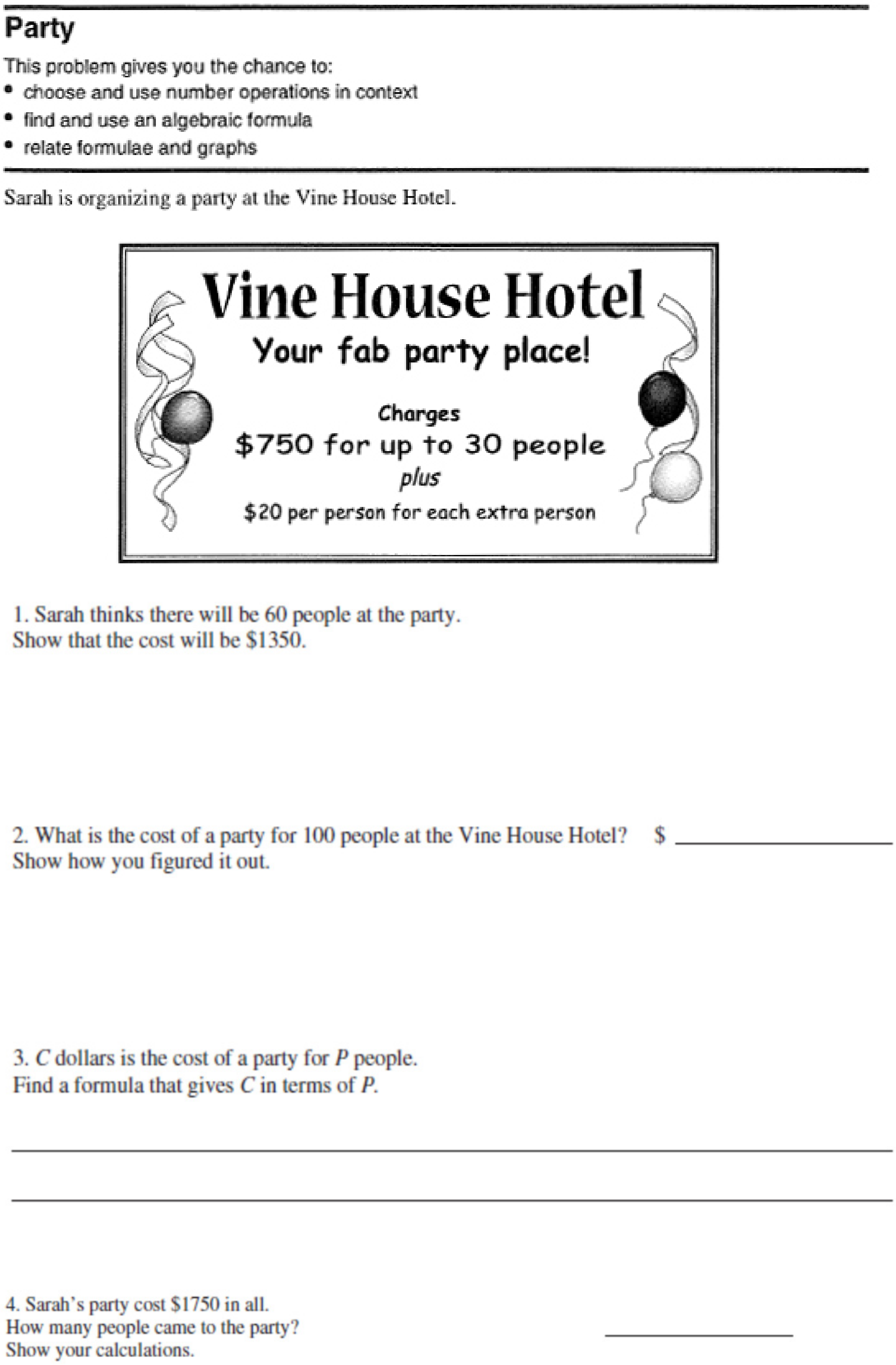
YJ000089



Item YJ000089.

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

For Dollars Item 3 (Level V, Theta 1.04), please see page 188.



Party

This problem gives you the chance to:

- choose and use number operations in context
- find and use an algebraic formula
- relate formulae and graphs

Sarah is organizing a party at the Vine House Hotel.

Vine House Hotel

Your fab party place!

Charges

$750 for up to 30 people

plus

$20 per person for each extra person

1. Sarah thinks there will be 60 people at the party.
Show that the cost will be $1350.

2. What is the cost of a party for 100 people at the Vine House Hotel? $ ____________________
Show how you figured it out.

3. *C* dollars is the cost of a party for *P* people.
Find a formula that gives *C* in terms of *P*.

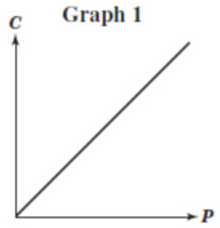
____

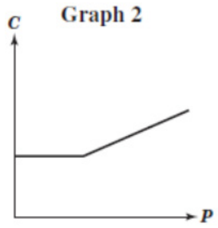
____

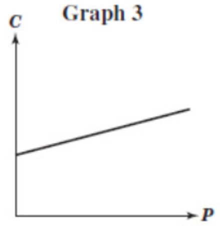
4. Sarah's party cost $1750 in all.
How many people came to the party? ____________________
Show your calculations.

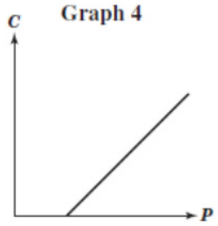
5. Which of these graphs shows the connection between the number of people at the party, P, and the cost, C?

Level V
Theta 1.53









Explain how you figured it out.

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

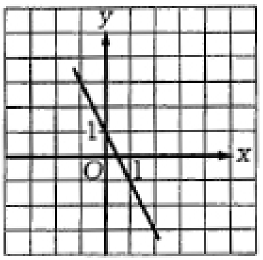
Note: Only Item 5 from Party pertains to this progression. The remaining items 1, 2, 3, and 4 do not occur in this progression.

11. **Which of the following is the graph of the line with equation $y = -2x + 1$?**

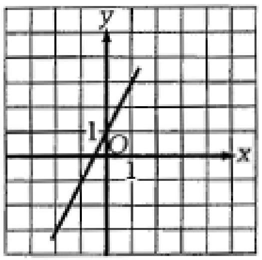
Level V
Theta 1.79

VB434934

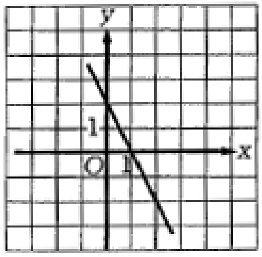
Ⓐ



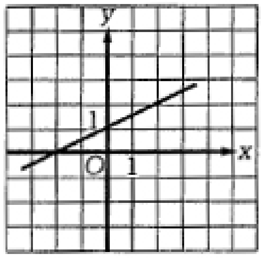
Ⓑ



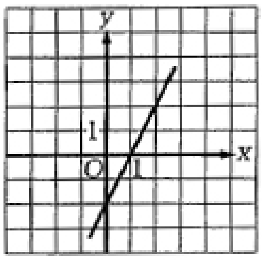
Ⓒ



Ⓓ



Ⓔ



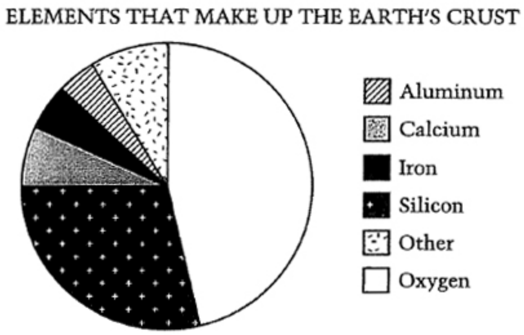
Item VB434934.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z2M11.

For A Swimming Race Item 4 (Level V, Theta 1.89), please see page 182.
For A Swimming Race Item 5 (Level V, Theta 2.19), please see page 182.
For A Swimming Race Item 3 (Level V, Theta 2.23), please see page 182.

Figure A-2. Statistics Learning Progression

Level I



8. According to the graph above, which element forms the second greatest portion of the earth's crust?

Ⓐ Oxygen

Ⓑ Silicon

Ⓒ Aluminum

Ⓓ Iron

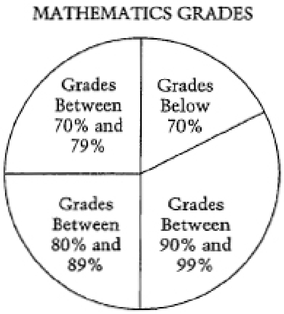
Ⓔ Calcium

VB335159

Level I
Theta -1.42

Item VB335159.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z2M12**.**



4. The circle graph above shows the distribution of grades for the 24 students in Shannon's mathematics class. Consider each of the following statements. Can the conclusion be made from the graph?

Fill in one oval to indicate YES or NO for each statement.

	Yes	No
(a) About $\frac{1}{2}$ of the class has a grade of 90% or better.	○	○
(b) Over $\frac{1}{2}$ of the class has a grade of 80% or better.	○	○
(c) There are no students with a grade of 60%.	○	○
(d) There are fewer students with a grade below 70% than there are between 70% and 79%.	○	○

HW000854

Level I
Theta 0.01

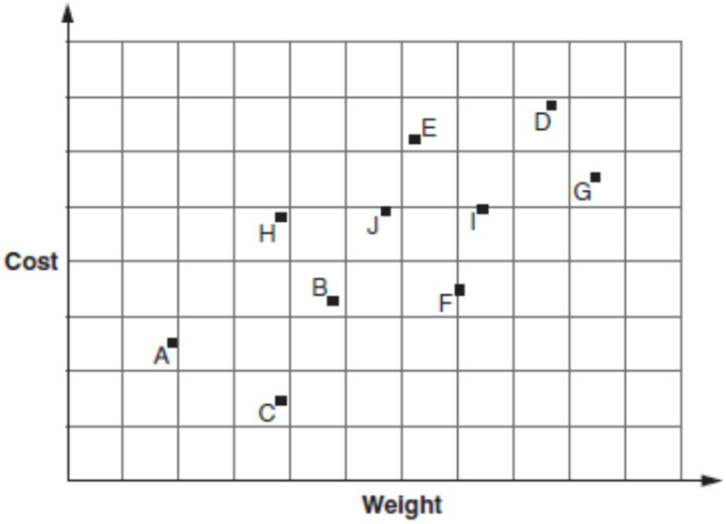
Item HW000854.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z23M8B.

Boxes of Candy

This problem gives you the chance to:
- interpret a scatter graph

This scatter graph shows the weights and the costs of 10 boxes of candy, A through J.



1. Which box of candy is the most expensive? ________________

2. Which two boxes of candy weigh the same? ________________

Level I
Theta 0.89

3. Which box of candy appears to be the best value for the money?

____

Explain how you found your answer.

____

____

4. What does the scatter graph show about the connection between the weights of the boxes of candy and how much they cost?

Level VII
Theta 1.21

____

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Level II

6. The results of a class survey on whether students liked a new television show are as follows.

25 students liked the new show.
15 students disliked the new show.
5 students had no opinion on the new show.

On the graph below, each ☺ represents 5 students. Draw the correct number of faces to illustrate the results of the class survey.

☺ = 5 students

Liked	
Disliked	
No Opinion	

IY002250

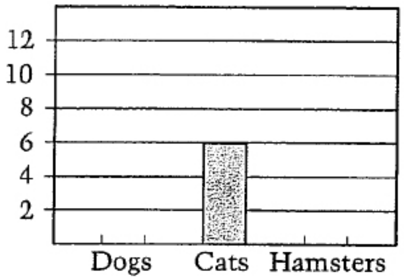
Item IY002250.

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

7. Draw bars on the graph below so that the number of dogs is twice the number of cats and the number of hamsters is one-half the number of cats.

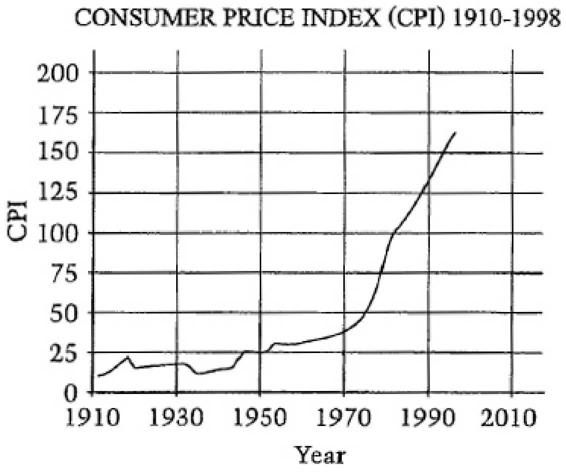
Level II
Theta -0.76

OM000557



Item OM000557.

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



8. The 1990 Consumer Price Index (CPI) was <u>about</u> how many times the 1950 CPI ?

Level II
Theta 0.65

Ⓐ 2

Ⓑ 5

Ⓒ 10

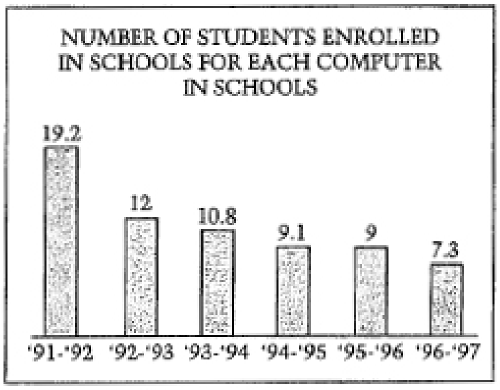
Ⓓ 25

Ⓔ 100

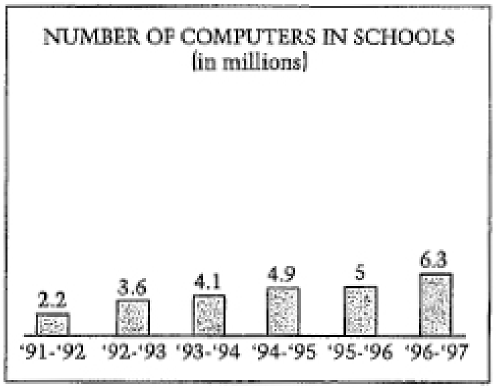
YJ000102

Item YJ000102.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z23M8B**.**



NUMBER OF STUDENTS ENROLLED IN SCHOOLS FOR EACH COMPUTER IN SCHOOLS



NUMBER OF COMPUTERS IN SCHOOLS (in millions)

13. Based on the information in the graphs above, how many students were enrolled in schools in '96-'97 ?

Level II
Theta 2.03

____

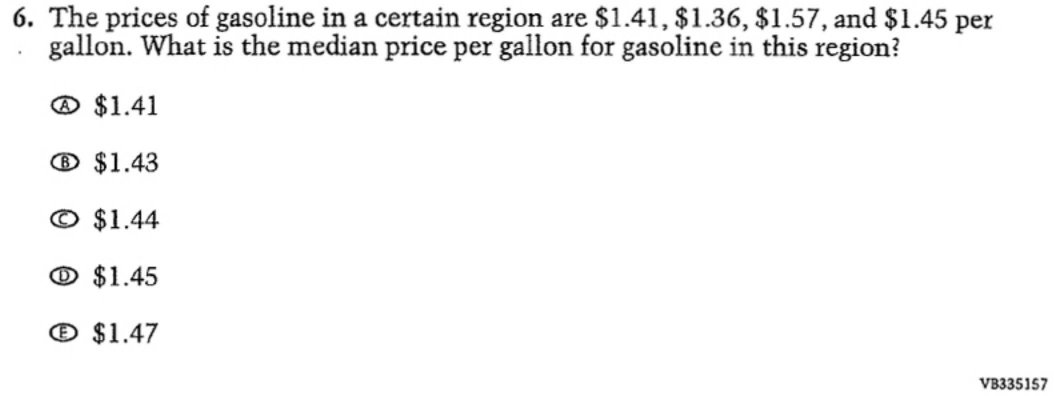
Show how you found your answer.

YJ000093

Item YJ000093.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z23M8B.

Level III



6. The prices of gasoline in a certain region are \$1.41, \$1.36, \$1.57, and \$1.45 per gallon. What is the median price per gallon for gasoline in this region?

Ⓐ \$1.41

Ⓑ \$1.43

Ⓒ \$1.44

Ⓓ \$1.45

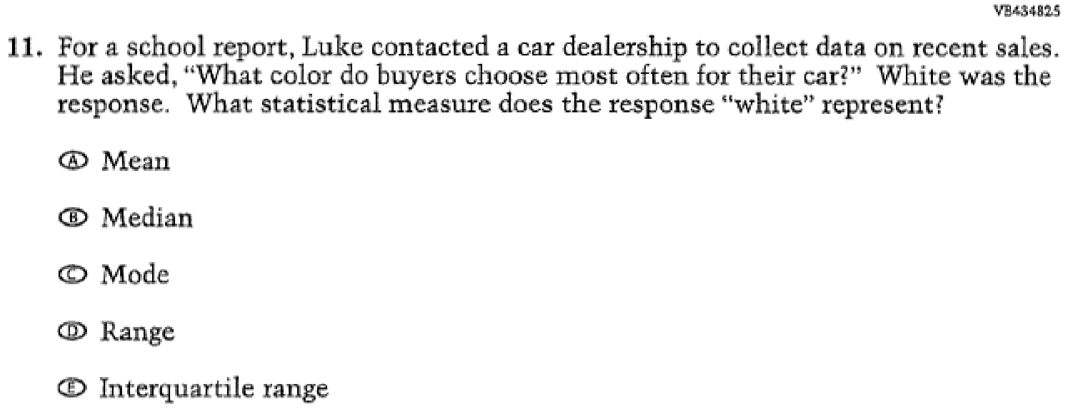
Ⓔ \$1.47

VB335157

Level III
Theta 0.99

Item VB335157.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z2M12**.**



VB434825

11. For a school report, Luke contacted a car dealership to collect data on recent sales. He asked, "What color do buyers choose most often for their car?" White was the response. What statistical measure does the response "white" represent?

Ⓐ Mean

Ⓑ Median

Ⓒ Mode

Ⓓ Range

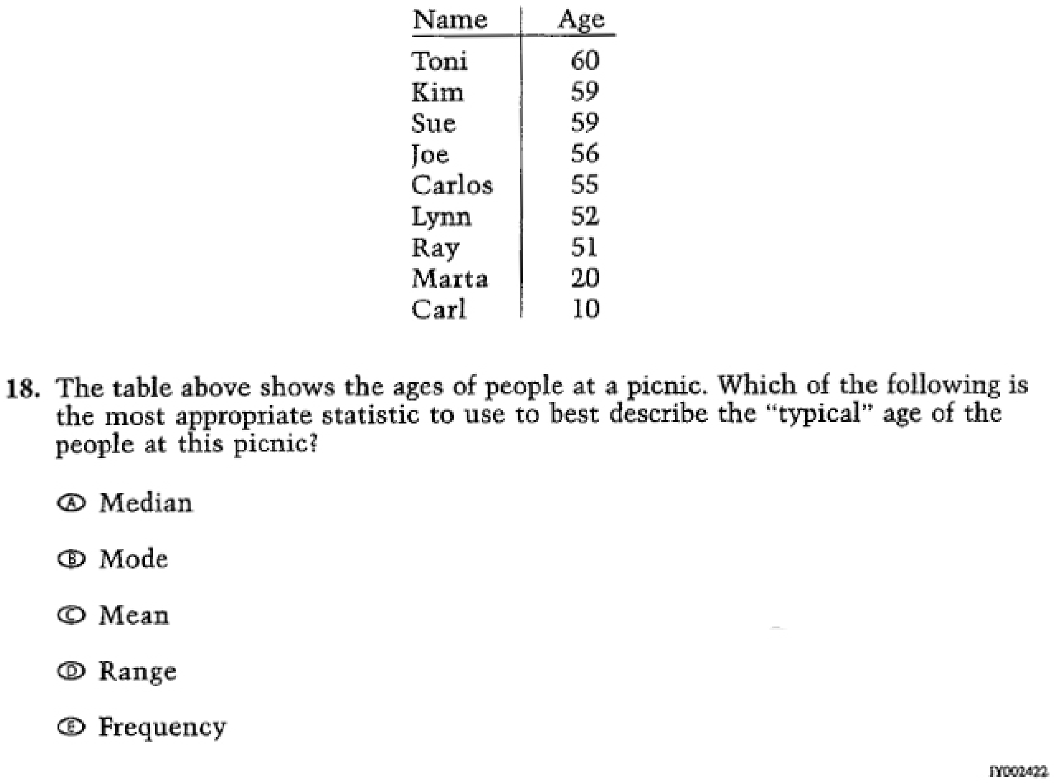
Ⓔ Interquartile range

Level III
Theta 1.07

Item VB434825.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M10.

Level IV



Name	Age
Toni	60
Kim	59
Sue	59
Joe	56
Carlos	55
Lynn	52
Ray	51
Marta	20
Carl	10

18. The table above shows the ages of people at a picnic. Which of the following is the most appropriate statistic to use to best describe the "typical" age of the people at this picnic?

Ⓐ Median

Ⓑ Mode

Ⓒ Mean

Ⓓ Range

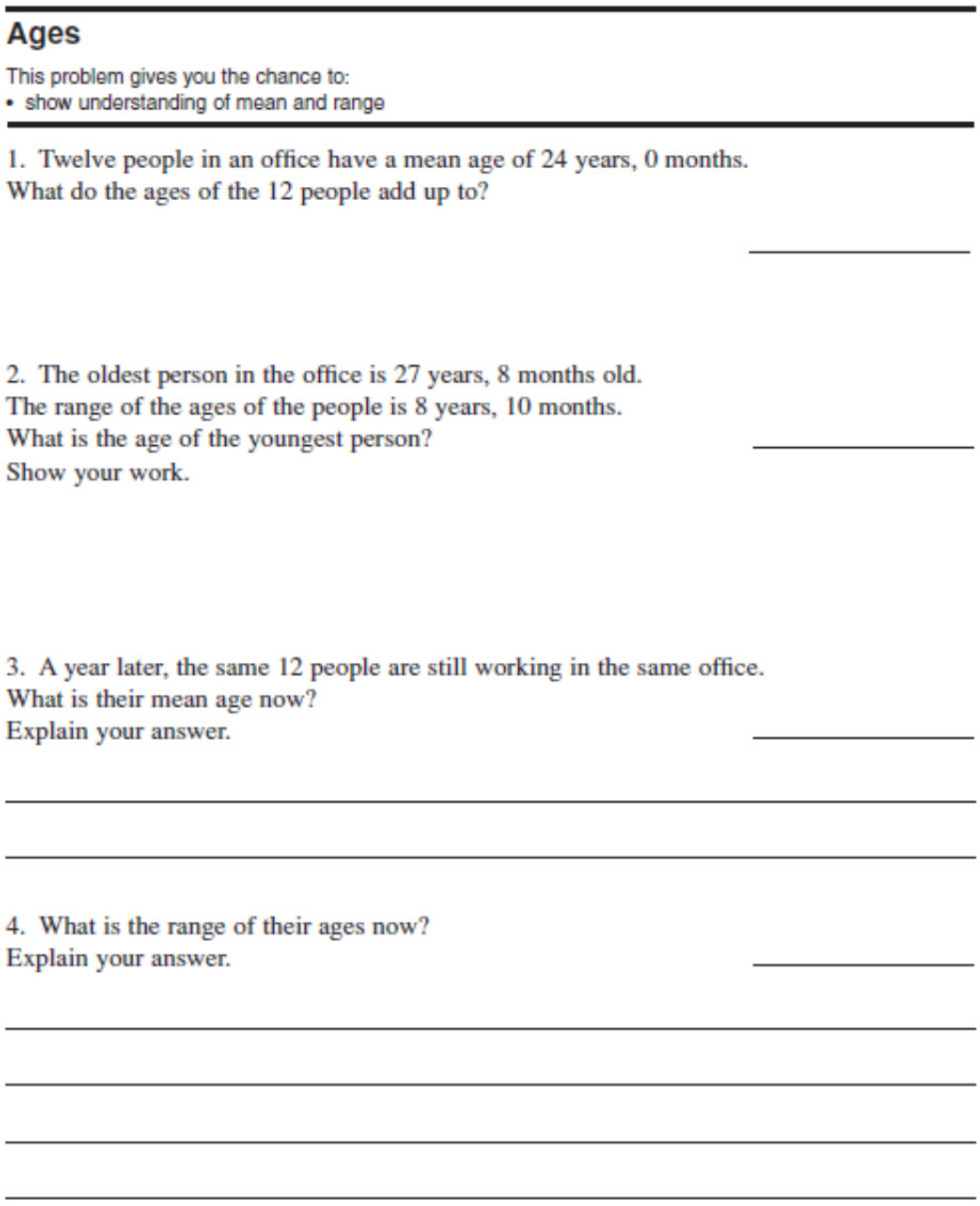
Ⓔ Frequency

IY002422

Level IV
Theta 1.1

Item IY002422.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z23M8B.



Ages

This problem gives you the chance to:
- show understanding of mean and range

1. Twelve people in an office have a mean age of 24 years, 0 months.
What do the ages of the 12 people add up to?

2. The oldest person in the office is 27 years, 8 months old.
The range of the ages of the people is 8 years, 10 months.
What is the age of the youngest person?
Show your work.

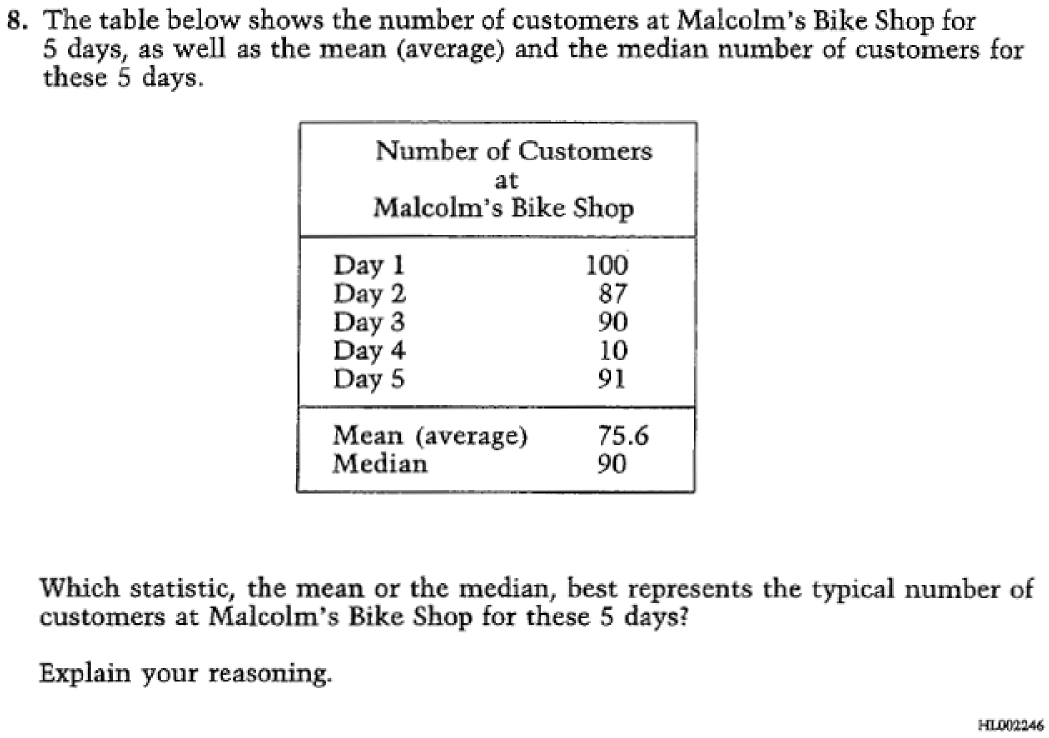
3. A year later, the same 12 people are still working in the same office.
What is their mean age now?
Explain your answer.

4. What is the range of their ages now?
Explain your answer.

Level IV
Theta 1.39

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Note: Only Item 3 from Ages pertains to this progression. The remaining items 1, 2, 4, and 5 do not occur in this progression.



8. The table below shows the number of customers at Malcolm's Bike Shop for 5 days, as well as the mean (average) and the median number of customers for these 5 days.

Number of Customers at Malcolm's Bike Shop	
Day 1	100
Day 2	87
Day 3	90
Day 4	10
Day 5	91
Mean (average)	75.6
Median	90

Which statistic, the mean or the median, best represents the typical number of customers at Malcolm's Bike Shop for these 5 days?

Explain your reasoning.

HL002246

Level IV
Theta 7.62

Item HL002246.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z23M9B.

Level V

VB417888

2. Which of the following types of graph would be best to show the change in temperature recorded in a city every 15 minutes over a 24-hour period?

Level V
Theta -0.86

Ⓐ Pictograph

Ⓑ Circle graph

Ⓒ Line graph

Ⓓ Box-and-whisker plot

Ⓔ Stem-and-leaf plot

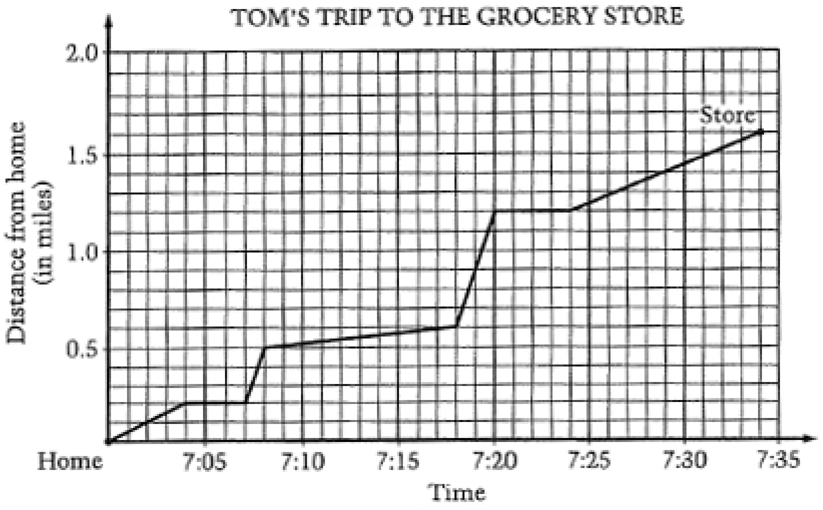
Item VB417888.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M7**.**

VB434849

10. Tom went to the grocery store. The graph below shows Tom's distance from home during his trip.

Level V
Theta 1.18



Tom stopped twice to rest on his trip to the store. What is the total amount of time that he spent resting?

Ⓐ 5 minutes

Ⓑ 7 minutes

Ⓒ 8 minutes

Ⓓ 10 minutes

Ⓔ 25 minutes

Item VB434849.

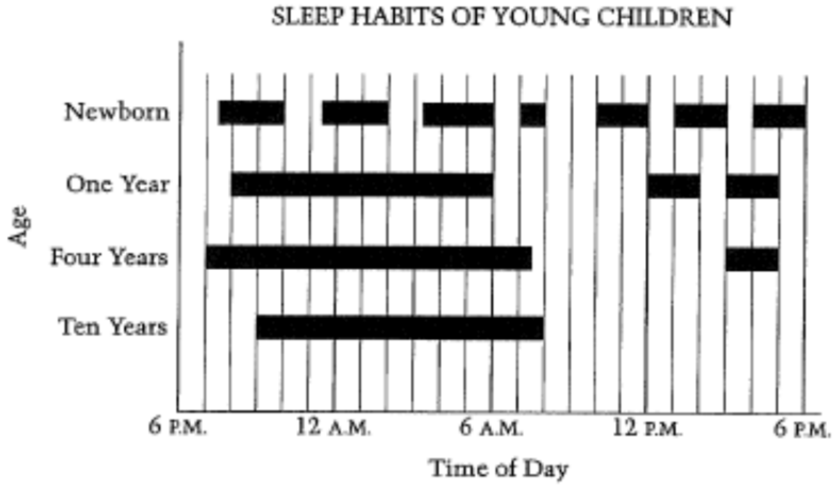
Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M10**.**

Level V
Theta 1.56

18. The graph below and written summary on the next page present information about the sleep habits of newborn babies, one year olds, four year olds, and ten year olds. Each solid bar represents a period of sleep.

Some of the information presented in the summary does not agree with the information in the graph.

For example, there is an error in sentence 1 that has already been identified and corrected for you.



- In sentences 2 and 3 below, underline the information that is not correct based on the graph. There is an error in each sentence.
- Then, write the correct information above the errors in sentences 2 and 3.

(1) According to research that has been done on sleep habits and patterns of sleep in children, the number of hours that a newborn baby sleeps in a 24-hour
more
period of time is less than that of a ten year old.

(2) From the time a child is born until it reaches age ten, the number of different time periods of sleep increases as the child grows older.

(3) Newborns need 2 more hours of sleep than ten year olds between 6 a.m. and 6 p.m.

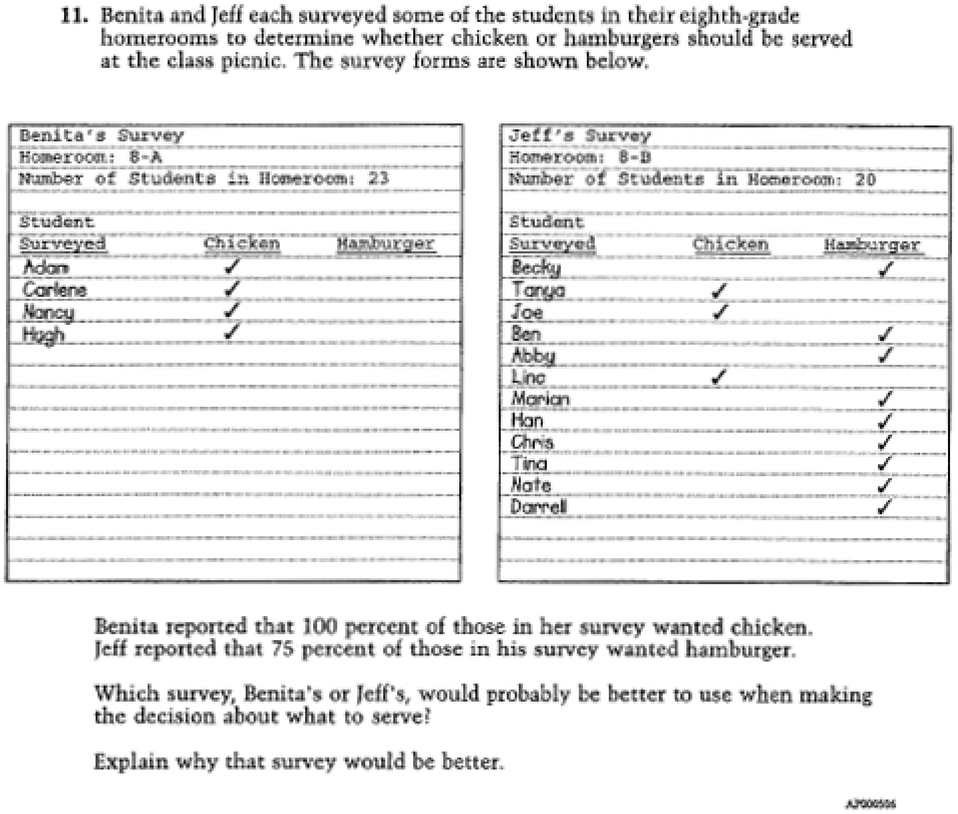
YJ000060

Item YJ000060.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z23M9B.

Level VI

Level VI
Theta 0.15



11. Benita and Jeff each surveyed some of the students in their eighth-grade homerooms to determine whether chicken or hamburgers should be served at the class picnic. The survey forms are shown below.

Benita's Survey
Homeroom: 8-A
Number of Students in Homeroom: 23

Student Surveyed	Chicken	Hamburger
Adam	✓	
Carlene	✓	
Nancy	✓	
Hugh	✓	

Jeff's Survey
Homeroom: 8-B
Number of Students in Homeroom: 20

Student Surveyed	Chicken	Hamburger
Becky		✓
Tanya	✓	
Joe	✓	
Ben		✓
Abby		✓
Lina	✓	
Marian		✓
Han		✓
Chris		✓
Tina		✓
Nate		✓
Darrell		✓

Benita reported that 100 percent of those in her survey wanted chicken.
Jeff reported that 75 percent of those in his survey wanted hamburger.

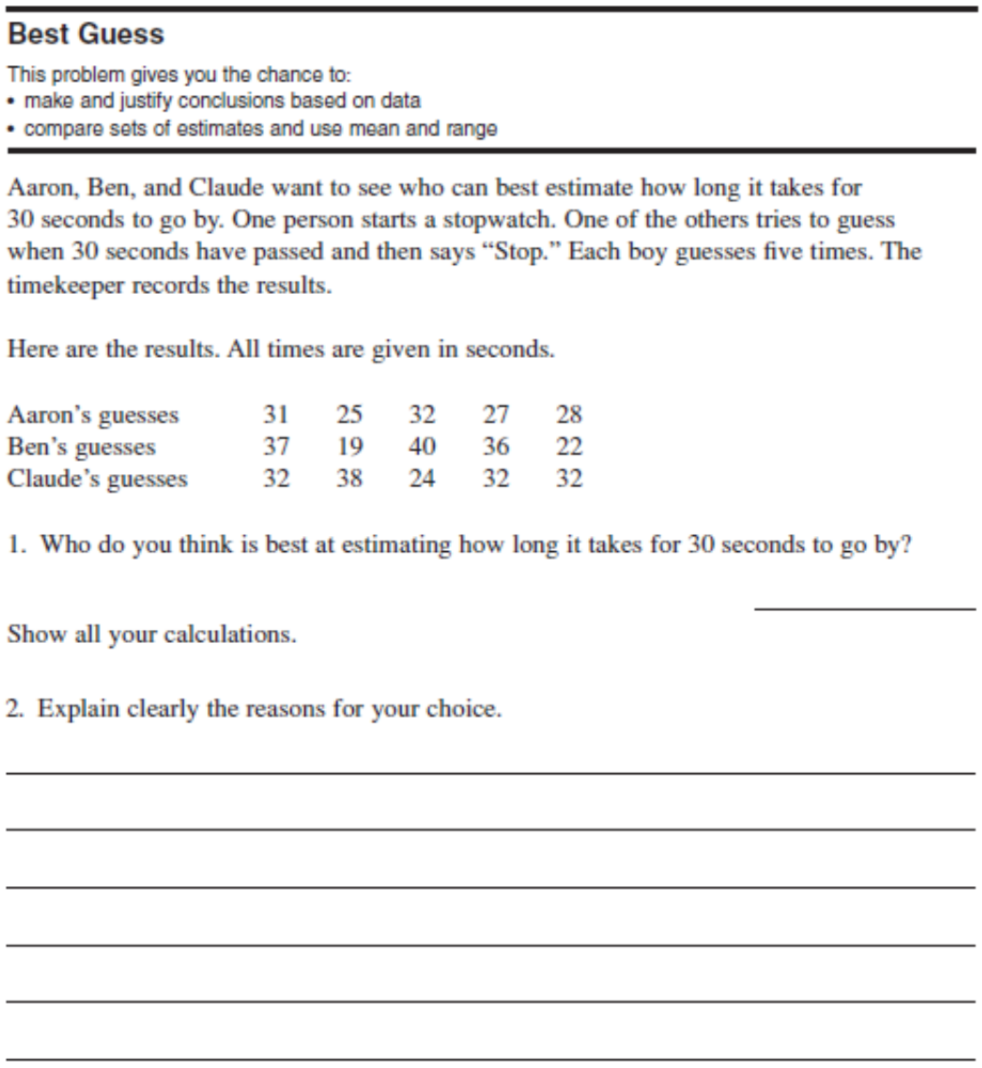
Which survey, Benita's or Jeff's, would probably be better to use when making the decision about what to serve?

Explain why that survey would be better.

AP000506

Item AP000506.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z23M9B.



Best Guess

This problem gives you the chance to:
- make and justify conclusions based on data
- compare sets of estimates and use mean and range

Aaron, Ben, and Claude want to see who can best estimate how long it takes for 30 seconds to go by. One person starts a stopwatch. One of the others tries to guess when 30 seconds have passed and then says "Stop." Each boy guesses five times. The timekeeper records the results.

Here are the results. All times are given in seconds.

Aaron's guesses	31	25	32	27	28
Ben's guesses	37	19	40	36	22
Claude's guesses	32	38	24	32	32

1. Who do you think is best at estimating how long it takes for 30 seconds to go by?

Level VI
Theta 3.97

____

Show all your calculations.

2. Explain clearly the reasons for your choice.

Level VI
Theta 3.11

____

____

____

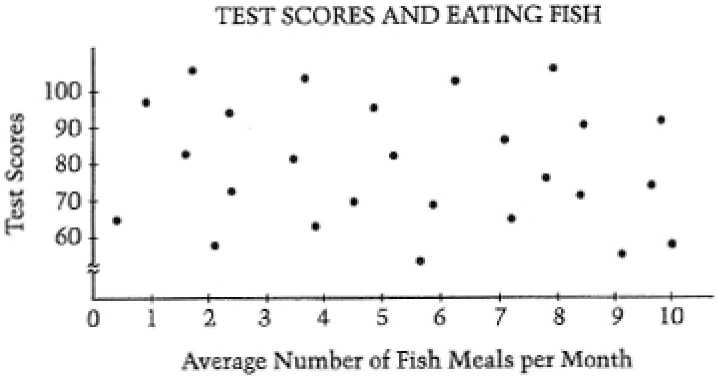
____

____

____

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Level VII



VB417891

Level VII
Theta 0.28

13. For a science project, Marsha made the scatterplot above that gives the test scores for the students in her math class and the corresponding average number of fish meals per month. According to the scatterplot, what is the relationship between test scores and the average number of fish meals per month?

Ⓐ There appears to be no relationship.

Ⓑ Students who eat fish more often score higher on tests.

Ⓒ Students who eat fish more often score lower on tests.

Ⓓ Students who eat fish 4-6 times per month score higher on tests than those who do not eat fish that often.

Ⓔ Students who eat fish 7 times per month score lower on tests than those who do not eat fish that often.

Item VB417891.

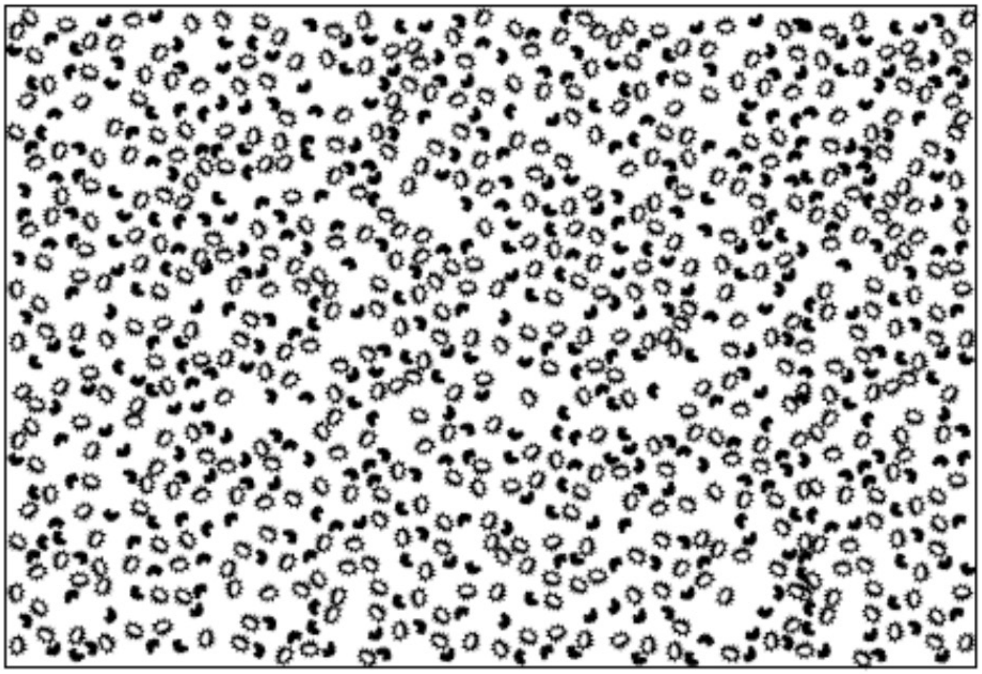
Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M7.

Bacteria

This problem gives you the chance to:
- use a sampling strategy to estimate a large number

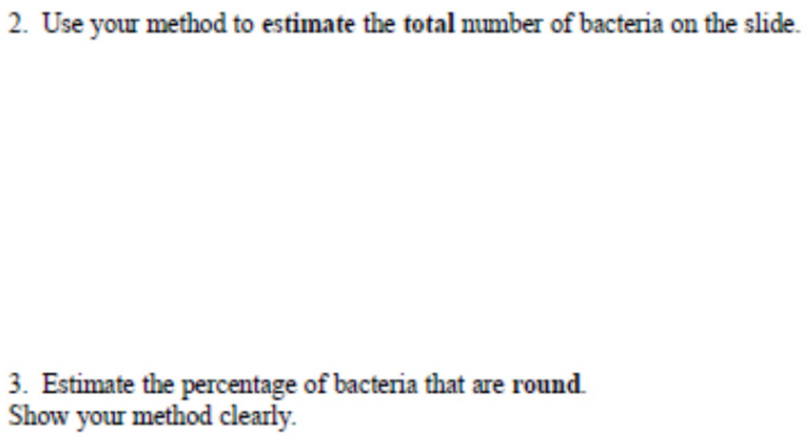
Two types of bacteria are shown on this microscope slide.

Some are **long,** with no "holes":  Some are **round,** with "holes": 



1. It would take a scientist a long time to count all these bacteria one by one. Describe a quicker method that could be used to **estimate** the **total** number of bacteria on the slide.

Level VII
Theta 0.67



2. Use your method to **estimate** the **total** number of bacteria on the slide.

Level VII
Theta 1.62

Level VII
Theta 2.26

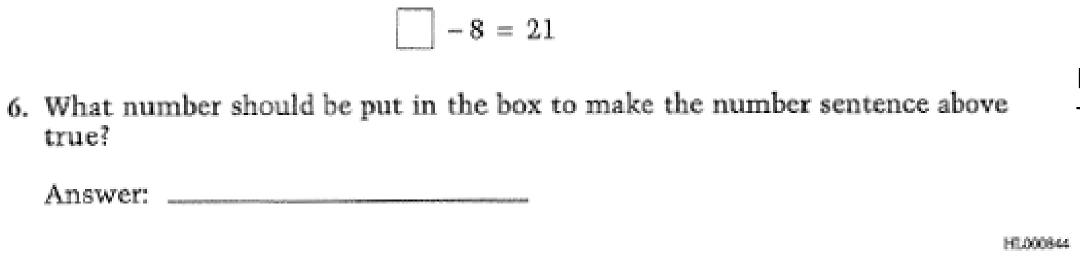
3. Estimate the percentage of bacteria that are **round**.
Show your method clearly.

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

For Boxes of Candy Item 4 (Level VII, Theta 1.21), please see page 196.

Figure A-3. Equations Learning Progression Branch 1

Level I



$\boxed{} - 8 = 21$

6. What number should be put in the box to make the number sentence above true?

Answer: ____________________

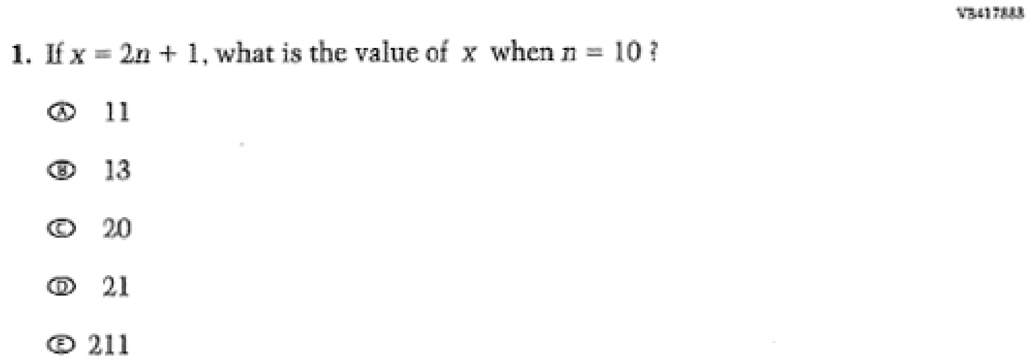
HL000844

Level I
Theta -1.58

Item HL000844.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z12M5B.

Level II



VB417883

1. If $x = 2n + 1$, what is the value of x when $n = 10$?

Ⓐ 11

Ⓑ 13

Ⓒ 20

Ⓓ 21

Ⓔ 211

Level II
Theta -0.45

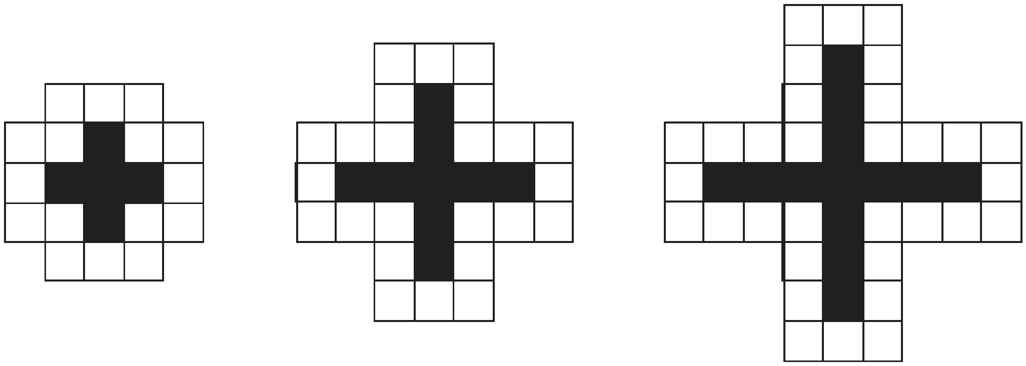
Item VB417883.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M7.

Tiling Squares

This problem gives you the chance to:
- extend and check patterns
- derive formulas connecting different pairs of variables

Marcia is using black and white square tiles to make patterns.



Pattern 1 Pattern 2 Pattern 3

1. How many black tiles are needed to make Pattern 4? _______________

Level II
Theta -0.29

Marcia begins to make a table to show the number of black and white tiles she is using.

Pattern number	1	2	3	4
Number of white tiles	16	24		
Number of black tiles	5	9		
Total	21	33		

2. Fill in the missing numbers in Marcia's table.

3. Marcia wants to know how many white tiles and black tiles there will be in the tenth pattern, but she does not want to draw all the patterns and count the squares. Explain or show another way she could find her answer.

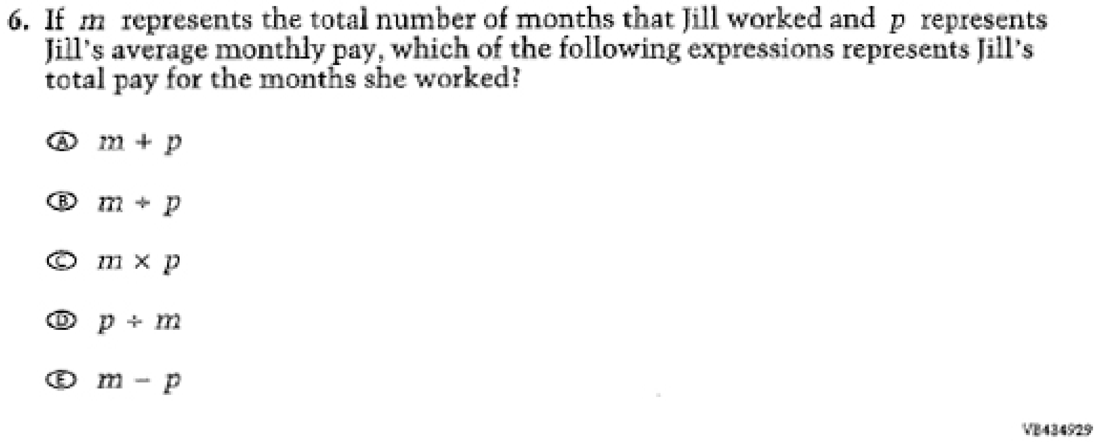
____

____

____

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Note: Only Item 1 from Tiling Squares pertains to this progression. The remaining items 2 and 3, as well as the omitted items 4, 5, and 6, do not occur in this progression.



6. If m represents the total number of months that Jill worked and p represents Jill's average monthly pay, which of the following expressions represents Jill's total pay for the months she worked?

Ⓐ $m + p$

Ⓑ $m \div p$

Ⓒ $m \times p$

Ⓓ $p \div m$

Ⓔ $m - p$

VB434929

Level II
Theta -0.14

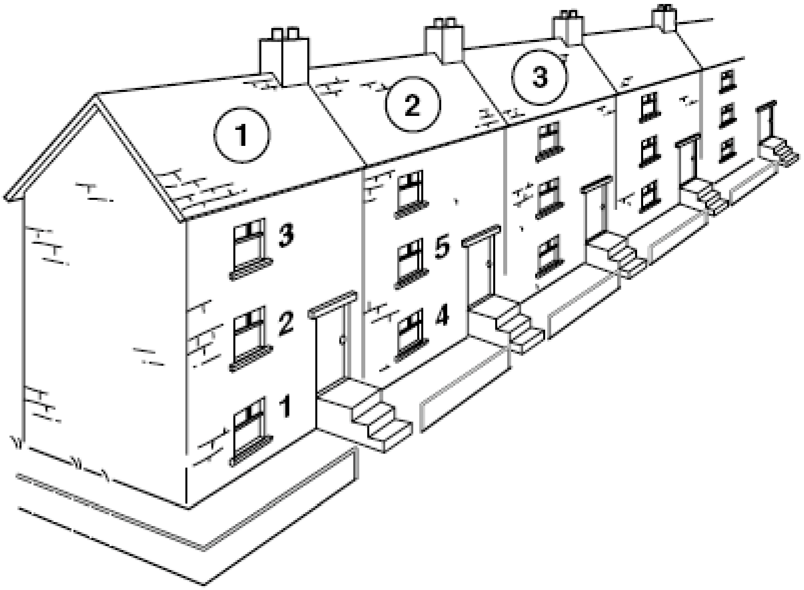
Item VB434929.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M7.

Apartment Numbers

This problem gives you the chance to:
- see and work with number patterns
- express number patterns in words and explain an error

A long row of houses has been changed into apartments.
Each house has been made into three apartments.



The apartments are numbered in order: basement, middle, and top, for each house in the row. Apartments numbered 1 to 5 are shown in the drawing.

Level II
Theta -0.05

1. Complete the following table to show the apartment numbers for the first five houses.

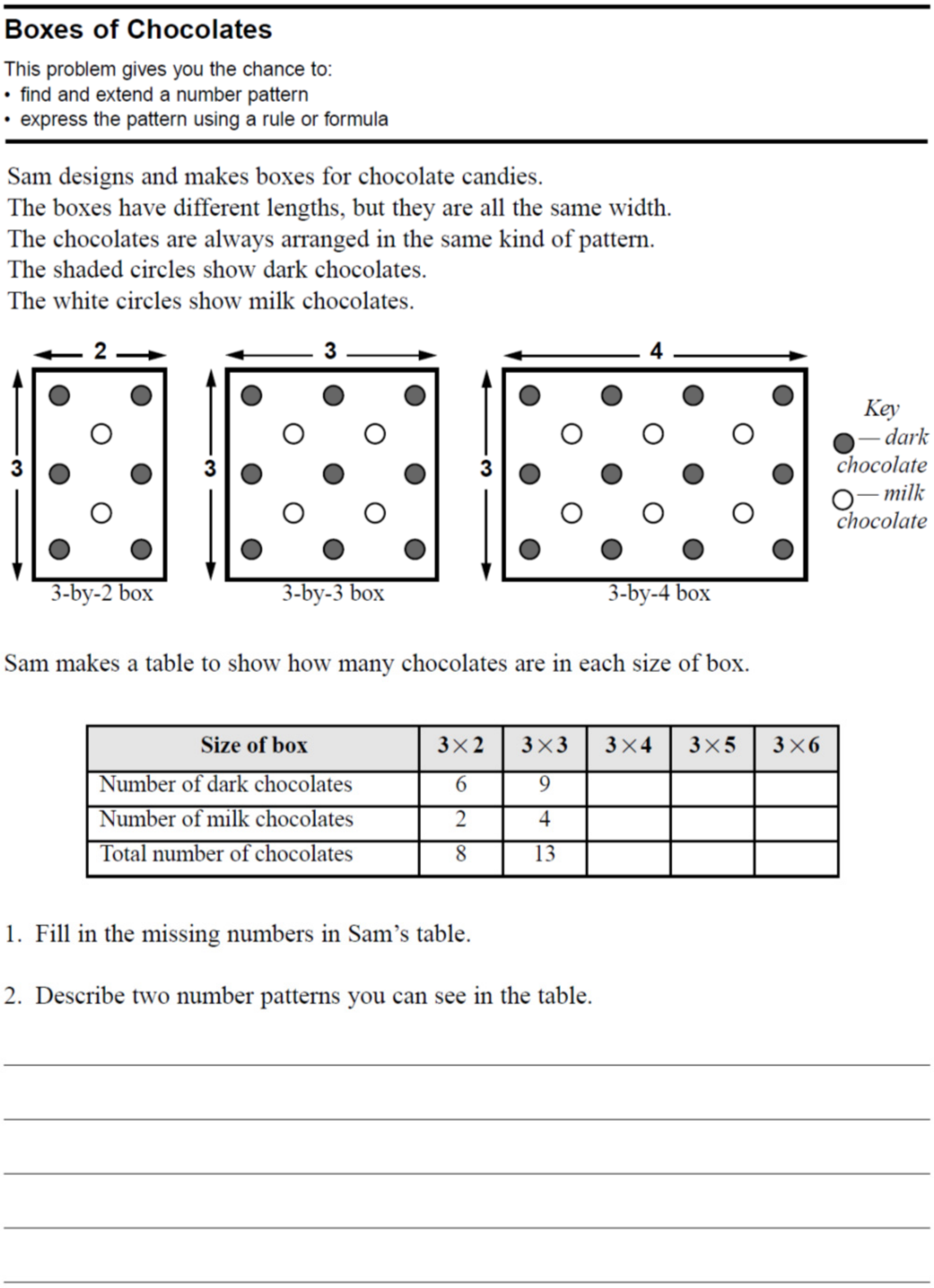
House	Basement	Middle apartment	Top apartment
1	1	2	3
2	4	5	
3			
4			
5			

2. Mrs. Smith lives in the top apartment in the tenth house.
What is the number of her apartment? ______________

3. Mr. Patel and Mr. Dobson are next door neighbors.
They both have basement apartments. Mr. Dobson lives in apartment 25.
What are the possible numbers of Mr. Patel's apartment? ______________

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Note: Only Item 1 from Apartment Numbers pertains to this progression. The remaining items 2 and 3, as well as the omitted items 4 and 5, do not occur in this progression.



Boxes of Chocolates

This problem gives you the chance to:
- find and extend a number pattern
- express the pattern using a rule or formula

Sam designs and makes boxes for chocolate candies.
The boxes have different lengths, but they are all the same width.
The chocolates are always arranged in the same kind of pattern.
The shaded circles show dark chocolates.
The white circles show milk chocolates.

Sam makes a table to show how many chocolates are in each size of box.

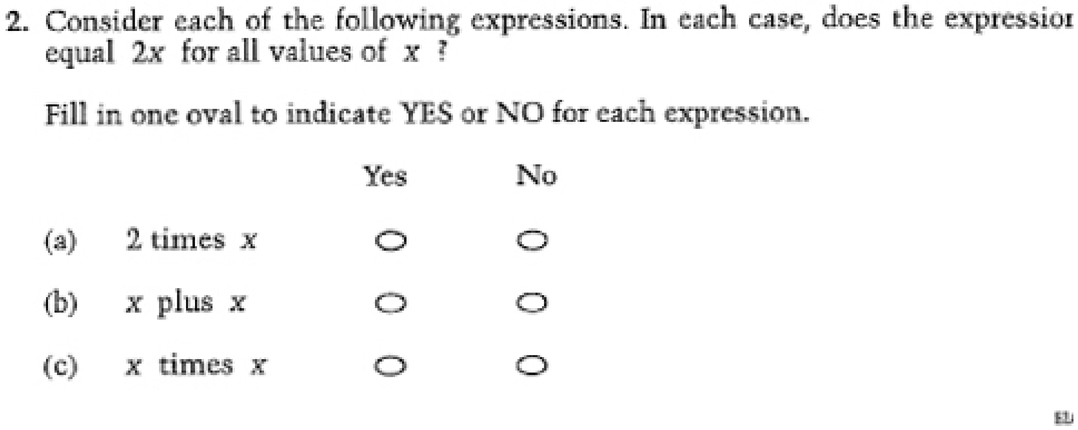
Size of box	3×2	3×3	3×4	3×5	3×6
Number of dark chocolates	6	9			
Number of milk chocolates	2	4			
Total number of chocolates	8	13			

1. Fill in the missing numbers in Sam's table.
2. Describe two number patterns you can see in the table.

Level II
Theta 0.42

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Note: Only Item 1 from Boxes of Chocolates pertains to this progression. The remaining item 2 as well as the omitted items 3, 4, and 5 do not occur in this progression.



2. Consider each of the following expressions. In each case, does the expression equal $2x$ for all values of x ?

Fill in one oval to indicate YES or NO for each expression.

Level II
Theta 1.69

		Yes	No
(a)	2 times x	○	○
(b)	x plus x	○	○
(c)	x times x	○	○

EL

Item EL001490.

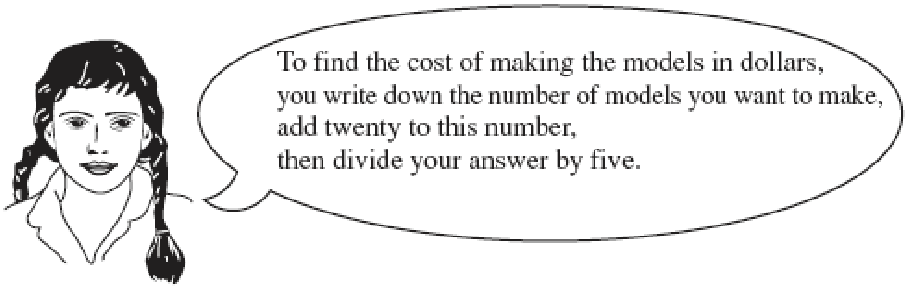
Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z23M9B.

Level III

Emma's Models

This problem gives you the chance to:
- use tables, graphs, and formulas to solve problems

Emma is making some clay models to sell at the school fair.

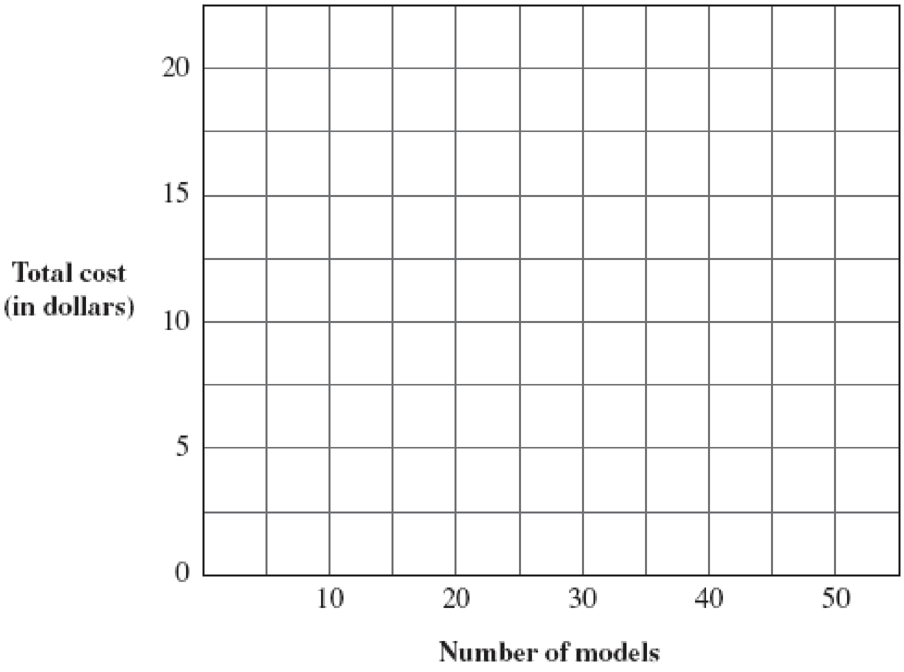


1. Complete the table below to show how the cost depends on the number of models Emma makes. The first value has been calculated and written in the table.

Number of models	10	20	30	40	50
Total cost (in dollars)	6				

Level III
Theta -0.33

2. Draw a graph that shows the information in the table above.



Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Note: Only Item 1 from Emma's Models pertains to this progression. The remaining item 2, as well as the omitted items 3 and 4, do not occur in this progression.

Party

This problem gives you the chance to:

- choose and use number operations in context
- find and use an algebraic formula
- relate formulae and graphs

Sarah is organizing a party at the Vine House Hotel.

Vine House Hotel

Your fab party place!

Charges

$750 for up to 30 people

plus

$20 per person for each extra person

1. Sarah thinks there will be 60 people at the party.
Show that the cost will be $1350.

Level III
Theta 0.31

2. What is the cost of a party for 100 people at the Vine House Hotel? $ ____________
Show how you figured it out.

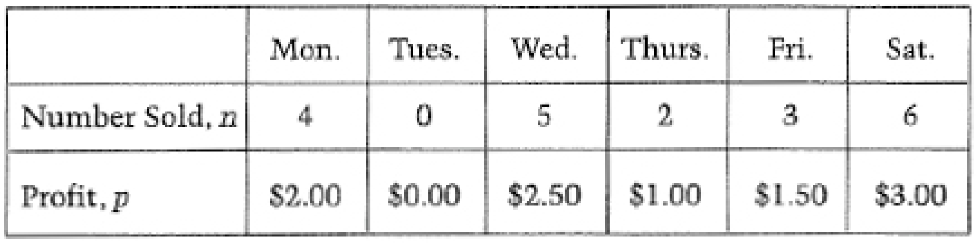
3. *C* dollars is the cost of a party for *P* people.
Find a formula that gives *C* in terms of *P*.

____

____

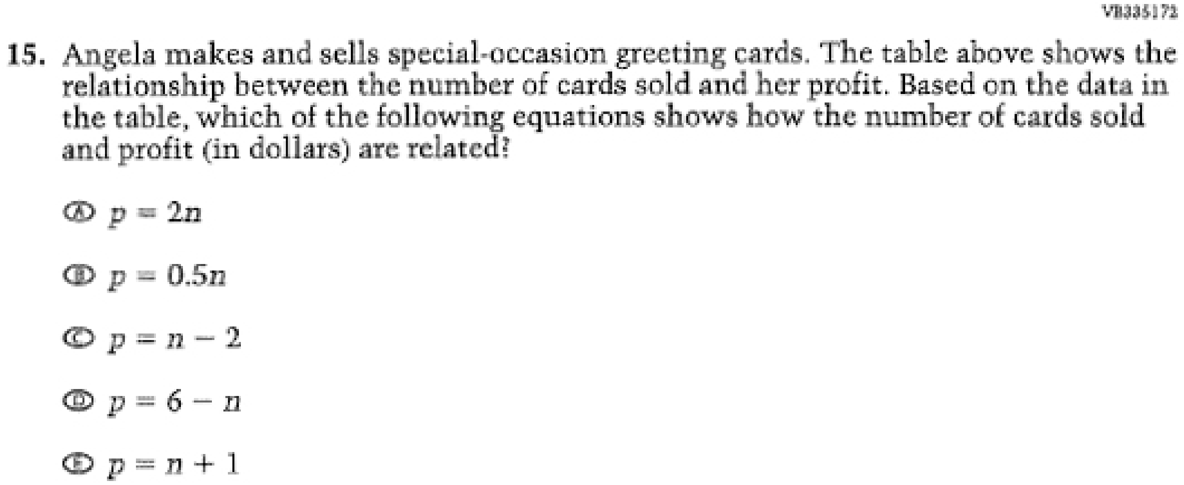
Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Note: Only Item 1 from Party pertains to this progression. The remaining items 2 and 3, as well as the omitted items 4 and 5, do not occur in this progression.



	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
Number Sold, n	4	0	5	2	3	6
Profit, p	\$2.00	\$0.00	\$2.50	\$1.00	\$1.50	\$3.00

Level III
Theta 0.60



VB335172

15. Angela makes and sells special-occasion greeting cards. The table above shows the relationship between the number of cards sold and her profit. Based on the data in the table, which of the following equations shows how the number of cards sold and profit (in dollars) are related?

Ⓐ $p = 2n$

Ⓑ $p = 0.5n$

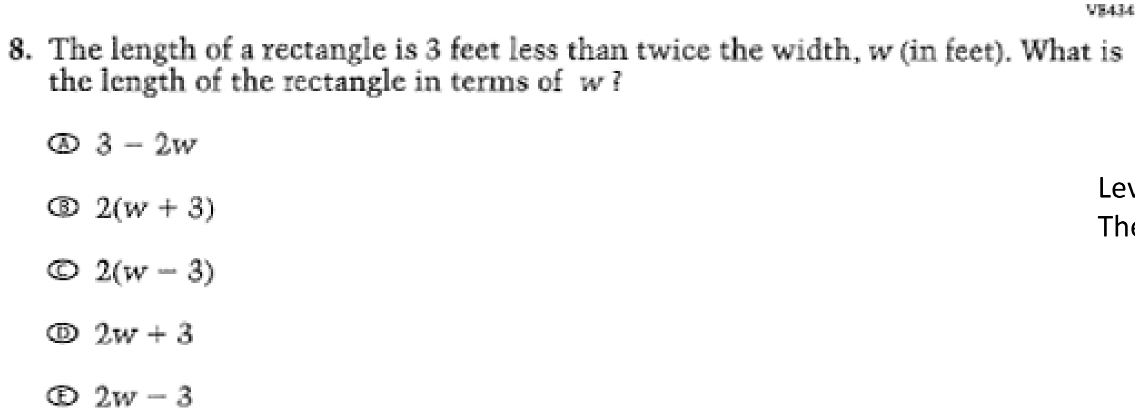
Ⓒ $p = n - 2$

Ⓓ $p = 6 - n$

Ⓔ $p = n + 1$

Item VB335172.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M7.



VB434848

8. The length of a rectangle is 3 feet less than twice the width, w (in feet). What is the length of the rectangle in terms of w?

Ⓐ $3 - 2w$

Ⓑ $2(w + 3)$

Ⓒ $2(w - 3)$

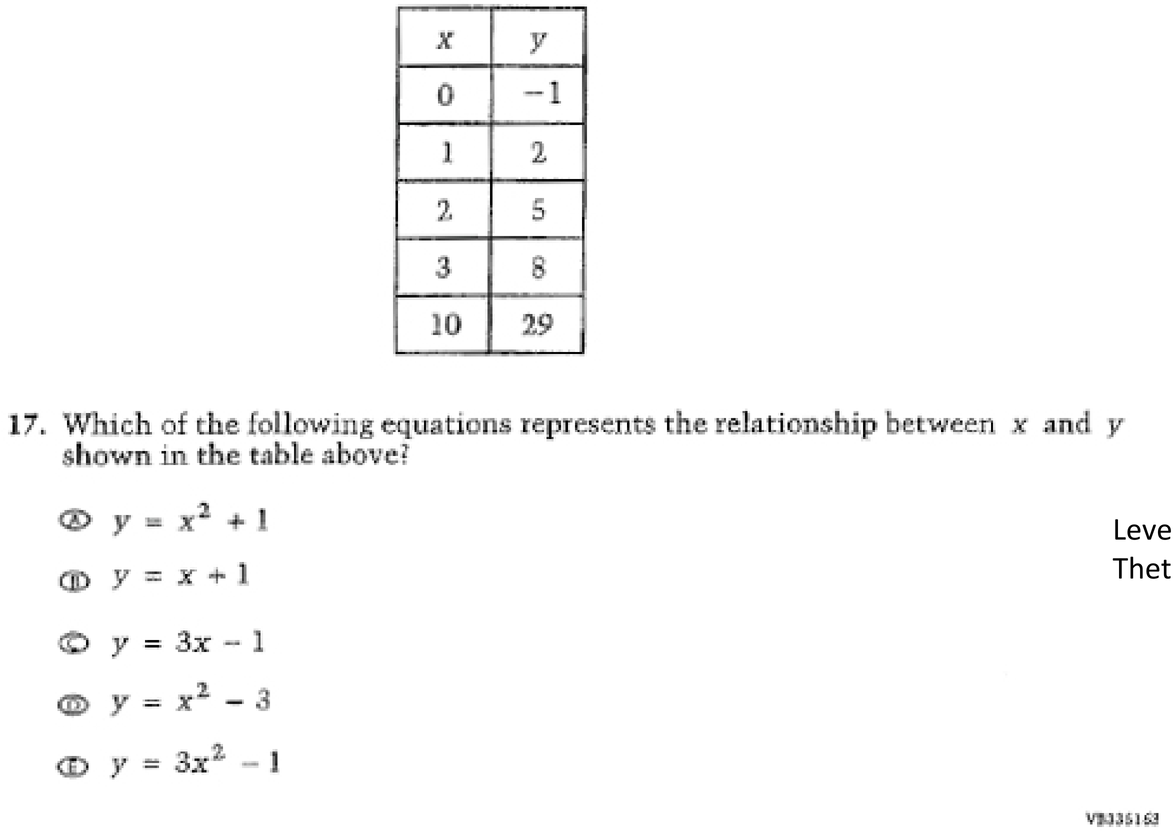
Ⓓ $2w + 3$

Ⓔ $2w - 3$

Level III
Theta 0.68

VB434848.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M10.



x	y
0	−1
1	2
2	5
3	8
10	29

17. Which of the following equations represents the relationship between x and y shown in the table above?

Ⓐ $y = x^2 + 1$

Ⓑ $y = x + 1$

Ⓒ $y = 3x - 1$

Ⓓ $y = x^2 - 3$

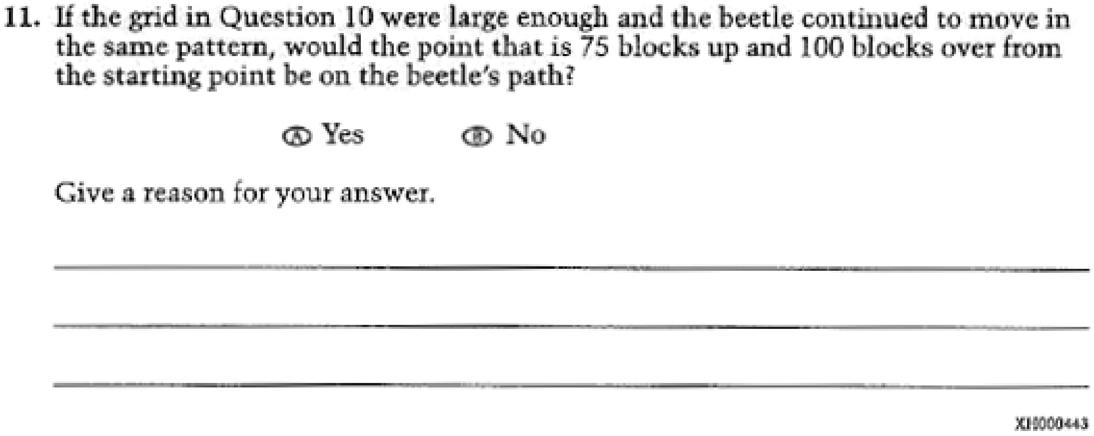
Ⓔ $y = 3x^2 - 1$

VB335163

Level III
Theta 0.71

Item VB335163.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z2M12.



11. If the grid in Question 10 were large enough and the beetle continued to move in the same pattern, would the point that is 75 blocks up and 100 blocks over from the starting point be on the beetle's path?

Ⓐ Yes Ⓑ No

Give a reason for your answer.

XH000443

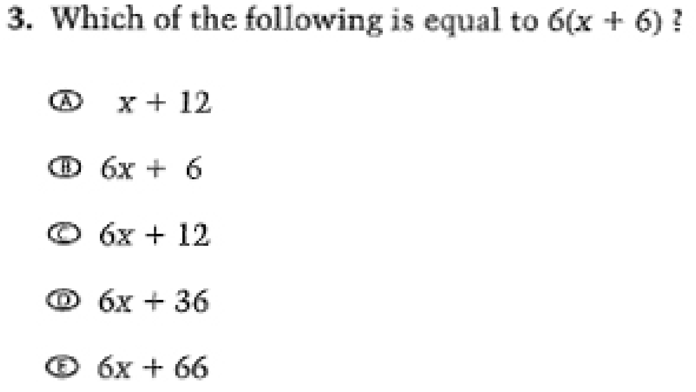
Level III
Theta 2.3

Item XH000443.

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

Note: See Item XH000442 on page 179 (first item in graphing learning progression) for the graph referenced in this question.

Level IV



3. Which of the following is equal to $6(x + 6)$?

Ⓐ $x + 12$

Ⓑ $6x + 6$

Ⓒ $6x + 12$

Ⓓ $6x + 36$

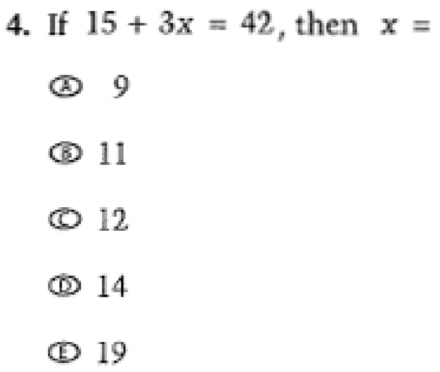
Ⓔ $6x + 66$

Level IV
Theta 1.02

Item VB335154.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z2M12.

Level V



4. If $15 + 3x = 42$, then $x =$

Ⓐ 9

Ⓑ 11

Ⓒ 12

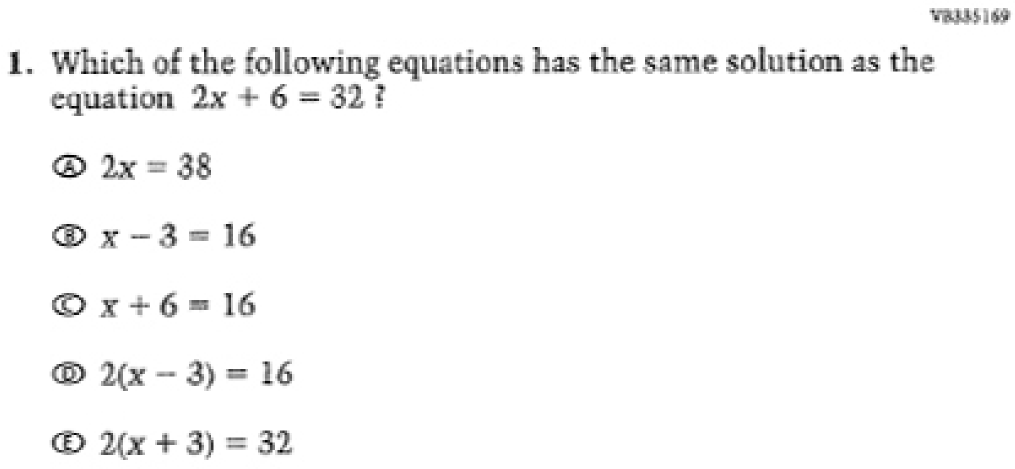
Ⓓ 14

Ⓔ 19

Level V
Theta -0.53

Item YJ000107.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z23M9B**.**



VB335169

1. Which of the following equations has the same solution as the equation $2x + 6 = 32$?

Ⓐ $2x = 38$

Ⓑ $x - 3 = 16$

Ⓒ $x + 6 = 16$

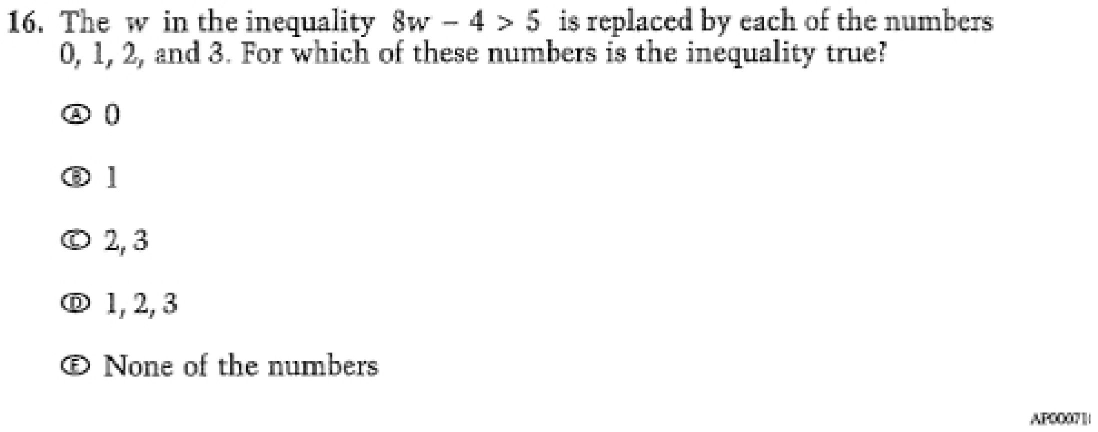
Ⓓ $2(x - 3) = 16$

Ⓔ $2(x + 3) = 32$

Level V
Theta 0.48

Item VB335169.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M10**.**



16. The w in the inequality $8w - 4 > 5$ is replaced by each of the numbers 0, 1, 2, and 3. For which of these numbers is the inequality true?

Ⓐ 0

Ⓑ 1

Ⓒ 2, 3

Ⓓ 1, 2, 3

Ⓔ None of the numbers

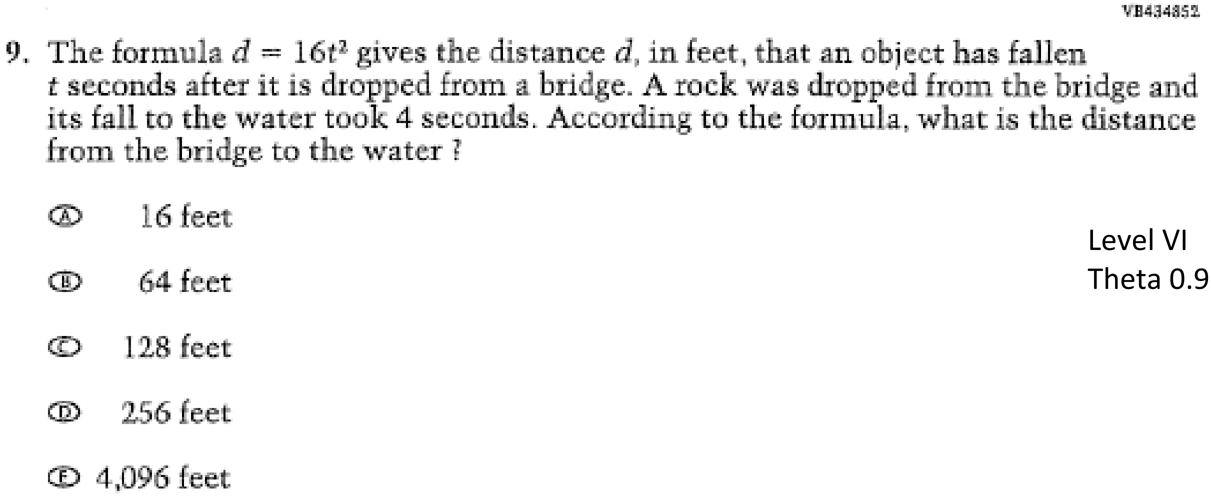
AP000710

Level V
Theta 0.69

Item AP000710.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z23M8B.

Level VI



VB434852

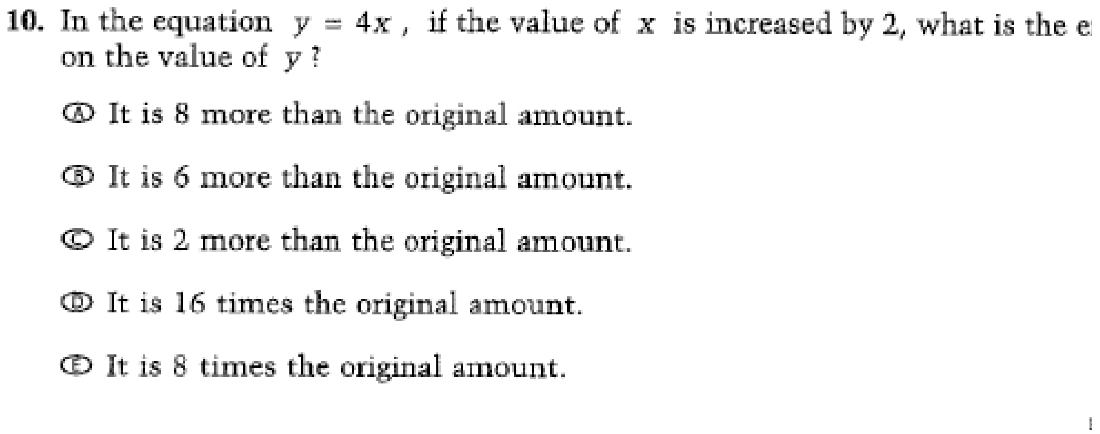
9. The formula $d = 16t^2$ gives the distance d, in feet, that an object has fallen t seconds after it is dropped from a bridge. A rock was dropped from the bridge and its fall to the water took 4 seconds. According to the formula, what is the distance from the bridge to the water ?

Ⓐ 16 feet
Ⓑ 64 feet
Ⓒ 128 feet
Ⓓ 256 feet
Ⓔ 4,096 feet

Level VI
Theta 0.93

Item VB434852.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M7.



10. In the equation $y = 4x$, if the value of x is increased by 2, what is the e on the value of y ?

Ⓐ It is 8 more than the original amount.
Ⓑ It is 6 more than the original amount.
Ⓒ It is 2 more than the original amount.
Ⓓ It is 16 times the original amount.
Ⓔ It is 8 times the original amount.

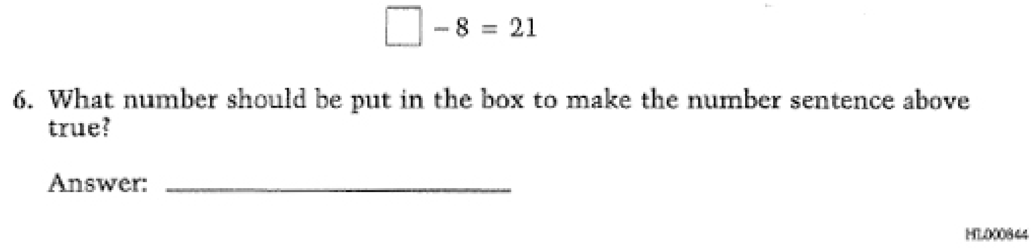
Level VI
Theta 1.71

Item HW000857.

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

Figure A-4. Equations Learning Progression Branch 2

Level I



$\boxed{} - 8 = 21$

6. What number should be put in the box to make the number sentence above true?

Answer: ______________________

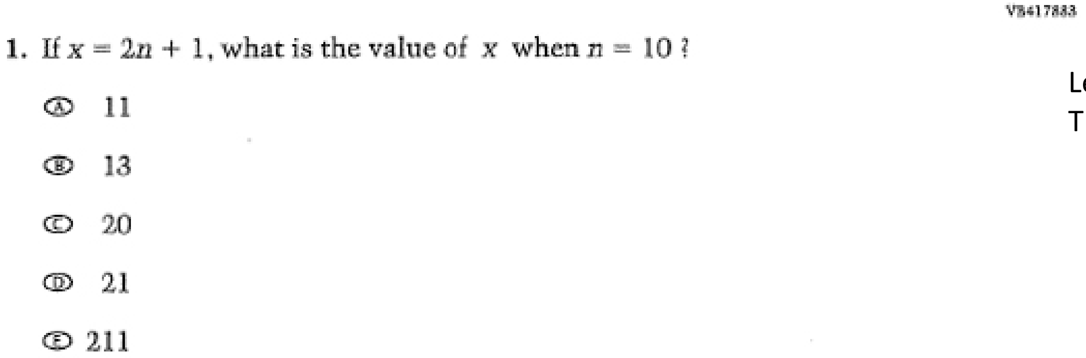
HL000844

Level I
Theta -1.58

Item HL000844.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z12M5B.

Level II



VB417883

1. If $x = 2n + 1$, what is the value of x when $n = 10$?

Ⓐ 11

Ⓑ 13

Ⓒ 20

Ⓓ 21

Ⓔ 211

Level II
Theta -0.45

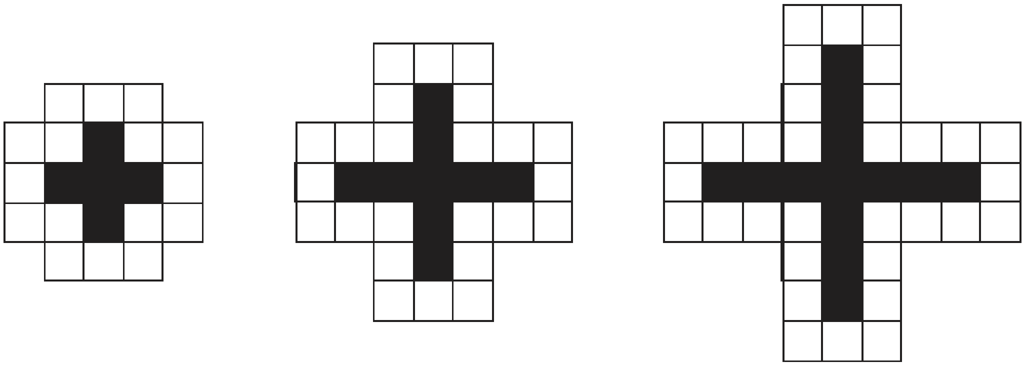
Item VB417883.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M7.

Tiling Squares

This problem gives you the chance to:
- extend and check patterns
- derive formulas connecting different pairs of variables

Marcia is using black and white square tiles to make patterns.



Pattern 1 Pattern 2 Pattern 3

1. How many black tiles are needed to make Pattern 4? ______________

Level II
Theta -0.29

Marcia begins to make a table to show the number of black and white tiles she is using.

Pattern number	1	2	3	4
Number of white tiles	16	24		
Number of black tiles	5	9		
Total	21	33		

Level IV
Theta 1.27

2. Fill in the missing numbers in Marcia's table.

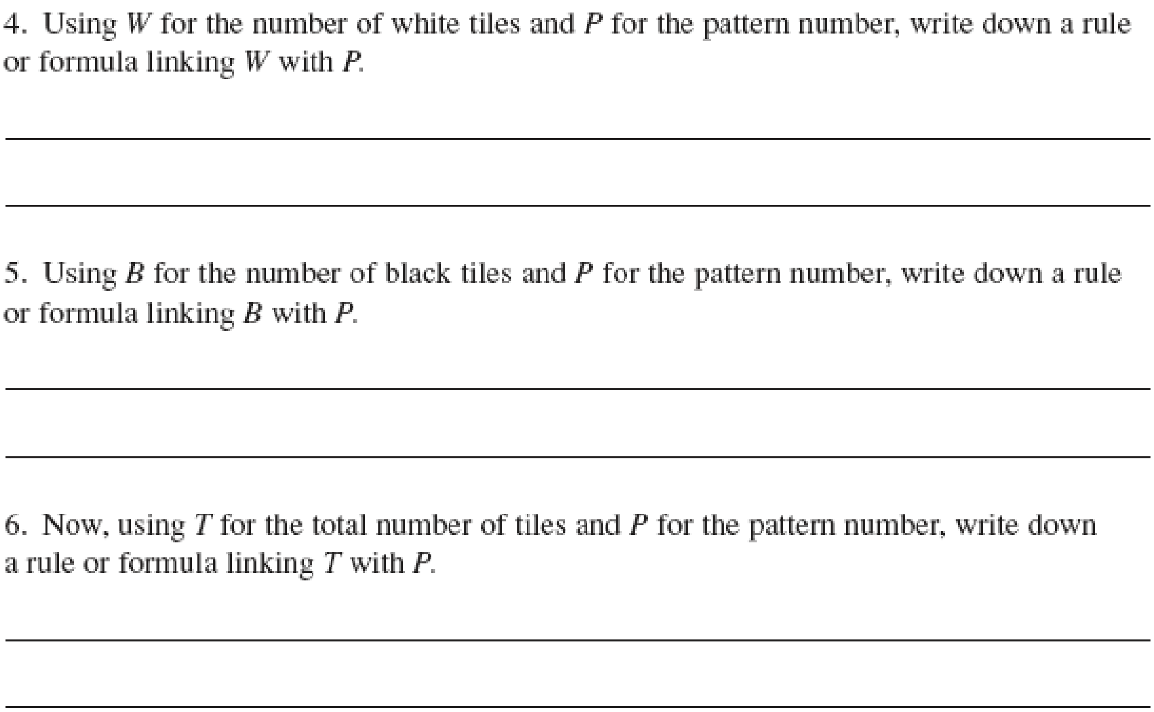
3. Marcia wants to know how many white tiles and black tiles there will be in the tenth pattern, but she does not want to draw all the patterns and count the squares.
Explain or show another way she could find her answer.

Level VII
Theta 1.56

____

____

____



4. Using W for the number of white tiles and P for the pattern number, write down a rule or formula linking W with P.

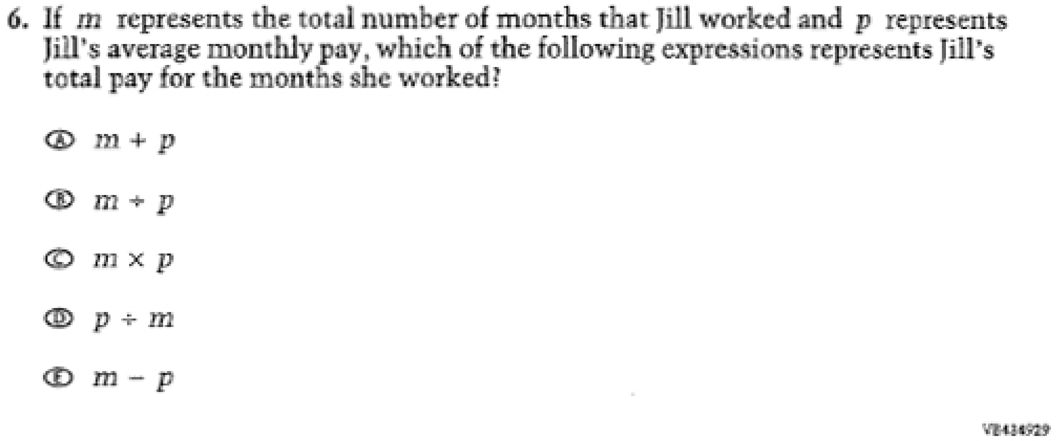
Level VII
Theta 1.5

5. Using B for the number of black tiles and P for the pattern number, write down a rule or formula linking B with P.

Level VII
Theta 1.62

6. Now, using T for the total number of tiles and P for the pattern number, write down a rule or formula linking T with P.

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.



6. If m represents the total number of months that Jill worked and p represents Jill's average monthly pay, which of the following expressions represents Jill's total pay for the months she worked?

Ⓐ $m + p$

Ⓑ $m \div p$

Ⓒ $m \times p$

Ⓓ $p \div m$

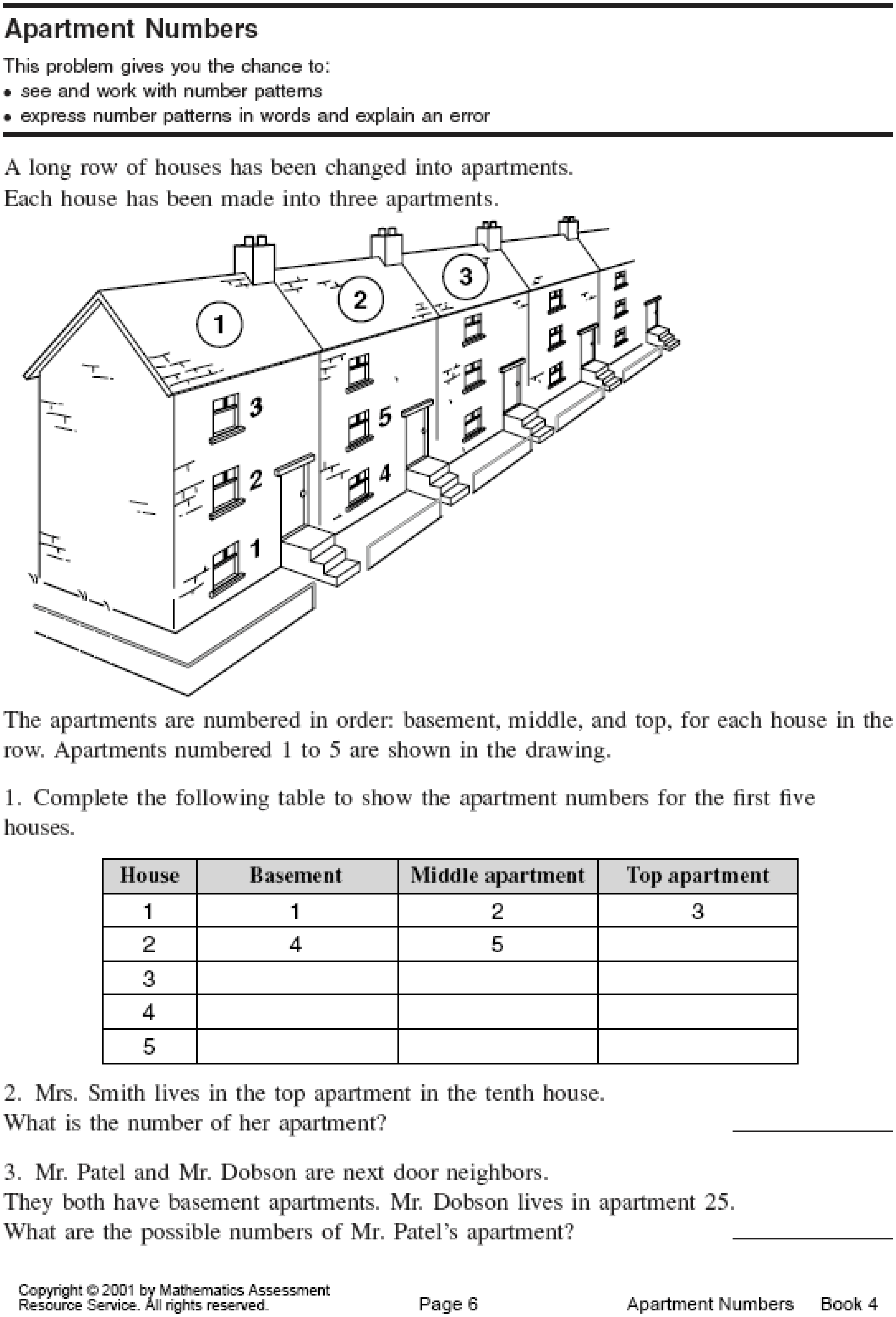
Ⓔ $m - p$

VB434929

Level II
Theta -0.14

Item VB434929.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z2M11.



Apartment Numbers

This problem gives you the chance to:
- see and work with number patterns
- express number patterns in words and explain an error

A long row of houses has been changed into apartments.
Each house has been made into three apartments.

The apartments are numbered in order: basement, middle, and top, for each house in the row. Apartments numbered 1 to 5 are shown in the drawing.

1. Complete the following table to show the apartment numbers for the first five houses.

Level II
Theta -0.05

House	Basement	Middle apartment	Top apartment
1	1	2	3
2	4	5	
3			
4			
5			

2. Mrs. Smith lives in the top apartment in the tenth house.
What is the number of her apartment? ____________

Level IV
Theta -0.49

3. Mr. Patel and Mr. Dobson are next door neighbors.
They both have basement apartments. Mr. Dobson lives in apartment 25.
What are the possible numbers of Mr. Patel's apartment? ____________

Level VI
Theta 1.10

4. Ms. Sanchez uses a rule to make a table that shows some of the house numb
the numbers of the middle apartments.

House number	Middle apartment number
2	5
3	8
6	17
10	29
12	35
25	74

Level VII
Theta 1.39

Write down what you think Ms. Sanchez's rule is.

____

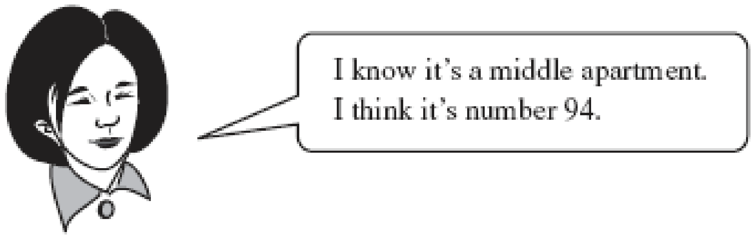
____

____

____

5. Miss Ling is going to visit her friend.

Level VII
Theta 1.52



Is Miss Ling correct? Explain your answer.

____

____

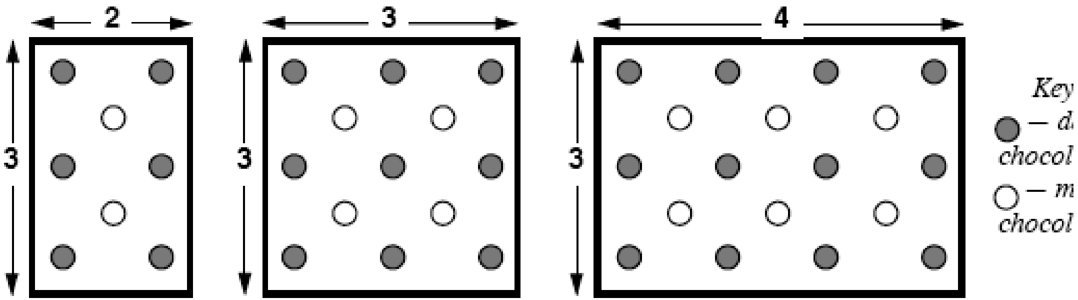
____

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Boxes of Chocolates

This problem gives you the chance to:
- find and extend a number pattern
- express the pattern using a rule or formula

Sam designs and makes boxes for chocolate candies.
The boxes have different lengths, but they are all the same width.
The chocolates are always arranged in the same kind of pattern.
The shaded circles show dark chocolates.
The white circles show milk chocolates.



3-by-2 box 3-by-3 box 3-by-4 box

Sam makes a table to show how many chocolates are in each size of box.

Size of box	3×2	3×3	3×4	3×5	3×6
Number of dark chocolates	6	9			
Number of milk chocolates	2	4			
Total number of chocolates	8	13			

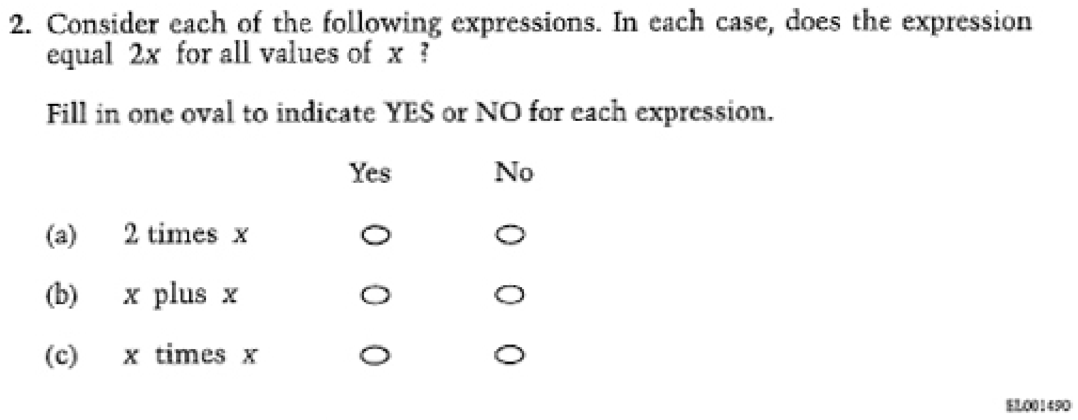
Level II
Theta 0.42

1. Fill in the missing numbers in Sam's table.

2. Describe two number patterns you can see in the table.

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Note: Only Item 1 from Boxes of Chocolates pertains to this progression. The remaining item 2, as well as the omitted items 3, 4, and 5, do not occur in this progression.



2. Consider each of the following expressions. In each case, does the expression equal $2x$ for all values of x ?

Fill in one oval to indicate YES or NO for each expression.

		Yes	No
(a)	2 times x	○	○
(b)	x plus x	○	○
(c)	x times x	○	○

EL001490

Level II
Theta 1.69

Item EL001490.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z23M9B.

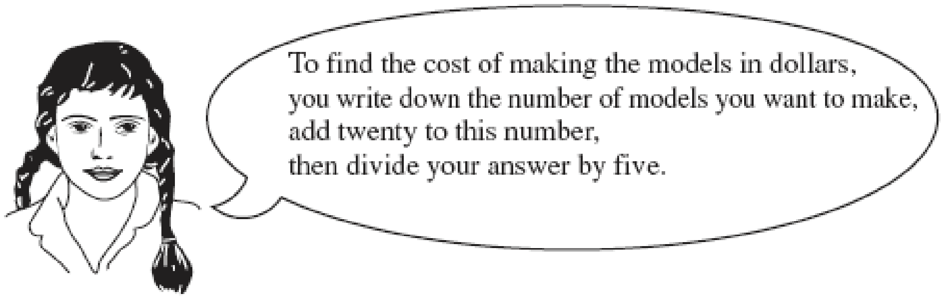
Level III

Emma's Models

This problem gives you the chance to:

- use tables, graphs, and formulas to solve problems

Emma is making some clay models to sell at the school fair.

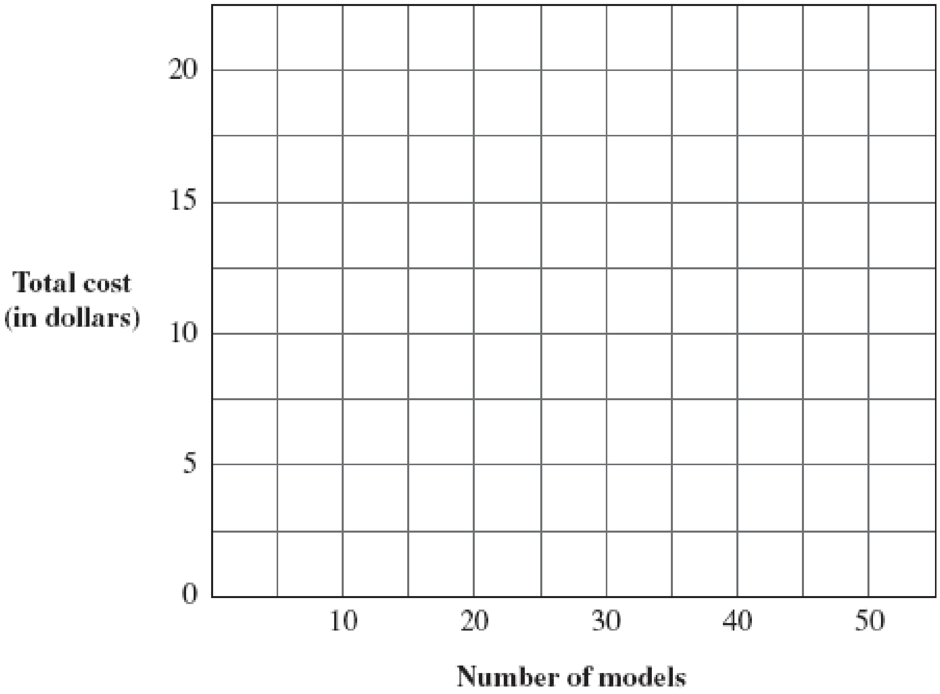


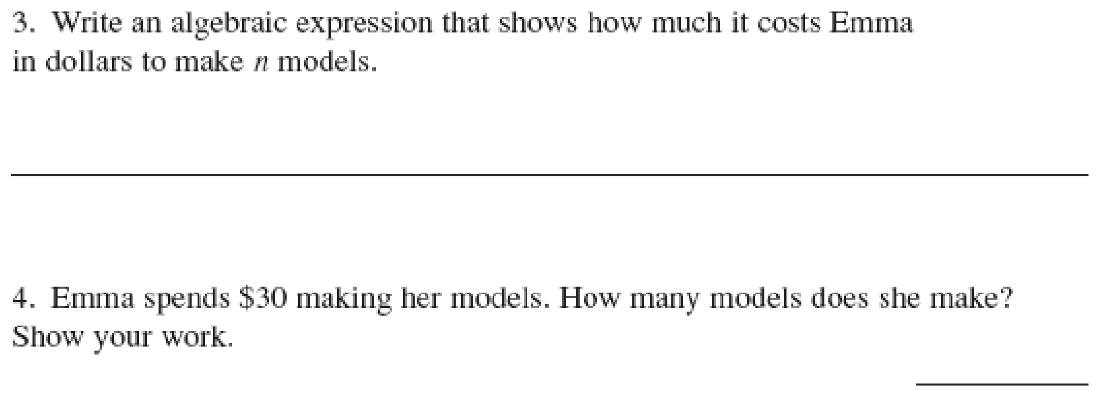
1. Complete the table below to show how the cost depends on the number of models Emma makes. The first value has been calculated and written in the table.

Number of models	10	20	30	40	50
Total cost (in dollars)	6				

Level III
Theta -0.33

2. Draw a graph that shows the information in the table above.





3. Write an algebraic expression that shows how much it costs Emma in dollars to make *n* models.

4. Emma spends $30 making her models. How many models does she make? Show your work.

Level VII
Theta 0.63

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Note: Only Items 1 and 4 from Emma’s Models pertain to this progression. Items 2 and 3 do not occur in this progression.

Party

This problem gives you the chance to:

- choose and use number operations in context
- find and use an algebraic formula
- relate formulae and graphs

Sarah is organizing a party at the Vine House Hotel.

Vine House Hotel

Your fab party place!

Charges

$750 for up to 30 people

plus

$20 per person for each extra person

1. Sarah thinks there will be 60 people at the party.
Show that the cost will be $1350.

Level III
Theta 0.31

2. What is the cost of a party for 100 people at the Vine House Hotel? $ __________
Show how you figured it out.

Level IV
Theta 0.73

3. *C* dollars is the cost of a party for *P* people.
Find a formula that gives *C* in terms of *P*.

Level VII
Theta 1.98

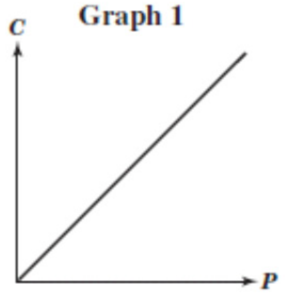
____

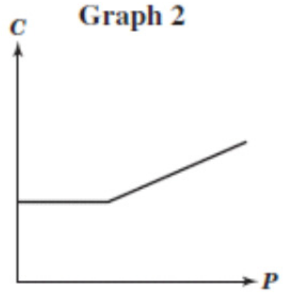
____

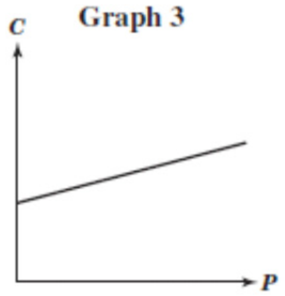
4. Sarah's party cost $1750 in all.
How many people came to the party?
Show your calculations. ______________

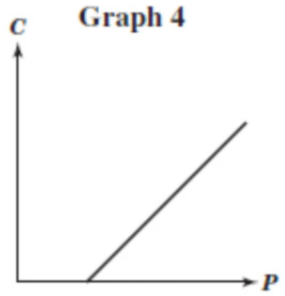
Level VI
Theta 1.20

5. Which of these graphs shows the connection between the number of people at the party, P, and the cost, C?


Graph 1


Graph 2


Graph 3


Graph 4

Explain how you figured it out.

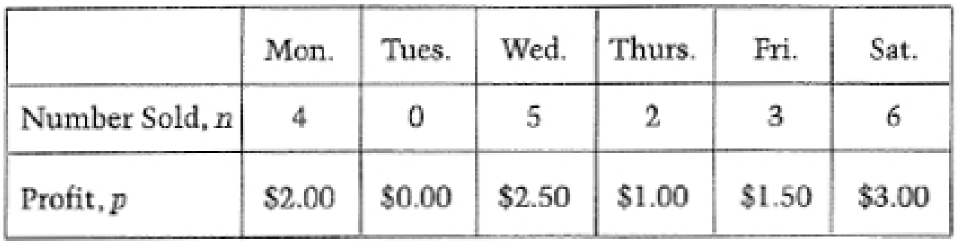
____

____

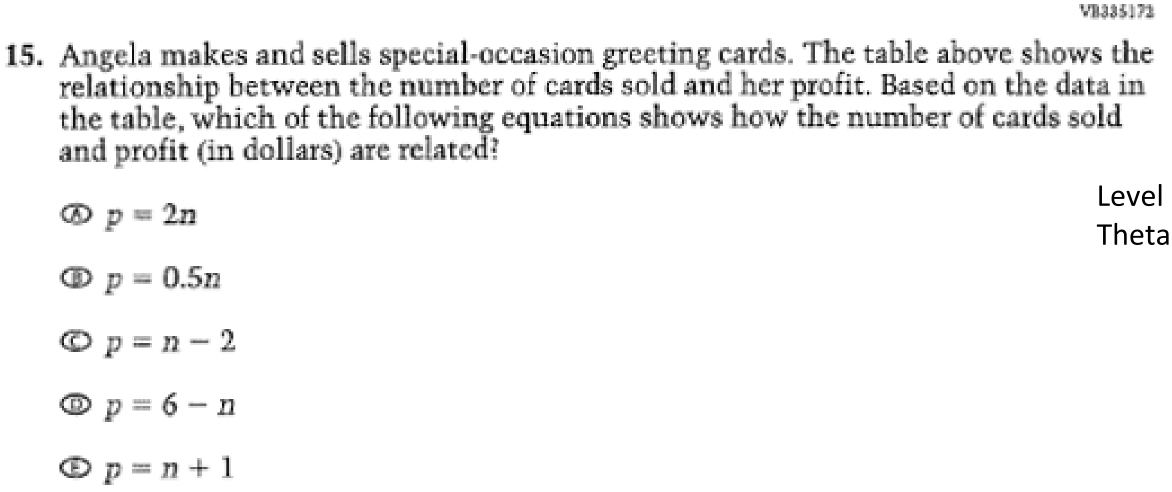
____

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Note: Only Items 1 through 4 from Party pertain to this progression. Item 5 does not occur in this progression.



	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
Number Sold, n	4	0	5	2	3	6
Profit, p	\$2.00	\$0.00	\$2.50	\$1.00	\$1.50	\$3.00



VB335172

15. Angela makes and sells special-occasion greeting cards. The table above shows the relationship between the number of cards sold and her profit. Based on the data in the table, which of the following equations shows how the number of cards sold and profit (in dollars) are related?

Ⓐ $p = 2n$

Ⓑ $p = 0.5n$

Ⓒ $p = n - 2$

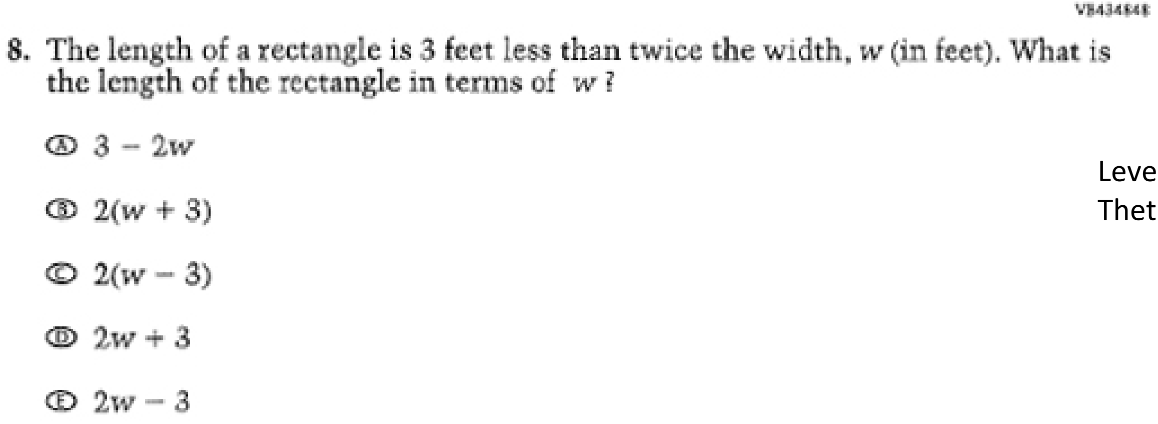
Ⓓ $p = 6 - n$

Ⓔ $p = n + 1$

Level III
Theta 0.60

Item VB335172.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M7.



VB434848

8. The length of a rectangle is 3 feet less than twice the width, w (in feet). What is the length of the rectangle in terms of w?

Ⓐ $3 - 2w$

Ⓑ $2(w + 3)$

Ⓒ $2(w - 3)$

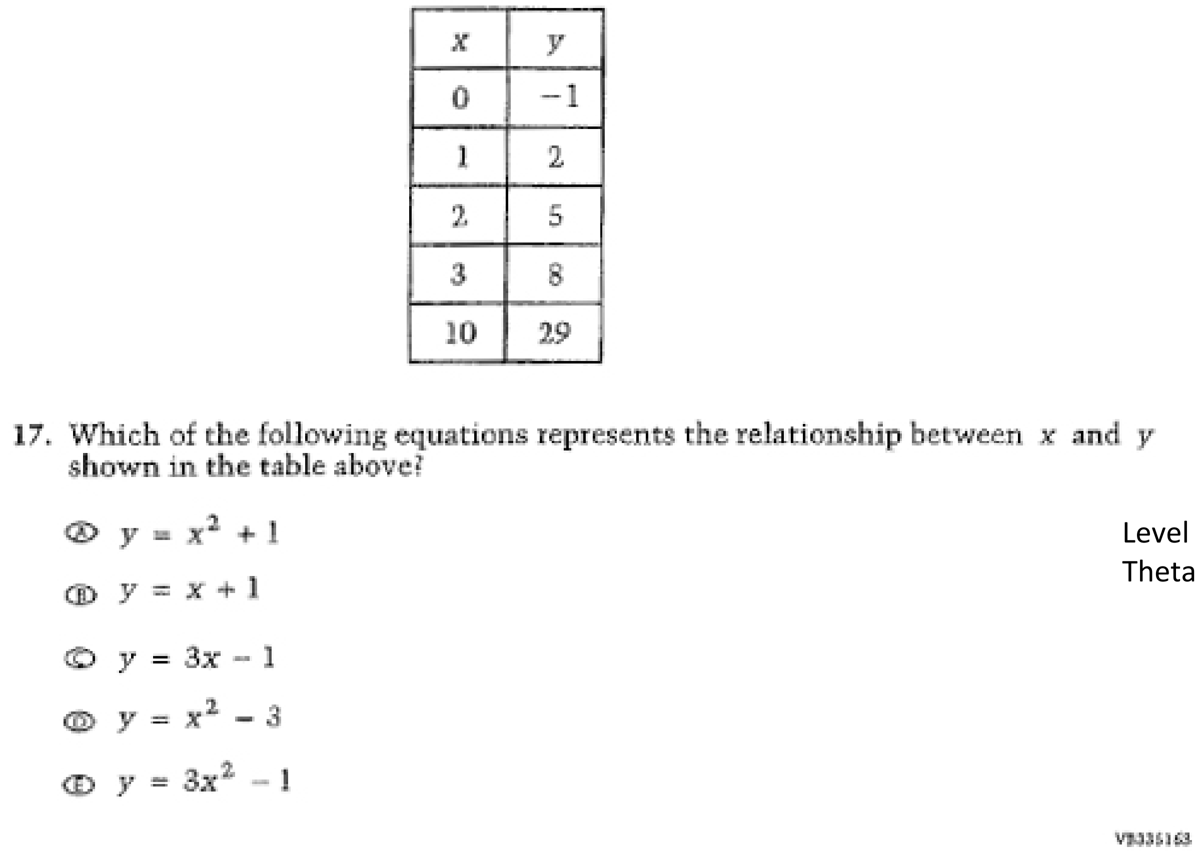
Ⓓ $2w + 3$

Ⓔ $2w - 3$

Level III
Theta 0.68

Item VB434848.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M10.



x	y
0	−1
1	2
2	5
3	8
10	29

17. Which of the following equations represents the relationship between x and y shown in the table above?

Ⓐ $y = x^2 + 1$

Ⓑ $y = x + 1$

Ⓒ $y = 3x - 1$

Ⓓ $y = x^2 - 3$

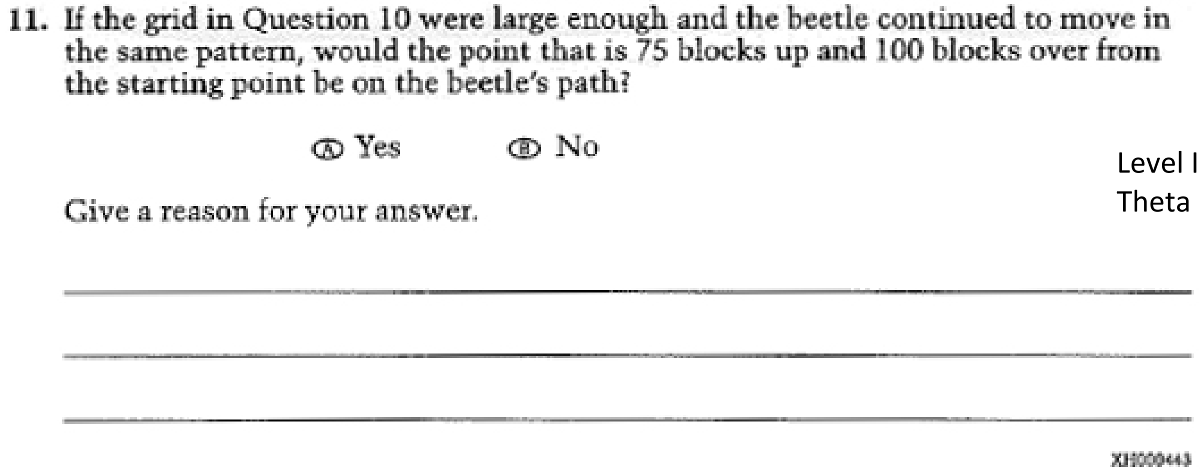
Ⓔ $y = 3x^2 - 1$

VB335163

Level III
Theta 0.71

Item VB335163.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z2M12.



11. If the grid in Question 10 were large enough and the beetle continued to move in the same pattern, would the point that is 75 blocks up and 100 blocks over from the starting point be on the beetle's path?

Ⓐ Yes Ⓑ No

Give a reason for your answer.

XH000443

Level III
Theta 2.3

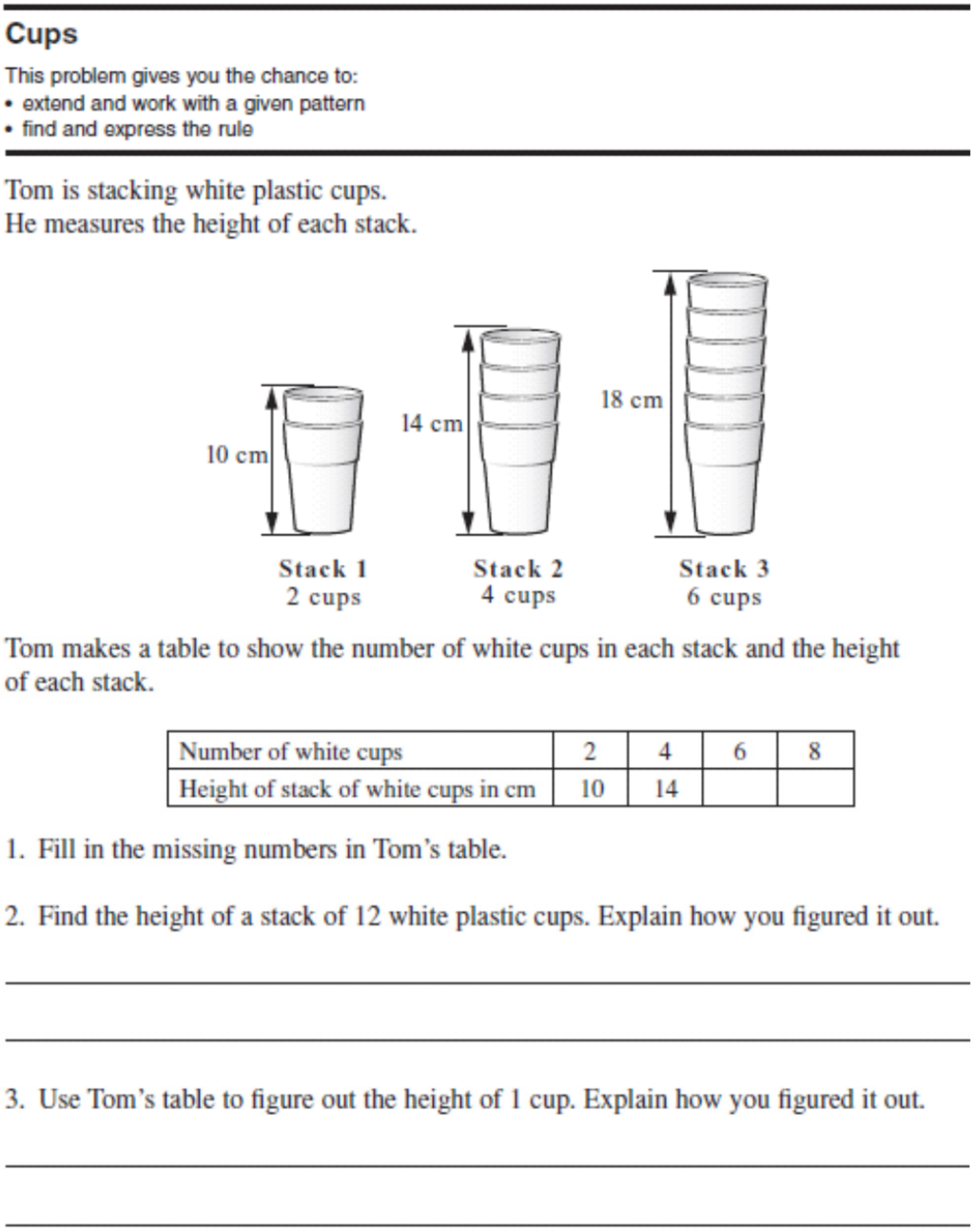
Item XH000443.

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

Note: See Item XH000442 on page 179 (first item in graphing learning progression) for the graph referenced in this question.

Level IV

For Apartment Item 2 (Level IV, Theta -0.49), please see page 230.



Cups

This problem gives you the chance to:
- extend and work with a given pattern
- find and express the rule

Tom is stacking white plastic cups.
He measures the height of each stack.

Tom makes a table to show the number of white cups in each stack and the height of each stack.

Number of white cups	2	4	6	8
Height of stack of white cups in cm	10	14		

1. Fill in the missing numbers in Tom's table.

2. Find the height of a stack of 12 white plastic cups. Explain how you figured it out.

____

____

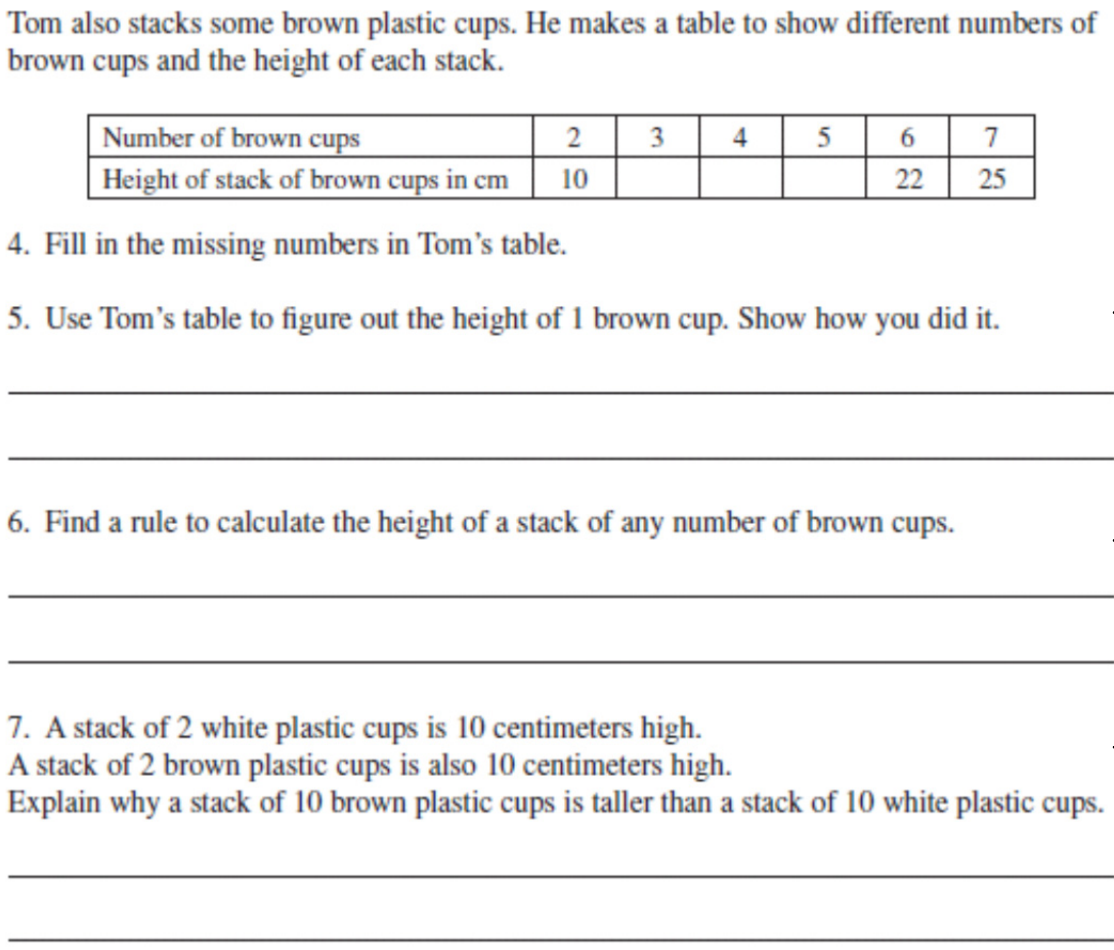
3. Use Tom's table to figure out the height of 1 cup. Explain how you figured it out.

____

____

Level IV
Theta 1.17

Level IV
Theta 0.96



Tom also stacks some brown plastic cups. He makes a table to show different numbers of brown cups and the height of each stack.

Number of brown cups	2	3	4	5	6	7
Height of stack of brown cups in cm	10				22	25

4. Fill in the missing numbers in Tom's table.

5. Use Tom's table to figure out the height of 1 brown cup. Show how you did it.

Level IV
Theta 0.32

6. Find a rule to calculate the height of a stack of any number of brown cups.

Level VII
Theta 1.65

7. A stack of 2 white plastic cups is 10 centimeters high.
A stack of 2 brown plastic cups is also 10 centimeters high.
Explain why a stack of 10 brown plastic cups is taller than a stack of 10 white plastic cups.

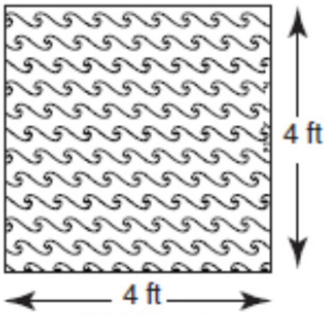
Level VIII
Theta 2.07

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

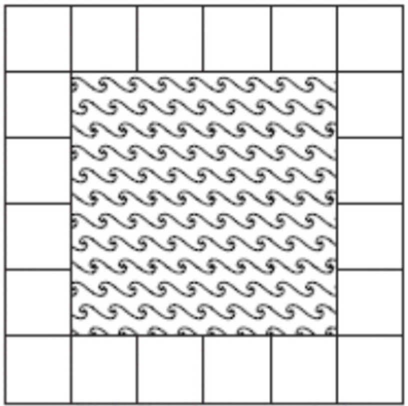
Fish Ponds

This problem gives you the chance to:

- find a number pattern in real spatial context and express the rule
- extend the rule to two variables



Fish Pond



Fish Pond surrounded by paving stones

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

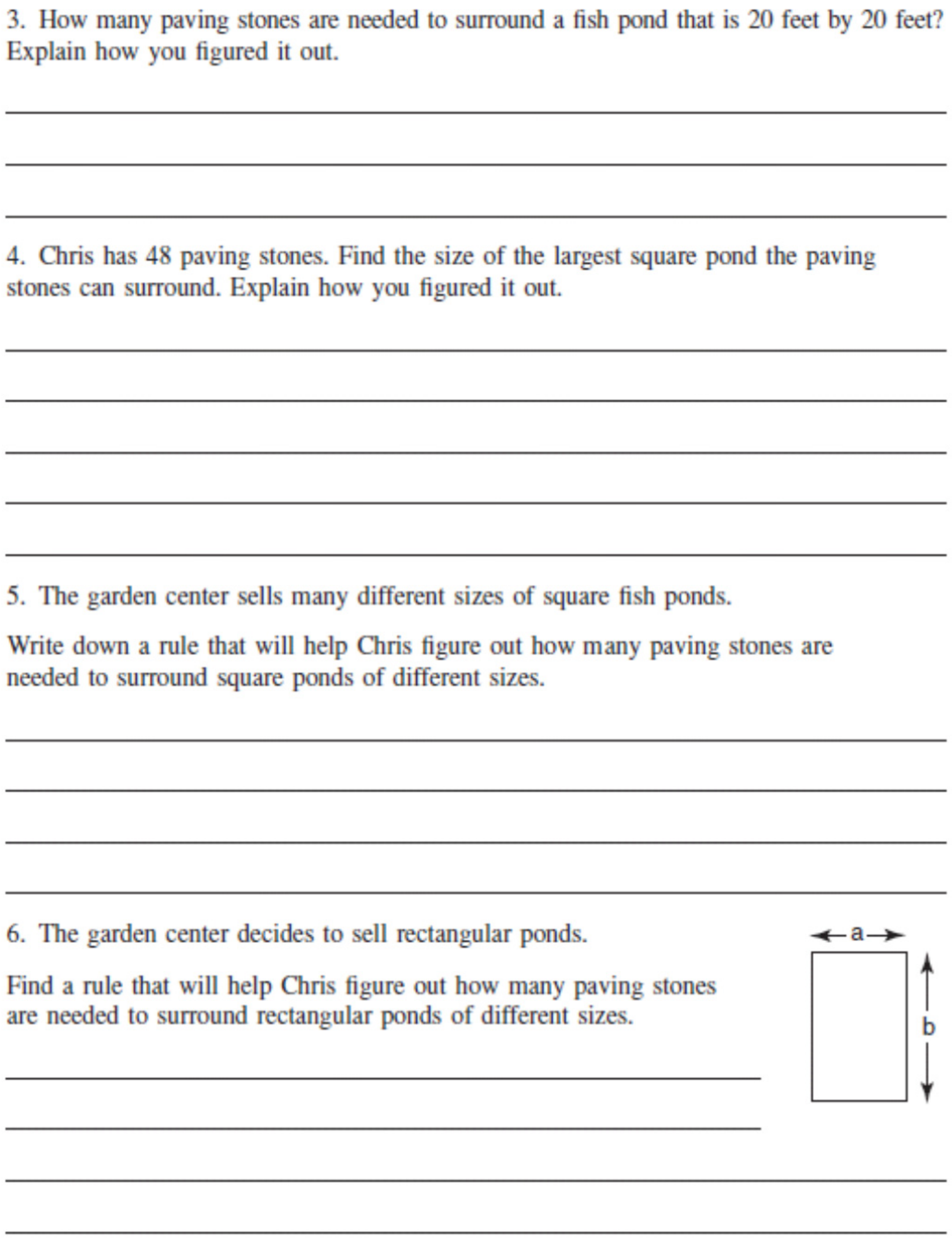
The paving stones are squares with sides one foot long.

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

Level IV
Theta 0.59

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

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



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

Level V
Theta 0.61

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

Level VI
Theta 0.79

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

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

Level VII
Theta 0.87

6. The garden center decides to sell rectangular ponds.

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

Level VII
Theta 1.22

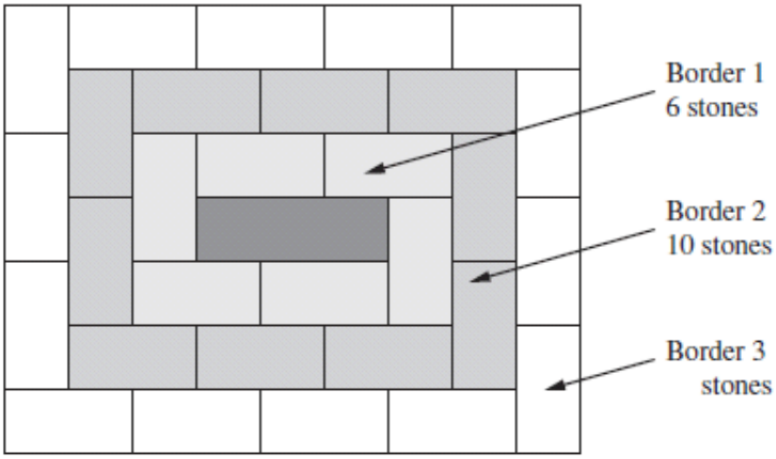
Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

For Party Item 2 (Level IV, Theta 0.73), please see page 236.

Design a Garden

This problem gives you the chance to:
- find and extend a number pattern
- find the rule of the pattern and express the rule in words or algebra

The diagram below is a scale drawing showing Dave's garden plan.



The rectangle across the center is for planting roses. It measures 3 feet by 1 foot.

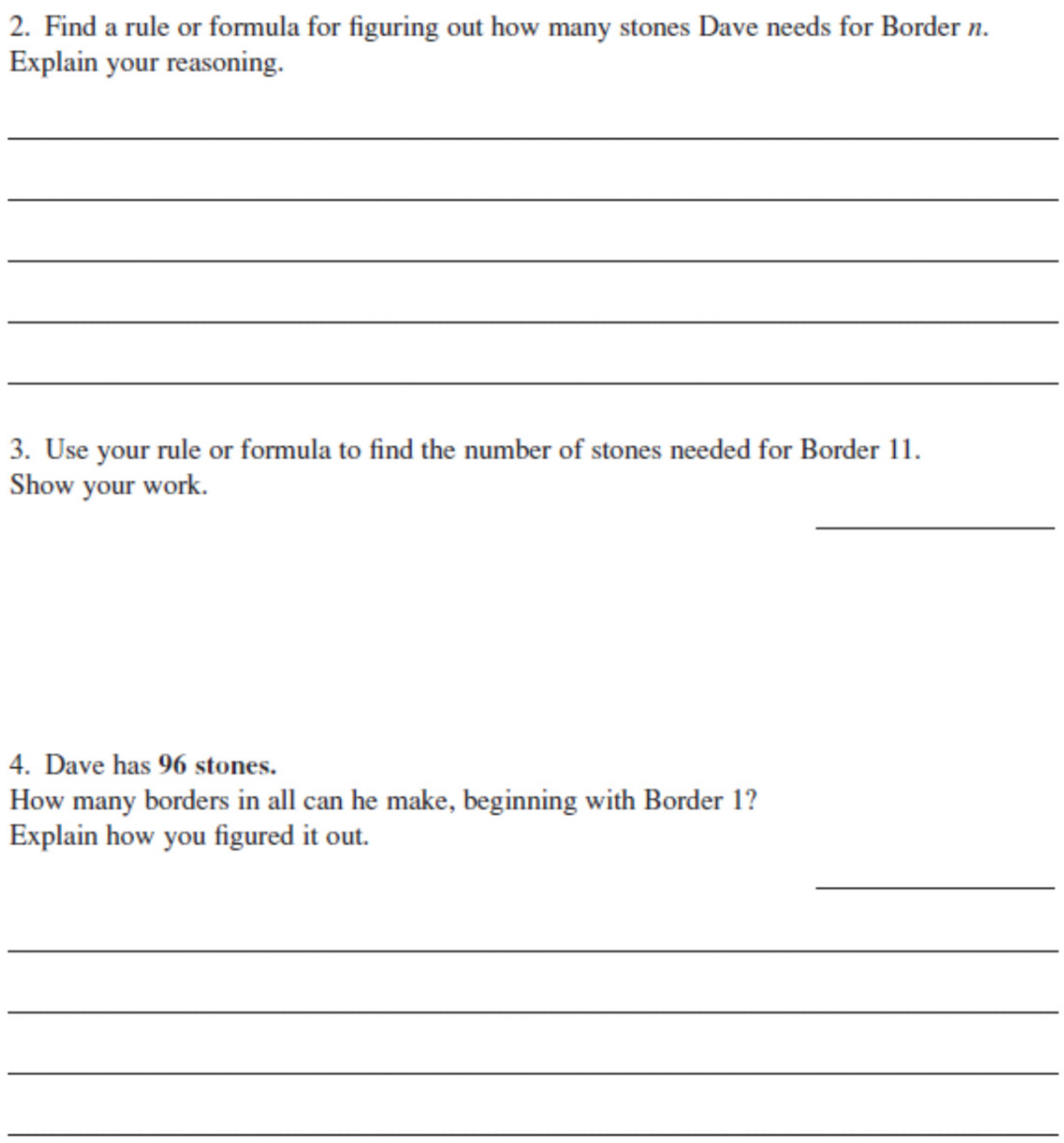
The borders will be made using colored rectangular stones. The stones measure 2 feet by 1 foot.

Dave decides to use this plan for three more borders.

He begins to make a table to find out how many stones he needs for each border.

Border	1	2	3	4	5
Number of stones	6	10			

1. Fill in the missing numbers in Dave's table. Explain how you figured out your answer.



2. Find a rule or formula for figuring out how many stones Dave needs for Border *n*.
Explain your reasoning.

3. Use your rule or formula to find the number of stones needed for Border 11.
Show your work.

Level IV
Theta 0.91

4. Dave has **96 stones.**
How many borders in all can he make, beginning with Border 1?
Explain how you figured it out.

Level VI
Theta 2.16

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

For Cups Item 3 (Level IV, Theta 0.96), please see page 240.

For Cups Item 2 (Level IV, Theta 1.17), please see page 240.

For Tiling Squares Item 2 (Level IV, Theta 1.27), please see page 227.

Level V

For Fish Ponds Item 3 (Level V, Theta 0.61), please see page 243.

Vacations

This problem gives you the chance to:
- analyze relationships using graphs and algebra

Here is some information about how some students are paying for their summer vacations.

Carla: Her mom gave her $100 in January and Carla has saved $25 every month since, starting in February.

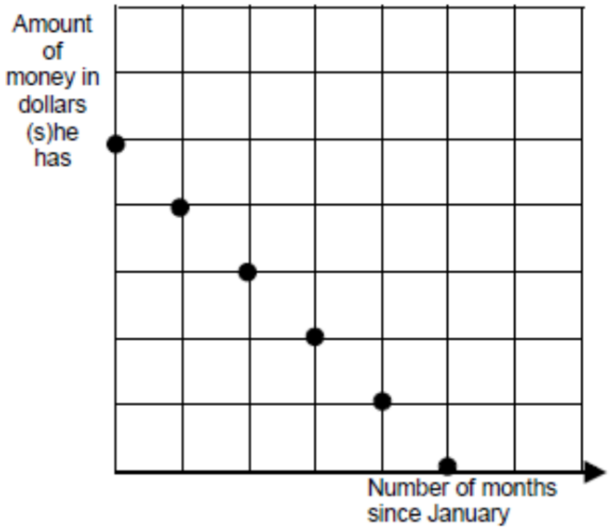
Arnie: Arnie put $150 in his piggy bank in January.

Sue: Sue booked her vacation in January. She had $250 in her piggy bank. Starting in February, she is paying $50 each month to the travel company.

Ben: Starting in February, Ben saves $30 every month.

Here are some graphs illustrating these situations.

1. Match each person with a graph and explain how you decided.

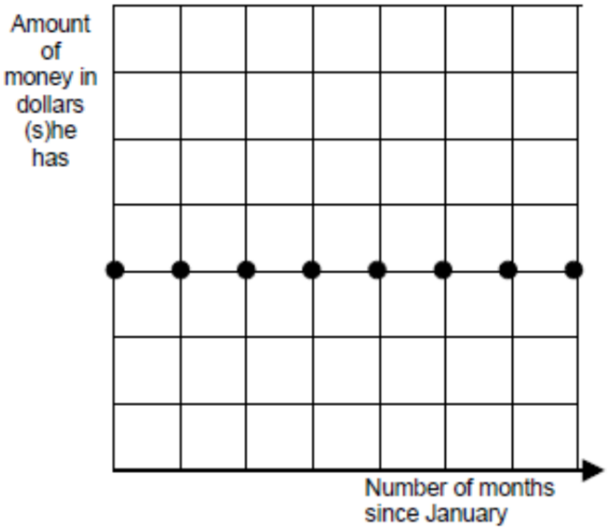


Name: ____________________

Reason: ____________________________

____

____

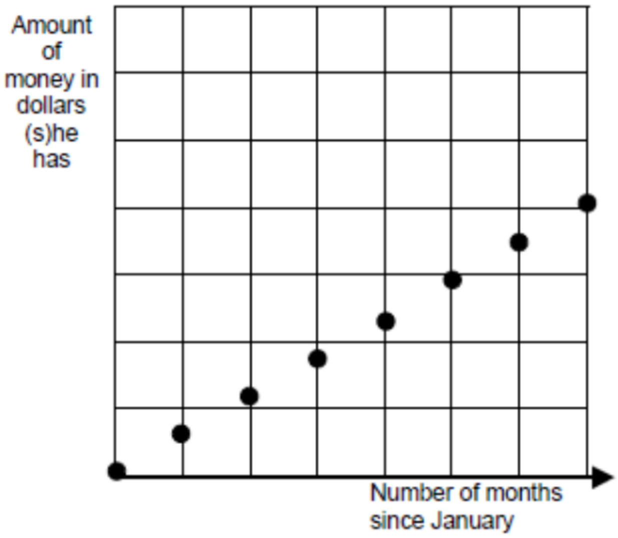


Name: ____________________

Reason: ____________________________

____

____

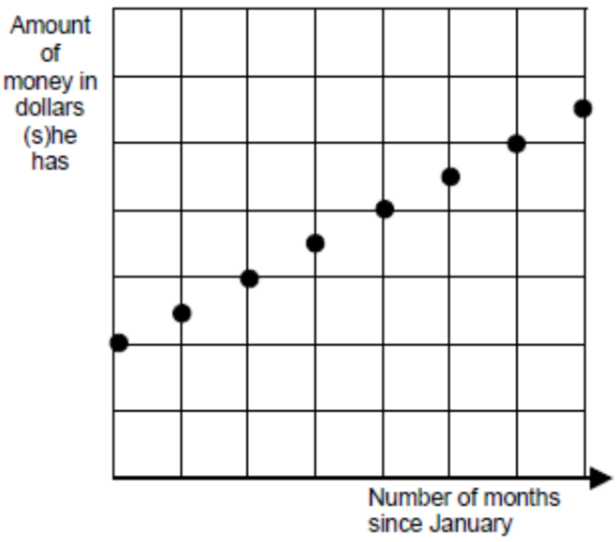


Name: ____________________

Reason: ____________________

____

____



Name: ____________________

Reason: ____________________

____

____

2. In these equations, $\$A$ is the amount of money and n is the number of months since January.

$A = 250 - 50n$
$A = 30n$
$A = 150$

a. Find the person for each of these equations.
b. Write a formula for the fourth person.

Carla ____________________

Arnie ____________________

Sue ____________________

Ben ____________________

Level V
Theta 0.66

3. Write a possible description for this formula: $A = 50n + 150$

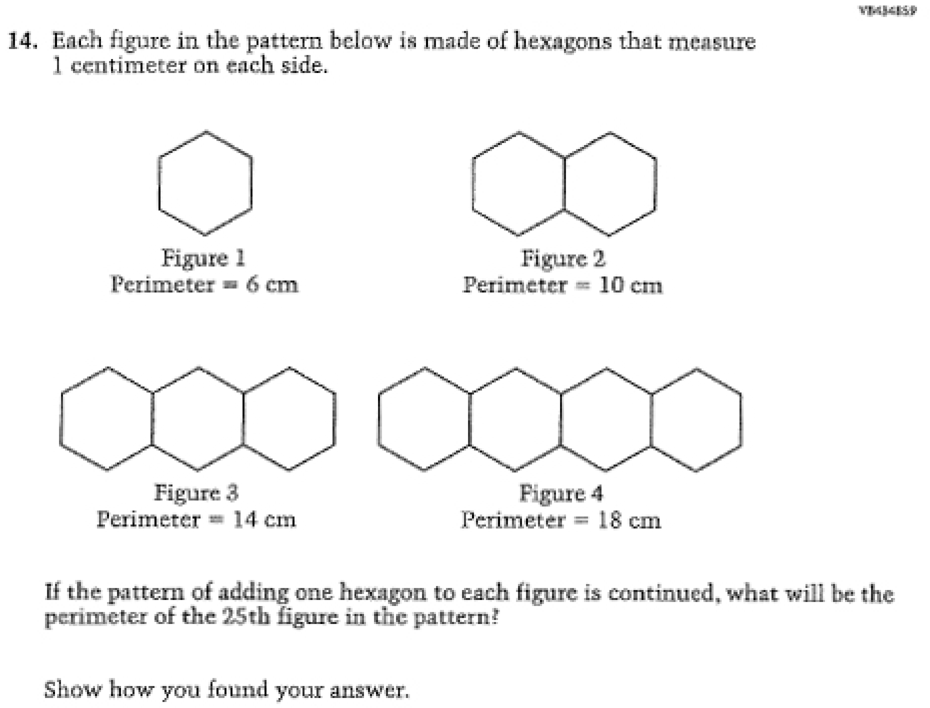
____

____

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Note: Only Item 3 from Vacations pertains to this progression. The remaining items 1 and 2 do not occur in this progression.

Level V
Theta 2.05



VB434859

14. Each figure in the pattern below is made of hexagons that measure 1 centimeter on each side.

Figure 1
Perimeter = 6 cm

Figure 2
Perimeter = 10 cm

Figure 3
Perimeter = 14 cm

Figure 4
Perimeter = 18 cm

If the pattern of adding one hexagon to each figure is continued, what will be the perimeter of the 25th figure in the pattern?

Show how you found your answer.

Item VB434859.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block B2M7.

Level VI

For Fish Ponds Item 4 (Level VI, Theta 0.79), please see page 243.

For Apartment Numbers Item 3 (Level VI, Theta 1.10), please see page 230.

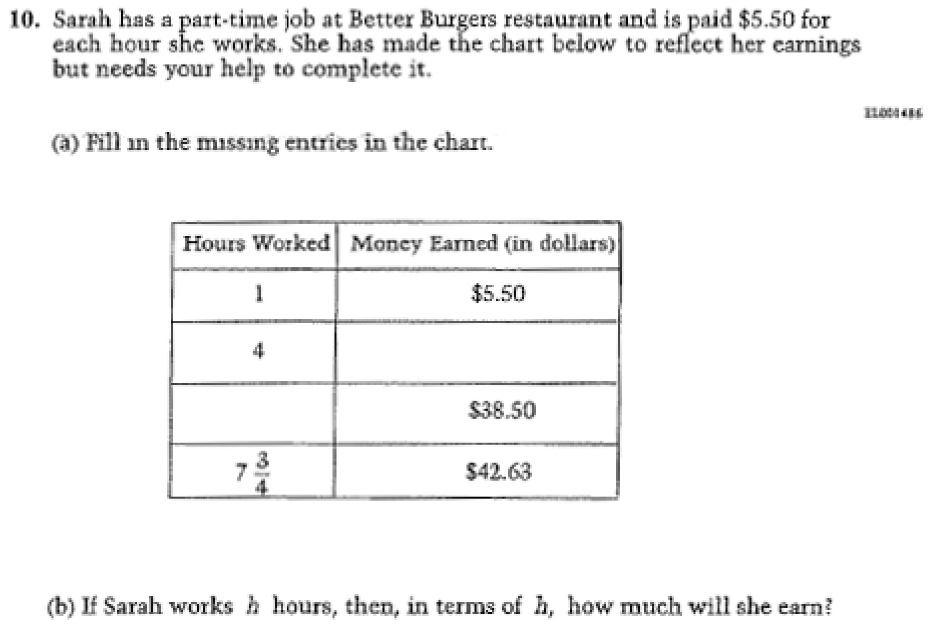
For Party Item 4 (Level VI, Theta 1.20), please see page 237.

For Design a Garden Item 4 (Level VI, Theta 2.16), please see page 245.

Level VII

For Emma's Models Item 4 (Level VII, Theta 0.63), please see page 235.

For Fish Ponds Item 5 (Level VII, Theta 0.87), please see page 243.



10. Sarah has a part-time job at Better Burgers restaurant and is paid $5.50 for each hour she works. She has made the chart below to reflect her earnings but needs your help to complete it.

(a) Fill in the missing entries in the chart.

Hours Worked	Money Earned (in dollars)
1	$5.50
4	
	$38.50
$7\frac{3}{4}$	$42.63

(b) If Sarah works h hours, then, in terms of h, how much will she earn?

Level VII
Theta 1.20

Item EL001486.

Source: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Mathematics Assessment, Grade 8, Block Z23M9B.

For Fish Ponds Item 6 (Level VII, Theta 1.22), please see page 243.

For Apartment Numbers Item 4 (Level VII, Theta 1.39), please see page 231.

For Tiling Squares Item 4 (Level VII, Theta 1.50), please see page 228.

For Apartment Numbers Item 5 (Level VII, Theta 1.52), please see page 231.

For Tiling Squares Item 3 (Level VII, Theta 1.56), please see page 227.

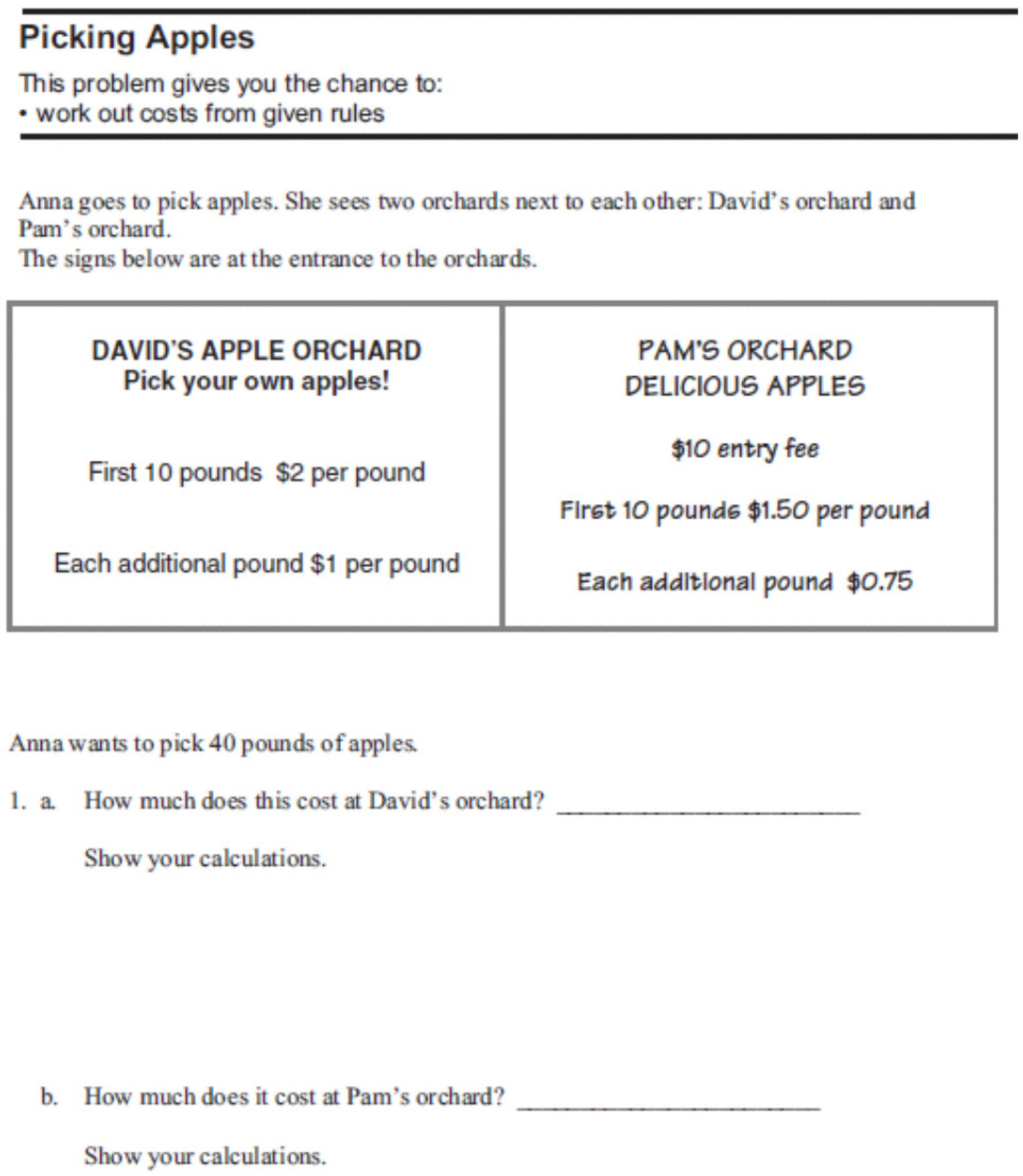
For Tiling Squares Item 5 (Level VII, Theta 1.62), please see page 228.

For Cups Item 6 (Level VII, Theta 1.65), please see page 241.

For Party Item 3 (Level VII, Theta 1.98), please see page 236.

Level VII

For Cups Item 7 (Level VIII, Theta 2.07), please see page 241.



Picking Apples

This problem gives you the chance to:
- work out costs from given rules

Anna goes to pick apples. She sees two orchards next to each other: David's orchard and Pam's orchard.
The signs below are at the entrance to the orchards.

DAVID'S APPLE ORCHARD **Pick your own apples!**	PAM'S ORCHARD DELICIOUS APPLES
	$10 entry fee
First 10 pounds $2 per pound	First 10 pounds $1.50 per pound
Each additional pound $1 per pound	Each additional pound $0.75

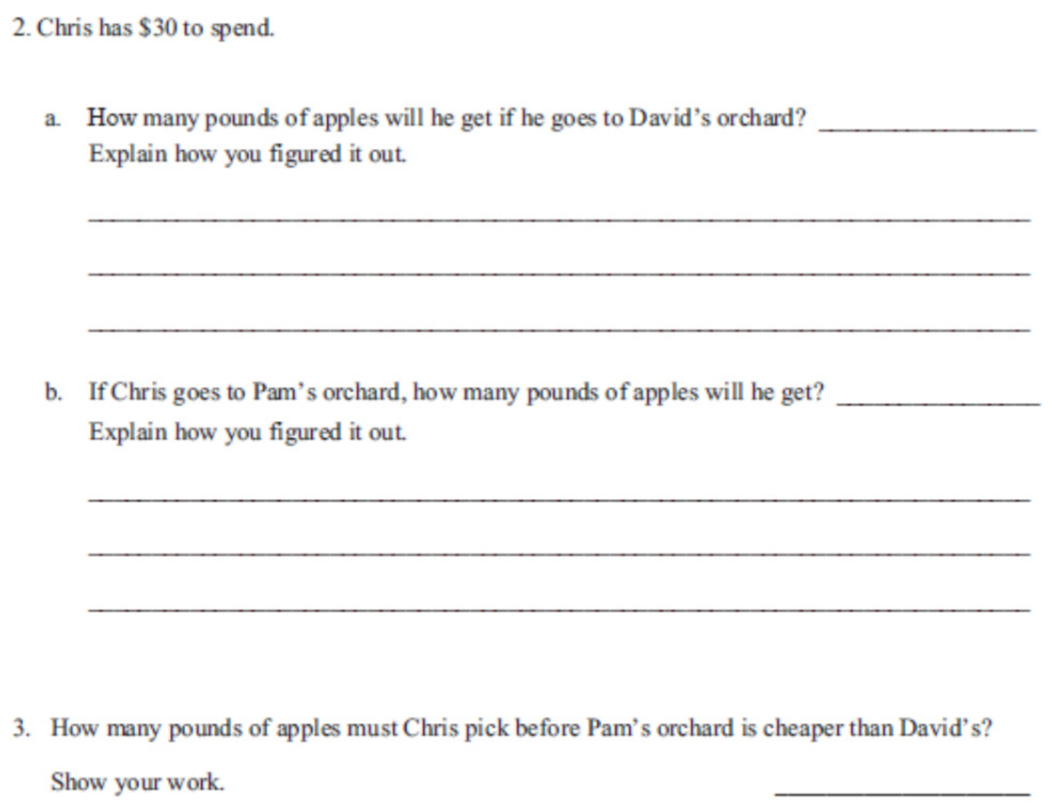
Anna wants to pick 40 pounds of apples.

1. a. How much does this cost at David's orchard? ______________________

Show your calculations.

b. How much does it cost at Pam's orchard? ______________________

Show your calculations.



2. Chris has $30 to spend.

a. How many pounds of apples will he get if he goes to David's orchard? ________________
Explain how you figured it out.

____

____

____

b. If Chris goes to Pam's orchard, how many pounds of apples will he get? ________________
Explain how you figured it out.

____

____

____

Level VIII
Theta 2.50

3. How many pounds of apples must Chris pick before Pam's orchard is cheaper than David's?

Show your work. ____________________

Source: Mathematics Assessment Resource Service (MARS). Reproduced with permission.

Note: Only Item 3 from Picking Apples pertains to this progression. The remaining items 1 and 2 do not occur in this progression.

What Might Changes in Psychometric Approaches to Statewide Testing Mean for NAEP?

David Thissen
University of North Carolina, Chapel Hill

Scott Norton
Council of Chief State School Officers

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

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

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

Acknowledgments

The authors wish to thank Peter Behuniak, Jim Chromy, Phil Daro, Richard Duran, Gerunda Hughes, Robert Linn, Ina Mullis, Karen Wixson, Frances Stancavage, and George Bohrnstedt for their helpful comments and suggestions. Any misstatements are, of course, our own.

Executive Summary

Development of the Common Core State Standards (CCSS), and the creation of the Smarter Balanced Assessment Consortium (Smarter Balanced) and the Partnership for Assessment of Readiness for College and Careers (PARCC), changes the pattern of accountability testing. These changes raise the question: "How should NAEP's validity and utility be maintained?" The assessments planned by the consortia may be different enough from current state assessments to raise questions as to whether NAEP can continue to play its historic role as an independent monitor or "check" on the validity of state assessments.

It is also clear is that computer-based assessment is coming to K–12 education, and both consortia plan to include more varied item types than have been commonly used in the past.

In considering the future of NAEP and state assessments over the next few years, three scenarios seem possible:

> (1) If most states use PARCC or Smarter Balanced assessments, NAEP would continue to have two roles: to monitor claims of improved achievement and to provide the "Rosetta Stone" (common metric) needed to compare performance across the consortia's boundaries.
>
> (2) If the two consortia merge, there would be a nearly national test. In the near term, NAEP would remain useful by serving two of its traditional purposes: to monitor and to provide historical context.
>
> (3) Even if the consortia do not continue indefinitely, their ideas are mostly likely the future of assessment. Questions about the validity of NAEP's results would arise if NAEP remained a paper-and-pencil assessment while statewide assessments were computerized.

NAEP as a Monitor

NAEP is widely regarded as a fair arbiter of results obtained from statewide assessments for the purpose of accountability. When statewide tests show improvement but NAEP results do not, questions are raised about the validity of the statewide test results.

For NAEP to continue to play this role, how similar must NAEP be to the new statewide tests? Statewide tests will soon be computer administered, with technology-enhanced item types. Should NAEP become a computerized test? Does it make any difference if the mode of administration of a test is paper-and-pencil or computerized?

Many studies examining mode effects in educational testing have reported inconsistent or mixed results. Comparability of results *can* often be maintained; however, computerization may have an effect on the results for some subgroups or subject areas.

Notable weaknesses in the literature on mode effects limit the extent to which it can be used to anticipate the effects that might be observed with NAEP. Most studies consider only a single point in time, and the literature is relatively silent on the question of whether gaps in scores among subpopulations may appear different. Examination of the pattern of results over time and among groups should be the foci of research on the effects of the computerization of NAEP.

Cross-Linkage Between the NAEP Scale and (Fewer) Statewide Tests

Efforts to link the scales of other assessments to the NAEP scale have only been moderately successful, and a large number of cautions have been offered about their usefulness. However, if most states use one of two assessments, the situation changes: More data collection options are practical for linking NAEP to the consortia assessments. The consortia assessments are in their planning stages, so a window of opportunity exists during which they might be designed to incorporate linking data collection.

It is strongly suggested that the scales of the assessments from the two consortia be linked to NAEP. In August 2011, the National Center for Education Statistics (NCES) convened a group of experts on the future of NAEP, followed by a second summit of stakeholders in January 2012. The report from those meetings made the same suggestion.

Conclusion

Computerization of NAEP is inevitable and already planned by the National Assessment Governing Board. Computerized NAEP assessments may appear more similar to future statewide assessments. Comparability of results can usually be maintained as a test makes the transition from paper-and-pencil to computerized administration, but computerization may have an effect on results for some subgroups of the population. Computerization of NAEP is best approached in the same way as other changes to NAEP assessments have been approached: A bridge study should insure the comparability of results across the transition unless an *a priori* decision is made to "break trend" regardless.

Assessments developed by Smarter Balanced and PARCC may reduce the number of statewide tests to the low single digits, thus making linkage feasible. Associations between the results of disparate educational assessments tend to change over time, so any linkage between the NAEP scale and the consortia statewide tests will need to be maintained regularly. A singular opportunity exists in a short window of time—essentially right now—to design the data collection for linkage between the NAEP scale and the consortia assessments while the latter are under development.

NAEP has a long history of implementing gradual change so that results remain comparable from year to year, while, at the same time, the assessments remain relevant in the presence of continuing educational and curricular change. We expect that spirit of gradual incremental change will continue to guide NAEP in its adaptation to the introduction of the Common Core State Standards assessments.

CONTENTS

Introduction

The development of the Common Core State Standards (CCSS), and the creation of two consortia of states—the Smarter Balanced Assessment Consortium (Smarter Balanced)[6] and the Partnership for Assessment of Readiness for College and Careers (PARCC)[7]—to develop assessments based on those standards, promises to change the pattern of K–12 accountability testing in the U.S. These changes raise the question "How should NAEP's validity and utility be maintained in the context of the CCSS?"

Crucial aspects of this question have to do with the relationship between the CCSS and the NAEP content frameworks, which will be examined in other studies. However, it is also possible that changes in the approach to testing planned by the two assessment consortia may induce changes in the ways that existing assessments, such as NAEP, are perceived, or may change how NAEP needs to be scored and maintained to provide an accepted "check" on the validity of the new statewide assessments. Furthermore, a few states have indicated that they will not be joining either of the consortia, further complicating the job of NAEP as a monitor of states' educational achievements.

---

[6] http://www.k12.wa.us/SMARTER/default.aspx

[7] http://www.parcconline.org/

Background

Common Core State Standards Initiative

At the time of this writing, 45 states and the District of Columbia have officially adopted the CCSS.[8] However, adoption of the standards does not necessarily mean state content standards for K–12 mathematics and English language arts (ELA) will become identical across the states. According to documentation from the Common Core State Standards Initiative, "adoption" means that a "State adopts 100% of the common core K–12 standards in ELA and mathematics (word for word), with option of adding up to an additional 15% of standards on top of the core."[9] Thus, even at the level of standards, there is likely to remain some variation among CCSS states' curricula, and possibly their assessments, while additional between-state variation will arise from the states that have not (yet) adopted the CCSS.

Although both consortia plan assessments that are based on the CCSS, they plan tests that differ in a number of respects. This will split states into three clusters—the Smarter Balanced states, the PARCC states, and the small number of states that are members of neither group and will presumably continue to operate their own assessment programs. Membership of states in the two consortia is listed in Appendix A. All of the states that are members of one or both consortia have adopted the CCSS; none of the five states that are not members of either consortium have done so. Utah adopted the CCSS, but has since withdrawn from Smarter Balanced, while the small number of states that are currently in both consortia will presumably settle on one or the other by the time of operational testing in 2014–2015.[10]

Smarter Balanced Assessment Consortium (Smarter Balanced)[11]

Features of the Smarter Balanced assessments include "Summative Assessments" that are planned to be "Mandatory comprehensive accountability measures that include computer adaptive assessments and performance tasks, administered in the last 12 weeks of the school year in Grades 3–8 and high school for English language arts (ELA) and mathematics." These "capitalize on the strengths of computer adaptive testing, i.e., efficient and precise measurement across the full range of achievement and quick turnaround of results" and "produce composite content area scores, based on the computer-adaptive items and performance tasks." Smarter Balanced also plans "Interim Assessments" that are "Optional comprehensive and content-cluster measures that include computer-adaptive assessments and performance tasks, administered at locally determined intervals." These are to be

---

[8] However, an article in the May 8, 2012, issue of the *Wall Street Journal* indicates that up to five states are reconsidering their commitment.

[9] Slide presentation "Common Core State Standards Initiative, March 2010," downloaded from http://www.corestandards.org/about-the-standards.

[10] Membership lists for states adopting the Common Core State Standards were obtained from http://www.corestandards.org/in-the-states.

[11] Quoted material in this section is from http://www.k12.wa.us/SMARTER/pubdocs/SBACSummary2010.pdf

"Grounded in cognitive development theory about how learning progresses across grades and how college- and career-readiness emerge over time." System features include "coverage of the full range of ELA and mathematics standards and breadth of achievement levels by combining a variety of item types (i.e., selected-response, constructed response, and technology-enhanced) and performance tasks, which require application of knowledge and skills."[12]

Partnership for Assessment of Readiness for College and Careers (PARCC)[13]

PARCC lists six "priority purposes" for their assessments:

"1. Determine whether students are college- and career-ready or on track
2. Assess the full range of the Common Core State Standards, including standards that are difficult to measure
3. Measure the full range of student performance, including the performance of high- and low-performing students
4. Provide data during the academic year to inform instruction, interventions, and professional development
5. Provide data for accountability, including measures of growth
6. Incorporate innovative approaches throughout the system"

PARCC plans an assessment system with four components. "Each component will be computer-delivered and will leverage technology to incorporate innovations."

Two summative, required assessment components will be designed to:

- "Make 'college- and career-readiness' and 'on-track' determinations,
- Measure the full range of standards and full performance continuum, and
- Provide data for accountability uses, including measures of growth."

Two nonsummative, optional assessment components will be designed to "generate timely information for informing instruction, interventions, and professional development during the school year. An additional third nonsummative component will assess students' speaking and listening skills."

"PARCC will also leverage technology throughout the design and delivery of the assessment system. The overall assessment system design will include a mix of constructed response items, performance-based tasks, and computer-enhanced, computer-scored items. The PARCC assessments will be administered via computer, and a combination of automated scoring and human scoring will be employed."

---

[12] In late 2012, the Smarter Balanced assessment design was revised to include only one performance task in each subject—mathematics and English/language arts (Gewertz, 2012).

[13] Quoted material in this section is from http://www.parcconline.org/parcc-assessment-design

The PARCC assessments are not, however, currently planned to be computer adaptive, as is the case with Smarter Balanced assessments.

Summary

The extent to which even consortium-member states will have identical assessments is not clear at the time of this writing; it is possible the consortia assessments will be locally augmented, or otherwise modified. In addition, there will probably be some states that use unique assessments. Nevertheless, it is clear that computer-adaptive (Smarter Balanced) or computer-based (PARCC) assessment is coming soon to K–12 testing. In addition, both consortia appear to plan to take advantage of computer administration by including much more varied item types than have been the norm in large-scale assessment.[14] This represents a potentially dramatic shift in assessment; while some states currently administer online tests, they are typically paper-and-pencil tests that have been transferred to the computer. Finally, documentation from Smarter Balanced specifically mentions the idea that some items may reflect learning progressions.

---

[14] It now appears that both consortia will have to provide paper-and-pencil versions of the test as not all schools will be able to support computer-based assessments. Such paper-and-pencil alternatives will not be the same as the computerized versions with respect to any technology-enhanced item types that the consortia develop or use, so the paper-and-pencil versions would probably be relatively short-term solutions to specific challenges in the initial implementation, rather than continuing alternate forms.

Questions for the Future of NAEP

Correctly anticipating future events is always a challenge. At the September, 2011, meeting of the NAEP Validity Studies Panel (NVS Panel), Peter Behuniak suggested that three scenarios might be considered for the next few years:

1. There might be minimal change from current commitments—i.e., most states become aligned with one of the two consortia, and a few states associate with neither. After the consortia assessments become available in academic year 2014–2015, the majority of the states will use one of those two assessments, with a small number of states using unique, state-specific tests.
2. The two consortia could conceivably merge to become one. There is some basis for such speculation in recent history: As the current consortia were being formed, several smaller exploratory groups merged to become PARCC. If a merger happens, there would be one nearly national assessment, although a few states would likely continue using unique, state-specific tests.
3. The consortia might fragment, become much smaller, or go out of existence entirely after the current Race to the Top federal funding ends. Race to the Top funding is being provided for assessment development only, so new structures will have to be established for Smarter Balanced and PARCC to administer operational assessments. Because it is not clear at the time of this writing what the mechanism might be to provide continued financing for the consortia, prudence demands that this possibility be considered.

Scenario 1: Two Consortia and Nonconsortium States

In a future that has approximately half the states using the PARCC assessments, approximately half the states using the Smarter Balanced assessments, and a few states using unique tests, an appropriately configured NAEP would continue to have two obvious roles. The first role would be to monitor claims of improved achievement. Even when created at the level of consortia (instead of individual states), statewide assessments would be vulnerable to "teaching to the test" and the possible appearance of inflated achievement gains, which would be identified, as they have been in the past, when statewide assessment scores appeared to rise faster than NAEP scores. A second role of NAEP, as the only assessment administered in all states, could be to provide the "Rosetta Stone" needed to compare performance across the consortia boundary (i.e., between the PARCC states and the Smarter Balanced states), possibly including the nonconsortium states. Without some linkage, each year there could be a stack of statewide averages on the PARCC assessment, an unconnected stack of statewide averages on the Smarter Balanced assessment,[15] and results from a few states comparable to neither group. Suitable linking *may* make it possible to compare PARCC and Smarter Balanced results. Of the two possible linking designs, common-population linking appears unlikely, because it seems improbable that any local authority would administer assessments from both

---

[15] It is conceivable that the consortia could cross-link their assessments without NAEP as an intermediary; however, no plan for this has been announced.

PARCC and Smarter Balanced. A common-item linking design might be feasible, using NAEP to supply the common items; this is discussed in a subsequent section, "NAEP as Lingua Franca: Cross-Linkage Between the NAEP Scale and (Fewer) Statewide Tests."

Scenario 2: Merged Consortia and Nonconsortium States

If the two consortia merge, the merger would produce a nearly national test. Setting aside for the moment the few nonparticipating states, a single merged consortium would displace NAEP from its unique role as the only national measure of achievement. In the very long term (i.e., decades), this development might render NAEP superfluous. However, in the nearer term, an appropriately configured NAEP would remain useful by serving two of its traditional purposes. NAEP would still perform a "monitor" function because the consortium's one nearly national test would still be vulnerable to "teaching to the test" and the possible appearance of inflated achievement gains. The latter would be identified, as they have been in the past, when statewide assessment scores appeared to rise faster than NAEP scores. In addition, NAEP would continue to provide historical context. It would take decades for a new assessment, even if it was national, to accrue the kind of trend data that NAEP possesses. Trend data have been important for policymakers for some time, and that would be expected to continue.

Scenario 3: No Consortia, but New Ideas Remain

Even if the consortia do not continue indefinitely, the ideas they plan to bring to large-scale assessment are most likely the ideas of the future. Specifically, the fact that both consortia, representing nearly all of the states, emphasize computerized assessment is a clear indicator that many statewide assessments may well use computerized administration within the next few years. In this scenario, NAEP's role as a monitor of fragmented statewide accountability systems could continue, but questions of the validity of NAEP's results would increasingly arise if NAEP remained an "old-fashioned" paper-and-pencil assessment while statewide assessments adopted computer administration and made use of technology-enhanced item types.

NAEP as a Monitor: Paper-and-Pencil NAEP in a World of Computerized Statewide Tests

NAEP is widely regarded as a fair arbiter of results obtained from statewide assessments for the purposes of accountability. When statewide tests show improvement but NAEP results do not, questions are raised about the validity of the statewide test results. Might the state results be the result of "teaching to the test" or "narrowing the curriculum" to obtain high scores?

How Similar Must NAEP Be?

The use of NAEP as a monitor depends on its acceptance as a widely respected measure of student achievement. NAEP's framework- and item-development processes and its data analysis procedures have been universally accepted as state of the art. For the less technically inclined, the paper-and-pencil format of NAEP is very similar to the paper-and-pencil format of most of the statewide assessments for which it serves a monitoring function.

However, this is about to change. Under any of the scenarios described above, within the next five (or very few more) years, statewide assessments will be computer administered, and, in many states, probably computer adaptive, with technology-enhanced item types. If NAEP remains as it has been, it will increasingly "look different."

If paper-and-pencil NAEP "looks different", and its results differ from computerized statewide tests with more varied item types, NAEP may cease to be accepted as the final arbiter, and NAEP results may be dismissed because "students were not as motivated on the old-fashioned paper test as they were on the attractive computerized test," or because "the old-fashioned paper test did not include the instructionally sensitive technology-enhanced item types that are on the computerized test."

Further, if linkages between the NAEP scale and those of the PARCC and Smarter Balanced assessments are proposed (see the next section on "NAEP as Lingua Franca: Cross-Linkage Between the NAEP Scale and (Fewer) Statewide Tests"), it may, as a practical matter, be necessary for NAEP to become a computer-administered test to perform its part in the linkage.

Should NAEP become a computerized test? There are three classes of considerations involved in answering this question.

- The first class of considerations is practical: Would computer administration make NAEP more or less expensive? If the answer is that it would make NAEP more expensive, is the cost acceptable? Another kind of practical difference between computerized and paper-and-pencil administration involves accommodations: Some accommodations (e.g., large type, audio presentation, some kinds of translation) are easier or less expensive to provide with a computerized test than with paper and pencil. Such practical questions are beyond the scope of this essay (and our expertise).

- The second class of considerations involves the need for NAEP to be computerized in order to administer questions that appropriately measure aspects of the CCSS. If (other) groups examining the CCSS frameworks and consortia assessment plans conclude that there are some objectives that can only be measured with technology-enhanced item types, it may be necessary for NAEP to computerize in order to provide measurement of those aspects of knowledge or skills.

- The third class of considerations can be summed up by the question "Does it make any difference if a test is administered in a paper-and-pencil or computerized format?" There is evidence in the psychometric literature on this question.

What Is Currently Known About Mode of Administration Effects?

Over the past three decades, a number of assessments have been converted from paper-and-pencil administration to computer-based or computer-adaptive administration, beginning with the transition of the Armed Services Vocational Aptitude Battery (ASVAB) in the 1980s–1990s (Sands, McBride, & Waters, 1997). NAEP is among the programs that have computerized some assessments: The 2011 NAEP writing assessment was administered as a computer-based test for Grades 8 and 12, and a pilot study of a Grade 4 computer-based writing assessment was in the field in early 2012.[16] The 2009 NAEP science assessment included interactive computer tasks,[17] and the NAEP Technology and Engineering Literacy (TEL) assessment will be computer-administered when it appears in 2014.[18]

Does computer administration in and of itself affect the results of an assessment?

Many research studies have examined the comparability of results obtained with paper-and-pencil and computerized tests. Appendix B summarizes some of the conclusions that can be drawn from studies over the past 15 years (since 1997); earlier studies were excluded because they would have involved computer administration very different from what would be used now.

The conclusion of Appendix B is that:

> Many studies examining mode effects in educational testing have shown inconsistent or mixed effects. The research is clear in demonstrating that comparability of results *can* often be maintained overall as a test makes the transition from paper-and-pencil to computerized administration. For example, most of the studies suggest that the structure of the test is likely to remain unchanged in moving from paper-and-pencil to computer-based administration. However, the evidence is mixed on the effects of mode on score comparability; computerization may have an effect on the results for some subgroups of the population and these can vary further as a function of

---

[16] http://nces.ed.gov/nationsreportcard/writing/cba.asp

[17] http://nces.ed.gov/nationsreportcard/science/whatmeasure.asp

[18] http://nces.ed.gov/nationsreportcard/techliteracy/

> the subject area being assessed. Schroeders and Wilhelm (2011) perhaps best summarize what is required when moving to computerized assessment when they write "… equivalence research is required for specific instantiation unless generalizable knowledge about factors affecting equivalence is available" (pg. 1).

Characteristics of assessments that have been shown to raise the possibility of different scores from computerized and paper testing include essay responses, which may be graded more or less stringently depending on mode, and items with graphics or manipulatives, which may be made either easier or more difficult in translation to computerized delivery. Participant characteristics that may interact with the relative difficulty of computerized presentation have included gender (in some studies) and special education status. Probably the most salient (unintended) individual differences variable that may be related to the results obtained with computerized assessments is computer familiarity, which, while not a very well defined term, includes skills with a keyboard and probably some other aspects of the idiom used in the computer interface. However, these effects have been rare historically, and can likely be eliminated with careful assessment design and thoughtful instructions and preparation. Indeed, given the ubiquity of a range of computerized devices in everyday life, from personal computers through tablets and smart phones, it may soon be the case that the question would be whether paper-and-pencil testing accurately or authentically measures what children know and can do.

For NAEP, the difference between computerization alone (making a computer-based test [CBT]) and adaptation (creating a computerized adaptive test [CAT]) should not be significant. NAEP is already "scored" (actually, aggregate summary statistics are computed) using an item response theory (IRT) model in the presence of planned missing data, due to the fact that each examinee responds only to the subset of items. Use of a CAT changes only the mechanism by which items are assigned to respondents. The assumption used in current NAEP IRT analysis—that the "missing" item responses are missing at random (MAR) (Rubin, 1976)—remains valid because in a CAT, the missingness mechanism depends only on observed data.

Two notable weaknesses in the literature on mode effects limit the extent to which it can be used to anticipate the effects that might be observed with NAEP. First, most studies consider only a single point in time, whereas NAEP is primarily an instrument to measure change. One might assume that a computerized test that appeared to measure the same constructs, in the same way, as an existing paper-and-pencil test at one point in time would also yield comparable trend results; however, there has been little, if any, empirical investigation of this question. A second weakness in the existing literature is that it is relatively silent on the question of whether gaps in scores among subpopulations may appear different, depending on whether computerized or paper-and-pencil tests are used. These two kinds of questions, on the pattern of results over time and between groups, should probably be the foci of research on the effects of the computerization of NAEP.

NAEP as Lingua Franca: Cross-Linkage Between the NAEP Scale and (Fewer) Statewide Tests

For the past two decades there has been continuing interest in linking the scales of other assessments to the NAEP scale in order to obtain more value from expensive data collection efforts by producing linked results that can be compared with data from additional sources. These efforts have been successful to varying degrees, and a large number of cautions have been offered about their usefulness (Thissen, 2007; Linn, McLaughlin, & Thissen, 2009).

However, especially under scenario 1 (described above)—in which the states are divided roughly into halves using one of two assessments—the linking landscape changes in two ways. First, although only limited practical strategies exist for linking NAEP to 50 statewide tests, more data collection options are practical for linking NAEP to a universe of two consortium assessments. Second, the two consortia assessments are still in their planning stages and a window of opportunity exists during which they might be designed to incorporate linking data collection.

The strong suggestion made here is that the scales of the assessments from the two consortia should be linked to NAEP. In August 2011, the National Center for Education Statistics (NCES) convened a group of experts in assessment, measurement, and technology for a summit on the future of NAEP, and this was followed by a second summit of state and local stakeholders in January 2012. NCES then assembled a panel of experts from the first summit, chaired by Edward Haertel, to consider and further develop the ideas from the two discussions, and make recommendations that were summarized in a report to the Commissioner of NCES (NCES Initiative on the Future of NAEP, 2012). That report proposed "the development of mechanisms for flexible linking of NAEP to other scales. This would include reweighting of content within NAEP if necessary, so as to maximize alignment with any of a range of large-scale assessment programs, including the Smarter Balanced and PARCC summative assessments as well as PISA [Program for International Student Assessment], the Progress in International Reading Literacy Study (PIRLS), TIMSS [Trends in International Mathematics and Science Study], and others" (p. 40). To facilitate linkage, the panel placed high priority on "studies of NAEP design changes to facilitate linkages between NAEP and other large-scale assessment programs, including the summative assessments developed by the PARCC and Smarter Balanced consortia at grades 4, 8, and possibly 12" (p. 47).

A Manageable Design Based on a Great Deal of Cooperation

In the past, linkages among disparate assessments have rarely been symmetrical efforts in which the linking data are collected in the naturally occurring context for both tests. However, attempts to link the scales of the PARCC and/or Smarter Balanced assessments with NAEP may be different. With PARCC and Smarter Balanced still in the planning stages, it may be possible to design linking data collections that symmetrically embed NAEP blocks or items within the PARCC and/or Smarter Balanced assessments, and embed PARCC and/or Smarter Balanced items within operational administrations of NAEP. We note that the *Future of NAEP* report (2012) suggested consideration of “main NAEP data collection with expanded slots for: (1) linking items; and (2) experimental item types” (p. 48) to facilitate such symmetrical linking.

If such symmetrical data were collected, questions about the effect of context on each assessment’s item responses could be resolved empirically, and threats to the validity of linkage would be subject to data analysis. The strategy would, moreover, be amenable to many (technical) forms of linking. If both PARCC and Smarter Balanced participate along with NAEP, not only might the scales of both consortia be linked to the NAEP scale, but the PARCC and Smarter Balanced scales may be (implicitly) linked to each other. Thus, such linkage could serve to align the two “stacks” of statewide results, one for the PARCC states and the other for the Smarter Balanced states.

The question of what to do with the states that participate in neither PARCC nor Smarter Balanced would remain. However, those states would have NAEP results and possibly greater motivation to participate in one of the consortia because comparability of scores would add value to the products of both consortia.

Conclusion

Computerization of NAEP is inevitable. Indeed, recent discussion of assessment schedules by the National Assessment Governing Board suggest that all NAEP assessments (with the possible exception of the long-term trend assessment) may be computer-administered by 2019; some will be computerized earlier and some have already been computerized. There are several reasons for computerization. NAEP assessments may be computerized so that technology-enhanced item types can be delivered when required by the frameworks, as has already happened with the science assessment in 2009 and is planned for the TEL assessment in 2014. NAEP assessments may be computerized so that they appear more comparable with statewide assessments being developed by the consortia or to facilitate linking with those assessments. They may be computerized simply because computer administration has become more cost effective—this will ultimately happen for all assessments as the cost of computing equipment decreases and the costs of printing and physical distribution and scoring of paper response sheets grow. Finally, all assessments will gradually become computerized as computer use becomes ubiquitous for real-world tasks, both within and outside schools.

From the literature on the computerization of other assessments, it is clear that comparability of results can usually be maintained as a test makes the transition from paper-and-pencil to computerized administration. It is also clear that, sometimes, some aspect of computerization may have an effect on results for some subgroups of the population. This suggests that the computerization of NAEP is best approached in the way that all other changes made to NAEP assessments since the advent of the "new design" in 1983 have been approached: careful consideration should be given to the design of the computerized administration, and a bridge study should be carried out to ensure comparability of results across the transition (unless an *a priori* decision is made to "break trend" regardless).

At an unlikely extreme, it is *possible* that in some subject matter areas a computerized NAEP might be found to measure the relevant constructs sufficiently differently that choices would have to be made between "breaking trend" and using the new assessment, continuing with the paper-and-pencil measure for the sake of continuity, or creating another parallel NAEP, with the old paper-and-pencil measure running alongside a new computerized assessment (much as the NAEP's long-term trend assessment has run in parallel with the new design for the past three decades). Although this possibility is not likely (given accumulated experience with computerizing existing assessments), it is best to avoid *a priori* rejection of any possibility.

Assessments developed by Smarter Balanced and PARCC may reduce the number of statewide tests in Grades 4 and 8 from nearly 50 to the low single digits, starting in the 2014–15 academic year. Assuming this happens, it will change the ways in which NAEP can serve as a monitor of progress, as reflected by statewide tests. With such a small set of tests to work with, linkage may become feasible, permitting close quantitative comparison between NAEP results and those obtained with the consortia tests, and providing a mechanism to link the consortia tests' scales with each other across the two groups of states. Historically, such linkage has been fraught with difficulties (Thissen, 2007; Linn, McLaughlin, & Thissen, 2009). However, linkage is better understood now than in previous decades, and there is agreement on the technical approaches required.

One result that is clear from the literature on linkage is that relations between the results of disparate educational assessments tend to change over time. This means that any linkage between the NAEP scale and the consortia statewide tests will need to be maintained regularly over the years of their use. However, we note that a singular opportunity exists in a short window of time—essentially right now—to design data collection for linkage between the NAEP scale and the consortia assessments while the latter are under development. At this time, central control remains possible, and cooperative agreements to collect suitable linking data may be more easily obtained than will be the case after the consortia tests branch and fork into two dozen statewide assessments. This opportunity is very attractive, and may spur computerization of some NAEP assessments so that parts of those assessments can be embedded by the consortia in item tryout or first operational administration, and vice-versa in NAEP in the 2014–15 time frame.

A useful side effect of embedded-block linkage of the new consortia tests with the NAEP scale during development may be that the process will help explain to policymakers any change that may arise in results reported by pre- and post-consortia statewide tests. The new tests, with associated new standards, may appear to suggest large changes in the proportions of students categorized as "proficient" in many states; such changes have, historically, been the reason that linkages have been found to change over time (Thissen, 2007; Linn, McLaughlin, & Thissen, 2009). Linkage of the results to some stable scale, like that of NAEP, could help consumers of the results distinguish between real change and artifactual "change" arising solely from new assessments or standards.

Looking ahead, we see that the only constant in educational assessment is change. NAEP has a long history of implementing gradual change so that results remain comparable from year to year, while, at the same time, the assessments remain relevant in the presence of continuing educational and curricular change. We expect that spirit of gradual incremental change will continue to guide NAEP's adaptation to the new environment of the second decade of the 21st century.

References

Gewertz, C. (2012). Test group rethinks questions. *Education Week*, *32*(13), 1–24.

Linn, R. L., McLaughlin, D., & Thissen, D. (2009). *Utility and validity of NAEP linking efforts*. A publication of the NAEP Validity Studies Panel. Washington, DC: American Institutes for Research.

NCES Initiative on the Future of NAEP. (2012). *NAEP: looking ahead—Leading assessment into the future*. A white paper commissioned by the National Center for Education Statistics. Washington, DC: U.S. Department of Education.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, *63*, 581–592.

Sands, W. A., McBride, J. R., & Waters, B. K. (Eds.) (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71*, 849–869.

Thissen, D. (2007). Linking assessments based on aggregate reporting: Background and issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.) *Linking and aligning scores and scales* (pp. 287–312). New York: Springer.

Appendix A. Membership in the PARCC and Smarter Balanced Consortia†

PARCC	Both	Smarter Balanced
Arizona*	North Dakota	Alaska
Arkansas*	Pennsylvania	California*
Colorado		Connecticut*
District of Columbia*		Delaware*
Florida*		Hawaii*
Georgia*		Idaho*
Illinois*		Iowa*
Indiana*		Kansas*
Kentucky		Maine*
Louisiana*		Michigan*
Maryland*		Missouri*
Massachusetts*		Montana*
Mississippi*		Nevada*
New Jersey*		New Hampshire*
New Mexico*		North Carolina*
New York*		Oregon*
Ohio*		South Carolina *
Oklahoma*		South Dakota*
Rhode Island*		Vermont*
Tennessee*		Washington*
		West Virginia*
		Wisconsin*
		Wyoming
	Neither	
	Alaska	
	Minnesota	
	Nebraska	
	Texas	
	Utah	
	Virginia	

* "Governing" states.

† Membership was compiled from the websites of the two consortia, http://www.parcconline.org/parcc-states for PARCC and http://www.k12.wa.us/SMARTER/States.aspx for Smarter Balanced on December 3, 2012. Membership in the two consortia has been somewhat fluid; these lists differ from the lists provided in the June 2010 Race to the Top applications.

Appendix B. Computer-Based Assessment: A Review of the Last 15 Years of Comparability Research

Sharyn Rosenberg, *American Institutes for Research*
Reanne Townsend, *American Institutes for Research*

As personal computers and other technologies become more advanced, and more prevalent among the U.S. population, it is becoming increasingly important to use these tools to improve and enhance educational assessment. The National Center for Education Statistics (NCES), which oversees the development of the National Assessment of Educational Progress (NAEP), recognizes this trend and plans to have NAEP fully computer-based by 2022. However, this transition cannot be made lightly; it is important to determine whether scores obtained from computer-based testing can be expected to be statistically comparable with those obtained from the previous paper-and-pencil based administrations, and whether meaningful comparisons can be made between the two modes. In other words, can trend be maintained in reporting?

In 1999, NCES commissioned two experimental studies—one for writing and one for mathematics—to examine potential mode effects when comparing paper-and-pencil based tests and computer-based administrations of NAEP. The writing online study, conducted in 2002 using nationally representative samples from Grade 8 main NAEP, found no differences in performance when comparing scores from the paper- and computer-based administrations overall or by subgroup, with one exception; students from urban schools performed significantly better on the paper test than the computerized test, with an effect size of 0.15 (Horkay, Bennett, Allen, Kaplan, & Yan, 2006). The mathematics online study, conducted in 2001 using nationally representative samples from Grade 8, found that overall scores were 4 points lower for the computer-based administration than for the paper version; several item difficulty parameters varied substantially across the two modes, indicating that the mathematics test did have score differences by mode (Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008).

The NCES experimental studies on mode effects are very informative, but they were performed on limited subjects and at a single grade, during a time period when computer use in schools for learning and assessment purposes was much less common. The purpose of this review is to examine research addressing the comparability of computer-based assessments and paper-and-pencil based tests as one way of informing expectations for a broader application of computerized NAEP.

The mode effects of computer-delivered tests and surveys have been the subject of investigation since the mid-1980s; however, the nature of interaction with computer-based technology has changed drastically since then. In light of this, and because of the great breadth of literature available on the subject, this paper examines 65 comparability studies of academic assessments during the last 15 years (since 1997).[19]

Investigating Measurement Equivalence by Mode

In the literature, there is substantial variation in the approach taken to define and measure comparability between paper-and-pencil and computer-based testing. Of the 65 journal articles and conference presentations reviewed here, 21 included an investigation of measurement equivalence (Vandenberg & Lance, 2000) across modes; factor analyses, item response theory analyses, and/or differential item functioning (DIF) analyses were used to determine whether or not an assessment was measuring the same construct in the paper-and-pencil version as in the computer-based version. The remaining 44 studies purported to measure mode effects by analyzing whether there were mean differences between scores produced by paper-and-pencil and computer-based versions of the same assessment. Importantly, in the latter approach, potential differences between constructs across modes may be confounded with differences in mean scores.

The literature review found 21 studies that evaluated potential mode effects by measuring the extent to which an assessment measured the same construct in paper-and-pencil and computer-based formats; these are listed in Tables B1–B2. The results of most of these studies (14 out of 21) found no threats to measurement equivalence. (See Table B1.) Six studies found mixed results, and one study concluded that the assessment generally was not measuring the same construct across modes. (See Table B2.) In general, the more holistic confirmatory factor analysis approach found that paper-based and computer-based versions of the same assessment typically were measurement invariant, at least at the level of configural invariance (where patterns of free and fixed factor loadings were similar across modes) and metric invariance (where item factor loadings were similar across modes) (Horn & McArdle, 1992). The item-by-item approach employed by differential item functioning (DIF) analyses generally led to mixed results, ranging from no evidence of DIF (Taherbhai, Seo, & Bowman, 2012) to 38 percent of items flagged for DIF across modes (Gu, Drake, & Wolfe, 2006).

In many cases, there was no relationship between whether a study found measurement equivalence of constructs across modes and whether there were significant score differences by mode. Of the 14 studies that found measurement equivalence across modes, five concluded that there were no statistically significant score differences by mode either overall or by subpopulation (Karkee, Kim, & Fatica, 2010; Lottridge, Nicewander, & Mitzel, 2011; Randall, Sireci, Li, & Kaira, 2012; Schroeders & Wilhelm, 2011; Staples & Luzzo, 1999). Five studies concluded

---

[19] A librarian performed the literature search in ERIC by searching for experimental studies related to "mode effects," "comparability," "computer-based assessments," "paper-pencil assessments," and several other variations of these terms. Articles were also added by searching the reference lists of existing studies. Included studies were limited to education and certification exams administered to students (up to and including the college level). The review was limited to studies in which the same or equivalent students took paper-pencil and computerized versions of an assessment; simulation studies, literature reviews, and thought pieces were excluded.

that the computer-based assessment was associated with significantly lower scores than the paper-based version (Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008; Pomplun, 2007; Pomplun & Custer, 2005; Rowan, 2010; Taberbhai, Seo, & Bowman, 2012), and one study found that the computer-based assessment was associated with significantly higher scores than the paper-based version (Pomplun, Frey, & Becker, 2002). The results of the remaining three studies (Choi, Kim, & Boo, 2003; Kim & Huynh, 2007; Kim & Huynh, 2008) were mixed.

Construct equivalence is a necessary condition for comparing mean scores across modes, and the majority of studies in the literature review did not include analyses of measurement equivalence. Given that the studies reviewed that focus on paper-based assessments were also administered by computer with minimal adaptation, and that 20 of the 21 measurement equivalence studies found full or partial measurement equivalence across modes, we extrapolate that the score differences of the remaining 44 studies likely can be analyzed by mode. Therefore, the remaining sections incorporate all 65 studies. The complete list of the studies (and capsule summaries of their findings) can be found in Tables B1–B7.

Investigating Score Differences by Mode

Of the 65 studies reviewed, 11 found consistent differences in scores between computer-based versions and paper-based versions of the same assessment; four studies found that the computer-based format was associated with higher scores than the paper-based format, and seven studies found that the computer-based format was associated with lower scores than the paper-based format. Nineteen studies found no significant score differences by mode, either overall or by subgroup. The majority of the studies reviewed (35 out of 65) found some score differences across mode, but the results varied by content area, ability, subgroup, and/or other dimensions of the assessment or students.

Despite the lack of consistent mode effects for all students in most of the research, the many studies that found significant mode effects under specific circumstances have important implications for NAEP. As NAEP transitions to computer-based testing, it is important to recognize that certain subjects or subpopulations may be more substantially affected than others by the change in delivery mode. For example, computer-based assessment introduces new possibilities for integrating testing accommodations into the main assessment, including some aspects of universal design that make certain features available to all students (Dolan, Hall, Banerjee, Chun, & Strangman, 2005; Lee, Osborne, & Carpenter, 2010). However, it is important to ensure that new features of the computer interface do not introduce construct-irrelevant variance. The literature uncovers several issues related to mode effects from computer-based administration that include aspects of the assessments and the participants, as well as interactions between the two.

Mode Effect and Assessment Characteristics

Although many innovative item formats are made possible through the use of computer-based assessment, most test developers with an interest in maintaining trend or investigating comparability with previous paper-based versions have simply chosen to transfer more traditionally formatted items to computer-based administration. Unfortunately, the literature shows that this does not completely eliminate mode effects associated with assessment characteristics. Transferring a test from a paper-based version to the computer involves changes to item formats, but also has mode-specific implications related to the tools that students access to answer the questions and the perceptual differences by human scorers across modes.

Many of the studies with mixed results for score differences found that scores varied by mode for only a subset of the subject areas and/or grades tested, but there were few clear patterns among the results. For most subject areas, there were no consistent findings in terms of whether the computer-based version or paper-based version was more difficult or whether they were equivalent. For mode comparisons of mathematics tests, the majority of studies found either that the computer-based version was associated with significantly lower scores than the paper-based version, or that there was no significant difference across modes. Only one study (Kingsbury, 2002) found significantly higher mathematics scores for the computer-based condition than the paper-based condition, after controlling for students' initial performance, and the difference was small (about one point).

Gu, Drake, and Wolfe (2006) found that mathematics items that involved equalities/inequalities and variables were most likely to exhibit DIF by being more difficult on paper than on a computer as compared to other item types. Johnson and Green (2006) found that participants' scores were significantly lower on computer-based items that required scratch paper than on paper-based versions of the same items. Similarly, another study found that scores on items involving graphic and geometric manipulation were negatively affected by computer-based administration (Keng, McClarty, & Davis, 2008).

Other assessment characteristics that were found to affect comparability include item format and whether the computer-based test was linear (i.e., fixed form) or adaptive. Russell and Haney (1997) found no significant score differences by mode for multiple-choice items but significantly higher scores for the computer-based version of performance writing tasks and short-constructed response items compared with the paper-based version. In a meta-analysis, Kim (1999) found that computer-adaptive tests were associated with significantly lower scores than paper-based tests, while computerized tests that were not adaptive were associated with significantly higher scores than paper-based tests. In a separate meta-analysis, Wang, Jiao, Young, Brooks, and Olson (2008) found that effect sizes between computer-based tests and paper-based tests were significantly larger when the computerized version was adaptive than when it was linear.

In addition to comparability issues related to the assessment content, the mode of administration has been shown to affect perceptions of human scorers. Systematic differences in how paper-based and computerized assessments are scored also can lead to differences in student performance across the two modes. Several studies have examined the effect of composition mode on scores for written essays and constructed-response items. In general, these studies found that human scorers, on average, assigned higher scores to handwritten papers compared with typed essays, although typed essays were longer, on average, and students generally preferred to work on computers. Researchers speculated that this difference was due to scorers being more lenient and forgiving of smaller errors when reading the handwritten essays (Russell & Tao, 2004; Way & Fitzpatrick, 2006). However, Russell and Tao (2004) found that this mode effect in scoring could be eliminated when scorers were made aware of the effect and given proper training. Further, Russell and Plati (2000a) conducted a study in which handwritten essays were later typed and provided to scorers blind to the original mode of composition. This study found that when scorers were blind to composition mode, essays originally written on computer were significantly longer and received significantly higher scores. Although the NAEP writing assessment transitioned to computer administration when it moved to a new framework in 2011, there are important implications for constructed-response items in other subject areas. Mode effects related to scoring will be particularly important for NAEP to examine given the large proportion of constructed-response items on NAEP assessments.

Mode Effects and Demographic Characteristics

Several studies also found that score differences between computer-based and paper-based tests varied by demographic characteristics, including gender, race, socioeconomic status, student ability, urbanicity, and SD/ELL (students with disabilities/English language learners) status.

Several studies have investigated mode effects by gender. A study by Gallagher, Bridgeman, and Calahan (2002) found that female performance was negatively affected by computers as the mode of test administration. In particular, the often-observed discrepancy between male and female performance on mathematics items grew significantly larger under the computer-administration condition. A similar effect was found in a study by Horne (2007) of a language arts and spelling test on which females performed significantly better than their male counterparts on a paper-based version; this score difference was eliminated in the computer-based version of the same assessment. However, several other studies (Bridgeman & Cooper, 1998; Clariana & Wallace, 2002; Fritts & Marszalek, 2010; Horkay, Bennett, Allen, Kaplan, & Yan, 2006; MacCann, 2006: Randall, Sireci, Li, & Kaira, 2012) found no consistent mode effects as a function of gender.

Results from surveys in a Hong Kong-based study showed that, when given the choice, male participants preferred to take their tests using computers, while females tended to opt for paper-and-pencil administered assessments (Coniam, 2006). Fritts and Marszalek (2010) found no significant difference by gender on measures of test anxiety, regardless of whether the test was taken by paper-and-pencil or computer administration.

It is not clear whether differential mode effects by gender indicate a disadvantage for females taking tests on computers, or whether the computer mode increases motivation and engagement for males, thus eliminating some of the construct-irrelevant variance in paper-based tests.

Several studies have investigated whether mode effects are more pronounced for students with low socioeconomic status (SES). Pomplun and Custer (2005) and

Pomplun, Ritchie, and Custer (2006) found that students eligible for free or reduced-price lunch had greater gaps between scores from paper-based and computer-based versions of an elementary reading assessment. On a computing skills test of high school students in Australia, MacCann (2006) found that although there were no score differences by mode for high SES students, low SES students performed significantly better on the paper-based version than the computer-based version of the test. Although not a study of SES directly, Horkay, Bennett, Allen, Kaplan, and Yan (2006) found a significant interaction between mode and school location, a variable often correlated with SES. Students from urban fringe/large town locations performed significantly better on the paper-based version than the computer-based version of a writing test On a state 10th- grade science assessment, Randall, Sireci, Li, and Kaira (2012) found no consistent mode effect between students who were eligible for free or reduced-price lunch and those who were not eligible.

Another population that has been the focus of several mode effect investigations is SD/ELL students. Despite the use of universal design elements that incorporated some accommodations into the general assessment, several studies found that SD/ELL students performed significantly better on paper-based versions than computer-based versions of language arts (Russell & Plati, 2000b) and reading and mathematics (Taberbhai, Seo, & Bowman, 2012) tests. Wolfe and Manalo examined TOEFL Writing results from nearly 134,000 English language learners and found that participants with lower English language ability scored higher on the paper-based version, and students with higher English language ability scored higher on the computer-based version. Bridgeman and Cooper (1998) found no significant interactions between mode and ELL status for the GMAT. Conversely, Dolan et al. (2005) used a small sample of 10 SDs and found no significant mode effect overall; however, scores were significantly higher in the computer-based version as compared with the paper-based version for items with reading passages that were more than 100 words. Finally, Kim and Huynh (2010) performed differential bundle functioning analyses on a statewide, end-of-course English assessment and found that "Researching items" significantly favored the paper mode for students without disabilities and "Building Vocabulary items" significantly favored the computer mode for SDs. Although there is not a clear pattern in these results, what stands out is the complexity of how mode of testing administration can interact with both SD/ELL status and other factors. It is not clear whether SD/ELL students generally have less experience with computers, which could also account for performance differences.

Mode Effects and Computer Familiarity

Perhaps the most important student characteristic to consider when examining the mode effect of computer-based assessment administration is computer familiarity. Although many studies have examined the impact of computer familiarity on mode effects in assessment, the relationship between computer experience and performance on computer-based assessments remains unclear. Several studies show that higher levels of computer familiarity correlate with higher scores on computer-based assessments (Bennett, Braswell, Oranje, Sandene, Kaplan, &Yan, 2008; Bridgeman & Cooper, 1998; Chen, White, McCloskey, Soroui, & Chun, 2011; Horkay, Bennett, Allen, Kaplan, & Yan, 2006); one found mixed results (Goldberg &

Pedulla, 2002); others have found no significant effect on mode by computer familiarity (Clariana & Wallace, 2002; Higgins, Russell, & Hoffman, 2005).

One possible reason for this inconsistency in results is that computer familiarity is still a vaguely defined construct that has yet to be operationalized consistently across studies. Because there is no standard measure of computer familiarity, the construct being measured as "computer familiarity" is not necessarily consistent between studies. For example, Horkay et al. (2006) developed their own, study-specific survey to determine participants' levels of computer familiarity. Clariana and Wallace (2002) measured computer familiarity using four previously developed questions from the Distance Learning Profile (Clariana & Moller, 2000). Higgins, Russell, and Hoffmann (2005) broke the construct into three parts: computer fluidity and computer literacy, for both of which they created their own metric; and frequency of computer use, for which they used a survey adapted from a fifth-grade USEIT (Use, Support, and Evaluation of Instruction Technology) study survey, developed by Russell, Bebell, and O'Dwyer (2003).

One specific aspect of computer familiarity is keyboarding skills. Mode effects on students with low keyboarding skill levels have been of particular concern recently, as NAEP pilots its new Writing Computer Based Assessment (WCBA) at the fourth-grade level. Studies by Russell (1999) and Russell and Plati (2000a and 2002) have found that keyboarding skills significantly affect student performance on writing tasks, but apparently only at the lower skill levels; there appears to be a skill level "threshold", above which keyboarding skills seem to have no significant effect.

A similar effect to the "threshold" described in Russell's (1999) investigation of keyboarding skills is observed in other equivalency studies investigating computer experience and familiarity. It is possible that computer familiarity is much more predictive of a mode effect for certain subpopulations and may account for some of the differential mode effects observed for certain subgroups. However, the majority of studies addressing computer familiarity were not performed within the past five years, and it is likely that computer familiarity has greatly increased during this time.

The results from studies examining computer familiarity highlight the confounding role of demographics, making it particularly difficult to isolate and confirm the myriad factors involved in mode effects in computer-administered assessments. Although the extent to which familiarity with computers affects performance on computer-based assessments is still unclear, there is enough evidence to suggest that familiarity should be taken into account when moving to computer-based assessments, and steps should be taken to mitigate these effects as much as possible.

Conclusion

Many studies examining mode effects in educational testing have shown inconsistent or mixed effects. The research is clear in demonstrating that comparability of results *can* often be maintained overall as a test makes the transition from paper-and-pencil to computerized administration. For example, most of the studies suggest that the structure of the test is likely to remain unchanged in moving from paper-and-pencil to computer-based administration. However, the evidence is mixed on the effects of mode on score comparability; computerization may have an effect on the results for some subgroups of the population and these can vary further as a function of the subject area being assessed. Schroeders and Wilhelm (2011) perhaps best summarize what is required when moving to computerized assessment when they write "… equivalence research is required for specific instantiation unless generalizable knowledge about factors affecting equivalence is available" (pg. 1).

This sentiment should also help guide and inform the move to computer-based assessment in NAEP. The computerization of an assessment should be treated as any other change one might make in NAEP: comparability of scores can be hoped for, but cannot be taken for granted. Research, including the use of bridge studies, is needed to evaluate the effects of moving assessments from paper-and-pencil to computer administration.

References for Appendix B

Anakwe, B. (2008). Comparison of student performance in paper-based versus computer-based testing. *Journal of Education for Business, 84*(1), 13–17.

Balizet, S., Treder, D., & Parshall, C. (1999, April). *The development of an audio computer-based classroom test of ESL listening skills.* Paper presented at the annual meeting of the American Educational Research Association, Quebec, Canada.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on a computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6, 9.*

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research, 31*(1), 51–60.

Bridgeman, B., & Cooper, P. (1998, April). *Comparability of scores on word-processed and handwritten essays on the Graduate Management Admissions Test.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing, 16*(1), 49–71.

Choi, I. C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20,* 295–320.

Clariana, R. B., & Moller, L. (2000). *Distance learning profile instrument: predicting on-line course achievement.* Paper presented at the annual convention of the Association for Educational Communications and Technology, Denver, CO.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology, 33*(5), 593–602.

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening test. *ReCALL, 18*(2), 193–211.

Coniam, D. (2009). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation, 15*(3), 243–263.

Dolan, R. P., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning and Assessment, 3*, 7.

Escudier, M. P., Newton, T. J., Cox, M. J., Reynolds, P. A., & Odell, E. W. (2011). University students' attainment and perceptions of computer delivered assessment: A comparison between computer-based and traditional tests in a "high-stakes" examination. *Journal of Computer Assisted Learning, 27*(5), 440–447.

Fritts, B. E., & Marszalek, J. M. (2010). Computerized adaptive testing, anxiety levels, and gender differences. *Journal Psychology of Education, 13*(3), 441–458.

Fulcher, G. (1999). Computerizing an English language placement test. *ELT Journal, 53*(4), 289–299.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. *Journal of Educational Measurement, 39*(2), 133–147.

Goldberg, A. L., & Pedulla, J. (2002). Performance differences according to test mode and computer familiarity on a practice Graduate Record Exam. *Journal of Educational and Psychological Measurement, 62*(6), 1053–1067.

Gu, L., Drake, S., & Wolfe, E. W. (2006). Differential item functioning of GRE mathematics items across computerized and paper-and-pencil testing media. *Journal of Technology, Learning, and Assessment, 5*(4), 1–25.

Higgins, J., Patterson, M. B., Bozman, M., & Katz, M. (2010). Examining the feasibility and effect of transitioning GED tests to computer. *Journal of Technology, Learning, and Assessment, 10*, 2.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning, and Assessment, 3,* 4.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117–144.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on a computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 5,* 2.

Horne, J. (2007). Gender differences in computerised and conventional educational tests. *Journal of Computer Assisted Learning, 23*(1), 47–55.

Johnson, M., & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and Assessment, 4,* 5.

Karkee, T., Kim, D. I., & Fatica, K. (2010). *Comparability study of online and paper and pencil tests using modified internally and externally matched criteria.* Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas Assessment of Knowledge and Skills. *Applied Measurement in Education, 21*(3), 207–226.

Kim, D. H., & Huynh, H. (2007). Comparability of computer and paper-and-pencil versions of algebra and biology assessments. *Journal of Technology, Learning, and Assessment, 6,* 4.

Kim, D. H., & Hyunh, H. (2008). Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course English test. *Educational and Psychological Measurement, 68*(4), 554–570.

Kim, D. H, & Huynh, H. (2010). Equivalence of paper-and-pencil and online administration modes of the statewide English test for students with and without disabilities. *Educational Assessment, 15*(2), 107–121.

Kim, J. P. (1999). *Meta-analysis of equivalence of computerized and P&P tests on ability measures.* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chicago, IL.

Kingsbury, G. G. (2002). *An empirical comparison of achievement level estimates from adaptive tests and paper-and-pencil tests.* Presented during the annual proceedings of the American Educational Research Association, New Orleans, LA.

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K–12 populations: A synthesis. *Applied Measurement in Education, 2*, 22–37.

Lee, K. S., Osborne, R. E., & Carpenter, D. N. (2010). Testing accommodations for university students with AD/HD: Computerized vs. paper-pencil/regular vs. extended time. *Journal of Educational Computing Research, 42*(4), 443–458.

Liao, C. H., & Kuo, B. C. (2011). A web-based assessment for phonological awareness, rapid automized naming (RAN) and learning to read Chinese. *Turkish Online Journal of Educational Technology—TOJET, 10*(2), 31–42.

Lottridge, S. M., Nicewander, W. A., & Mitzel, H. C. (2011). A comparison of paper and online tests using a within-subjects design and propensity score matching study. *Multivariate Behavioral Research, 46*(3), 544–566.

MacCann, R. (2006). The equivalence of online and traditional testing for different subpopulations and item types. *British Journal of Educational Technology, 37*(1), 79–91.

Mason, B. J., Patry, M., & Bernstein, D. J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. *Journal of Educational Computing Research, 24*(1), 29–39.

Minnick, J. R. (2009). *Motherboards cutting razor wire: Assessment and incarcerated youth.* Doctoral dissertation, Capella University.

Mogey, N., Paterson, J., Burk, J., & Purcell, M. (2010). Typing compared with handwriting for essay examinations at university: Letting the students choose. *ALT-J: Research in Learning Technology, 18*(1), 29–47.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3,* 6.

Pommerich, M. (2002, April). *The effect of administration mode on test performance and score precision, and some factors contributing to mode differences.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2,* 6.

Pommerich, M., & Burden, T. (2000, April). *From simulation to application: Examinees react to computerized testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Pomplun, M. (2007). A bifactor analysis for a mode-of-administration effect. *Applied Measurement in Education, 20*(2), 137–152.

Pomplun, M., & Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K–3 reading tests. *Journal of Educational Computing Research, 32*(2), 153–166.

Pomplun, M., Frey, S., & Becker, D. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*(2), 337–354.

Pomplun, M., Ritchie, T., & Custer, M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment, 11*(2), 127–143.

Randall, J., Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the comparability of paper- and computer-based science tests across sex and SES subgroups. *Educational Measurement: Issues and Practice, 31*(4), 2–12.

Puhan, G., Boughton, K., & Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing. *Journal of Technology, Learning and Assessment, 6*, 3.

Rowan, B. E. (2010). *Comparability of paper-and-pencil and computer-based cognitive and non-cognitive measures in a low-stakes testing environment.* Doctoral dissertation, James Madison University.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*, 20.

Russell, M., Bebell, D., & O'Dwyer, L. (2003). *Use, support, and effect of instructional technology study: An overview of the USEIT study and the participating districts.* Boston: Technology and Assessment Study Collaborative.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing students' performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5*(3), 1–20.

Russell, M., & Plati, T. (2000a). Effects of computer versus paper administrations of a state-mandated writing assessment. *The Teachers College Record: Technology and Assessment Study Collaborative, 103*, 1–34.

Russell, M., & Plati, T. (2000b). *Mode of administration effects on MCAS composition performance for grades four, eight, and ten.* Chestnut Hill, MA: National Board on Educational Testing and Public Policy.

Russell, M., & Plati, T. (2002). Does it matter with what I write? Comparing performance on paper, computer, and portable writing devices. *Current Issues in Education, 5*, 4.

Russell, M., & Tao, W. (2004). The influence of computer-print on rater scores. *Practical Assessment, Research & Evaluation [Online], 9*, 10.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71*(5), 849–869.

Schwarz, R. D., Rich, C., & Podrabsky, T. (2003, April). *A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Staples, J. G., & Luzzo, D. A. (1999). Measurement comparability of paper-and-pencil and multimedia vocational assessments. *ACT Research Report Series 99-1.* Iowa City, IA: American College Testing Program.

Taherbhai, H., Seo, D., & Bowman, T. (2012). Comparison of paper-pencil and online performances of students with learning disabilities. *British Educational Research Journal, 38*(1), 61–74.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K–12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219–238.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5–24.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). *Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Way, W. D., & Fitzpatrick, S. (2006). *Essay responses in online and paper administrations of the Texas Assessment of Knowledge and Skills (*Research Report 06-04). Pearson Educational Measurement.

Whiting, H., & Kline, T. J. B. (2006). Assessment of the equivalence of conventional versus computer administration of the Test of Workplace Essential Skills. *International Journal of Training and Development, 10*(4), 285–290.

Wolfe, E. W., & Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning & Technology, 8*(1), 53–65.

Zandvliet, D., & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education, 29*(4), 423–445.

Table B1. Fourteen studies that investigated measurement equivalence between modes of assessment, and failed to find a lack of equivalence.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Bennett, Braswell, Oranje, Sandene, Kaplan, Yan	2008	National Assessment of Educational Progress (NAEP) PPT, 2001 Mathematics Online (MOL)	Independent *t*-test; item response theory analyses	1,970 Grade 8 students (nationally representative)	Computer facility predicted MOL performance (controlled for performance on paper-based test). Eighth-grade performance was significantly lower for those taking the computerized test, with an effect size of 0.15. At the item level, the difficulties for the computer test were generally greater and item discrimination differences estimates suggested minimal effects.
Choi, Kim, Boo	2003	Test of English Proficiency by Seoul National University (TEPS); listening comprehension, grammar, vocabulary, reading comprehension	Correlational analyses; analysis of variance (ANOVA); confirmatory factor analyses	971 university students in Korea	Statistically significant score differences were found among the listening comprehension, vocabulary, and reading comprehension subtests, but not for the grammar subtest. The factor structure for the four subtests was consistent across test administration modes. Correlations of subtests, disattenuated correlations, and confirmatory factor analyses support that the computer-based and paper-based subtests measure the same constructs.
Karkee, Kim, Fatica	2010	End-of-instruction social studies assessment	Item response theory and differential item functioning (DIF) analyses	50,000 participants	No statistically significant mode effect was found based on model fit, DIF, or student performance.
Kim, Huynh	2007	End-of-course assessments in algebra and biology	Counter-balanced repeated-measures ANOVA; item response theory analyses and comparison of information functions; confirmatory factor analyses	Students from 15 middle and high schools in a southeastern state (788 algebra students and 406 biology students); Black and Hispanic students were underrepresented.	No evidence was found to suggest that mode changed the constructs measured. Results suggest the comparability of computer-based and paper-based assessments at the item-, subtest- and whole test-levels. For algebra, scores were significantly higher for the paper-based assessment than the computer-based assessment, with an effect size of 0.17. For biology, there were no significant score differences by mode.

Table B1 (continued). Fourteen studies that investigated measurement equivalence between modes of assessment, and failed to find a lack of equivalence.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Kim, Huynh	2008	NC end-of-course English test	Two-way repeated measures ANOVA; item response theory analyses; confirmatory factor analyses	439 middle- and high-school students; Black students were under-represented.	Students scored significantly higher on the paper-based assessment than the computer-based assessment overall with a small effect size. Results from the confirmatory factor analyses suggest the mode does not alter the test constructs. Analysis at the content domain level indicates that students perform worse in reading comprehension in a computer mode; however, there were no differences by mode in the other content domains.
Lottridge, Nicewander, Mitzel	2011	End-of-course algebra and English Assessments	Comparison of within-subjects design and propensity score matching; confirmatory factor analyses	3,628 students in Grades 8, 9	The study showed that the online and paper tests appeared to be measuring the same underlying constructs with the same level of reliability. The computer mode was slightly more difficult than the paper mode, but it is not clear whether the difference was statistically significant.
Pomplun	2007	Initial-Skills Analysis (part of the Basic Early Assessment of Reading)	Single-group counterbalanced design; bifactor model to test equivalence of paper-based and computer-based formats	About 2,000 students in K–3 across 12 states	Mean scores were significantly higher for the paper-based assessment compared with the computer-based assessment for all grades, with effect sizes ranging from .27 to .48. At each grade level, the model with the method factors included led to significant improvement in fit. There were some minor differences in the item factor loadings across formats. The authors concluded that score equivalence was found between the two modes but that the increased difficulty of the computerized version would require test equating to use results from the two modes interchangeably.

Table B1 (continued). Fourteen studies that investigated measurement equivalence between modes of assessment, and failed to find a lack of equivalence.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Pomplun, Custer	2005	Initial-Skills Analysis (part of the Basic Early Assessment of Reading)	Single-group counterbalanced design; dependent *t*-tests, confirmatory factor analyses	About 2,000 students in K–3 across 12 states	Mean scores were significantly higher for the paper-based assessment compared with the computer-based assessment for all grades, with effect sizes ranging from .27 to .48. At three out of four grades, the test variance was significantly different across modes. Free/reduced-price lunch students had greater gaps between paper-based assessment and computer-based assessment scores, though it is not clear whether the differences are statistically significant. Confirmatory factor analyses found equivalence between the modes.
Pomplun, Frey, Becker	2002	Nelson-Denny reading test	Counter-balanced design; dependent *t*-tests; coefficient alpha; linear and equipercentile equating; predictive validity with grades	215 college students	Computer-based assessment generally produced higher scores compared with the paper-based assessment, though not all score differences were significant. The variance of the two forms was equivalent. The predictive validity of scores was comparable between the two modes.
Randall, Sireci, Li, Kaira	2012	State science assessment	Confirmatory factor analyses; Rasch item response theory DIF analyses	1,439 students (computer condition) and 10 random samples of 1,439 students drawn without replacement from 95,422 students (paper condition) in Grade 10	Confirmatory factor analyses found partial measurement invariance by mode, sex, and socioeconomic status (SES). DIF analyses found a few items with possible DIF. There were no consistent differences across modes, either overall or by sex or SES.

Table B1 (continued). Fourteen studies that investigated measurement equivalence between modes of assessment, and failed to find a lack of equivalence.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Rowan	2010	Archival data from mandatory university assessments: Natural World, Ver. 9 (NW9): cognitive scientific knowledge and reasoning, computer-based and paper-based versions; Attitude Toward Learning (ATL): noncognitive, computer-based and paper-based versions.	Confirmatory factor analysis; Mantel-Haenszel DIF analyses	About 4,000 college students	The paper-based assessment and computerized versions of the test were found to be tau-equivalent. Mean differences between test administration modes were found to exist with higher scores on the paper version than the computer version, with an effect size of .26. The author noted that scores would need to be rescaled to be equivalent across the two modes. Three items exhibited C-level DIF across modes.
Schroeders, Wilhelm	2011	English Reading and Listening Comprehension (dichotomous items): English as a second language	Multigroup confirmatory factor analysis	442 German high school students, Grades 9, 10, English language learners (high ability)	Scores were measurement invariant across modes for both reading comprehension and listening comprehension.
Staples, Luzzo	1999	Unisex Edition of the American College Testing Inventory (UNIACT), Inventory of Work-Relevant Abilities (IWRA)	Scale correlations; coefficient alpha; exploratory factor analyses	1,022 students, Grades 9, 11	Factor loadings and internal consistency appeared similar across modes. There were no differences in mean scores by mode.
Taherbhai, Seo, Bowman	2012	Modified Maryland School Assessment (mod-MSA) in reading and mathematics	Analysis of covariance (ANCOVA); DIF	About 5,500 students with disabilities in Grades 7, 8	Students with disabilities who took the paper-based assessment performed significantly higher than the students with disabilities who took the computer-based assessment in reading and mathematics across grades, with effect sizes ranging from 0.06 to 0.12. No C-level DIF items were found.

Table B2. Seven studies that investigated measurement equivalence between modes of assessment, and found some lack of equivalence.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Gu, Drake, Wolfe	2006	60 quantitative (mathematics) items, similar to GRE, Original items created using POWERPREP (ETS 1999)	*t*-test; differential item functioning analyses	165 first-year graduate students; high computer familiarity	No significant score differences were found between paper-based assessment and computer-based assessment groups; 38% of items were flagged for cross-medium DIF. Of the assessment characteristics examined, mathematical notation and content appeared to contribute most significantly to DIF across media.
Keng, McClarty, Davis	2008	Texas Assessment of Knowledge and Skills	*t*-tests; DIF analyses	Grades 8 and 11: 2,546 for mathematics; 3,680 for reading; 2,898 for social studies; statewide	Several items showed evidence of DIF. The paper-based assessment group significantly outperformed the computer-based assessment group on selected mathematics (e.g., Spatial Relationships and Geometric Relationships) and reading objectives (e.g., Basic Understanding, Applying Critical Thinking Skills) at Grades 8 and 11. No significant differences were found for social studies or science at Grade 11.
Kim, Huynh	2010	Statewide End-of-Course English Assessment	*t*-tests; confirmatory factor analyses; differential item/bundle functioning analyses; quasi-experimental design using propensity score matching	~15,000 participants, (~1,000 SD), Grade 9	There were some significant interactions between disability status and mode for some of the content areas, though the effect sizes were very small (less than 0.1). The confirmatory factor analyses found measurement equivalence by mode at the weak, strong, and strict levels. The DIF analyses found no items with C-level DIF. The differential bundle functioning analyses did find a significant result favoring the paper-based mode for Researching items for students without disabilities and Reading III —Building vocabulary items for students with disabilities.

Table B2 (continued). Seven studies that investigated measurement equivalence between modes of assessment, and found some lack of equivalence.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Poggio, Glasnapp, Yang, Poggio	2005	Kansas Computerized Assessment (large-scale state test) and parallel paper-based version	Descriptive statistics; hierarchical linear modeling; item response theory analyses	2,861 students in 7th grade	No meaningful statistically significant difference was found in performance between computer-based assessment and paper-based assessment scores (less than 1 percentage point); 9 of the 204 items were flagged as having mode effects, but no common factors were identified to account for this.
Puhan, Boughton, Kim	2007	Praxis—reading, writing, and mathematics	Cohen's *d*; DIF analyses	About 7,000 participants entering teaching programs	Based on Cohen's *d*, results indicated no substantial difference between computer-based and paper-based scores. DIF analyses revealed all reading and mathematics items were comparable for both versions. DIF analyses indicated item-level differences exist across the paper-based and computer-based versions of the writing test, with the three items favoring examinees who took the paper-based version.
Schwarz, Rich, Podrabsky	2003	InView (norm-referenced aptitude test); Test of Adult Basic Education (TABE) (norm-referenced)	DIF)	1. Grades 4–9; 2. Adults in large-scale, matched samples	Several items in each assessment did exhibit cross-medium DIF. On the TABE, differences by mode were largest at the lower end of the ability distribution.
Way, Davis, Fitzpatrick	2006	Texas Assessment of Knowledge and Skills (TAKS)—Mathematics, Reading, Science and Social Studies	Random-groups equating; matched-samples comparability analysis	Students in Grades 8, 11	Mixed results across subjects, with the largest difference for TAKS 8th-grade reading.

Table B3. Eight studies that evaluated effects of assessment characteristics, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Johnson, Green	2006	Selected mathematics items from UK's Mathematics National Curriculum	ANOVA	104 students ages 10–11	No statistically significant differences between overall performance on paper and computer.
Kim	1999	Meta-analysis, various subjects, mostly mathematics and reading	Various	Age range: Grade 3–adult, (about 50% university students)	The type of computer-based assessment was the most important variable when evaluating the equivalence between computer-based and paper-based tests. For adaptive tests, mathematics, source, and sampling age were significant variables. For nonadaptive computer-based tests, the analysis did not find significant moderators. Computer-based testing was significantly more advantageous for the high school-aged population.
Kingsbury	2002	ALT and Measure of Academic Progress (MAP) state tests in Idaho—reading, mathematics, language use	ANCOVA	8,560 students in 4th and 5th grades	Language usage and mathematics scores were significantly higher for computer-based tests than paper-based tests after controlling for initial performance (by about 1 point); there was no significant difference for reading scores.
Russell, Haney	1997	NAEP items (multiple-choice and short constructed-response language arts, mathematics, science, and reading items); unspecified open-ended writing items	Independent *t*-tests	114 students in Grades 6–8	No difference in multiple-choice test results by mode of administration. For the performance writing tasks, scores were significantly higher for computer-based tests than paper-based tests, with an effect size of .94. When scores of open-ended items were used as a covariate, there was a significant mode effect for short constructed-response items in science and language arts.
Russell, Plati	2000a	Massachusetts Comprehensive Assessment System (MCAS) Language Arts	Independent *t*-tests; Welch's *t*-tests	Students in Grades 8 (144) and 10 (145)	Scores were significantly higher for computer-based tests than paper-based tests at both Grades 8 and 10, regardless of keyboarding skills.

Table B3 (continued). Eight studies that evaluated effects of assessment characteristics, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Russell, Tao	2004	MCAS Composition items	ANOVA	Grade 8, 60 responses	Composition scores produced on computer (typed) received significantly lower scores than on paper (handwritten). Study found that upon training scorers using both modes, especially noting problems with mode effect, the presentation effect was eliminated.
Wang, Jiao, Young, Brooks, Olson	2008	Various mathematics assessments	Meta-analysis of mean score differences by mode (11 studies with 42 independent effects)	K–12	Meta-analysis found that overall there was no difference between scores from paper-based testing and computer-based testing. Effect sizes across the studies did vary, however, as a function of study design, sample size, computer practice, and computer delivery algorithm.
Way, Fitzpatrick	2006	Texas Assessment of Knowledge and Skills—Writing	Rater agreement; logistic regression; ANCOVA	1,340 Grade 11 lower performing students	Computer-based essays were scored more stringently than those completed on paper (handwritten). There was a positive relationship between essay score and the use of computers for language arts classes in the school. The paper-based test had higher interrater reliability of essay scoring than the computer-based test.

Table B4. Eleven studies that evaluated effects of demographic characteristics, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Bridgeman, Cooper	1998	Graduate Management Admissions Test (GMAT), two 30-minute essay items	Within-subjects; ANOVA	3,470	Significantly higher paper-based test scores compared with computer-based test scores for people with relatively low word-processing experience. No significant differences between paper-based test scores and computer-based test scores based on gender, race/ethnicity, or ELL status. Mode effect was smallest for participants with the most computer experience. Found higher interrater reliability for word-processed essays. Found no interaction of score differences by gender, race/ethnicity, or ELL status.
Clariana, Wallace	2002	100-item teacher-made multiple-choice course tests for introductory university class on computer fundamentals; Distance Learning Profile (Clarianna & Moller, 2000)	ANOVA (posttest only)	105 freshman university students	Overall, the computer-based testing group scored significantly higher than the paper-based testing group. Gender, competitiveness, and computer familiarity were not significantly related to performance difference between modes. There was a significant interaction between the administration mode and content familiarity. Low-attaining students had similar performance in both modes, while high-attaining students performed better on the computer-based test than the paper-based test.
Coniam	2006	English Language Listening Test	Posttest survey on preferences	Grade 11, 12 students in Hong Kong	Significantly higher scores for Grade 11 computer-based assessment than paper-based assessment; no significant score differences for Grade 12. Survey found males preferred computer-based tests and females preferred paper-based tests.

Table B4 (continued). Eleven studies that evaluated effects of demographic characteristics, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Dolan, Hall, Banerjee, Chun, Strangman	2005	Released items from NAEP U.S. history and civics	Matched-samples *t* -tests	10 students with specific learning disabilities from Grades 11 and 12	There were no significant differences overall between scores in the two modes. Scores were significantly higher in the computer-based test condition for items with reading passages more than 100 words. Usability interviews indicated that participants preferred the computer-based test.
Fritts, Marszalek	2010	Measure of Academic Progress assessment (MAP), ALT	Regression analyses; *t* – tests	132 students (mean age: 13.36)	There was no difference between the two groups in the standardized mathematics score or standardized reading score. The computer-based test was found to produce less test anxiety than the linear paper-based test. No significant mode effect was found by gender.
Gallagher, Bridgeman, Calahan	2002	Graduate Record Examination (GRE), SAT I, GMAT, Praxis	Standardized mean differences; repeated-measures ANOVA; *t* – tests	Several hundred thousand high school and college students	Mode effects varied by gender, race/ethnicity, and gender by race/ethnicity interactions across the different tests.
Horkay, Bennett, Allen, Kaplan, Yan	2006	Main NAEP— Writing	Repeated-measures ANOVA	1,313 8th-grade students, nationally representative	No significant mean score differences between paper-based and computer-based modes. Computer familiarity significantly related to online writing test performance after controlling for paper writing skill. Subpopulation analysis indicated a significant interaction effect of delivery mode with school location (specifically, students from urban/large town locations performed significantly higher on paper as compared with computer).

Table B4 (continued). Eleven studies that evaluated effects of demographic characteristics, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Horne	2007	Lucid Assessment System for Schools (LASS) Secondary and LASS Junior (Language Arts, Spelling)	*t*-tests	242 students, ages 9–15	In the paper-based test, females scored significantly higher than males on the reading and spelling tests. In the computer-based test, there was no significant difference by gender.
MacCann	2006	Computing skills test	Regression analyses; repeated-measures ANOVA	14,248 volunteer students ages 15–16 (New South Wales, Australia)	There was no significant interaction between gender and mode of administration. There was a significant score difference by mode found for socioeconomic status (SES), where low-SES students performed better on the paper-based mode than the computer-based mode. There was no significant interaction between item format and mode of administration.
Pomplun, Ritchie, Custer	2006	Initial-Skills Analysis (part of the Basic Early Assessment of Reading)	Single-group counterbalanced design; omit rates by mode; regression analyses	2,000 students in Grades K–3, (23% free/reduced-price lunch eligible, 78% white)	Mean scores were significantly higher for the paper-based test compared with the computer-based test for all grades, with effect sizes ranging from .27 to .48. More items were omitted in the paper form than the computer form, though the difference was significant for only two of the four grades. Deferring, omitting items, and free/reduced-price lunch status were significant predictors of computer-based test scores after controlling for paper-based test scores.
Russell, Plati	2000b	MCAS Language Arts	Independent *t*-tests; regression analyses	Students in Grades 4 (152), 8 (228), 10 (145)	Scores were significantly higher for computer-based test scores than paper-based test scores. At Grades 8 and 10, special education students had significantly higher midterm grades when performing composition items on paper. There was no significant difference for special education students in Grade 4 by mode.

Table B5. Five studies that evaluated effects of computer familiarity, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Chen, White, McCloskey, Soroui, Chun	2011	Functional Writing, items from 2008 National Assessment of Adult Literacy (NAAL)	Between-subjects; within-subjects; ANOVA and repeated *t*-tests	1,607 subjects, ages 16+	Scoring bias analysis: When handwritten essays were transcribed, there were no statistically or practically significant scoring differences between handwritten and transcribed computer responses to the three writing tasks. Regarding the effects of administration mode, the analyses showed a consistent advantage for the paper mode over computer mode for the overall tasks scores and individual scoring criteria. For the length of writing, there was no significant difference. Some significant effects were found in individual tasks by race/ethnicity, age, education, word-processor experiences, and employment status. None of these showed consistent effects across all three tasks.
Goldberg, Pedulla	2002	Practice GRE	Multivariate analysis of covariance (MANCOVA)	222 3rd- and 4th-year university students (28% male)	Positive main effect of computer familiarity on Analytical and Quantitative subtests (not on Verbal). Performance differences were statistically significant among test modes on each of the subtests: Analytical Verbal and Quantitative. There was a statistically significant interaction effect between test mode and computer familiarity on the Quantitative subtest performance.

Table B5 (continued). Five studies that evaluated effects of computer familiarity, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Higgins, Russell, Hoffmann	2005	Writing items from NAEP, Progress in International Reading Literacy Study (PIRLS), and New Hampshire State Assessment	Computer fluidity test, computer literacy test, computer use survey	219 participants, 4th grade	No differences in reading comprehension across testing modes (paper-based test, computer-based test with scrolling, computer-based test whole page); No statistically significant differences in reading comprehension based on computer fluidity (use of mouse and keyboard) and computer literacy; Computer anxiety levels did not significantly affect scores.
Russell	1999	MCAS, NAEP open-ended items in Language Arts, Science, and Mathematics	Independent *t*-tests; multiple regression	229 middle school students	The study found that computer-based testing led to higher scores in Science and lower scores in Mathematics subtests. In the English and Language Arts subtest, there was no overall effect, but there was a significant effect found by keyboarding skills.
Russell, Plati	2002	Writing items from MCAS	Independent *t*-tests; regression analyses	Grades 4, 8	Keyboarding skills were positively correlated with test scores in 4th grade; however, there appears to be a threshold above which keyboarding skills have no significant effect.

Table B6. Ten other studies that found score differences between computer-based and paper-and-pencil administration, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Escudier, Newton, Cox, Reynolds, Odell	2011	Undergraduate dental school course assessments; attitude survey	Repeated-measures ANOVA; focus-group discussions	132 year 3 and 134 year 5 dental undergraduates	For year 3 students, there was a significant interaction between test order (whether the paper-based test or computer-based test was administered first) and performance. For year 5 students, computerized scores were higher than paper test scores regardless of the test order. The attitude survey revealed that participants felt the online test format did not disadvantage students, even in a high-stakes situation.
Fulcher	1999	English-Language Placement Test, 80 items: all multiple choice	Within-subjects; ANCOVA	57 university students	Computer-based test scores were higher than paper-based test scores. There is a possible practice/order effect because students took paper-based test first.
Kingston	2009	K–12 Assessments in Mathematics, Reading, English Language Arts, Social Studies, and Science	Meta-Analysis	K–12	The study found that computer-based assessment led to higher scores for English language arts and social studies, but lower scores for mathematics. No significant effect by grade level was found.

Table B6 (continued). Ten other studies that found score differences between computer-based and paper-and-pencil administration, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Liao, Kuo	2011	Four Assessments on Chinese Language Ability: One-Minute Word Reading; Onset Detection; Rhyme Detection; Rapid Automatized Naming (RAN) (e.g., reading fluency). Paper-based assessment: In-person read-aloud of audio tasks. Computer-based assessment: Computer-delivered audio.	Hierarchical multiple regression	93 students, Grade 6	Results showed that the two modes for RAN are highly correlated, but not for Rhyme detection and onset detection. The results showed that conventional and Web-based versions were equally predictive of Chinese reading measures.
Pommerich	2002	Fixed-form tests in English, Mathematics, Reading, Science Reasoning	Two different computer interfaces were used; *t*-tests	Large scale (about 20,000)	Levels of comparability were inconsistent. A variety of factors appeared to be related to mode effects. Changes to computer interface seemed to have significant effect on cross-mode differences.
Pommerich	2004	English, Reading, and Science Reasoning assessments	Two different computer interfaces were used; *t*-tests	12,000 students from 61 schools in Grades 11, 12	Results varied by computer interface condition and subject area.
Pommerich, Burden	2000	20-minute content area tests in English, Mathematics, Reading, Science	Within-subjects, nonrandom assignment; *t*-test	36 students, Grades 11, 12	Assessment factors that were found to be the most likely to lead to construct-irrelevant effects were pages and line length, layout features, highlighting, and item characteristics.
Wang, Jiao, Young, Brooks, Olson	2007	Various mathematics assessments	Meta-analysis of mean score differences by mode (14 studies with 44 independent effects)	K–12	Meta-analysis found that overall there were few small differences between modes, with effect sizes ranging from -.28 to .08. There was a significant difference in the effect size by delivery algorithm (linear versus adaptive computer-based assessments). The paper-based test had larger variance than the computer-based test. No differences were found by grade or computer practice.

Table B6 (continued). Ten other studies that found score differences between computer-based and paper-and-pencil administration, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Wolfe, Manalo	2004	TOEFL Writing	Generalized linear model (GLM)	133,906 English language learners ranging from 15 to 55	The paper-based test had higher essay scores than the computer-based test but mode explained only a small amount of variation (r^2=.01). Participants with lower English language ability scored slightly better on paper (interaction). Participants with higher English language ability scored slightly better on computer (interaction).

Table B7. Ten other studies that found no score differences between computer-based and paper-and-pencil administration, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Anakwe	2008	University Accounting course assessments (3 courses)	*t*-tests	54 university students	No statistically significant score differences across modes in any of the three courses.
Balizet, Treder, Parshall	1999	Study-specific tests of Academic Listening Comprehension and Vocabulary; PPT: Audio-cassette, Computer-based test: Computer-delivered audio	*t*-tests; descriptive statistics	28 high-intermediate level English as a second language students	No significant score difference between the two administration modes.
Bodmann, Robinson	2004	Undergraduate Educational Psychology Course Assessments	Dependent *t*-tests	113 undergraduate students in an educational psychology class	Computer-based assessments were completed faster than paper-based assessments with no significant differences in scores.
Coniam	2009	2007 Hong Kong Certificate of Education Examination (HKCEE) Year 11 English Language Writing Paper (Hong Kong Public Exam)	Scoring modes compared: “Onscreen Marking” and “Paper-Based Marking” scoring methods; metric: inter-rater reliability (IRR); chi-square tests; *t*-tests	30 raters (scorers) in Hong Kong	Scores awarded by “Onscreen Marking” and “Paper-Based Marking” were comparable.
Higgins, Patterson, Bozman, Katz	2010	25 General Educational Development (GED) mathematics practice items	Regression analyses	216 participants	There was no significant difference between paper-based test scores and computer-based test scores after controlling for initial performance.
Mason, Patry, Bernstein	2001	Introductory psychology course assessments	One-way ANOVA	27 university students (mean age: 20.2)	There were no significant differences by mode.
Minnick	2009	Tests of Adult Basic Education (TABE)	*t*-tests	150 male prison inmates ages 14–18	There were no significant differences by mode.
Mogey, Paterson, Burk, Purcell	2010	Essay test, mock course exam	Responses were transcribed so that each response was scored in both modes; ANCOVA	70 first-year divinity school students (nonrandom: participants chose condition)	No significant differences (including length of essay, overall scores, and some qualitative measures designed to indicate essay quality) found by mode.

Table B7 (continued). Ten other studies that found no score differences between computer-based and paper-and-pencil administration, without explicitly checking measurement equivalence between modes of assessment.

Authors	Year	Assessment	Design/Metrics Used	Participants	Main Findings
Whiting, Kline	2006	Test of Workplace Essential Skills (TOWES), Test of adult literacy skills, Subscales: Reading test, Document skills, Numeracy	Computer-based test scores and archived paper-based test scores matched based on years of education, age, gender; rank order equivalency; *t*-tests	73 undergraduate university students	Scores on all three subscales were equivalent based on their means and variances. In posttest survey, participants rated the computer-based test as easy to use.
Zandvliet, Farragher	1997	Three tests adapted from instructors' guide in an introductory college-level computer course.	*t*-tests	50 students in introductory computer classes	No significant mode effect on assessment scores was found.