Validity of the National Assessment of Educational Progress to Evaluate Cutting-Edge Curricula

Lorrie A. Shepard

Sami Kitmitto

Phil Daro

Gerunda B. Hughes

David C. Webb

Fran Stancavage

Natalie Tucker-Bradway

January 2020
Commissioned by the NAEP Validity Studies (NVS) Panel

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
Criterion Consulting, LLC

Jack Buckley
American Institutes for Research

James R. Chromy
Research Triangle Institute (retired)

Phil Daro
Strategic Education Research Partnership (SERP) Institute

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Ina V.S. Mullis
Boston College

Scott Norton
Council of Chief State School Officers

James Pellegrino
University of Illinois at Chicago

Gary Phillips
American Institutes for Research

Lorrie Shepard
University of Colorado Boulder

David Thissen
University of North Carolina, Chapel Hill

Gerald Tindal
University of Oregon

Sheila Valencia
University of Washington

Denny Way
College Board

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Grady Wilburn
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS) Panel
American Institutes for Research
2800 Campus Drive, Suite 200
San Mateo, CA 94403
Email: fstancavage@air.org

CONTENTS

INTRODUCTION

Common sense tells us that test results depend to a large extent on test content. If a test – used as an outcome measure to compare two curricula – favors the content of one curriculum over the other, it should be no surprise that the test-favored, or *test-aligned*, curriculum will most likely be judged to be the more effective curriculum. Because of the importance of test content specifications, the growth of high-stakes accountability testing over the last few decades resulted in a corresponding increase in methodologies to evaluate *instructional sensitivity*, *opportunity to learn* (OTL), and *alignment*. These methodologies are in wide use to determine, for example, whether a given state assessment does a good job of assessing state content standards.

The National Assessment of Educational Progress (NAEP) is similar to state assessments in that items for assessments in each subject matter domain must be well aligned with an agreed upon assessment framework. Unlike state assessment programs, however, NAEP's assessment frameworks cannot privilege one state's standards over another's. In this respect, NAEP is more like international assessments, because it must provide a fair, *cross-jurisdictional* assessment of learning outcomes. Moreover, as implied by the "educational progress" in its title, NAEP has a commitment to measure trends in student achievement over time. The National Assessment Governing Board's General Policy (2013) states that NAEP's first priority is "to serve as a consistent external, independent measure of student achievement by which results *across education systems* can be *compared at points in time and over time*" (p. 5) (emphasis added). Thus, to serve in its role as an independent monitor, NAEP must be broader than the typical state assessment, and it must anticipate the future.

This idea – that measuring progress over time means measuring well what students are currently able to do and at the same time reaching to measure expanded learning goals that are likely to be normal expectations in the near future – is sometimes referred to as NAEP's "lead and reflect" design principle (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007). The test design challenge posed by responding to new knowledge and new learning goals is reflected in the Governing Board's second goal statement, focused on technically sound assessment development.

> For NAEP to measure trends in achievement accurately, the frameworks (and hence the assessments) must remain sufficiently stable. However, as new knowledge is gained in subject areas, the information and communication technology for testing advances, and curricula and teaching practices evolve, it is appropriate for NAGB to consider changing the assessment frameworks and items to ensure that they support valid inferences about student achievement. (2013, p. 6)

The present study addresses this issue of new knowledge and future directions for NAEP's Mathematics Assessment at grades 4 and 8 by examining whether NAEP has the reach to assess the learning outcomes for cutting-edge curricula already in use. Before presenting the study's specific research questions and methodology, we provide two additional background summaries focused, respectively, on the cognitive science research behind deeper learning and 21st century skills and on prior studies addressing

the relationship between NAEP and Common Core State Standards for Mathematics (CCSS-M). CCSS-M represent one but not the only framework for understanding how learning goals and expectations for knowledge use are changing over time.

Deeper Learning and 21st Century Skills

At the turn of the 21st century, policymakers, politicians, and business leaders became keen on expanding the definition of outcomes for school learning. Primarily, they were concerned about international competitiveness and the need for a workforce with technological and analytical thinking skills (Murnane & Levy, 1996; National Alliance of Business, 2002). In 2009, when the Common Core State Standards (CCSS) Initiative was launched by the National Governors Association and the Council of Chief State School Officers, Jay Mathews (2009) of the *Washington Post* called "21st Century Skills" the new buzz phrase. That same year, the National Research Council (NRC) Committee on Defining Deeper Learning and 21st Century Skills was convened to examine research evidence as to how such skills are developed. The committee acknowledged that these types of goals for learning are not new; skills and abilities such as critical thinking, reasoning and argumentation, innovation, flexibility, initiative, self-reflection, collaboration, and communication have always been valued in society, but what may be new is the expectation that *all students* develop these abilities (National Research Council, 2012).

The NRC consensus report provides a comprehensive and accessible distillation of relevant studies from the learning sciences over the past three decades. The committee defined "'deeper learning' as the process through which an individual becomes capable of taking what was learned in one situation and applying it to new situations" (National Research Council, 2012, p. 5). Thus, "the product of deeper learning is *transferable knowledge*, including content knowledge in a domain and knowledge of how, why, and when to apply this knowledge to answer questions and solve problems" (p. 6). The committee emphasized that this conception of competencies that enable adept and flexible knowledge use is quite different from traditional learning goals focused on discrete facts and procedures.

Key to understanding how deeper learning occurs is the idea that deep mastery of disciplinary content and various reasoning, problem-solving, and communication skills involving that content are *jointly developed*. Fifty years ago, prevalent learning theories assumed that content knowledge had to be mastered *before* it could be applied. In contrast, learning scientists today have found that deep learning occurs as students are engaging with content in applications that are authentic to the "everyday activities" of professionals who work in a discipline (Sawyer, 2006). As part of their review, the NRC committee examined the CCSS documents and the Next Generation Science Standards (NGSS; National Research Council, 2013). They found that the inclusion of a "practices" dimension in both mathematics and science, and the English language arts requirement that students be able to synthesize and apply evidence to create and effectively communicate an argument, are consistent with research on deeper learning. The integrated development of knowledge and "fundamental mathematical capabilities" is also reflected in the 2015 PISA Mathematics Framework (Organisation for Economic Co-operation and Development, 2017), which calls for competencies such as communicating,

representation, reasoning and argument, using tools, and devising strategies for solving problems.

NAEP and the Common Core

As stated previously, the CCSS-M are not the only standards framework that reflect a research-based conceptualization of deeper learning goals for mathematics. Nonetheless, the CCSS-M are highly salient across many states in the United States, either because they have been adopted directly or because standards developed by individual states closely emulate the Common Core (Usiskin, n.d.). As a result, a number of studies have already been done to examine the relationship between NAEP and the CCSS-M. Comparisons between and among standards, frameworks, and assessments can be undertaken using a number of different methodologies. Conceptual or judgmental alignment methods (see review by Martone & Sireci, 2009) involve training disciplinary content experts to apply well-specified criteria for judging similarities and differences. The current study as well as recent comparisons reviewed here are all examples of expert-judgment studies.

When data are available, it is also possible to conduct empirical studies to examine similarities in psychometric structure as well as any variation in assessment outcomes associated with differences in assessment content. Under the direction of David Thissen, the NAEP Validity Studies (NVS) Panel has begun an empirical study linking NAEP with state assessments, including one of the consortium assessments. This study will evaluate how NAEP's items in reading and mathematics, designed to measure complex, advanced content, compare with items that are part of the consortium's assessments designed to do the same thing (Thissen, 2016, pp. 10–11).

Expert judgment methods differ in the level of specificity at which comparisons are made. Intended learning goals can be examined at the level of frameworks by comparing standards to standards, or the adequacy of item pools for "covering" or representing standards can be examined by comparing items to frameworks. An even more fine-grained comparison between two assessments can be undertaken by comparing item pools to item pools. The first study summarized here, by Hughes, Daro, Holtzman, and Middleton (2013), is a framework-level analysis, comparing the NAEP Mathematics Assessment framework to the CCSS-M. The analyses considered both the grade 4 and grade 8 frameworks and were conducted in both directions, thus identifying learning objectives that were common to both as well as objectives that were unique to each.

Hughes et al. (2013) found substantial overlap between the content in the CCSS-M and the NAEP Mathematics Framework. However, the study also identified four types of discrepancies that could have serious implications for valid interpretation of NAEP results. Compared to the NAEP framework, the CCSS-M have

- more rigorous content in eighth-grade algebra and geometry.
- more extensive and systematic treatment of mathematical expertise (found in the Standards for Mathematical Practice).
- a more conceptual perspective on many mathematical topics, explicitly stating the mathematics to be understood rather than the type of problem to be solved.

- 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 in the CCSS-M, and data sets and probability are taught in grades 6 and 7, respectively (Hughes et al., 2013, p. 58).

Points 1, 2, and 3 are instances of more challenging and rigorous standards in CCSS-M than are called for in the NAEP framework. To the extent then that the CCSS-M are actually being taught and students are mastering this material, NAEP will underestimate student achievement because more advanced levels of expertise and conceptual understanding will not be tapped by NAEP. Conversely, the discrepancy identified in point 4 refers to content in NAEP that is not intended to be taught under CCSS-M until a later grade. This may result in a "real" decline in NAEP results at grade 4 simply because students whose instruction follows the CCSS-M will not yet have been taught data and probability ideas that had previously been taught by fourth grade.

In addition to the framework-to-framework comparison above, Daro, Hughes, and Stancavage (2015) conducted a study to examine the alignment of the 2015 NAEP Mathematics *items* at grades 4 and 8 to the CCSS-M. Again, the analyses were conducted in two directions to examine both the fit of NAEP items within the CCSS-M framework and the coverage of CCSS-M by NAEP. At grade 4, "79% of NAEP items clearly matched to the CCSS standards at or below grade 4, and 77% of grade 3 and 4 CCSS standards [are] assessed by at least one NAEP item" (Daro et al., 2015, p. 14). Consistent with the framework comparisons, mismatches at grade 4 in geometry and data analysis, statistics, and probability occurred because objectives covered in the grade 4 NAEP assessments have been moved to higher grade levels in the CCSS-M. At the same time, NAEP at grade 4 underrepresents algebraic thinking, which has a more prominent place at grades 3 and 4 in the CCSS-M.

At grade 8, the similarities are quite strong when NAEP is viewed as measuring a subset of CCSS-M, with 87% of NAEP items aligned with seventh- or eighth-grade CCSS-M standards. However, when viewed from the other direction – how well CCSS-M is covered or represented by NAEP – only 58% of CCSS-M standards for grades 6, 7, and 8 are tapped by at least one grade 8 NAEP item. Thus, Daro et al. (2015) concluded that "there appears to be a notable amount of middle-school mathematics content recommended by the CCSS-M that is not part of the current NAEP assessment" (p. iii).

Both framework-to-framework and item-to-framework analyses summarized thus far focus on the similarities and differences in coverage of the assessments' respective content domains. Another important dimension – especially considering the significance of research on deeper learning and 21st century skills – has to do with the cognitive complexity and mathematical practices represented by assessment items. The most recent study undertaken by Daro et al. (2019) has updated the content analyses of NAEP items based on the 2017 assessments, comparing these item pools to item sets from two non-consortium states and the two consortia. The study employed teams of expert judges to compare items from NAEP and these four state assessments using four broad indicators of complexity and mathematical practices:

- Problem Solving and Modeling Challenge
- Depth and Robustness of Conceptual Understanding

- Procedural Fluency
- Demand for Argumentation or Communicating Reasoning

Expert judges also rated items from all five assessments using a rubric representing the "Construct Centrality" of items based on grade-level appropriateness, centrality of the mathematics assessed, the avoidance of construct-irrelevant sources of difficulty, and the integration of mathematical practices with content. The Daro et al. (2019) NAEP-state-comparison study was carried out jointly with the cutting-edge curriculum study reported here. The shared methodology for the two studies is described in greater detail in the next section. In the case of the current study, the "items" include both instructional tasks and assessment items from end-of-unit summative tests sampled from cutting-edge curricula.

With respect to construct centrality, the Daro et al. (2019) study found that nearly all items used in both NAEP and the various state assessments (SA1–SA4) were rated at levels 3 and 4 of the 4-point rubric. A level 3 rating indicates that grade-appropriate mathematics is assessed without interference from construct-irrelevant factors. Level 4-rated items satisfy these criteria and, in addition, assess *important* mathematics and engage at least one *mathematical practice*. Two state assessments did notably better than NAEP at grade 8, with 32% (SA4) and 40% (SA3) of their item score points rated as 4. On the mathematical practices dimensions, NAEP was similar to the state assessments, except that NAEP had substantially lower percentages of score points calling for higher levels of Argumentation or Communicating Reasoning (rated 3 or 4) compared to SA3 (4% compared to 13%, at grade 4; 5% compared to 17% at grade 8). Higher demands for engagement with mathematical practices mean that this state assessment is more consistent with the "fundamental mathematical capabilities" called for in the 2015 PISA Mathematics Framework (Organisation for Economic Co-operation and Development, 2017) referenced previously.

STUDY PURPOSE AND RESEARCH QUESTIONS

To remain true to its mission as an independent monitor of "educational progress" over time, NAEP content and item types must include adequate representations of likely future learning targets. Although the CCSS per se, and the two assessment consortia linked to the Common Core, have experienced considerable political backlash and erosion of participation, there has at the same time been widespread adoption of curricula that attend to college and career readiness standards, deep learning, and 21st century skills (Usiskin, n.d.). *The important question for NAEP, then, is whether its assessments are able to validly assess student learning in learning environments where curriculum and instruction are tied to new, ambitious standards. Or, are students in these schools developing knowledge and skills that NAEP is not assessing?*

The process for identifying and refining a list of exemplary curricula is described in the Methodology section. As a shorthand, these ambitious curricula aimed at deeper learning of content and 21st century competencies came to be called *cutting-edge curricula (CEC),* and this study was dubbed the "cutting-edge curricula study." The study addressed two specific research questions:

RQ1 How does the content of cutting-edge curricula (intended knowledge and cognitive competencies) as represented by *instructional tasks* compare with the content of NAEP?

RQ2 How does the content of cutting-edge curricula (intended knowledge and cognitive competencies) as represented by *assessment items* compare with the content of NAEP?

The current study is envisioned as part one of a two-part investigation. If important differences are found by this judgmental review comparing instructional tasks and assessment items from CEC with NAEP, then a second, empirical study will be conducted to test further how much the inclusion of ambitious assessment items currently missing from NAEP might alter assessment results.

METHODOLOGY

This study uses an expert-judgment alignment methodology to compare NAEP item pools with the instructional tasks and assessment items from two CEC selected for each of grades 4 and 8. Consistent with well-known alignment methodologies, such as those developed by Webb (1997) and Achieve (Rothman, Slattery, Vranek, & Resnick, 2002), evaluation criteria were developed to examine dimensions of cognitive demand as well as the distribution of items across content categories (Martone & Sireci, 2009). As described below, the Consolidated Content Framework and the Mathematical Practices Rubrics were jointly developed with the Daro et al. (2019) study. Both studies were projects overseen by the NVS panel. In addition to the development of the Consolidated Content Framework, initial sorting of items into content domains and subdomains, identification of mathematics education experts to serve as judges, training of judges, convening of an in-person meeting to review items, management of the ratings data base, and analyses by AIR staff were all carried out jointly as if it were one study, not two.

Curriculum Identification

Preliminary work to inform the design of the CEC study began with an effort to identify instances where curriculum and instruction are informed by new college- and career-ready standards. The focus was on two subject areas, mathematics and science, and two grade levels, grade 4 and grade 8. Grade 12 was not considered because curricula at grade 12 are more subject-specific, making potential comparisons to a curriculum-general assessment like NAEP more difficult.

For inclusion in this study, curricula would be required to meet the following criteria:

a. **Be cutting-edge in terms of the skills and knowledge being taught and assessed.** Being tied to newer standards, such as NGSS, is not required but is a useful indicator for this criterion. Incorporating advanced digital technology (e.g., virtual laboratories, software-based tools) would be another useful indicator of a curriculum that meets this criterion, but it is also not essential.

b. **Be in use in schools on more than just an experimental basis.** For the second part of this study, we will eventually need to administer test items in locations where CEC are being used in practice, and therefore we would want to choose schools where the curriculum meets the first criterion above and is, in addition, an established part of the schools' approach to teaching and learning.

The study team conducted informal interviews with experts on math and science curricula, reviewed publicly available materials from a wide range of nominated curricula, and presented an overview of these materials to the NVS panel. Because the science curricula that were nominated, for example, those funded by the National Science Foundation (NSF), were not in widespread use, and because publicly available sample tasks were not appreciably different from NAEP items, the decision was made to focus on mathematics curricula at grades 4 and 8.

The second requirement – that a candidate CEC be already in use on a sufficient scale to support later empirical work – turned out to be determinative in selecting the curricula to be studied. This constraint imposes a conservative bias on the study comparisons, given that innovative curricula still in development were necessarily excluded. Based on use in a number of jurisdictions and initial evidence of potential content and practices demands beyond NAEP, the following curricula were identified for the study:

- Engage New York Mathematics (Eureka Math) at grades 4 and 8
- Investigations Mathematics at grade 4
- Connected Mathematics (CMP) at grade 8

Note that Investigations and CMP are in use in large numbers of districts in part because they are updates of innovative curricula funded by NSF as part of curricular reforms in the 1990s.

As part of this review process, it also became apparent to the study team that assessment items provided with these curricula were often not so rich as their instructional tasks. In fact, CEC assessment items frequently looked like traditional test items. The study was, therefore, expanded to examine both instructional tasks and assessment items from each of the CEC, with greater attention being paid to the problem types used as part of instruction.

Sampling of Instructional Tasks and Assessment Items

As summarized in Table 1, each of the identified CEC had unique terminology and organizational structures by which lessons were organized within content domains. Instructional tasks were sampled for the study from the lowest level of the curricular structure, which best represented instructional activities. Assessment items were sampled from the respective end-of-unit or end-of-module assessments.

Table 1. Structure of Curricula

Engage NY – Grade 4	Investigations – Grade 4
Instructional Materials 1. Module 2. Topic 3. Lesson a. Problem Set (sampled tasks) b. Exit Ticket (sampled tasks) c. Homework d. Sprint *Assessment Materials* 1. End-of-Module Assessment (sampled items)	*Instructional Materials* 1. Unit 2. Investigation 3. Session a. Named Activities (sampled tasks) *Assessment Materials* 1. Unit Test (sampled items)
Engage NY – Grade 8	**CMP – Grade 8**
Instructional Materials 1. Module 2. Topic 3. Lesson a. Classwork (sampled tasks) b. Problem Set* (sampled tasks) *Assessment Materials* 1. End-of-Module Assessment (sampled items)	*Instructional Materials* 1. Unit 2. Investigation 3. Problem Set (sampled tasks) *Assessment Materials* 1. Unit Test (sampled items)

* Tasks for Engage NY grade 8 were sampled only from Classwork except for one lesson where tasks from the Problem set were sampled because they better characterized the lesson.

The sampling strategy used to select *instructional tasks* from the curricular materials was designed so that the total number of tasks chosen for analysis was approximately equal to 42 (ranging from 36 to 48 tasks across the four curricula). This sample size was agreed upon as a manageable number to be rated by judges and is roughly equivalent to the size of one form of a state assessment being rated in the concurrent Daro et al. (2019) study. The process for sampling was as follows:

a. Lessons/Sessions/Problem Sets ("Level 3" in the structure) and then tasks were sampled broadly across all the content domains covered in proportion to the number of lessons allotted to that domain. (In this process, all the "Level 1" Modules/Units were sampled and almost all the "Level 2" Topics/Investigations were represented in the study.)

b. Within each "Level 2" Topic/Investigation, either the last or the most representative "Level 3" Lesson/Session/Problem Set was selected so as to best reflect the culminating learning targets for each Topic or Investigation. The

Investigations curriculum had fewer "Level 2" Investigations; therefore, in some instances, two "Level 3" Sessions were selected from the longer Investigations.

c. Within each selected "Level 3" Lesson/Session/Problem Set, tasks were then sampled at random. Note that instructional tasks were often arranged in multipart problems, which were discretized prior to the random sampling so that the task being rated would be more parallel to the items being rated in NAEP and the state assessments – that is, the discretized elements were each counted as a separate task. For Engage NY; two to five tasks were selected for each sampled Lesson at grades 4 and 8, either from the Problem Set or the Exit Ticket, again to best characterize the Lesson. For Investigations, two to six tasks were chosen from one Named Activity within each sampled Session, and for CMP, tasks were sampled from the last Problem Set in each Investigation.

Assignment of Score Points

As described previously, multipart problems in the instructional materials were segmented to more closely resemble "items" on NAEP and state assessments. In addition, to better ensure comparability between instructional materials and assessments, score points were assigned to each instructional task following rules derived from NAEP rubrics. Multiple-choice and short constructed response items received one point. More extended response questions received two points.[1]

Weighting of the Instructional Tasks

The original sampling strategy for instructional tasks was intended to represent content domains in proportion to their occurrence in the overall curriculum. However, some disproportion could have been introduced by variability in the number of lessons per topic or investigation. In addition, because the segmenting of selected tasks and assignment of score points occurred after the random sampling of tasks within lessons, some disproportion could have been introduced if there were any within-curriculum interactions between content domain and task format. To ensure that findings could be reported in terms of the amount of instructional time devoted to each content domain, a sampling frame was constructed by assigning each of the full population of Lessons/Sessions/Problem Sets to the appropriate content domain of the Consolidated Framework described in the next section. Sampling weights were then determined so that the content distribution of sampled tasks in each curriculum matched the content distribution for the population of Lessons/Sessions/Problem Sets in that curriculum. The weighted results differed only slightly from the unweighted results. Only the weighted (and hence proportional) analyses are reported.

The sampling strategy to select *assessment items* from the curricular materials was more straightforward.

a. The population of possible items was defined by the full set of items from all the end-of-module assessments/unit tests in a given curriculum.

b. A 50% sample of items was drawn at random, resulting in samples of 40 to 79 items per curriculum.

[1] In six instances, there were multipart problems sampled that were not segmented and they received 3 to 6 points according to their complexity.

Consolidated Content Framework

To facilitate expert review of assessment items and instructional tasks across all the assessments and curricula in the combined studies, items and tasks were presorted into content categories by mathematics education experts at the University of Colorado Boulder. NAEP, each of the state assessments (consortia and single-state), and CEC all use slightly different content categories. Therefore, a consolidated content framework was developed that showed the correspondence between content domains (and subdomains) across the various materials to be rated. A more complete explanation of the reasoning behind the development of the consolidated content framework is provided in Daro et al. (2019).

At the domain level, there was, for the most part, a clear correspondence in the organization of content across the various frameworks. When differences occurred, it was primarily because of differences in level of aggregation. These differences, and the ways in which we addressed them, are described below.

Grade 4. The consolidated content domains for grade 4 are as follows:

- Numbers, Operations, & Algebraic Thinking (NOAT)
- Calculations & Place Value
- Fractions
- Measurement
- Data
- Geometry

At this grade level, topics in the NAEP Number Properties and Operations strand distribute across three CCSS-M domains: Operations & Algebraic Thinking, Number & Operations in Base Ten, Number & Operations – Fractions. As a rule, our consolidated content framework uses the more disaggregated categories to permit more apples-to-apples comparisons. Consequently, in the previous example, the CCSS-M categories were preferred. However, for Measurement and Data, two separate categories were preserved corresponding to two strands in the NAEP framework, although these two categories are combined in CCSS-M. Exceptions to the general rule arose in some instances because of the structure of the NAEP framework. Because NAEP uses the same content strands across grade levels, there are instances in which one of the strands contains only a few subobjectives at a particular grade. This occurs in the NAEP Algebra strand at grade 4, and consequently, these few Algebra subobjectives are incorporated into the NOAT category in the consolidated framework.

Grade 8. Consolidated content categories for grade 8 are as follows:

- The Number System
- Expressions & Equations
- Functions
- Geometry
- Statistics and Probability

At this grade level, content domains are quite similar between NAEP and CCSS-M, except that topics in NAEP's Algebra strand break out into Expressions & Equations and Functions for CCSS-M. Again, we use the more disaggregated categories. Also, at grade 8, the two subobjectives in the NAEP Measurement strand are subsumed under Geometry in the consolidated framework.

Complexity and Mathematical Practices Rubrics

In the same way that a Consolidated Content Framework had to be developed to enable comparisons across assessments and curricula that used different content categories, it was also necessary to determine what aspects of cognitive complexity could reasonably be evaluated across different assessment and curricular programs. As noted by Martone and Sireci (2009), well-known alignment methodologies typically have both a content and a cognitive demand dimension, which attends to the type of thinking required by each assessment item. The Surveys of Enacted Curriculum (SEC) methodology (Porter & Smithson, 2001), for example, has five cognitive levels: memorize facts; perform procedures; demonstrate understanding; conjecture, generalize, prove, and solve nonroutine problems; and make connections. For purposes of the Daro et al. (2019) and this CEC study, it was important that the dimensions of cognitive demand be encompassing of the thinking skills required by the various assessments and curricula without favoring any one conceptualization over others.

NAEP's Mathematics Framework (National Assessment Governing Board, 2017) has a general Mathematical Complexity dimension defined as follows:

> *Low Complexity*
> Low-complexity items expect students to recall or recognize concepts or procedures specified in the framework. Items typically specify what the student is to do, which is often to carry out some procedure that can be performed mechanically. The student is not left to come up with an original method or to demonstrate a line of reasoning. (p. 38)
>
> *Moderate Complexity*
> Items in the moderate-complexity category involve more flexibility of thinking and choice among alternatives than do those in the low-complexity category. The student is expected to decide what to do and how to do it, bringing together concepts and processes from various domains…. Students might be asked to show or explain their work but would not be expected to justify it mathematically. (p. 43)
>
> *High Complexity*
> High-complexity items make heavy demands on students, because they are expected to use reasoning, planning, analysis, judgment, and creative thought. Students may be expected to justify mathematical statements or construct a mathematical argument. Items might require students to generalize from specific examples. Items at this level take more time than those at other levels due to the demands of the task, not due to the number of parts or steps. (p. 46)

Many states organize the kinds of thinking and problem-solving skills required by assessments according to the eight Standards for Mathematical Practice from the CCSS-M (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010):

- Make sense of problems and persevere in solving them.
- Reason abstractly and quantitatively.
- Construct viable arguments and critique the reasoning of others.
- Model with mathematics.
- Use appropriate tools strategically.
- Attend to precision.
- Look for and make use of structure.
- Look for and express regularity in repeated reasoning.

The two consortia have their own organizational structures. The Partnership for Assessment of Readiness for College and Careers (PARCC; n.d., p. 2) identifies three task types:

- Type I tasks assess concepts, skills and procedures.
- Type II tasks assess mathematical reasoning through "written arguments/justifications, critique of reasoning, or precision in mathematical statements."
- Type III tasks call for modeling or applications in a real-world context or scenario.

While Smarter Balanced (n.d., p. 1) gathers evidence with regard to four claims:

- Claim #1: Concepts & Procedures
- Claim #2: Problem Solving
- Claim #3: Communicating Reasoning
- Claim #4: Modeling and Data Analysis

Authors Daro et al. (2019), in collaboration with the leadership team from the panel of experts used in both studies, identified the four mathematical practices most widely shared across the studied assessments and curricular materials and developed the following rubric to evaluate levels of increasing complexity on each (Table 2).

Table 2. Rubrics for Evaluating Items for Mathematical Practices

Expertise/ Practices Domains	Level of Complexity			
	1	2	3	4
Problem Solving and Modeling Challenge	Little or no problem solving demanded, execute an indicated calculation. E.g., 12X4 = ?; mark 3/8 on the number line; solve 2n +3 = 9	Involves application of mathematics, but the math is indicated or routine for grade level	Decide what to do in nonroutine situation; make sense of quantities or figures and their relationships implied by posed problem Strategic thinking	Quantities or figures and their relationships are not explicit and a mathematical model must be formulated, or a model is evaluated against its purpose
Depth and Robustness of Conceptual Understanding	No grade-level conceptual understanding demanded	Recall or recognize a concept, routine use; match terminology to examples to which the term refers	Adapt or extend a concept; or apply in an unfamiliar/ nonroutine setting	Use or explain a relationship among multiple concepts, and/or show the conceptual basis for a strategy or procedure
Complexity of Procedural Fluency/ Demand	Little or no procedural demand or procedural demand is well below grade l e v e l	Common or grade-level procedure(s), with friendly numbers	Common or grade-level procedure(s), with unfriendly numbers; unconventional combination of procedures; or requires unusual perseverance or organizational skills in the execution of a procedure	N/A
Demand for Argumentation or Communicating Reasoning	Little or no argumentation demanded	(a) Show work or explain how (b) Respond to given reasons	Generate reasons; justify statement(s); explain why a solution or method makes sense and/or provide evidence for reasoning and/or explain analysis	Construct a viable argument about the truth or generality of a mathematical statement that employs mathematical principles and/or logical argumentation

The authors and leaders from the math education expert panel also developed a rubric for evaluating the Construct Centrality of each assessment item and instructional task. The construct for any given item is the union of the mathematical content and mathematical practice(s) that an item is intended to assess. The Construct Centrality of an item expresses how closely the item hits the priorities of the intended mathematics content and practices; it does not mean merely fitting into a topical category (Table 3).

Table 3. Construct Centrality Rubric

Construct Centrality	Ratings			
	1	2	3	4
	(a) The mathematical content is either far above/below grade span or is not in the content standards or not central to the priorities in the standards OR (b) Major construct-irrelevant challenges overwhelm any construct-relevant challenges. Many students are likely to get it wrong (or right) for irrelevant reasons.	What is assessed is marginal for some combination of the following reasons: (a) Low-priority mathematics or (b) Irrelevant features likely to affect performance. Some students are likely to get it wrong or right for irrelevant reasons. (c) Lack of relevant practices.	Grade-span mathematics is assessed (may be less important), construct irrelevant features are not likely to diminish or enhance performance for most students, practices may be lacking or present.	Addresses all four of the following: 1. Aims at grade-span important math; 2. Hits what it aims for; 3. Avoids construct-irrelevant challenges; 4. Engages at least one mathematical practice.

Expert Panel and Rating Process

Authors Daro et al. (2019) first identified a core, leadership team of mathematics education experts to aid in the development of the study rubrics described previously and the recruitment of a full panel of expert judges. The core group collaborated first by attending a multiday in-person meeting and then followed up with a series of webinars to finalize the rubrics, exemplars, and rating processes. In addition to the narrative descriptors for each level of the rubric, the leadership team selected multiple *anchor items* from the various assessments and curricular materials to illustrate the type of item warranting a score of 1, 2, 3, or 4 on each dimension.

An expert panel of 31 mathematics content experts was recruited representing a variety of professional roles, including K–12 teachers, mathematicians, mathematics education researchers, state supervisors, and consultants. Members of the leadership team and expert panel are listed in Appendix A. Prior to an in-person rating meeting, all of the experts participated in a training webinar to become familiar with (a) the study purpose and study methodology, (b) the rating rubrics and exemplar items/tasks, (c) the technology interfaces needed to access each set of secure assessment items as well as the selected samples of curricular materials, and (d) the use of Excel spreadsheets for recording individual ratings.

A 2-day, in-person meeting was held to develop greater shared understandings about how to apply the rubrics while also rating and resolving differences for as many items and tasks as possible. At the start of the meeting, panel members met as a group to once again be oriented to the purpose of the study and to the rubrics for each of the four math practices and Construct Centrality.

To allow for greater specialization panelists were then divided in four 7- or 8-person teams based on grade level and content domains:

1. Fourth-grade Number (NOAT, Calculations & Place Value)

2. Fourth-grade Measurement, Data, & Geometry
3. Eighth-grade Algebra (Expressions & Equations, Functions)
4. Eighth-grade Number, Geometry, Statistics & Probability

These specialist teams reviewed and discussed grade-specific anchor items and tasks associated with each point on the rubrics. Based on these discussions, the rubrics were edited to incorporate any necessary refinements and clarifications and finalized. The final rubrics are the basis for the analyses reported here.

The remainder of the meeting was devoted to work within the four specialist teams. Each team was assigned a calibration set of 73 to 85 items/tasks representative of the grade and content area in which they were to specialize and taken from each of the sources (NAEP, three state assessments, and two CEC[2]). Their instructions were to complete as many as possible of these items/tasks during the in-person meeting in collaboration with their team members. The calibration sets of items/tasks included 12 items from each state assessment, 12 items from NAEP, and 6 instructional tasks plus 6 assessment items from each curriculum. Any of the calibration items/tasks not rated during the in-person meeting were to be completed at home, and panelists were encouraged to continue to engage their team members through email or teleconference during this process.

Following completion of the full calibration set, each panelist received an individualized set of items/tasks to complete on their own. Individually assigned item/task sets included a random sample of the remaining, not-yet-rated assessment items and instructional tasks from the panelist's designated grade and content area. Each item/task was randomly assigned to four team members such that each combination of team members had an approximately equal number. Table 4 provides a summary of the calibration and individually-assigned rating sets by content and grade level team.

[2] One state assessment did not make materials available in time, so only three were included in the in-person meeting. For the CEC, two curricula were included at each grade.

Table 4. Summary of Item/Task Assignments by Grade Level & Content Team

	Calibration Items/Tasks			Individually Assigned Items/Tasks		
	Number of Items/Tasks Completed	Number of Panelists Assigned to Each Item/Task	Number of Panelists Who Completed a Rating for Each Item/Task	Number of Items/Tasks Completed	Number of Panelists Assigned to Each Item/Task	Number of Panelists Who Completed a Rating for Each Item/Task
Grade 4 Number Team	89	8	7	268	4	3–4
Grade 4 MDG Team	73	7	6	159	4	4
Grade 4 Total	**154**			**463**		
Grade 8 Algebra Team	85	8	8	177	4	4
Grade 8 NGS Team	76	8	7	186	4	3–4
Grade 8 Total	**161**			**562**		

After all of the individual item ratings had been received, a two-stage review process was developed to arrive at a final rating for each item/task on each of the five rubrics. Ratings for most of the items/tasks rated in-person had already been resolved by group discussion during the meeting. For items/tasks rated at home or for which the in-person discussions had not produced a final consensus rating, a computer algorithm was used initially that assigned the *modal* rating under conditions where (a) only one panelist out of three or four disagreed with the modal rating or (b) where only two or three panelists out of eight disagreed with the mode. Items/tasks were flagged for secondary review on specific rubrics when there were greater levels of disagreement among panelists or when there were only three raters and they had not all given the same rating. The secondary review was conducted by authors Daro et al. and members of the leadership team by webinar. To make final decisions about whether an item/task should be rated as a 1, 2, 3, or 4 on a given rubric, the review team relied on the ratings provided by the panelists, any notes from the panelists that explained their ratings, specific requirements of the rubric, and on their knowledge of how other items with similar attributes had been rated.

FINDINGS FROM CONTENT ANALYSES AND CONSTRUCT CENTRALITY

Findings for the proportional allocation of NAEP items, CEC instructional tasks, and CEC assessment items to each of the content domains in the Consolidated Framework are presented in this section, along with findings for the ratings of Construct Centrality. Although studies have already been reported comparing both the NAEP framework and the 2015 NAEP items to CCSS-M, this study and the concurrent Daro et al. (2019) study are not tied specifically to the CCSS. The Daro et al. study looked at state assessments; then, by examining instructional tasks and assessment items from curriculum materials, this study provides a comparison that is one step closer to what is actually being taught, at least in those districts where these CEC are currently in use. The comparisons are based on 2017 NAEP items. Because the underlying validity question is whether students might be learning content different from what NAEP is assessing, our discussion of findings focuses primarily on the distributions for *instructional tasks* compared to NAEP.

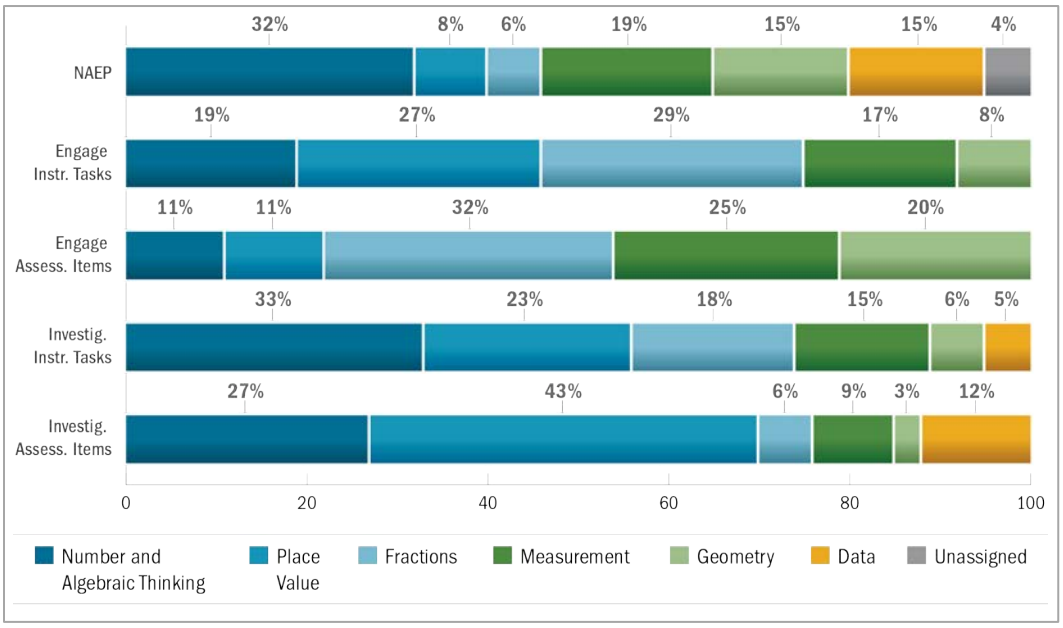
As noted in the Methodology section, assessment items and instructional tasks from the selected CEC were classified into the domains of the consolidate framework by the same team of mathematics experts that sorted NAEP and state assessment items. When items/tasks tapped more than one content domain, these experts classified them according to the most demanding mathematics needed for the item/task. For instructional tasks, the classifications almost always corresponded with the named topic of the lesson.

For assessment items, all findings in this and following sections are presented in terms of the contribution of an item to the total assessment score (i.e., the percentage of total score points allocated to an item). There is no "total score" associated with curricula. Therefore, to create a common metric for comparisons, "score points" were assigned to instructional tasks, as described in the Methodology section, using a logic derived from NAEP scoring rubrics.

Grade 4

At grade 4, consistent with Common Core–influenced standards, both Engage NY and Investigations have a much higher proportion of instructional task score points devoted to arithmetic than is true for NAEP. As shown in Figure 1, the first three content categories – *Number & Algebraic Thinking, Place Value*, and *Fractions* – account for 75% of Engage NY and 74% of Investigations instructional task score points. These high proportions are in comparison to only 46% of NAEP allocated to these arithmetic categories. The difference is especially pronounced in the treatment of *Fractions*, with only 6% of NAEP score points addressing *Fractions* at grade 4 compared to 29% for Engage NY and 18% for Investigations instructional tasks.

Figure 1. Distribution of Score Points Across Content Domains: NAEP, CEC Instructional Tasks, and CEC Assessment Items, Grade 4

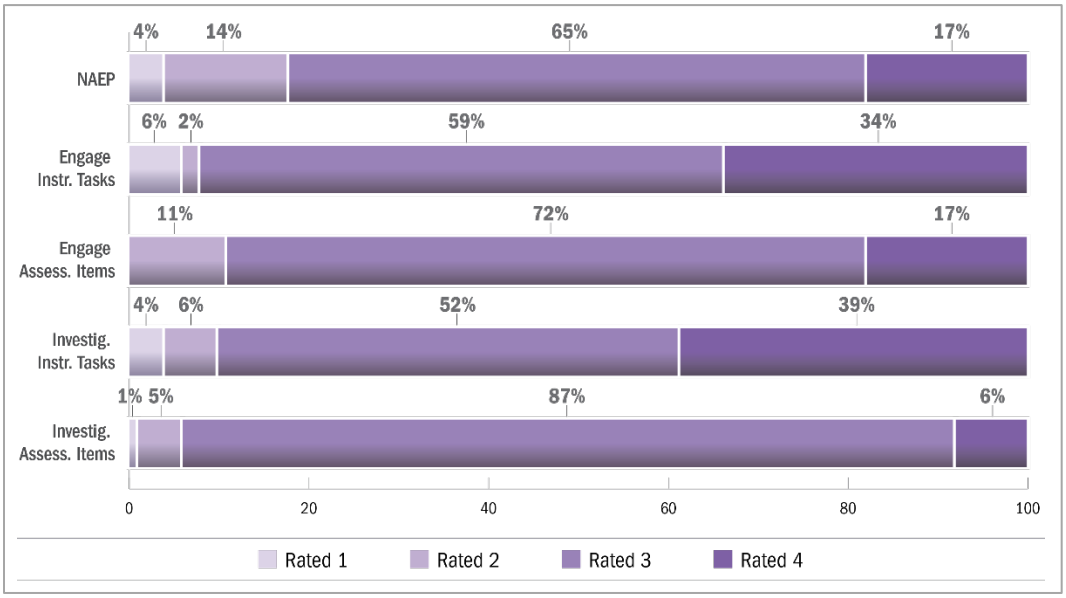


NAEP and instructional tasks for the two CEC give similar attention to *Measurement*, with 19% of NAEP score points addressing *Measurement* compared to 17% for Engage NY instructional tasks and 15% for Investigations instructional tasks. The areas underrepresented in the CEC instructional tasks compared to NAEP are *Geometry* and *Data*. Four percent of NAEP score points were unassigned to a content domain; these points derived from items that were either multidomain or assessed topics that were below grade level.

We do not comment on the allocation of assessment items across content domains for each CEC because they roughly track the proportional allocation of instructional tasks. Where there are departures from this pattern, as with the overrepresentation of Place Value items in the Investigations assessment items, they appear to be distortions of intended curricular emphases.

Figure 2 presents the distributions of grade 4 score points for Construct Centrality. The Construct Centrality rubric requires judgments about both the importance or centrality of the mathematics (content and practices) assessed and the validity of that assessment (i.e., whether an item or task is free of construct-irrelevant challenges). In general, the vast majority of score points for NAEP, CEC instructional tasks, and CEC assessment items at grade 4 derive from items that were judged by the expert panel to be measuring appropriate grade-level mathematics and doing so in a way that was not confounded by irrelevant sources of difficulty (rated Level 3 and above). However, Engage NY and Investigations instructional tasks have much higher percentages of score points rated at Level 4 than either NAEP or the CEC assessment items. A Level 4 rating means that the instructional tasks not only involved important mathematics content in a valid way but also engaged students in at least one mathematical practice.

Figure 2. Distribution of Score Points Across Levels of Construct Centrality: NAEP, CEC Instructional Tasks, and CEC Assessment Items, Grade 4

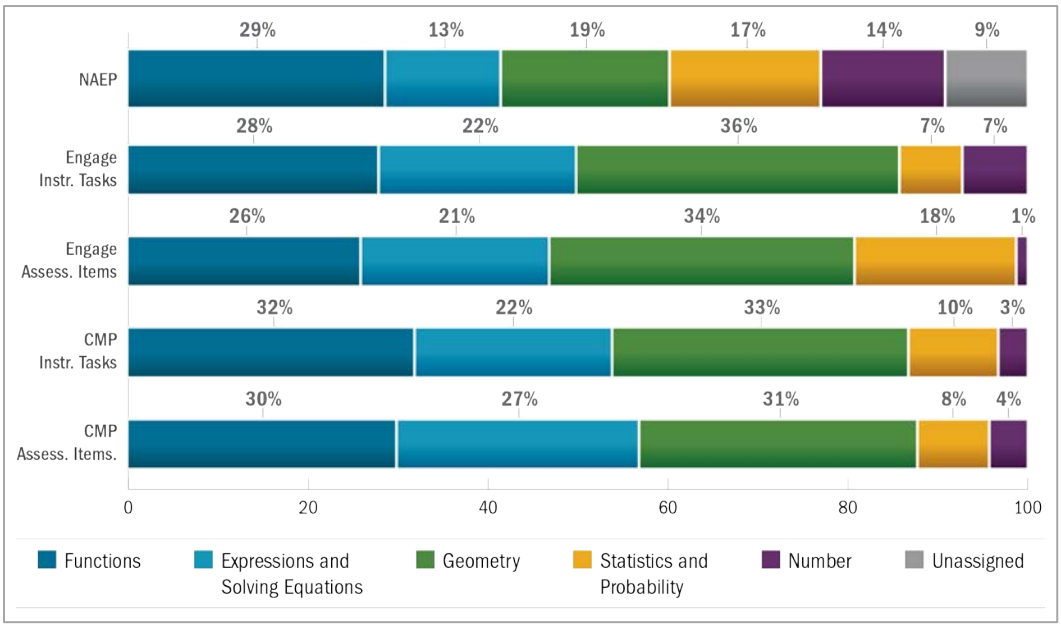


The Construct Centrality dimension reflects the overall quality of items or tasks used to engage students in mathematical thinking. As such, and especially at Level 4, the rubric for Construct Centrality is not independent of later rubrics used to evaluate mathematical practices. The differential results seen here for CEC instructional tasks are not surprising given results reported in the next section, where CEC instructional materials are found to be consistently better on mathematical practices – especially Problem Solving and Modeling and Argumentation. We should also note that 14% of NAEP score points are rated at Level 2 for Construct Centrality, which means either that "low priority mathematics" is assessed or that "irrelevant features likely affect performance." An additional 4% of NAEP grade 4 score points (and 7% of grade 8 score points) come from items rated at Level 1; however, Level 1 is difficult to interpret because it occurs primarily as the result of content judged to be below grade level. We, therefore, focus only on Level 2 in this discussion.

Grade 8

In Figure 3, the distribution of score points across content domains is shown at grade 8 for NAEP, CEC instructional tasks, and CEC assessment items. Similar to the content distributions at grade 4, the content emphases in the grade 8 CEC instructional tasks are different from NAEP in ways that would be predictable from Common Core–influenced standards. Engage NY and CMP instructional tasks reflect a greater emphasis on the two algebra categories (*Functions* and *Expressions & Solving Equation*s) than NAEP – 50% and 54%, respectively, compared to NAEP's 42%. The CEC instructional tasks also devote considerable attention to *Geometry*, accounting for 36% and 33% of instructional task score points compared to NAEP's 19%. Conversely, NAEP assigns greater weight to *Statistics & Probability* and to *Number* than do the two curricula. Nine percent of NAEP score points were not assigned to content domains because they assess below-grade-level content.

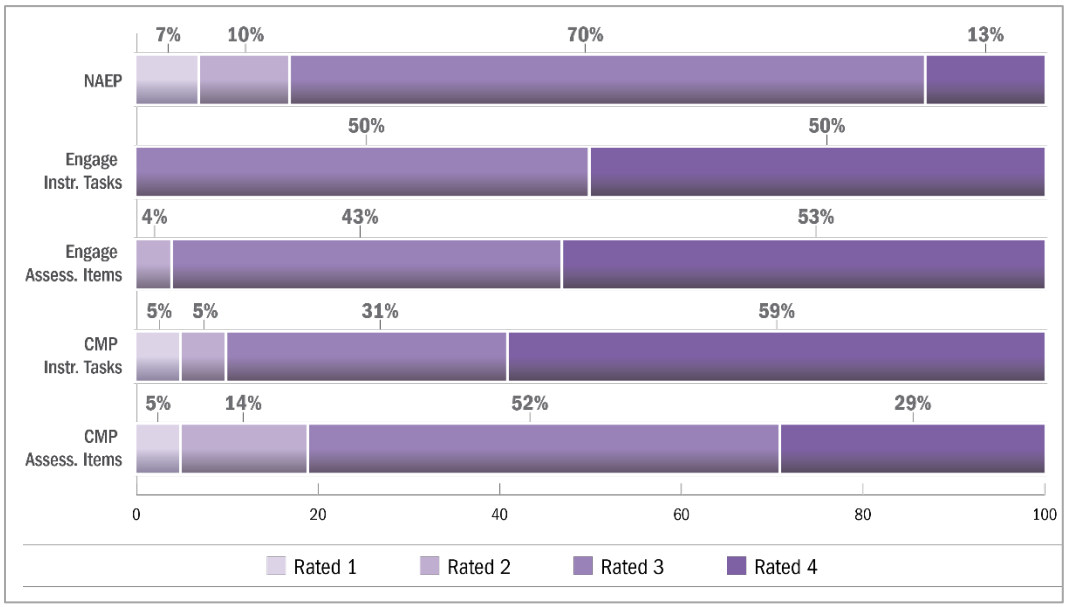
Figure 3. Distribution of Score Points Across Content Domains: NAEP, CEC Instructional Tasks, and CEC Assessment Items, Grade 8



Although it is frequently the case in this study that CEC assessment items do not map closely with the instructional tasks from their own curriculum, in this case there is a close correspondence between the distribution of assessment items and their respective instructional tasks.

As shown in Figure 4, the vast majority of score points for NAEP, CEC instructional tasks, and CEC assessment items at grade 8 derived from items that were judged by the expert panel to be measuring appropriate grade-level mathematics and doing so in a straightforward way, unconfounded by irrelevant sources of difficulty (rated Level 3 and above on Construct Centrality). Engage NY and CMP instructional tasks and assessment items all had much higher percentages of score points rated at Level 4 than NAEP, ranging from 29% of CMP assessment score points to 59% of CMP instructional task score points. Level 4 ratings mean that these instructional and assessment tasks not only involved important mathematics content in a valid way but also engaged students in at least one mathematical practice.

Figure 4. Distribution of Score Points Across Levels of Construct Centrality: NAEP, CEC Instructional Tasks, and CEC Assessment Items, Grade 8



FINDINGS REGARDING COMPLEXITY FROM MATHEMATICAL PRACTICES ANALYSES

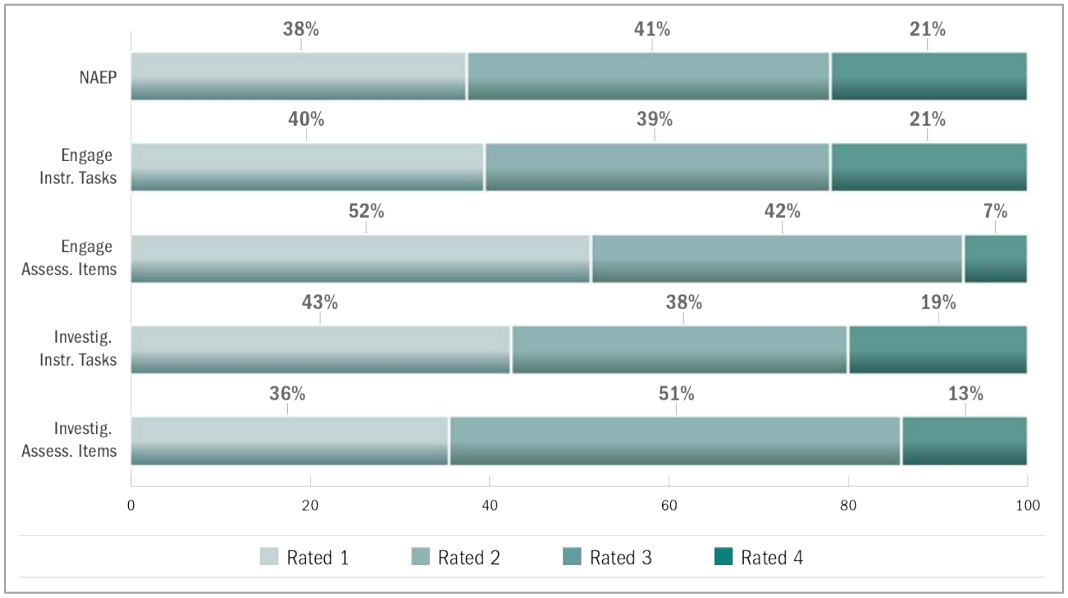
As described in the Daro et al. (2019) study, the Mathematical Complexity dimension in NAEP calls for students to engage in low, moderate, or high levels of mathematical reasoning and analysis to be able to solve problems. For this study, in parallel to the Daro et al. study, complexity (or cognitive demand) was decomposed into four domains of mathematical practice: Problem Solving and Modeling Challenge, Depth and Robustness of Conceptual Understanding, Complexity of Procedural Fluency, and Argumentation or Communicating Reasoning. Each practice domain, with the exception of Procedural Fluency, was rated on a 4-point scale, where 1= little or no cognitive demand or complexity, 2= low complexity or grade-level expectation, 3= moderate complexity, and 4= high complexity. Procedural Fluency was rated on a 3-point scale, with 1= little or no cognitive demand or complexity, 2= low complexity or grade-level expectation, and 3= moderate/high complexity. The complete scoring rubrics appear in the Methodology section.

Expert judgments of assessment items and instructional tasks are translated into score points and reported here as percentages for each of the mathematical practices, first for grade 4 and then for grade 8.

Grade 4

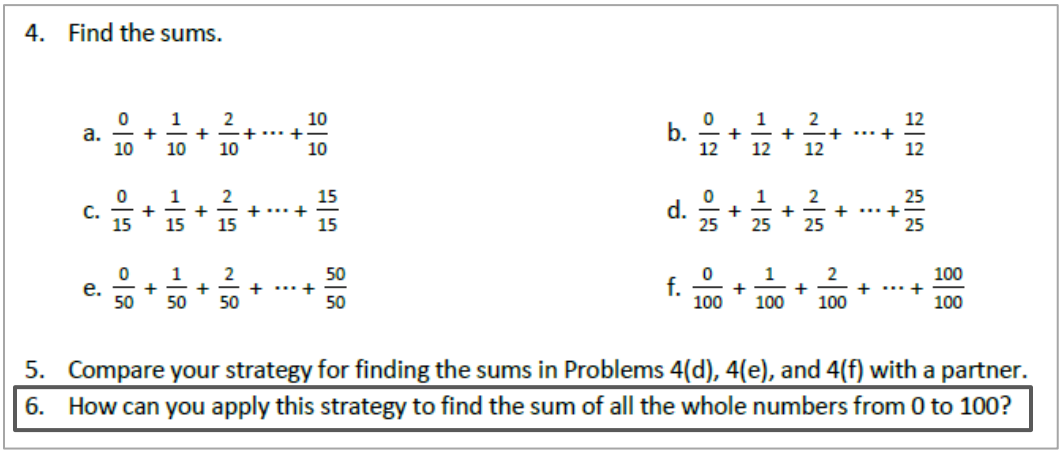
Problem Solving and Modeling Challenge. Figure 5a shows the distribution of score points across levels of Problem Solving and Modeling Challenge for grade 4. As was noted in the Methodology section, the assessment items attached to CEC materials are not necessarily exemplary in their representation of higher levels of complexity called for by mathematical practices. Consequently, the pattern observed in Figure 5a is one that repeats across many (though not all) of the analyses of mathematical practices. Only 7% of the Engage NY assessment score points and only 13% of the Investigations assessment score points derived from items that were rated at Level 3 on Problem Solving and Modeling Challenge. Higher percentages of points from instructional tasks were at Level 3 – 21% for Engage NY and 19% for Investigations. NAEP is similar to the instructional tasks, with 21% of its score points deriving from items rated as Level 3.

Figure 5a. Distribution of Score Points Across Levels of Problem Solving and Modeling Challenge: NAEP, CEC Instructional Tasks, and CEC Assessment Items, Grade 4



Note that none of the grade 4 materials evaluated received a rating of 4 in this practice domain, although some (apparently rare) examples do exist in the instructional materials. In the three-task sequence shown in Figure 5b, task 6 was selected by the leadership team as a Level 4 anchor item for the Problem Solving and Modeling Challenge rubric.

Figure 5b. An Anchor Item Selected to Represent Level 4 of the Problem Solving and Modeling Challenge Rubric



4. Find the sums.

a. $\frac{0}{10} + \frac{1}{10} + \frac{2}{10} + \cdots + \frac{10}{10}$

b. $\frac{0}{12} + \frac{1}{12} + \frac{2}{12} + \cdots + \frac{12}{12}$

c. $\frac{0}{15} + \frac{1}{15} + \frac{2}{15} + \cdots + \frac{15}{15}$

d. $\frac{0}{25} + \frac{1}{25} + \frac{2}{25} + \cdots + \frac{25}{25}$

e. $\frac{0}{50} + \frac{1}{50} + \frac{2}{50} + \cdots + \frac{50}{50}$

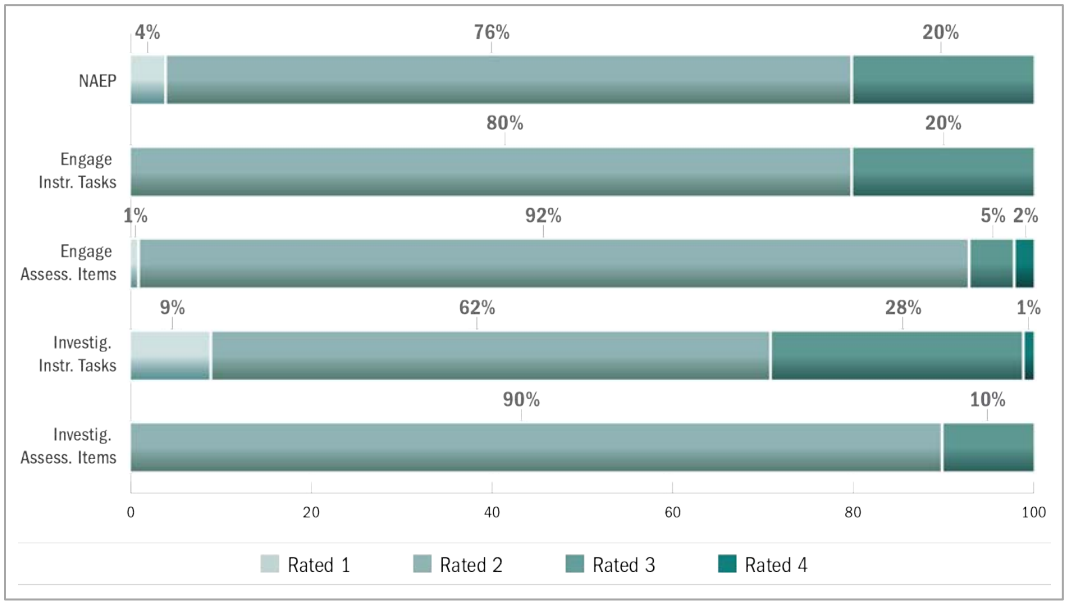
f. $\frac{0}{100} + \frac{1}{100} + \frac{2}{100} + \cdots + \frac{100}{100}$

5. Compare your strategy for finding the sums in Problems 4(d), 4(e), and 4(f) with a partner.
6. How can you apply this strategy to find the sum of all the whole numbers from 0 to 100?

SOURCE: EngageNY.org of the New York State Education Department. Grade 4, Module 5, Topic H, Lesson 41.

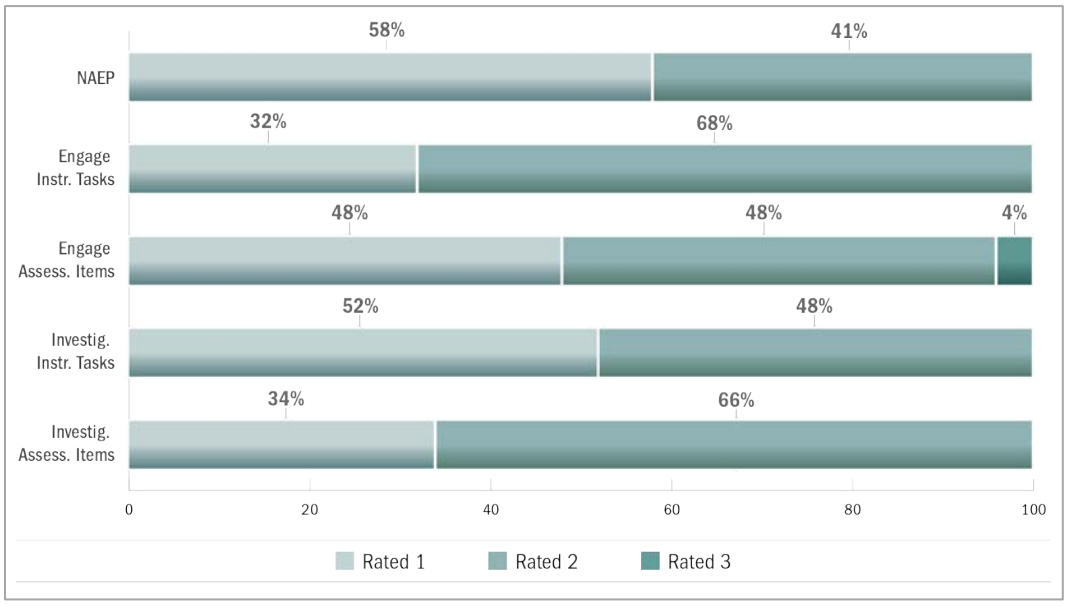
Depth of Conceptual Understanding. As seen in Figure 6, the distribution of Conceptual Understanding ratings at grade 4 is similar to the pattern for Problem Solving. Level 3 ratings occur at a much higher rate for CEC instructional tasks than for CEC assessment items, from 20% to 29% as compared to 7% to 10%. NAEP does nearly as well as the instructional tasks, with 20% of score points rated at Level 3. At the low end of the continuum, both NAEP and Investigations instructional tasks have some score points at Level 1, indicating that they derived from items that did not require the application of grade-level conceptual understanding. Level 4 ratings occurred rarely.

Figure 6. Distribution of Score Points Across Levels of Depth of Conceptual Understanding: NAEP, CEC Instructional Tasks, and CEC Assessment Items, Grade 4



Procedural Fluency. Procedural Fluency is different from the other mathematical practices in terms of the type of complexity required. The rating rubric was constrained to a 3-point scale, in which Level 1 was assigned to items with little or no procedural demand, often by design. The distinction between Level 2 and Level 3 depends on whether the items require common or grade-level procedures carried out with friendly numbers (Level 2), or, at Level 3, require one or more of the following – common or grade-level procedure(s), with unfriendly numbers; unconventional combination of procedures; or unusual perseverance or organizational skills in the execution of a procedure. The high percentages of Level 2 or 3 Procedural Fluency score points (ranging from 48% to 68%) for the CEC instructional tasks and assessment items show that these materials provide plenty of practice with grade-level procedures (Figure 7). Note, however, that this emphasis on Procedural Fluency is not at the expense of the other mathematical practices because the vast majority of items requiring Procedural Fluency also required Problem Solving and Modeling or Conceptual Understanding, or both.

Figure 7. Distribution of Score Points Across Levels of Procedural Fluency: NAEP, CEC Instructional Tasks, and CEC Assessment Items, Grade 4



Argumentation and Communicating Reasoning. Figure 8a presents the distribution of score points across levels of Argumentation and Communicating Reasoning. Both Investigations and Engage NY have significant percentages of score points derived from instructional tasks that call for higher levels of this mathematical practice. Ninety-two percent of NAEP score points are at Level 1, which means that little or no argumentation or communicating is required. Only 4% of NAEP score points are at Level 2, requiring that students "show work, explain how, or respond to given reasons;" another 4% are at Level 3, indicating that students are asked to "generate reasons" or "explain why a solution makes sense." This is in contrast to 19% of Engage NY instructional task score points, 20% of Engage NY assessment score points, and 18% of Investigations instructional tasks score points at Level 3 or above. Examples of Level 3 and Level 4 items are shown in Figures 8b and 8c, respectively.

Figure 8a. Distribution of Score Points Across Levels of Argumentation or Communicating Reasoning: NAEP, CEC Instructional Tasks, and CEC Assessment Items, Grade 4

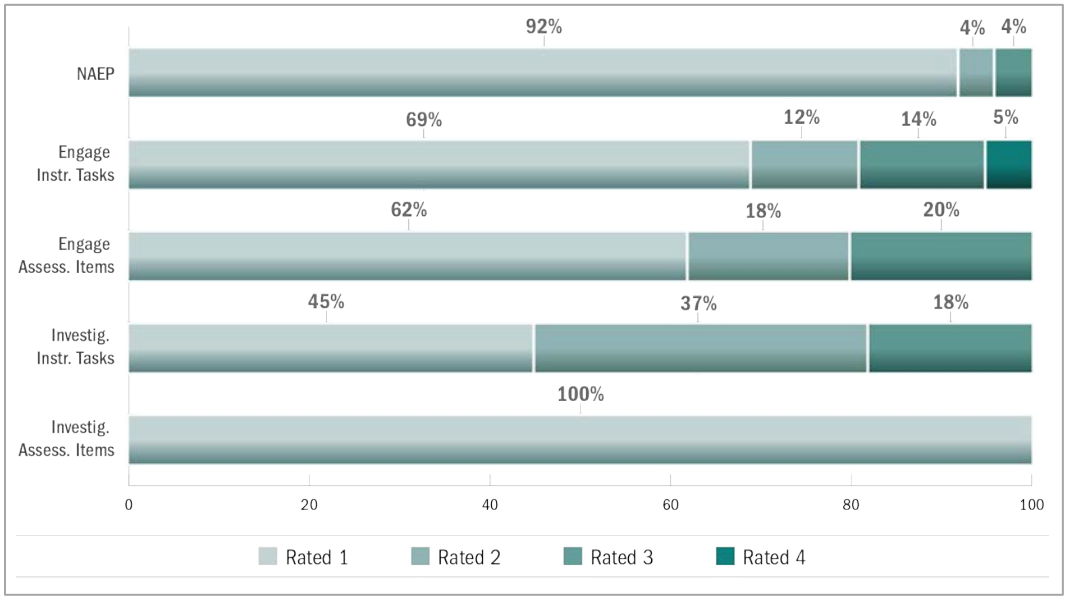
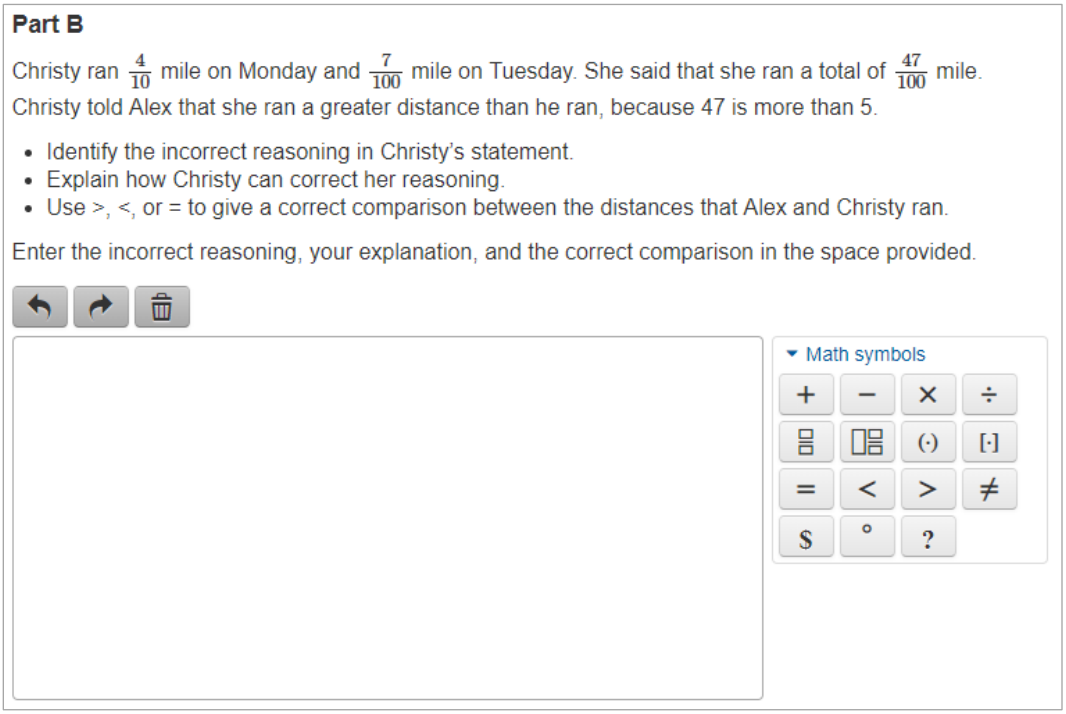


Figure 8b. An Anchor Item Selected to Represent Level 3 of the Argumentation or Communicating Reasoning Rubric



SOURCE: ©CCSSO, ILC 10/9. Reprinted with permission. All rights reserved. For more information, contact New Meridian Corporation.

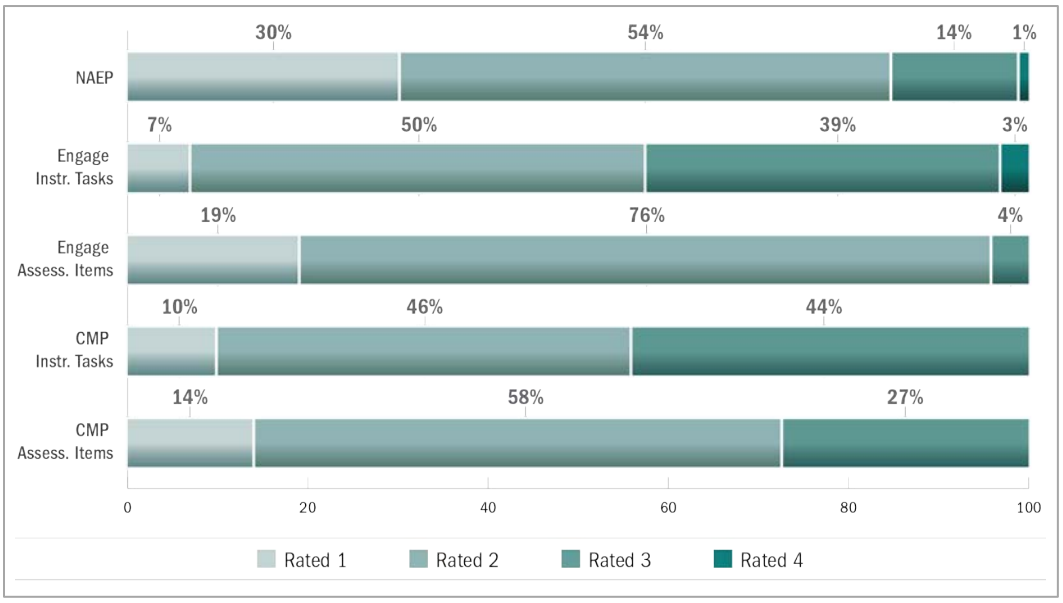
Figure 8c. An Anchor Item Selected to Represent Level 4 of the Argumentation or Communicating Reasoning Rubric

SOURCE: EngageNY.org of the New York State Education Department. Grade 4, Module 4, Topic A, Lesson 4.

Grade 8

Problem Solving and Modeling Challenge. As shown in Figure 9a, the pattern at grade 8 for Problem Solving and Modeling Challenge is very similar to that seen at grade 4. Instructional tasks from both CEC (Engage NY and CMP) are much better at assessing higher levels of complexity than are assessment items, with 42% to 44% of instructional task score points rated at Level 3 or higher. Level 3 of the Problem-Solving rubric specifies that students "decide what to do in non-routine situations and make sense of quantities or figures and their relationships implied by the problem posed."

Figure 9a. Distribution of Score Points Across Levels of Problem Solving and Modeling Challenge: NAEP, CEC Instructional Tasks, and CEC Assessment Items, Grade 8



Furthermore, 3% of Engage NY instructional task score points derive from items rated at Level 4, which means that students must actually be able to formulate a model. One example of model building would be developing an equation to represent a relationship, as called for in Exercise 3, highlighted in Figure 9b. This instructional task was scored at Level 4 by the study panelists.

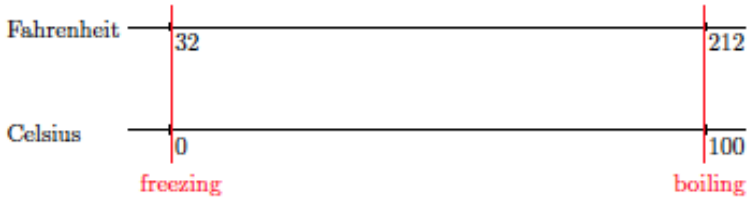
Twenty-seven percent of CMP assessment score points are also rated at Level 3 compared to only 15% of NAEP items that were rated 3 or higher. A major difference between the respective grade 8 and grade 4 ratings on Problem Solving is that many fewer grade 8 items are rated at Level 1. This is true for both instructional tasks and assessment items. (At grade 4, 36% to 52% of items were rated 1 on this mathematical practice, whereas the corresponding percentages for grade 8 are 7% to 30%.)

Figure 9b. A Sample Instructional Task Rated as Level 4 on the Problem Solving and Modeling Challenge Rubric

Mathematical Modeling Exercise

(1) If t is a number, what is the degree in Fahrenheit that corresponds to t°C?

(2) If t is a number, what is the degree in Fahrenheit that corresponds to $(-t)$°C?



Exercises

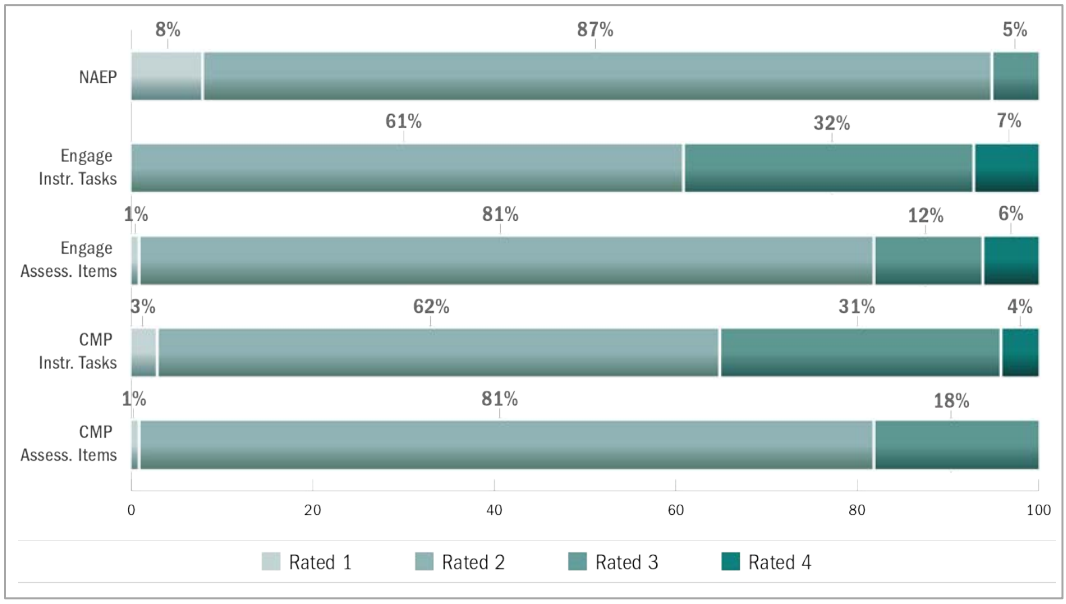
Determine the corresponding Fahrenheit temperature for the given Celsius temperatures in Exercises 1–5.

1. How many degrees Fahrenheit is 25°C?
2. How many degrees Fahrenheit is 42°C?
3. How many degrees Fahrenheit is 94°C?

SOURCE: EngageNY.org of the New York State Education Department. Grade 8, Module 4, Topic D, Lesson 30.

Depth of Conceptual Understanding. Figure 10 shows the distribution of score points across levels of Conceptual Understand for grade 8. As at grade 4, score points derived from items rated at Level 1 on this math practice are rare in both instructional tasks and assessment items. For higher levels of Conceptual Understanding, the pattern mirrors the pattern for Problem Solving. That is, compared to assessment items, instructional tasks have a greater percentage of score points at Level 3 or higher. Engage NY has 39% of its instructional task score points derived from items rated at Level 3 or 4, and CMP has 35%, whereas NAEP has only 5% of its score points rated a Level 3 or above.

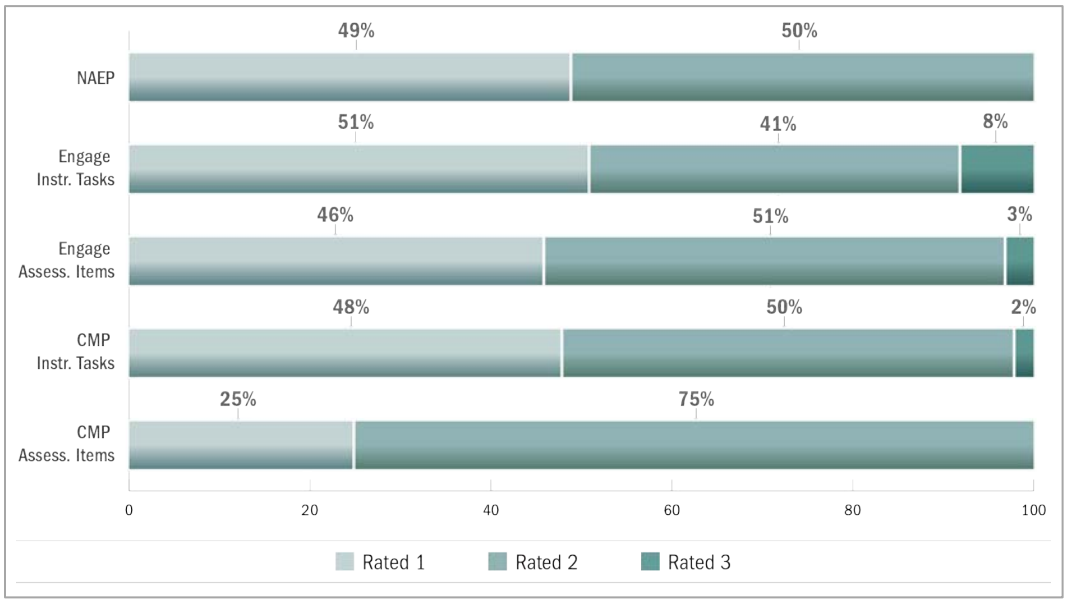
Figure 10. Distribution of Score Points Across Levels of Depth of Conceptual Understanding: NAEP, CEC Instructional Tasks, and CEC Assessment Items, Grade 8



Whereas items rated at Level 2 require that students "recall or recognize a concept, use it in a routine way, or match terminology to examples," Level 3 asks students to "adapt or extend a concept or apply it in an unfamiliar/nonroutine setting," and at Level 4 students must "use or explain a relationship among multiple concepts and/or show the conceptual basis for a strategy or procedure."

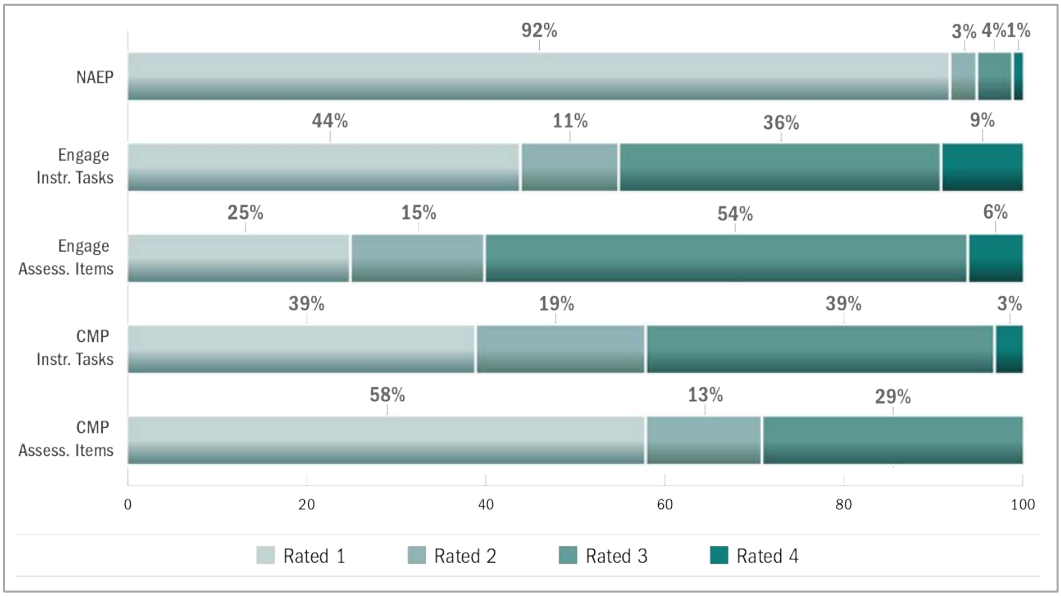
Procedural Fluency. At grade 8, a high percentage of score points from both instructional and assessment items demand Procedural Fluency (Figure 11), with 49% to 75% of score points derived from items rated at Level 2 or higher. Level 2 involves "common or grade-level procedures with friendly numbers"; 8% of Engage NY instructional task score points derive from items that go beyond this – asking students to handle "unconventional combinations of procedures" or to organize responses involving more than one set of procedures. As was mentioned in the discussion of grade 4, these percentages do not imply attention to procedures at the expense of other practices because the majority of items are rated above Level 1 on more than one practice dimension.

Figure 11. Distribution of Score Points Across Levels of Procedural Fluency: NAEP, CEC Instructional Tasks, and CEC Assessment Items, Grade 8



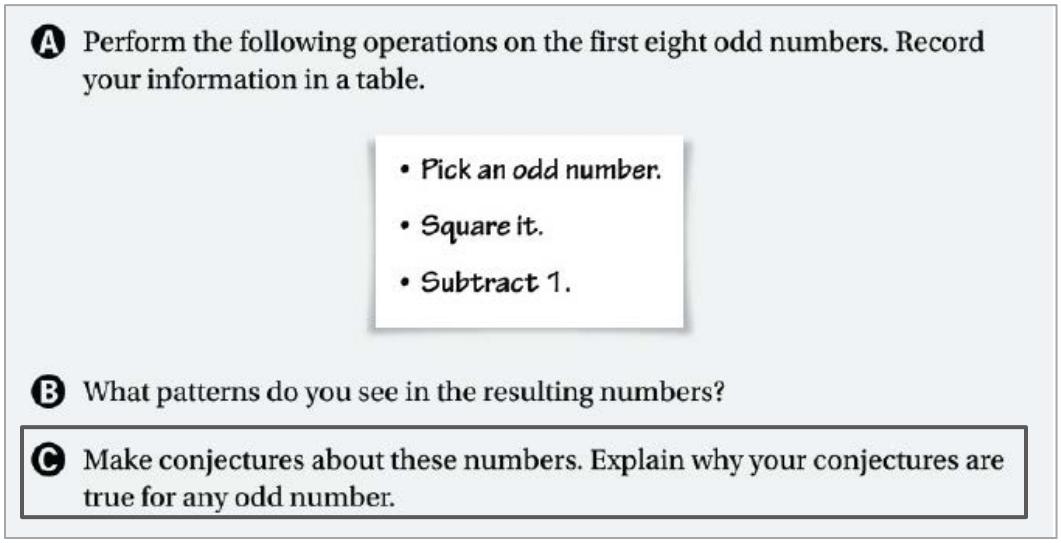
Argumentation or Communicating Reasoning. As was also true at grade 4, grade 8 NAEP has the greatest discrepancy compared to CEC instructional tasks in the mathematical practice area of Argumentation or Communicating Reasoning (Figure 12a). Only 8% of NAEP score points derive from items that ask students to show their work or explain how they solved a problem (Levels 2, 3, and 4 on Argumentation), and only 5% are at Level 3 or 4. By contrast, all the CEC instructional tasks and assessment items have substantial percentages of items at Level 3 or higher – ranging from 29% of CMP assessment score points to 60% of Engage NY assessment score points. (We have no explanation as to why, in this particular comparison, Engage NY's assessment items do better than its instructional tasks.)

Figure 12a. Distribution of Score Points Across Levels of Argumentation or Communicating Reasoning: NAEP, CEC Instructional Tasks, and CEC Assessment Items, Grade 8



Part C, highlighted in the instructional task shown in Figure 12b, is an example of an instructional task rated at Level 4 on Argumentation or Communicating Reasoning by the expert panel. As stated in the Argumentation rubric, Level 4 asks students to "construct a viable argument about the truth or generality of a mathematical statement that employs mathematical principles and/or logical argumentation."

Figure 12b. A Sample Instructional Task Scored as Level 4 of the Argumentation or Communicating Reasoning Rubric



SOURCE: From *A Guide to Connected Mathematics 3: Understanding, Implementing & Teaching* by Glenda Lappan, Elizabeth Difanis Phillips, James T. Frey, and Susan N. Friel © 2014 by Pearson K12 Learning, LLC, or its affiliates. Used by permission. All rights reserved.

CONCLUSIONS AND RECOMMENDATIONS

This study began with the acknowledgment that NAEP cannot swing wildly in response to new curricular developments because measuring change over time requires that assessment frameworks remain stable. In addition, and similar to international assessments, NAEP cannot favor one particular curriculum over another. However, the General Policy of the National Assessment Governing Board (2013) also recognizes that "as new knowledge is gained in subject areas, the information and communication technology for testing advances, and curricula and teaching practices evolve, it is appropriate for NAGB to consider changing the assessment frameworks and items to ensure that they support valid inferences about student achievement" (p. 6).

The NAEP item data presented in this CEC study are the same data that are presented by Daro et al. (2019). The difference is that, here, NAEP items are compared to CEC instructional tasks and assessment items; whereas four state assessments (two of which are consortium tests) form the basis for comparison in the Daro et al. study. With regard to the relative emphases given to mathematics content domains and subdomains by NAEP, the Daro et al. study provides the more authoritative basis for comparison because *state assessments are much more likely to determine what is taught* than are the apparent weights given to particular content domains in curricular materials. Curricula are built so that teachers can pick and choose units of instruction, so the percentage allocations by content domain in the full curricula, as published, do not necessarily reflect the proportion of teaching time allotted. *The nature of CEC instructional tasks and the extent to which they engage mathematical practices, however, is representative of current instructional practices*, given that teachers typically teach whole lessons, not just parts of lessons. Both the Daro et al. state assessment comparisons and this study's curricular comparisons go further in capturing likely instructional practices and opportunity to learn than prior studies focused only on the Common Core and only on intended standards.

Daro et al. (2019) identified a few important differences in content allocations between NAEP and state assessments that could lead to underestimates of student performance. At grade 4, NAEP has less emphasis on Fractions and Calculations by Place Value than state assessments, and more emphasis on Data. At grade 8, NAEP gives less weight to the two algebra domains than state assessments and more weight to Statistics & Probability and Number. These findings are consistent with findings from the prior standards-to-standards and items-to-standards comparisons (Daro et al., 2015; Hughes et al., 2013), and are understandable given the topics that are being emphasized earlier versus those delayed until later grades by Common Core–influenced curricula. These same patterns of divergence between NAEP and current instructional emphases are seen in this CEC study as well. At grade 4, consistent with Common Core–influenced standards, both Engage NY and Investigations have a much higher proportion of instructional-task score points devoted to arithmetic than is true for NAEP. Number & Algebraic Thinking, Place Value, and Fractions account for 75% of Engage NY and 74% of Investigations instructional task score points, compared to only 46% of NAEP score points allocated to these arithmetic domains. The differences are not quite so pronounced at grade 8 but are still consistent with

what would be predictable from Common Core influences. Engage NY and Connected Math instructional tasks place a greater emphasis on the two Algebra domains than NAEP, 50% and 54% respectively, compared to NAEP's 42%.

The greatest contribution of the CEC study is the insights it provides regarding the complexity dimension of NAEP and the extent to which NAEP is or is not assessing mathematical practices that are likely being taught in classrooms using CEC materials. This study found more serious discrepancies between NAEP and CEC instructional tasks on mathematical practices dimensions at grade 8 than were found between NAEP and state assessments in the Daro et al. (2019) study. Daro et al. found NAEP to be less sensitive than two of the state assessments in representing Argumentation or Communicating Reasoning at grade 4 and grade 8. This same discrepancy was found for CEC instructional tasks at both grade levels, but at grade 8 there were other substantial differences as well.

At grade 8:

- NAEP substantially underrepresented Problem Solving compared to CEC instructional tasks, with only 15% of score points rated at Level 3 or higher, while the two curricula had 42% to 44% instructional task score points at Level 3 or higher.
- NAEP substantially underrepresented Conceptual Understanding compared to CEC instructional tasks, with only 6% of score points rated Level 3 or higher, compared to 35% to 39% of CEC instructional task score points at these higher levels.
- For Argumentation or Communicating Reasoning, NAEP had only 5% of its score points derived from items rated at Level 3 or 4, compared to 42% to 45% of CEC instructional tasks score points.

At grade 4, NAEP underrepresented Problem Solving and Conceptual Understanding compared to CEC instructional tasks, but not to the same degree as was found at grade 8. At grade 4, the discrepancy was somewhat greater for Argumentation than for the other practices: NAEP had only 4% of score points at Level 3 (with none at Level 4), compared to 18% to 19% at Level 3 or 4 for CEC instructional tasks.

The challenging thinking, reasoning, and communication skills taught in the cutting-edge curricula selected for this study are consistent with findings from cognitive and learning sciences research documenting how deeper learning and 21st century skills are developed. PISA has for a decade been relying on this same research to inform development of its assessments, and a majority of states have adopted new standards that call for these more conceptual and strategic ways of engaging with mathematics. To the extent that NAEP does not include items that tap these more challenging levels of understanding, then NAEP will not be able to detect learning gains on these dimensions.

In parallel to the Daro et al. (2019) study, it is our hope that the findings provided here will be given serious attention as part of the current review of the NAEP mathematics framework and that the Governing Board will seriously consider creating a new framework and beginning a new trend.

REFERENCES

Daro, P., Hughes, G., Stancavage, F., Shepard, L. A., Kitmitto, S., Webb, D. C., & Tucker-Bradway, N. (forthcoming). *A Comparison of the 2017 NAEP Mathematics Assessment With Current-Generation State Assessments in Mathematics: Expert Judgment Study*. A publication of the NAEP Validity Studies Panel. San Mateo, CA: American Institutes for Research.

Daro, P., Hughes, G. B., & Stancavage, F. (2015). *Study of the alignment of the 2015 NAEP Mathematics items at grades 4 and 8 to the Common Core State Standards for Mathematics*. A publication of the NAEP Validity Studies Panel. San Mateo, CA: American Institutes for Research.

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.

Hughes, G. B., Daro, P., Holtzman, D., & Middleton, K. (2013). *A study of the alignment between the NAEP Mathematics Framework and the Common Core State Standards for Mathematics (CCSS-M)*. A publication of the NAEP Validity Studies Panel. San Mateo, CA: American Institutes for Research.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research*, *79*(4), 1332–1361.

Mathews, J. (2009, January 5). The rush for '21st-Century Skills.' *The Washington Post.*

Murnane, R. J., & Levy, F. (1996). *Teaching the new basic skills: Principles for educating children to thrive in a changing economy*. New York, NY: Free Press.

National Alliance of Business. (2002). *A nation of opportunity: Building America's 21st century workforce*. Washington, DC: Department of Labor.

National Assessment Governing Board. (2013). *General policy: Conducting and reporting the National Assessment of Educational Progress*. Washington, DC: Author.

National Assessment Governing Board. (2017). *Mathematics framework for the 2017 National Assessment of Educational Progress*. Washington, DC: Author.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards for mathematics*. Washington, DC: Author. Retrieved from http://www.corestandards.org/Math/

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Committee on Defining Deeper Learning and 21st Century Skills, J. W. Pellegrino & M. L. Hilton (Eds.). Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences in Education. Washington, DC: National Academies Press.

National Research Council. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: National Academies Press.

Organisation for Economic Co-operation and Development. (2017). PISA 2015 Mathematics Framework. In *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. Paris: Author.

Partnership for Assessment of Readiness for College and Careers. (n.d.). *PARCC high level blueprints – mathematics*. Retrieved from https://parcc-assessment.org/content/uploads/2017/11/PARCCHighLevelBlueprints-Mathematics_08.25.15.pdf

Porter, A. C., & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators* (CPRE Research Report Series, No. RR-048). Philadelphia, PA: Consortium for Policy Research in Education.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* (CSE Technical Report No. CSE-TR-566). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing

Sawyer, R. K. (Ed.). (2006). *The Cambridge handbook of the learning sciences*. New York, NY: Cambridge University Press.

Smarter Balanced. (n.d.). *Mathematics summative assessment blueprint*. Retrieved from https://portal.smarterbalanced.org/library/en/mathematics-summative-assessment-blueprint.pdf

Thissen, D. (2016, September). *Examining whether NAEP items are measuring the complexity of content measured by the items found in the Common Core consortia assessments*. Notes on the NAEP Validity Studies Panel linkage study. San Mateo, CA: American Institutes for Research.

Usiskin, Z. (n.d.). *Should the current NAEP Mathematics Framework be changed – And, if so, why and how*? Washington, DC: National Assessment Governing Board. Retrieved from https://www.nagb.gov/content/nagb/assets/documents/publications/reports-papers/frameworks/mathematics/usiskin.pdf

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 6). Washington, DC: Council of Chief State School Officers.

APPENDIX A. EXPERT JUDGES

Name	Affiliation
Christine Avila	Comprehensive Education Partners
Jessica Balli	Callahan Consulting
Hyman Bass	University of Michigan
Mary Bouck	Michigan State University
Diane Briars	Mathematics Education Consultant
Patrick Callahan	University of California, Los Angeles
Shelbi Cole	Student Achievement Partners
Jennifer Curtis	Emerald Education
Linda Ruiz Davenport	Boston Public Schools
Brad Findell	Ohio State University
Kaye Forgione	Independent consultant
David Foster	Silicon Valley Mathematics Initiative
Lawrence Gray	University of Minnesota
Meghan Hearn	Age of Learning & Notre Dame of Maryland University
Roger Howe	Texas A&M University and Yale University
Raymond Johnson	Colorado Department of Education
Grace Kelemanik	Fostering Math Practices
David Kirshner	Louisiana State University
Nancy Kress	University of Colorado Boulder
Solana Ray	Callahan Consulting
William McCallum	University of Arizona
Valerie L. Mills	Michigan State University
Frederick Peck	University of Montana
Morgan Polikoff	University of Southern California, Rossier School of Education
Mary Lynn Raith	Independent consultant; Pittsburgh Public Schools (Retired)
Diana Suddreth	Utah State Board of Education
Mary Jo Tavormina	University of Illinois at Chicago
Shawn Towle	Falmouth Schools
Kristin Umland	Illustrative Mathematics
Nicola Vitale	NYC Department of Education
Tad Watanabe	Kennesaw State University

NOTE: Affiliations are for individual identification purposes only and do not imply institutional endorsement or sponsorship of participation in this study.