

Preliminary Evidence on Measurement Characteristics for the Foundational Assessment of Competencies for Teaching Performance Tasks

ETS RR–20-27

Geoffrey Phelps
Brent Bridgeman
Fred Yan
Jonathan Steinberg
Barbara Weren
Jiawen Zhou

December 2020



ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

John Mazzeo
Distinguished Presidential Appointee

ASSOCIATE EDITORS

Beata Beigman Klebanov
Senior Research Scientist

Heather Buzick
Senior Research Scientist

Tim Davey
Research Director

John Davis
Research Scientist

Marna Golub-Smith
Principal Psychometrician

Priya Kannan
Managing Research Scientist

Sooyeon Kim
Principal Psychometrician

Anastassia Loukina
Senior Research Scientist

Gautam Puhan
Psychometric Director

Jonathan Schmidgall
Research Scientist

Jesse Sparks
Research Scientist

Michael Walker
Distinguished Presidential Appointee

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Gontz
Senior Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

Preliminary Evidence on Measurement Characteristics for the Foundational Assessment of Competencies for Teaching Performance Tasks

Geoffrey Phelps, Brent Bridgeman, Fred Yan, Jonathan Steinberg, Barbara Weren, & Jiawen Zhou

Educational Testing Service, Princeton, NJ

In this report we provide preliminary evidence on the measurement characteristics for a new type of teaching performance assessment designed to be combined with complementary assessments of teacher content knowledge. The resulting test, which we refer to as the Foundational Assessment of Competencies for Teaching (FACT), is designed for use as part of initial teacher licensure. Twenty elementary FACT performance tasks (10 for mathematics [MATH] and 10 for reading language arts [RLA]) were developed and then administered to 59 teacher candidates. The results from the pilot indicate that the performance tasks function as designed with candidates completing the tasks on average in approximately 3.5 min. Human raters were able to score the tasks quickly and accurately. All score points were well represented for all the scored tasks. Total scores for all tasks combined and subscores for reading language arts RLA and MATH had respective alpha reliabilities of .86, .77, and .79, with the scores well distributed across the scale. A large majority of teacher candidates participating in the study strongly endorsed the FACT tasks as authentic, assessing valuable competencies, and suitable for use as part of teacher licensure. These preliminary results indicate that the FACT performance tasks show great promise for use in large-scale, high-stakes testing programs that seek to provide evidence of both the knowledge and skills needed for effective teaching.

Keywords Teacher licensure; teacher assessment; teaching quality; reading language arts; mathematics

doi:10.1002/ets2.12310

Teacher candidates have long completed a licensure test to ensure that they can demonstrate the critical competencies needed to be a classroom teacher. Historically, licensure assessments have focused on the content that students are expected to master as part of the K–12 curriculum. Even today, most teacher licensure assessments are largely made up of questions that are remarkably similar to those found on student tests. This focus on K–12 content recognizes the simple reality that teachers themselves need to know the content that students are expected to learn.

However, contemporary research has demonstrated that teaching demands additional types of content proficiency that go far beyond a basic understanding of the student curriculum (see for example, Ball et al., 2008; Phelps, 2009). A focus only on the student curriculum leaves unassessed critical content knowledge used in the work of teaching. On a day-to-day and moment-to-moment basis, teachers engage in a wide range of content teaching practices, such as evaluating student thinking, selecting appropriate material for use in instruction, and deciding how to represent concepts to best support student learning, to name just a few. A new generation of content knowledge assessments, widely referred to as “content knowledge for teaching” (CKT), has emerged in the last few decades with a focus on assessing the full range of content knowledge that is needed to carry out the content-intensive work of teaching a school subject. CKT test designs start by identifying the foundational content practices or tasks that make up the work of teaching a subject. Assessment questions are then developed that focus on the content knowledge that is needed to carry out a particular content teaching practice or task (Phelps et al., 2020).

Although research on CKT has led to fundamental changes in the assessment of teacher quality, this approach has been limited to providing evidence of what teachers “know” about the content knowledge used in teaching. CKT assessments provide at best only indirect evidence of what teachers can actually “do” when carrying out the content practices that make up the interactive work of teaching a subject. This lack of attention to teaching performance leaves unassessed many critical content competencies that characterize effective teaching. These include practices or tasks such as modeling concepts to

Corresponding author: G. Phelps, E-mail: gphelps@ets.org

support student learning, explaining procedures, questioning to elicit student thinking, and so on. More comprehensive assessments of teaching competence, ones that go beyond assessing knowledge to also assess performance, need to provide evidence of both the content knowledge and teaching skills used in effective teaching.

In this report, we present results from the initial development of a new type of standardized performance task. These new performance tasks were designed to be administered along with CKT tasks as part of a new licensure test design, referred to as the Foundational Assessment of Competencies for Teaching (FACT). When administered together, the new performance tasks and the CKT tasks provide more complete evidence of the content competencies needed to carry out the work of teaching a school subject. This new design for licensure has a number of advantages over extant assessments that have historically characterized licensure testing. Because traditional licensure tests have focused on the content that students are learning, they provide little to no evidence of the professional knowledge and skill used in teaching a subject (Phelps & Sykes, 2020). FACT is designed to address these limitations with an assessment that provides more comprehensive evidence of the foundational content knowledge and teaching skills required for safe and effective beginning teaching.

The results presented in this report focus on the initial development of 20 FACT performance tasks. The main purpose of this work is to determine if these tasks show promise for use as part of licensure testing. To provide context and explain the motivation for the FACT design, we begin with a brief review of CKT, with an emphasis on the CKT assessment framework and the need for a new type of complementary performance task.

Practice-Based Assessment of Teaching Competence

Content Knowledge for Teaching

In the mid-1980s, Lee Shulman and his colleagues proposed a framework to serve as the basis for a series of tests making up the National Board of Professional Teaching Standards. Adding to the existing understanding that teachers need knowledge grounded in the discipline, Shulman (1987) introduced the notion of pedagogical content knowledge (PCK), a distinct domain of content knowledge that represents “the blending of content and pedagogy into an understanding of how particular topics, problems, or issues are organized, represented, and adapted to the diverse interests and abilities of learners and presented for instruction” (p. 8). The domain of PCK caught the attention of policy makers, teacher educators, and others interested in the standing of teaching as a profession with its own associated professional knowledge base.

Subsequent work on teacher content knowledge built off and refined the conceptual distinctions proposed by Shulman and his colleagues and eventually led to efforts to assess different components of teacher content knowledge (see, e.g., Gitomer *et al.*, 2014; Herbst & Kosko, 2014; Hill *et al.*, 2004; Kersting, 2008; Krauss *et al.*, 2008; Mikeska *et al.*, 2018; Phelps & Schilling, 2004; Sadler *et al.*, 2013; Smith & Banilower, 2015; Tatto *et al.*, 2008). CKT, one of the most widely referenced frameworks, was developed by Ball *et al.* (2008). The CKT framework is meant to encompass the full range of content knowledge used in the work of a teaching a school subject. The framework includes two main domains: content knowledge and PCK.

The idea that received the most attention was a subdomain of content knowledge referred to as “specialized content knowledge.” Ball *et al.* (2008) argued that teachers need to use a type of pure mathematical knowledge that is specialized to the work of teaching. Individuals who use mathematics in their regular lives, but do not teach mathematics, neither need nor use this type of specialized mathematics. Like PCK, specialized content knowledge is a form of professional knowledge that is only needed and regularly used by teachers. It follows that these are types of knowledge that are critically important to emphasize in both teacher preparation (because nonteachers will not have had opportunity to learn this content) and in licensure (because this is knowledge used in teaching a subject).

These efforts to conceptualize teacher content knowledge, and its components, were followed by a number of projects that set out to develop assessments of CKT. All of these projects shared a common focus on assessing types of content knowledge unique to the work of teaching a school subject—what might be generally labeled as “professional content knowledge.” These efforts to develop new assessments have been located across the K–12 education space. They have included elementary or middle school mathematics (see for example, Hill *et al.*, 2004; Kersting, 2008; Phelps *et al.*, 2014; Tatto *et al.*, 2008), science (Mikeska *et al.*, 2017; Mikeska *et al.*, 2018; Sadler *et al.*, 2013; Smith & Banilower, 2015) and RLA (Carlisle *et al.*, 2009; Phelps *et al.*, 2014; Phelps & Schilling, 2004). While the work at the secondary level has been more limited, CKT assessments have also been developed for algebra and geometry (Herbst & Kosko, 2014; Howell

et al., 2016; Krauss et al., 2008; McCrory et al., 2012; Mohr-Schroeder et al., 2017; Phelps et al., 2014), physics (Phelps et al., 2020), and English (Phelps et al., 2014).

The majority of these assessments were developed to investigate different components of CKT, how CKT develops, or how teachers' CKT contributes to effective teaching or student learning. Again, there is a large body of research that spans subjects and grade levels. For example, CKT scores for contrasting groups (including prospective teachers, practicing teachers, and nonteachers) have been compared to support claims that CKT is a form of professional knowledge (Hill et al., 2007; Iaconangelo et al., 2020; Kleickmann et al., 2013; Krauss et al., 2008; Phelps, 2005, 2009; Phelps et al., 2019). Assessments of CKT have been used to study and evaluate professional development to show that CKT assessments are sensitive to professional learning opportunities (Goldschmidt & Phelps, 2010; Hill & Ball, 2004; Liu & Phelps, 2019; Phelps et al., 2016; Tröbst et al., 2018; van Driel et al., 1998).

Researchers have also used CKT assessments to explore whether there is empirical evidence that CKT is composed of different knowledge types (Copur-Gencturk et al., 2018; Hill et al., 2004; Mikeska et al., 2018; Phelps, 2009; Phelps & Schilling, 2004). A large number of studies have investigated whether teachers' CKT contributes to either teaching quality or student learning outcomes (Baumert et al., 2010; Carlisle et al., 2009; Correnti & Phelps, 2010; Hill et al., 2005; Hill et al., 2008; Kersting et al., 2012; Phelps et al., 2012). Although the results are somewhat mixed across studies, when taken together this body of research provides strong evidence that different components of CKT can be assessed and that CKT assessments can measure a type of professional content knowledge that is associated with teaching quality.

Using Evidence Centered Design to Assess CKT

The research cited in the previous section provides compelling evidence to support interest in CKT for a variety of uses including foundational research on content knowledge, evaluating teacher education and professional development, studying and supporting teacher learning, and assessing teachers for licensure. Although these advances in both conceptualization and assessment of CKT represent a clear step forward from prior work that simply assessed teachers' knowledge of the student curriculum, the actual assessment approaches that have been used are quite variable and have not yet coalesced around a unified theory of teacher competency or any one assessment approach. Indeed, many of the assessment development efforts have taken somewhat of an ad hoc approach to test design, with the main emphasis placed on sharing interesting assessment tasks that illustrate types of CKT and less attention on carefully articulated design principles (Phelps et al., 2020).

This general lack of attention to systematic principles of test design limits the potential for both interpreting the results across the types of studies referenced above and for future efforts to develop programs of assessment that can be applied consistently across subjects and grade levels. To address these issues, Phelps et al. (2020) recommended an evidence centered design (ECD) approach (e.g., Mislevy et al., 2003). In this approach, "assessment design, item development, scoring, and reporting are all considered with reference to specified knowledge, skills, and abilities in order to develop assessments that support explicit interpretations about score meaning" (Phelps et al., 2020, p. 2).

In their argument for using ECD as the basis for assessing CKT, Phelps et al. (2020) distinguished between *componential* and *integrated* approaches to assessing teacher knowledge:

In a componential model, relevant aspects of knowledge are identified, and specific assessment tasks are designed to assess discrete knowledge components ... Adopting such an approach to assessing teaching breaks complex performances into discrete skills, and the associated knowledge can fall short in capturing the coordination among skills and the range of knowledge that is necessary to carry out an integrated task. ... [In an integrated approach] teachers must use their knowledge (CKT) to engage in multiple and coordinated tasks such as asking questions, interpreting student responses, providing explanations and helping students develop their own explanations, asking follow-up questions, promoting discussions, etc. ... Rather than starting with a taxonomy of discrete knowledge types, this approach starts with a taxonomy of knowledge as it is combined to execute the more complex tasks encountered in teaching. (p. 3)

To support an integrated approach to assessing CKT, Phelps et al. (2020) argued for assessment frameworks that specify both tasks of teaching and student learning targets.

Tasks of teaching describe how teachers work with content in their moment-to-moment interactions with students. Tasks of teaching are the core features of content teaching that are frequent and recurrent across lessons, content areas and grade levels. They provide a common structure to describe CKT generally and are then elaborated in ways specific to subject matter teaching that supports students in developing the core concepts ... [Learning] targets then provide a starting point for modeling CKT by identifying the core concepts that students should come to understand through instruction in a domain. (Phelps *et al.*, 2020, p. 6)

This practice-based approach provides a means to directly link student learning, tasks of teaching, and teacher content knowledge. CKT and the associated assessment tasks are defined at the intersection of learning targets and tasks of teaching.

Complementing CKT With Performance Tasks

To make the underlying framework for the assessment of CKT more concrete, it is useful to refer directly to a CKT framework that was developed using the ECD approach described previously. The Measures of Effective Teaching (MET) project (Phelps *et al.*, 2014) took this approach and began by specifying both higher level and subject specific tasks of teaching (Appendix A). The MET framework helps to illustrate two important features that illustrate the need for developing a new type of performance assessment.

Unlike the majority of CKT assessment projects that have focused only on a single subject, with assessment frameworks that only refer to the topical organization of that subject, the MET project set out to develop a common or shared framework that can be used across subjects (Gitomer *et al.*, 2014; Phelps *et al.*, 2014). As a first step, general tasks of teaching that are shared across subjects are identified. In a second step, these higher level tasks of teaching are elaborated with subject-specific examples. While these subject-specific examples guide actual test development, inferences can be drawn up to the more general level and then interpreted in similar ways across subjects. Given the potential to support the development and use of coordinated programs of assessment, these cross-subject frameworks are of particular interest in developing licensure assessments that need to account for multiple different grade levels and subjects.

The MET framework also serves to illustrate that assessments of content knowledge alone do not provide adequate evidence of the content competencies called on in the practices or tasks that make up the work of teaching a school subject. Even though the MET framework and associated tasks of teaching were explicitly developed to support the assessment of teacher content knowledge, the framework still demonstrates the need for task types that can assess teaching performance. Consider, for example, the following three tasks of teaching from the MET framework: (a) explaining concepts, procedures, representations, models, examples, definitions, and hypotheses; (b) creating and adapting resources for instruction; and, (c) developing questions, activities, tasks, and problems to elicit student thinking (see Appendix A). Each of these tasks of teaching invoke competencies (e.g., explaining, creating, developing) that are directly tied to a teaching performance. Although assessment tasks that focus on knowledge can provide, for example, evidence on whether a teacher candidate can evaluate an explanation, such knowledge tasks cannot provide strong evidence of whether a candidate can actually deliver a high-quality explanation. To provide strong evidence of teaching, performance tests need to include tasks that not only focus on, but also elicit evidence of teaching performance.

The FACT design represents a more comprehensive approach to assessing the teaching competencies (i.e., knowledge and skill) that are used in the work of teaching school subjects. The new performance tasks complement CKT by sharing the same basic task of teaching framework. But they also serve to provide evidence of currently unassessed performance skills or competencies.

FACT Pilot

Research Questions

The FACT pilot was designed to provide information on the administration, scoring, and measurement characteristics of the newly developed FACT performance tasks. The overarching goal of the pilot was to test the viability of the FACT performance tasks for use in licensure testing and to collect information to guide platform and task revision for use in future large-scale development efforts and associated research studies. In this pilot, we investigated the following six research questions.

1. Can the tasks be administered as intended?
2. Can the tasks be scored accurately and efficiently?
3. Is there a reasonable distribution of scores within and among tasks?
4. Do the tasks provide reliable measures of the competencies assessed?
5. Are the measures associated with participants' academic and professional attainment?
6. How do participants view the tasks and the testing experience?

FACT Performance Task Development

A number of assumptions about large-scale licensure testing informed the design criteria for FACT performance tasks. Some of the criteria are purely practical and are ultimately tied to the cost of administering an assessment at scale at a price point that is affordable for a teacher candidate. Other considerations focus on the types of candidate performance that the tasks need to elicit and the types of construct-relevant evidence that they need to provide. The design criteria are listed here along with brief explanations.

- *Developed quickly and at scale.* Our immediate goal is to use FACT tasks as part of licensure testing. In order to be suitable for this use, the tasks need to be based on design criteria that support developing many similar tasks. The development process needs to be relatively straightforward, accessible, and learnable so that assessment development can be accomplished by teams with somewhat diverse expertise. Finally, task development needs to be relatively quick without adding many layers of review or revision beyond those that are necessary (e.g., to promote fairness, accessibility, and appropriateness of content), so that the development costs are reasonably constrained.
- *Administered in short amount of time.* It is important that tasks are short enough to allow for assessing a sufficient range of the domain topics. Including many tasks also increases the potential for creating tests that have sufficiently high reliability and decision accuracy for high-stakes uses such as a licensure decision.
- *Performance components scored quickly and accurately.* Human scoring is a major cost driver for any testing program. In order to contain the cost of a performance test with human scoring, it is critical to keep scoring time for each task as short as possible. For performance tasks that require viewing a video performance, the length of the performance and associated viewing time creates a lower bound on scoring time. Therefore, to constrain overall scoring time, the performance time needs to be relatively short. The other main factor that determines scoring time is the complexity of the scoring rules. Simple scoring rules not only reduce scoring time but can also reduce scoring error and improve associated test reliability and decision accuracy.
- *Authentic performance like actual teaching.* The FACT construct, as defined by the tasks of teaching framework, calls for eliciting a teaching performance. The short performances assessed by FACT should represent valid decompositions and approximations of teaching practice that are suitable for eliciting test taker performances that provide evidence of the desired teaching skills.
- *Test center administration is feasible.* Large-scale, high-stakes assessments, such as those used in teacher licensure, are typically administered in proctored environments. FACT tasks need to be designed to assess critical performance competencies, such as the use of a white board as part of a teaching performance, while placing as little additional burden as possible on the cost of administration.

The performance tasks that were developed for the FACT pilot included a number of standard design features. All tasks are presented on a single page (or computer display) that is broken into two panels. The left panel presents a teaching situation (which sometimes also includes a short video) and directions for the performance. The right panel is a virtual white board that allows the candidate to record written work along with a spoken performance. In order to ensure that tasks are completed within the allocated time, both the preparation and performance are timed. These design components are illustrated and described in Figure 1. All 20 tasks developed for the FACT pilot follow this basic design logic (see Appendix B for screenshots of the 20 FACT performance tasks).

FACT performance tasks require a touch-sensitive monitor to support stylus entry on the virtual white board. The pilot administration discussed in this report made use of touch screen-enabled laptops placed in tablet mode (Figure 2). FACT tasks were loaded onto the laptops and then performances were captured on the laptops and later uploaded to a secure project website. FACT tasks can also be delivered online using a web-browser delivery platform and a touch screen-enabled tablet device such as an iPad.

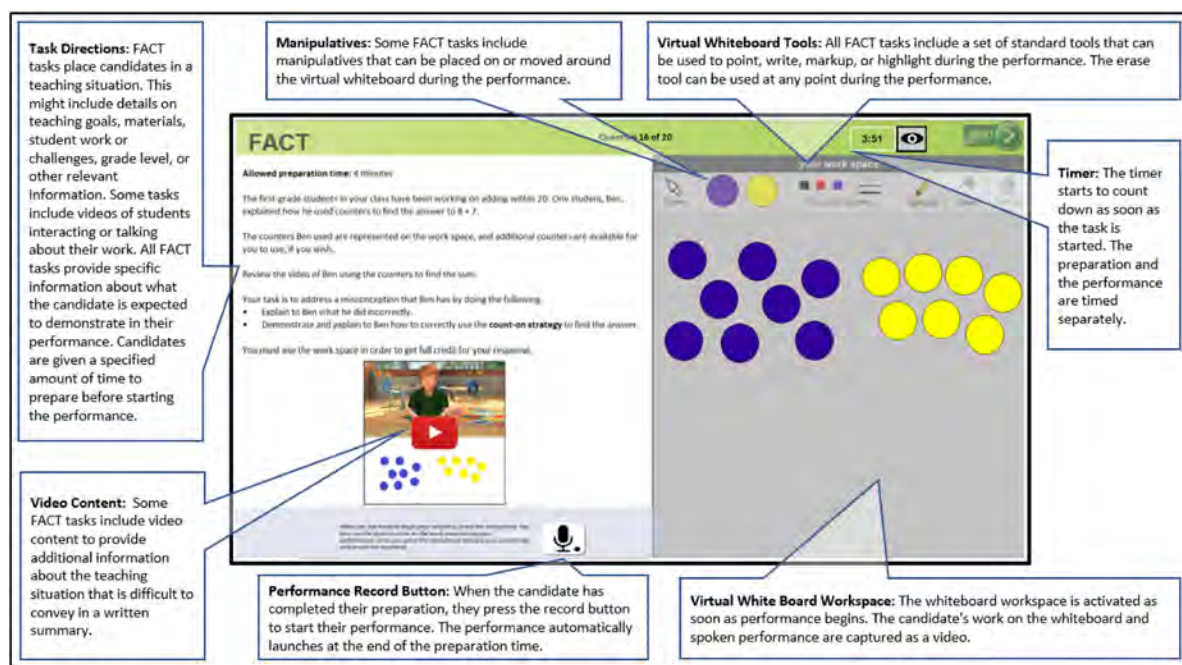


Figure 1 FACT task components. For additional examples of FACT tasks see Appendix B.

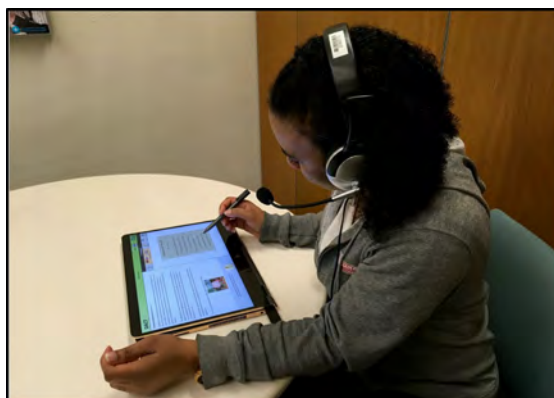


Figure 2 FACT delivery on a touch screen laptop in “tablet mode.”

The 20 FACT tasks administered in the pilot included 10 tasks that focused on elementary (RLA) and 10 tasks that focused on elementary mathematics (MATH). Table 1 summarizes the tasks of teaching and the content topics that were the primary focus of the 20 FACT tasks administered in the pilot.

The 20 FACT tasks included some variable design features. For example, some FACT tasks included a video that presented additional information on the teaching situation. Based on small-scale usability trials and the judgment of the development team, different amounts of time were allocated for preparation. Tasks also differed in characteristics such as whether the digital white board was required for the performance and whether manipulatives were offered. The variable design characteristics for the 20 tasks are summarized in Table 2.

Pilot Study Instruments

Each participant in the FACT pilot completed the 20 FACT tasks described previously and presented in Appendix B. Four different form variants were created. These form variants administered blocks of five RLA and five MATH tasks in different orders within and across the two subjects. The form blocks use the item numbering conventions as follows: RLA Block 1 = RLA 11, RLA 12, RLA 13, RLA 14, RLA 15; RLA Block 2 = RLA 21, RLA 22, RLA 23, RLA 24, RLA 25;

Table 1 Tasks of Teaching and Content Topics

Task	Task of teaching	Content topic
RLA 11	Model reading or writing process	Sentence structure and punctuation
RLA 12	Explain process/concept using student work	Reading fluency
RLA 13	Explain process/concept using student work	Poetry and evidence for theme
RLA 14	Explain process/concept using student work	Persuasive writing and use of examples
RLA 15	Model or explain reading process or concept	Graphophonemic segmentation
RLA 21	Model or explain reading process or concept	Word roots and affixes
RLA 22	Model how to improve writing	Sentence structure and punctuation
RLA 23	Explain process/concept using student work	Main idea and theme
RLA 24	Model how to improve writing	Narrative writing sensory details
RLA 25	Model reading or writing process	Developmental spelling and segmenting
MATH 11	Explain and model mathematical concepts	Line of symmetry
MATH 12	Question to elicit student understanding	Decimal addition
MATH 13	Represent concepts using mathematical model	Equivalent fractions and area models
MATH 14	Explain student error and model correct method	Two-digit addition with regrouping
MATH 15	Represent concepts using mathematical model	Partitive and measurement models of division
MATH 21	Explain student error and model correct method	One-digit addition
MATH 22	Explain problems using a representation	Fraction word problems and number line
MATH 23	Compare student methods	Two-digit addition with regrouping
MATH 24	Explain mathematical concepts	Composite numbers and factors
MATH 25	Represent concepts using mathematical model	Fraction set models

Note. MATH = mathematics; RLA = reading language arts.

Table 2 Task Design Characteristics

Task	Grade level	Video stimulus	Time allowed ^a		Virtual white board workspace		
			Preparation	Performance	Required	Manipulatives	Material provided ^b
RLA 11	1		2	2.5	Yes		Yes
RLA 12	3	Yes	3	2.5			Yes
RLA 13	6	Yes	4	2.5	Yes		Yes
RLA 14	4		3	2.5	Yes		Yes
RLA 15	K		2	2.5	Yes		Yes
RLA 21	4		2	2.5	Yes		Yes
RLA 22	2		2	2.5	Yes		Yes
RLA 23	5		4	2.5	Yes		Yes
RLA 24	4		3	2.5	Yes		Yes
RLA 25	1		3	2.5	Yes		Yes
MATH 11	4		3	2.5	Yes		
MATH 12	5	Yes	4	2.5			Yes
MATH 13	3		4	2.5	Yes		Yes
MATH 14	2	Yes	5	2.5	Yes		Yes
MATH 15	3	Yes	4	2.5	Yes		
MATH 21	1	Yes	4	2.5	Yes	Yes	Yes
MATH 22	4		4	2.5	Yes		Yes
MATH 23	2	Yes	4	2.5			Yes
MATH 24	4		4	2.5	Yes		
MATH 25	3		4	2.5	Yes	Yes	Yes

Note. MATH = mathematics; RLA = reading language arts. ^aPreparation and performance time are in minutes. The total time allowed for completing a task is the sum of preparation and performance time. ^bThis category includes student work, classroom assignments, text passages, mathematics problems, teaching tools, or any other material that is provided in the workspace for the candidate to interact with during their performance.

Table 3 Participating Educator Preparation Programs and Number of Participants

Program type	Location	Participants
State College/University	NJ	24
State College/University	CT	15
State College/University	NJ	9
Historically Black College/University	PA	7
Historically Black College/University	MD	4

MATH Block 1 = MATH 11, MATH 12, MATH 13, MATH 14; and, MATH Block 2 = MATH 21, MATH 22, MATH 23, MATH 24, MATH 25. Forms 1 and 2 started with RLA and Forms 3 and 4 started with MATH alternating the order of the five item blocks within subjects. The form variants were randomly assigned when participants were scheduled for an administration to help account for order and fatigue effects.

Participants also completed a preadministration background survey and postadministration perception survey. The background survey asked participants to provide information on their educational attainment, minor and major degree field(s), the type of teacher education program attended, their intended area(s) of certification, whether they had completed student teaching, grade and subject they planned to teach in the future, any teaching experiences, and their race/ethnicity and gender. The postassessment perception survey focused on the test and testing experience (e.g., value, relevance, and authenticity of the FACT tasks, how well the administration platform functioned, adequacy of the time provided).

Administration and Study Sample

Recruitment for the FACT tryout was conducted in two phases. First, five partner education preparation programs (EPPs) were recruited. The EPPs agreed to distribute information about the study to teacher candidates attending their program and to provide space at the EPP where the FACT assessment could be administered. Once candidates received information about the FACT study, they reached out directly to ETS via email. Candidates were then provided additional information about the requirements and the honorarium provided for participation. Candidates who wished to enroll in the study completed an online eligibility survey with their contact information, gender, race/ethnicity, current year in the teacher preparation program, and dates and times they were available to complete the FACT assessment. Once their enrollment in an appropriate teacher preparation program was confirmed, candidates received consent materials and an appointment for the administration. The administration took place at the candidate's EPP and was proctored by FACT research staff. A two-hour block was allowed for administration of the background survey, 20 FACT tasks, and the postassessment perception survey. After completing the surveys and assessment tasks, participants received a \$100 gift card.

FACT tasks and surveys were administered on a touch screen enabled laptop.¹ The performance records were then stored on the laptop and later processed into a video that combined the spoken and white board performance. Each administration generated 20 separate performance video files and a file that recorded task preparation time, performance time, and total time taken to complete each task. The surveys were captured independently from the video performance and timing files. Video and timing files for each performance were uploaded to a secure server at the end of each testing day.

A total of 63 candidates were enrolled. Four candidates canceled and were not assessed. A sample of 59 candidates attending five different EPPs in four different states participated in the study (Table 3).

The final sample was predominantly female (93%) and White (58%). Over 20% of the sample was African American, and there was also substantial representation from Asian and Hispanic candidates. While the majority of the sample attended an undergraduate teacher preparation program and had yet to receive a bachelor's degree, a substantial proportion of the sample had already completed a bachelor's degree (42.4%) and were attending a postbaccalaureate preparation program (44.1%). Additional information on the study sample is provided in Table 4.

Scoring Procedures

It is important to recognize that scoring materials and procedures are an integral component of the FACT performance tasks. It is easy to consider the tasks presented during testing as the whole of the developed product. However, the real

Table 4 Demographics of Study Participants ($n = 59$)

Demographic	Percent
Gender	
Female	93
Male	7
Race/Ethnicity	
Asian or Asian American	12
Black or African American	20
Hispanic/Latinx	7
White	58
Other	3
Highest education level	
Master's or greater	5
Bachelor's	37
Senior	36
Junior	17
Sophomore	5
Undergraduate major(s) ^a	
Early childhood education	14
Elementary education	73
Special education	10
Arts/Humanities	15
Math	9
Social sciences	14
Teacher preparation program	
Master's degree program	44
Undergraduate degree program	56
Intended certification(s)	
Pre-K	14
K–6	98
5–8	24
9–12	14
Completed student teaching	
Yes	83
No	15
Not responding	2

Note. For a number of demographic categories, participants were asked to select all that apply, and for these categories, the percentage for the groups may total to more than 100%. ^a Arts/Humanities includes English, art and foreign languages; social sciences includes social studies, history, psychology and sociology.

product for any testing program also includes the training procedures for raters and the scoring rules for assigning scores to performances. The score is the result of the task as delivered, the performance it elicits, and the scoring routines that are applied to score the performance. Because scoring accuracy is typically a substantial source of error variance in an overall test score, the components of the task that support accurate human scoring are among the most vital in developing a reliable assessment. Similar to assessment tasks, where development often occurs over multiple pilots with associated task revisions, there are associated development processes for scoring materials that also occur over multiple pilots. A companion report (Weren et al., 2020) provides a detailed account of the formative scoring process used to learn about and improve the scoring procedures and materials for the FACT performances. Although the research methods and findings from the formative scoring study are beyond the scope of this report, they are important to consider in order to fully understand the FACT performance task design, development process, and subsequent task revisions.

Task Selection

Out of the 20 tasks administered in the pilot, 14 were selected for scoring. This decision was made primarily to reduce the costs associated with reviewing and selecting performances for use in rater training, with developing task specific scoring materials, and with compensating raters. In order to identify 14 tasks for scoring, each of the 20 administered tasks were

reviewed by four members of the relevant subject matter task development teams and the four proctors who administered FACT to the teacher candidates. The following criteria were used in the review process to identify six tasks (three RLA and three MATH) to be removed from the scoring pool.

- *Similar content to other tasks.* When two tasks have similar content and are likely to provide somewhat overlapping evidence, select the less effective or otherwise weaker task.
- *Construct irrelevant features.* Select tasks that might measure some irrelevant competencies (e.g., higher reading load), present some undesired cognitive challenge, or compared to other tasks, lack some clarity around what the task is asking for.
- *Idiosyncratic, rarified or potentially contested competency.* Select tasks that assess teaching competencies that are rarely used or are only associated with a particular teaching ideology (i.e., an approach or view that is not widely endorsed).
- *Authenticity.* Select tasks where the performances do not feel as authentic as other tasks which clearly depict work that teachers do.
- *Not a beginning competency.* Select tasks that focus on competencies that may be too advanced (i.e., not reasonable for beginning teaching).
- *Scoring complexity.* Select tasks that might be more difficult to score consistently compared to other tasks.

Each reviewer recommended three tasks from their relevant subject matter area of focus to not score based on one or more of the criteria listed above. These recommendations were then collated, and comments from all reviewers were considered in making final decisions on the six tasks to not score. The remaining 14 tasks were selected for scoring (seven RLA and seven MATH).

Performance Selection Procedures

The steps described below were repeated for each of the 14 tasks to develop task-specific rater training materials. First, two members of each subject matter team were each assigned roughly half of the performances for a given task to review. For this initial review each team member selected seven performances for potential use as training cases. The seven performances were selected to represent a 1 (*benchmark*); a 1 (*high rangefinder*); a 2 (*low rangefinder*); a 2 (*benchmark*); a 2 (*high rangefinder*); a 3 (*low rangefinder*); and a 3 (*benchmark*). Benchmarks were defined as performances that are clear illustrations of the evidence that characterizes a particular score point. Rangefinders were defined as performances that provide evidence of more than one score level and help illustrate the boundaries between adjacent score levels (e.g., a 2 (*high rangefinder*) would contain some evidence for a score point 3 and some evidence for a score point 2).

Once the two members of each review team selected seven performances that fit these criteria, those performances were traded among the two reviewers for another round of scoring. The 14 selected performances were also scored by one or two other members of the development team. This process resulted in at least three independent scores for each of the performances identified through the initial performance review. The team members then met and discussed the ratings for the identified performances and assigned a consensus score for each performance. The team also noted changes needed to the task specific scoring guide and/or comments to be included in the rater training. After completing these steps, seven performances were selected to serve as the training responses for the target scoring points noted above. The selected training responses were then annotated indicating relevant evidence from the performance and how it aligned with elements of the rubric. The reasoning used to reach the consensus score was also documented (e.g., considering the preponderance of evidence for rangefinder performances). This process was repeated for all 14 tasks.

In addition to the above steps used to develop the rater training materials, the seven identified performances for each task not selected for use as training cases were then reviewed and, for each task, four to seven consensus-scored responses were selected for inclusion as validity responses. This selection needed to include at least one response at each of the three score points (1, 2, and 3). These validity responses were later mixed in with the unscored responses and were assigned to all raters to score (raters were blind to whether a performance was a “score” or validity” case). Rater scores and consensus scores for the validity response were later compared to gauge how well raters were conforming to the developers’ conceptions of what constitutes different levels of performance for each of the tasks.

Table 5 Performance Sample for Each Task

Task	Training ($n = 107$)	Validity ($n = 74$)	Score ($n = 611$)	Other ($n = 6$)
RLA 11	8	6	43	
RLA 12	7	6	44	
RLA 13	8	6	43	
RLA 14	7	5	45	
RLA 15	7	5	45	
RLA 21	8	6	43	
RLA 24	9	7	41	
MATH 11	8	5	44	
MATH 13	9	4	44	
MATH 14	7	7	37 ^a	6 ^b
MATH 21	8	4	45	
MATH 22	7	5	45	
MATH 23	7	4	46	
MATH 24	7	4	46	

Note. MATH = mathematics; RLA = reading language arts. ^a Six performances for task MATH 14 intended for use as score cases, were unintentionally left out of the scoring pool. These six tasks were consensus scored and included in all subsequent task and score analysis.

^b Six training cases for task MATH 14 were unintentionally also assigned to the score cases. These six performances were used as training cases as intended. However, these 6 training cases are also included in the agreement analyses. Therefore, for the agreement analyses, there are 43 scored performances shown for task MATH 14.

Scoring

Fourteen raters with experience scoring for one or more of the ETS elementary teacher licensure tests were recruited to score FACT tasks. The scoring event started with a general training session, which included information about the FACT project, the design of the FACT performance tasks, use of the Constructed Response Item Scoring Platform (CRISP), and a general bias training. For the task specific training, the raters reviewed the task and task-specific scoring guide, watched each benchmark and rangefinder training video together, and discussed the annotated descriptions provided for that response. Following completion of the task-specific training, the raters used the CRISP system to watch each assigned response video and record their score. Training for and scoring of responses for a task took approximately 2.5 h, with the task-specific training taking approximately 1.5 h and scoring of the responses taking approximately 1 h. Because the training was designed to inform scoring procedures, raters also completed a brief survey and group debrief after scoring for each task and a general survey about the overall FACT scoring after scoring for all 14 tasks was complete. Additional detail on the formative rating, survey analysis, and main lessons learned is provided in a supplemental research report (Weren et al., 2020).

For each of the tasks, raters were randomly assigned to the performances. Each performance was independently scored by three raters. The validity performances were scored by all 14 raters. Table 5 provides a count of the number of training, validity, and score performances for each of the 14 tasks.

Results

Can the FACT Performance Tasks Be Administered as Intended?

The FACT tasks were designed to be administered in a short amount of time, ideally in 5 min (300 s) or less. Figure 3 presents three panels showing the distribution of time spent on preparation, planning, and total time for each of the administered tasks. As illustrated in the first panel, preparation time was relatively consistent across the RLA tasks with two exceptions. Task RLA 13 and RLA 23 required substantially more preparation time than the other RLA tasks. This could be due to a combination of factors. These were the only two RLA tasks that allowed 4 min of prep time (all other tasks were 3 min or less). It is also possible that task complexity and higher reading or comprehension load compared to the other RLA tasks contributed to longer preparation time. There was greater variation in preparation time for the MATH tasks. Some of the variability in preparation time was likely due to differences in the amount of time allowed for preparation (see Table 2 above for allowed preparation time for each task).

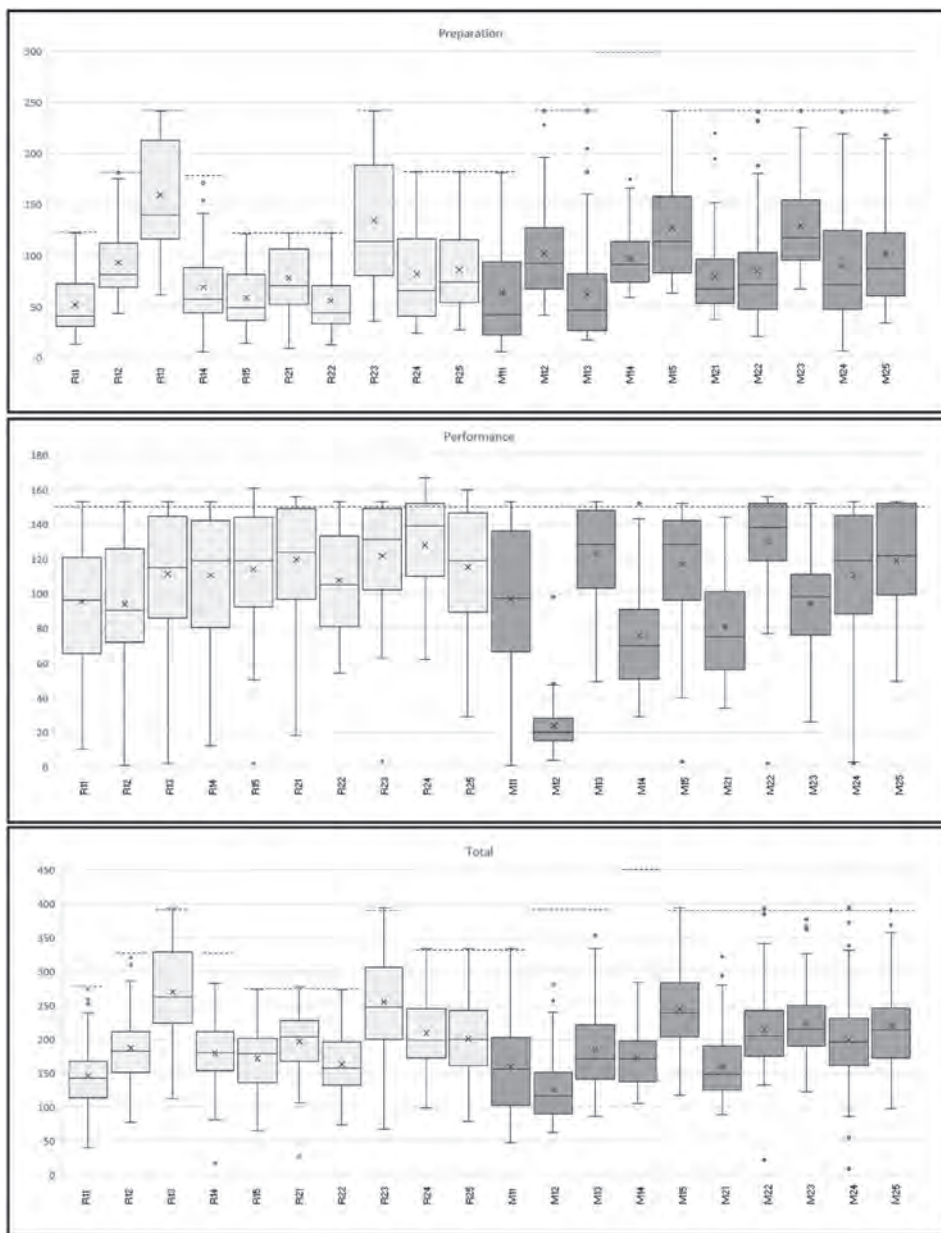


Figure 3 Variation in task administration preparation, performance, and total time in seconds. The dotted line represents the allocated time for the task. A programming issue allowed some task performances of slightly more than the allocated 150 s.

The performance portions of the FACT tasks were designed to have time limits of 2.5 min each. However, due to a programming issue, the timing clock was stopped briefly when participants executed particular types of markup on the workspace. This allowed some participants to take as much as 17 additional seconds on particular tasks. For all but three tasks (RLA 15, RLA 24, and RLA 25), the additional allowed time was no more than 6 s. For RLA, performance times were, in general, consistent across the tasks. However, the distribution for tasks RLA 23 and RLA 24 were skewed toward the time limit of 2.5 min (150 s), indicating that a substantial number of the participants used all (or nearly all) the time available and may have had to rush their performances. Once again, the MATH tasks were more variable. A number of MATH tasks (e.g., MATH 13, MATH 15, MATH 22, MATH 24, and MATH 25) had a large proportion of participants taking nearly the full 150 s to complete the task. MATH 12 stands out for the very short performance time. This task differed from all other RLA and MATH tasks in that participants were instructed to simply ask a single question of the students before stopping the performance.

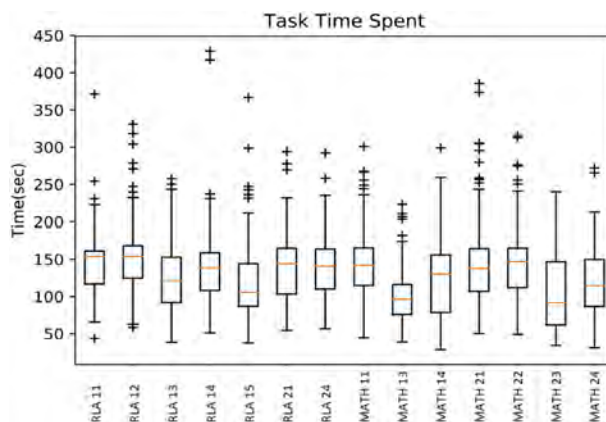


Figure 4 Variation in rating time across tasks in seconds.

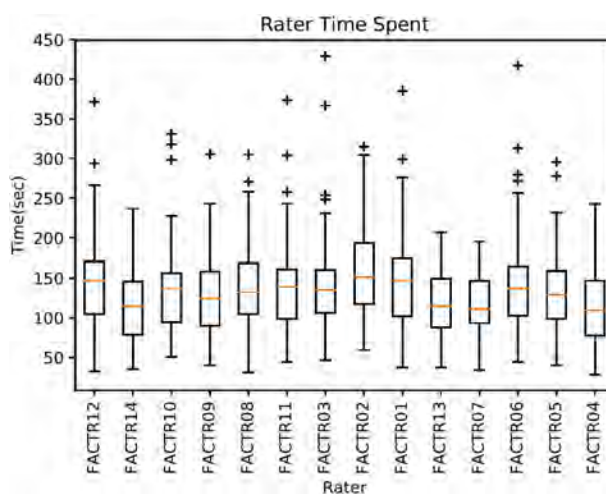


Figure 5 Variation in rating time across tasks by raters in seconds.

The mean completion time (combining preparation and performance times) across all tasks was 3.2 min. Looking across all performances for all tasks, a total of 90 performances took longer than 5 min (8% of the total number of performances). With the exception of task RLA 13, all tasks were completed in under 5 min by 75% or more of the participants. In fact, with the exclusion of a few outlier candidates, nearly half of the tasks (9 out of 20) were completed by all candidates in less than 5 min.

Can the FACT Performance Tasks Be Scored Accurately and Efficiently?

Scoring Efficiency

The FACT performance tasks were designed to be scored quickly and accurately. The average time taken to score a given task is important because it is directly related to the hours of human rating and ultimately the cost of administering an assessment such as FACT. The boxplots shown in Figures 4 and 5 summarize the time spent scoring within and among tasks and raters. Across all tasks and raters, the average time spent on scoring a performance was 130 s, or just over 2 min (for more detail on rating time, see Weren et al., 2020).

Scoring Accuracy and Agreement

A small number of the score performances could not be rated due to issues with the technical quality of the performance (e.g., sound quality of the participant performance or technical issues capturing the performance on the virtual white

Table 6 Score and Validity Case Rating Sample for Each Task

Task	Score performances (3 raters for each performance)			Validity performances (14 raters for each performance)	
	Assigned (<i>n</i> = 602)	Not scoreable ^a (<i>n</i> = 15)	Ratings (<i>n</i> = 1806)	Assigned (<i>n</i> = 74)	Ratings (<i>n</i> = 1,036)
RLA 11	42	1	126	6	84
RLA 12	43	1	129	6	84
RLA 13	42	1	126	6	84
RLA 14	44	1	132	5	70
RLA 15	44	1	132	5	70
RLA 21	42	1	126	6	84
RLA 24	41	0	123	7	98
MATH 11	41	3	123	5	70
MATH 13	42	2	126	4	56
MATH 14	43 ^b	0	129	7	98
MATH 21	45	0	135	4	56
MATH 22	44	1	132	5	70
MATH 23	46	0	138	4	56
MATH 24	43	3	129	4	56

Note. RLA = reading language arts; MATH = mathematics. ^aOne or more raters determined that a performance was not scoreable. These performances were all reviewed by the research team and the performance for RLA 21 and MATH 22 were later consensus scored and included in subsequent and task and score analyses. ^bSix training case for task M14 were also assigned to the score cases. These six performances were included in agreement analyses and scored by three raters.

board). In total, there were 15 performances that were not rated by all three raters. In addition, there were six performances for task MATH 14 that were not assigned as planned for rating. All assigned validity performances were scored by all 14 raters (Table 6).

Rating Agreement for Validity Cases

Validity responses (sometimes referred to as validity papers or validity cases) are used in large-scale operational scoring programs to monitor rater performance and the potential need for retraining. Typically, validity responses are chosen because they represent specific scoring decisions that have been addressed in rater training. For more complex scoring, a range of validity responses are selected to ensure that raters recall, notice, and accurately score important differences in performance characteristics. Validity responses are then seeded into the scoring pool. When raters score a validity response, their score is compared to the consensus score. When raters do not agree with the consensus score, they can be identified for retraining (or these instances can be used to guide the design of new training procedures).

The validity response selection for the FACT pilot did not follow the fully developed operational design outlined above. Instead, the FACT pilot was designed to identify the types of critical differences in performances that could support revisions to training and targeted selection of validity responses. This formative use of validity cases to inform task and scoring revisions is discussed in the companion scoring report (Weren et al., 2020). For the purposes of the results reported here, the validity responses offer the opportunity to directly compare rater scores with consensus scores and, therefore, provide a direct measure of how well raters are conforming to the developers' conceptions of what constitutes different levels of performance for the selected responses. Logically, the consensus score is treated as a "true" score and lack of agreement is interpreted as scoring error. Validity case results provide a basis for evaluating the extent to which ratings for tasks and raters on average diverge from the consensus score.

Agreement results are presented in Table 7. These are presented by various levels and types of aggregation (e.g., all ratings, ratings summarized by subject and by task). While the overall agreement rate of 62% might appear low, it is important to recognize that for this study, the selected validity cases did not always reflect specific rating decisions that were addressed in training. We anticipate that the validity case agreement rate will increase substantially in subsequent pilots where validity cases are deliberately aligned to training.

Table 7 Validity Case Rating Accuracy

Task name	Exact		Adjacent		Discrepant	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
All	643	62	372	36	21	2
Subjects						
RLA	307	53	249	44	18	3
MATH	336	72	123	27	3	1
Tasks						
RLA 11	31	37	53	63	0	0
RLA 12	43	51	38	45	3	4
RLA 13	33	39	41	49	10	12
RLA 14	43	61	24	34	3	4
RLA 15	49	70	21	30	0	0
RLA 21	46	55	38	45	0	0
RLA 24	62	63	34	35	2	2
MATH 11	49	70	21	30	0	0
MATH 13	44	79	12	21	0	0
MATH 14	77	79	18	18	3	3
MATH 21	41	73	15	27	0	0
MATH 22	58	83	12	17	0	0
MATH 23	33	59	23	41	0	0
MATH 24	34	61	22	39	0	0

Note. MATH = mathematics; RLA = reading language arts.

There was a noteworthy difference in the rates of agreement for RLA and MATH. RLA has both a substantially lower proportion of raters who agree with the consensus scores and a higher proportion of raters who give a discrepant rating compared to MATH. The task level breakdown shows substantial variation among tasks. Although the MATH tasks had higher level of agreement compared to the RLA tasks, there was also substantial variation among the tasks within each subject. For RLA tasks, RLA 11 and RLA 13 stand out with agreement rates below 40% (i.e., more raters disagree with the consensus score than agree). Task RLA 13 stands out from all other tasks for the very high proportion of discrepant ratings (12%). Although agreement was more consistent among MATH tasks, there was still some notable variation. For example, agreement rates for MATH 22 and MATH 23 differ by 24%.

It is important to recognize that the agreement findings listed above are sensitive to differences in the quality and score level of selected performances. For RLA 13, nine out of 10 of the discrepant ratings were from one validity case performance. Also, the tasks vary in the number of validity cases assigned at each score point. This will influence the probability of discrepant ratings, because a discrepant rating can only occur for performances with a consensus score of 1 or 3. Readers interested in more nuanced discussion of the validity results can refer to the FACT formative scoring report (Weren et al., 2020).

Table 8 describes the agreement between rater scores and consensus scores. Compared to the descriptions above by task, raters were much more consistent in their rates of agreement with consensus scores. However, Rater 14 stands out from the rest of the raters with a lower agreement rate (51%) and a higher discrepancy rate (5%).

When rater disagreement was broken out by ratings that are higher and lower than the consensus score, there was some variation across raters (Figure 6). Some raters were more lenient and some more stringent. For example, both Rater 1 and Rater 2 have similar levels of agreement with the consensus score. However when examining ratings where they do not agree, Rater 1 tends to rate higher than the consensus score, whereas Rater 2 tends to rate lower than the consensus score. There appears to be a slight tendency across all raters to rate lower than the consensus score. It is important, again, to emphasize that agreement with validity cases is likely to look different when the validity cases are selected to represent rating decisions addressed in the training. Not only will the overall agreement likely increase, but the variation among raters will be more directly linked to how they learn from training and apply the associated decision rules.

Table 9 summarizes the agreement between raters and consensus scores for each of the score points. This analysis provides insight into whether agreement rates differ by the level of the performance. When summarizing across all performances, there was a small trend toward higher levels of agreement for lower performances (i.e., 1 score point). The breakdown by subject shows a different pattern for RLA and MATH. Although there was no trend for RLA by the level of

Table 8 Validity Case Rating Accuracy by Rater

Rater	Exact		Adjacent		Discrepant	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Rater 1	50	68	22	29	2	3
Rater 2	43	58	30	41	1	1
Rater 3	50	68	24	32	0	0
Rater 4	44	59	28	38	2	3
Rater 5	46	62	26	35	2	3
Rater 6	49	66	24	32	1	1
Rater 7	45	61	27	36	2	3
Rater 8	43	58	29	39	2	3
Rater 9	49	66	24	32	1	1
Rater 10	42	57	30	41	2	3
Rater 11	50	68	23	31	1	1
Rater 12	49	66	24	32	1	1
Rater 13	45	61	29	39	0	0
Rater 14	38	51	32	43	4	5

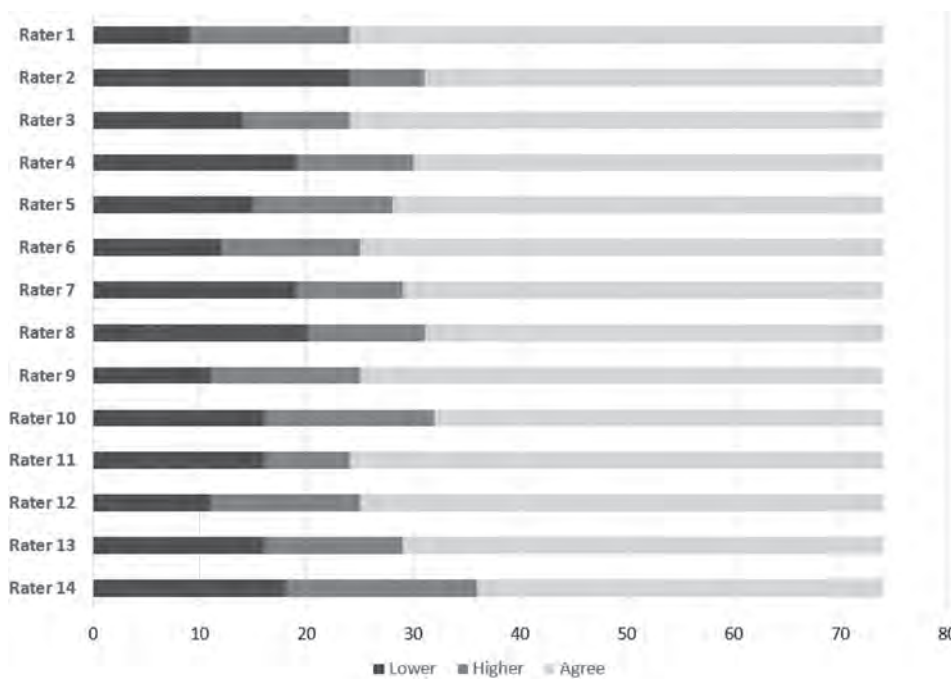


Figure 6 Validity case count of ratings that are higher, lower, and exact by rater.

the performance, for MATH there was a large difference in agreement (24%) between performances that were consensus scored 1 and consensus scored 3. This difference may be due to the relative clarity in scoring low performances that had a mathematical error.

Rating Agreement for Score Cases

Agreement rates for score cases are summarized in Table 10. Although there was little difference in agreement rates across RLA and MATH, there was more variation at the task level. The task level agreement rates vary from a low of around 55% to a high of around 70%. Tasks RLA 13 and MATH 11 stand out as lower than the rest of the tasks with agreement rates of 46% and 51% respectively.

Table 9 Validity Case Proportion of Exact, Adjacent, and Discrepant Ratings by Score Point for Each Subject

Task name	Exact		Adjacent		Discrepant	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
All tasks						
Score 1	172	68	75	30	5	2
Score 2	293	62	183	38		
Score 3	178	58	114	37	16	5
RLA tasks						
Score 1	75	54	60	43	5	4
Score 2	134	53	118	47		
Score 3	98	54	71	39	13	7
MATH tasks						
Score 1	97	87	15	13	0	0
Score 2	159	71	65	29		
Score 3	80	63	43	34	3	2

Note. MATH = mathematics; RLA = reading language arts. When a performance is given a score of 2 by either rater, it is not possible to have discrepant agreement among two raters.

Table 10 Score Cases Rating Agreement by Task

Task name	Exact		Adjacent		Discrepant	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
All Subjects	1,111	62	660	37	35	2
RLA	541	61	340	38	13	1
MATH	570	63	320	35	22	2
Tasks						
RLA 11	83	66	42	33	1	1
RLA 12	71	55	56	43	2	2
RLA 13	58	46	64	51	4	3
RLA 14	77	58	54	41	1	1
RLA 15	92	70	36	27	4	3
RLA 21	91	72	34	27	1	1
RLA 24	69	56	54	44	0	0
MATH 11	63	51	54	44	6	5
MATH 13	77	61	46	37	3	2
MATH 14	88	68	38	29	3	2
MATH 21	94	70	40	30	1	1
MATH 22	85	64	42	32	5	4
MATH 23	90	65	46	33	2	1
MATH 24	73	57	54	42	2	2

Note. MATH = mathematics; RLA = reading language arts.

Differences in agreement rates among raters for score cases were relatively small ranging from a low of roughly 60% to a high of 70% (Table 11). Rater 14 stands out as having a lower agreement rate of 47%.

A number of statistics can be used to summarize agreement among raters. The quadratic weighted kappa (QWK) adjusts agreement for the likelihood of guessing. The intra-class correlation coefficient (ICC) breaks down the proportion of variance that is within and among the ratings for the task performances. Both the QWK and ICC statistics should be evaluated with caution because both statistics are sensitive to sparsely populated data matrices that result from a small number of administrations, such as was the case with the FACT pilot data. To help with interpretation, the exact agreement from Table 11 above is presented again along with the ICC and the QWK results (Table 12).

Comparisons across the agreement descriptives and ICC and QWK results indicate that the descriptive and summary results are not always consistent. Consider, for example, MATH 13 and MATH 22. Both had similar agreement rates (MATH 13 = 61% and MATH 22 = 64%), but MATH 13 has a low ICC (.26) and QWK (.25), while MATH 22 has a

Table 11 Score Cases Rating Agreement by Rater

Rater	Exact		Adjacent		Discrepant	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Rater 1	143	58	96	39	7	3
Rater 2	153	60	97	38	4	2
Rater 3	171	63	98	36	1	0
Rater 4	178	67	80	30	6	2
Rater 5	160	61	100	38	4	2
Rater 6	153	59	98	38	7	3
Rater 7	158	63	86	34	8	3
Rater 8	168	63	95	36	3	1
Rater 9	185	63	104	36	3	1
Rater 10	143	58	97	39	8	3
Rater 11	147	60	94	38	5	2
Rater 12	168	69	73	30	3	1
Rater 13	176	69	77	30	1	0
Rater 14	119	47	125	49	10	4

Table 12 Summary Agreement Statistics

Task	% Exact	ICC ^a	QWK ^b
RLA 11	66	.68	.67
RLA 12	55	.55	.55
RLA 13	46	.44	.43
RLA 14	58	.58	.58
RLA 15	70	.73	.73
RLA 21	72	.64	.63
RLA 24	56	.66	.66
MATH 11	51	.29	.31
MATH 13	61	.26	.25
MATH 14	68	.41	.41
MATH 21	70	.66	.66
MATH 22	64	.64	.63
MATH 23	65	.66	.66
MATH 24	57	.48	.47

Note. MATH = mathematics; RLA = reading language arts. ^aThe intraclass correlation (ICC) is based on the following modeling assumptions: each subject is rated by multiple raters; raters are randomly assigned to subjects; and, all subjects have the same number of raters. ^bThe quadratic weighted kappa (QWK) is calculated by taking the average of the pairwise QWK for each of the three rating pairs.

relatively high ICC (.64) and QWK (.63). As indicated in Figure 7, MATH 13 was the easiest math task, with few scores of 1. Easy items, like MATH 13, tend to have high levels of chance agreement, and therefore, lower values for QWK, which corrects for chance agreement. A similar effect occurs for the ICC because of the resulting restriction of range. In contrast, Figure 7 indicates that MATH 22 had the most scores of 1 and the most even distributions of scores over the three score points.

Because the goal of this pilot and the associated analysis was to guide revisions to both tasks and associated task scoring, we decided not to conduct generalizability (G) and decision (D) studies. Instead, we intend to conduct G and D studies after revising tasks and conducting a larger scale pilot study. However, we recognize that one important question that a G and D study could explore is how the number of raters influences the reliability of a test score. We ran a number of different ICC estimations to provide some insight into how the number of raters influences the reliability of a test score. As reported just above, we have already estimated the ICCs for each task including all three raters. However, in a typical operational administration, each performance is likely to be scored by no more than two raters. To evaluate the potential impact on overall test reliability of dropping from three rating pairs to two raters, we also estimated the ICCs separately for each of the three sets of rating pairs and then calculated a pairwise average ICC (Table 13). These results indicate that

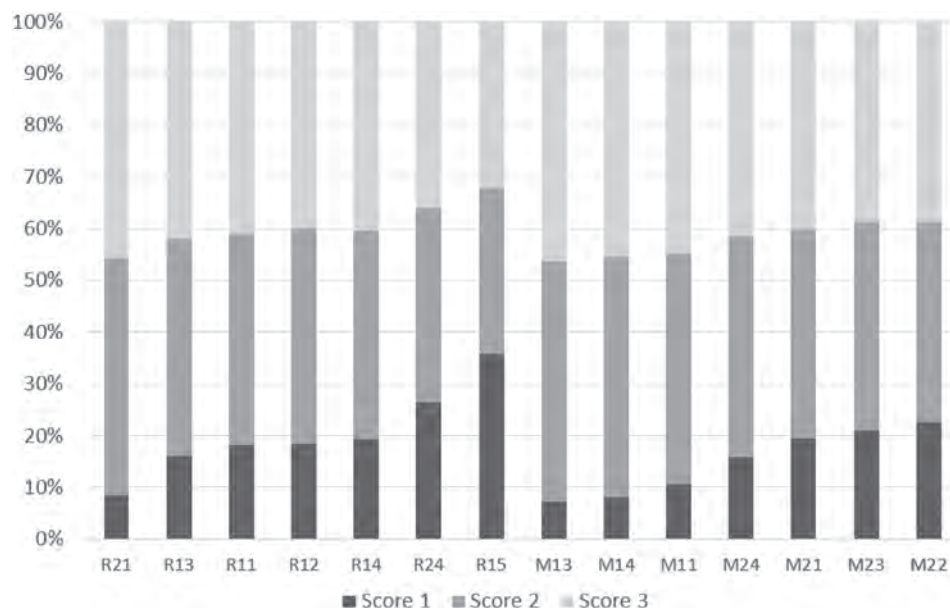


Figure 7 Proportion of 1, 2, and 3 ratings for each task.

Table 13 Impact on Intraclass Correlations (ICC) of Reducing the Number of Raters

Task	ICC Two Raters				ICC Three Raters
	Pair 1–2	Pair 1–3	Pair 2–3	Pairwise Average	
RLA 11	.68	.62	.72	.67	.68
RLA 12	.46	.64	.55	.55	.55
RLA 13	.50	.45	.38	.44	.44
RLA 14	.62	.61	.52	.58	.58
RLA 15	.70	.75	.74	.73	.73
RLA 21	.53	.73	.66	.64	.64
RLA 24	.66	.66	.66	.66	.66
MATH 11	.27	.37	.20	.28	.29
MATH 13	.30	.17	.31	.26	.26
MATH 14	.55	.34	.33	.41	.41
MATH 21	.67	.68	.63	.66	.66
MATH 22	.60	.73	.57	.63	.64
MATH 23	.56	.68	.75	.66	.66
MATH 24	.37	.61	.44	.47	.48

Note. MATH = mathematics; RLA = reading language arts. Intra-class correlation is based on the following modeling assumptions: each subject is rated by multiple raters; raters are randomly assigned to subjects; and, all subjects have the same number of raters. All averages of triple scored performances are rounded to the nearest whole unit to facilitate presentation and interpretation.

there was substantial variation among the three sets of rater pairs. For RLA 24 there was no difference among rating pairs and for RLA 15 and MATH 21 the difference was minimal ($<.05$). However, for RLA 21, MATH 14, and MATH 24, there was a substantial difference ($>.20$) among the rating pairs. These larger differences among the rating pairs suggest that dropping to two raters could have an impact on overall test reliability.

Is There a Reasonable Distribution of Scores Within and Among FACT Performance Tasks?

Distribution of Task Scores

The task analyses presented in this section combine the scores for the training, validity, and score cases. For the training and validity cases the consensus score from the task developers was used. For the score cases, the average of the three rater

Table 14 Correlation Between All Reading Language Art (R) and Mathematics (M) Tasks

Task	R11	R12	R13	R14	R15	R21	R24	M11	M13	M14	M21	M22	M23
R12	.57 ***												
R13	.50 ***	.52 ***											
R14	.43 **	.38 **	.56 ***										
R15	.49 ***	.57 ***	.40 **	.40 **									
R21	.44 **	.40 **	.39 **	.41 **	.50 ***								
R24	.34 *	.22	.22	.14	.30 *	.23 *							
M11	.15	.44 **	.30 *	.26	.29 *	.15	.32 *						
M13	.15	.27 *	.12	.32 *	.13	.40 **	.08	.49 ***					
M14	.35 *	.48 ***	.32 *	.38 **	.33 *	.35 *	.33 *	.50 ***	.06				
M21	.47 ***	.28 *	.37 **	.48 ***	.47 ***	.39 **	.57 ***	.42 **	.26	.35 *			
M22	.32 *	.47 ***	.49 ***	.36 **	.40 **	.34 *	.37 **	.30 *	.28 *	.33 *	.47 ***		
M23	.36 **	.57 ***	.42 **	.46 ***	.50 ***	.38 **	.31 *	.54 ***	.47 ***	.40 **	.55 ***	.48 ***	
M24	.19	.23	.09	.27 *	.30 *	.20	.37 **	.33 *	.45 **	.16	.43 **	.09	.44 **

*** $p < .001$. ** $p < .01$. * $p < .05$.

scores was used. We found only minor variation in mean performance score across the tasks ranging from a low of 1.94 for MATH 21 to a high of 2.26 for task RLA 14, and we found a similar consistency in the standard deviations of these task means which ranged from .48 to .78.

Another way to examine variation in task scores is by the number of performances that received a score at each of the three score points. Figure 7 breaks down each task by performances that fall into each score category. Of particular interest are the number or proportion of performances for each task that were scored at the low or 1 score point. Tasks with a high proportion of 1 ratings can be thought of as more difficult than tasks with lower proportions of 1 scores. A test made up of FACT tasks should have some variation in the “difficulty” of the tasks to ensure that the test provides information on test takers with varying levels of competence. To illustrate the variation of task difficulty, the results presented in Figure 7 are sorted within RLA and MATH from the lowest proportion of performances that received a 1 (*easiest*) to the tasks with the highest proportion of performances that received a 1 (*hardest*). In addition to illustrating variation in “difficulty” across the tasks, this display also highlights another way to consider scoring for FACT tasks. A score of 1 not only represents a low performance, but it can be thought of as indicating an unacceptable level of competence for beginning teaching. From this perspective, the tasks vary in the proportion of unacceptable performances from a low of under 10% to a high of over 30%.

Intertask Correlations

We also examined correlations among the different RLA and MATH tasks (Table 14). Most of the tasks were significantly correlated with most other tasks. There was a minor tendency for MATH tasks to have higher correlations with other MATH tasks compared to RLA tasks. And the same was true for RLA tasks, which had a minor tendency to have higher correlations with other RLA tasks and lower correlations with MATH tasks. Perhaps of more interest are the instances where tasks had very low correlations with other tasks, suggesting that the two tasks may be measuring competencies that are somewhat distinct (e.g., RLA 14 and RLA 24, MATH 22 and MATH 24).

Task Score and Administration Time

We were interested in better understanding whether the amount of time that participants took to prepare for and carry out a performance was related to their task score. For the vast majority of tasks there was no correlation between preparation time and task score (Table 15). The one exception was for MATH 11 where a longer preparation time was associated with a lower score. For roughly half of the tasks, there was a significant positive correlation between performance time and task score. It is important to recognize that the positive association between performance time and task score could be driven by a number of factors. For example, this could be due to lower scoring participants taking less time, higher scoring participants taking more time, or both.

Table 15 Correlation of Task Score With Preparation and Performance Time

Task	Preparation	Performance
RLA 11	.01	.35**
RLA 12	.18	.09
RLA 13	.12	.25
RLA 14	-.04	.54**
RLA 15	-.08	.38**
RLA 21	-.01	.19
RLA 24	-.09	.05
MATH 11	-.35*	.33*
MATH 13	-.06	.23
MATH 14	-.07	.24
MATH 21	-.19	.11
MATH 22	-.26	-.03
MATH 23	-.02	.41**
MATH 24	-.06	.48**

Note. MATH = mathematics; RLA = reading language arts. ** $p < .01$. * $p < .05$.

Table 16 Score Characteristics for All, Reading Language Arts (RLA) and Mathematics (MATH) Scores ($n = 50$)^a

	Tasks	Score						Alpha	SEM	Intrataask correlation ^b
		Range	Min	Max	Median	Mean	SD			
All	14	14–42	15.3	37.7	30.7	29.6	5.31	.86	1.99	.35
RLA	7	7–21	7.7	19.3	15.3	15.0	3.07	.77	1.33	.40
MATH	7	7–21	7.7	19.7	14.8	14.6	2.80	.79	1.40	.38

^aOf the 57 participants, seven did not complete one or more tasks. Therefore, the total number of participants with complete score data for All, RLA, and MATH was 50. ^bThe intrataask correlation is the average of all task correlations for tasks contributing to the relevant score.

Do the FACT Performance Tasks Provide Reliable Measures of the Competencies Assessed?

It is useful to create scores for all of the FACT tasks and for the subject areas of RLA and for MATH. These scores can provide preliminary insight into how the piloted FACT tasks might function to assess each of these constructs. Specifically, it is helpful to look at whether there was variation across candidate scores and whether the tasks when combined provide a reliable measure. In this section, we first consider sum scores that were created by simply adding scores for all tasks, for just RLA tasks, and for just MATH tasks. Second, we consider a different scoring model that counts the number of low or 1 scores for All, RLA, and MATH tasks.

Sum Score

Summary Statistics

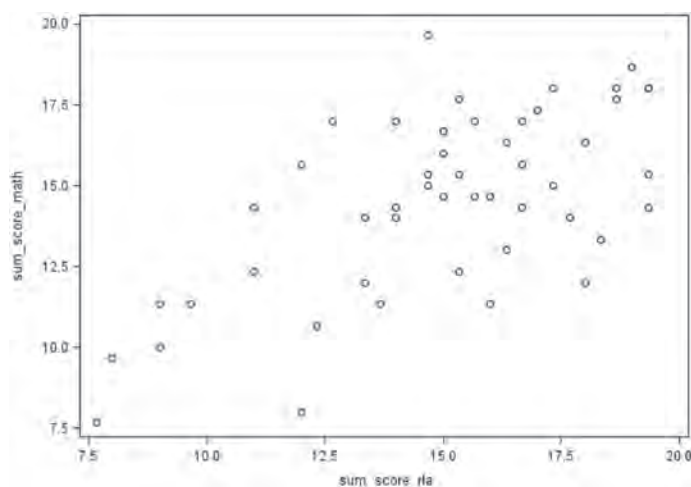
The score results presented in Table 16 suggest that FACT performance tasks are assessing an underlying construct and that the measure of this construct is reliable at a level that approaches or surpasses what is typically achieved for licensure tests. In fact, the reliability for all 14 tasks is .86, and this surpasses a value of .85 which is typically considered as “reasonable” for licensure tests (Luecht, 2017, p. 144). The intrataask correlations are also in the general recommended range of .15–.50 (Briggs & Cheek, 1986; Clark & Watson, 1995).

Sum Score and Task Correlations

The correlation of each task score and the sum scores for All, RLA, and MATH is shown in Table 17. The polyserial correlations for All vary across the tasks from a low of $r = .48$ for MATH 13 to a high of $r = .77$ for MATH 17. With only a few exceptions, the polyserial correlation was higher when the items were correlated to the subject specific score. This suggests that the tasks are assessing competencies that are specific to either RLA or MATH.

Table 17 Polyserial Correlation With All, Reading Language Arts (RLA) and Mathematics (MATH) Scores

Tasks	Polyserial Correlations		
	All	RLA	MATH
RLA 11	.60	.75	—
RLA 12	.70	.72	—
RLA 13	.62	.73	—
RLA 14	.66	.66	—
RLA 15	.71	.78	—
RLA 21	.58	.65	—
RLA 24	.49	.43	—
MATH 11	.61	—	.76
MATH 13	.48	—	.65
MATH 14	.57	—	.62
MATH 21	.77	—	.75
MATH 22	.68	—	.67
MATH 23	.74	—	.84
MATH 24	.53	—	.64

**Figure 8** Reading language arts (RLA) sum score by MATH sum score scatter plot.

RLA and MATH Score Correlation

To provide initial evidence on whether subscores for RLA and MATH assess different constructs, we correlated the sum scores for RLA and MATH. We found that the two scores were significantly correlated $r(48) = .64, p < .001$. However, unadjusted correlations of test scores can be misleading because the constructs are measured with error. When the correlations were disattenuated for measurement error ($r = 1.0$) the MATH and RLA sub-scales provide nearly identical information about participant competencies. Although these results strongly suggest that the RLA and MATH tasks assess closely related constructs, basic descriptives from the scatter plot shown in Figure 8 indicate that there were participants with relatively higher scores on one subscore compared to the other.

Count of Performances Receiving a 1

We also considered an alternative method for generating scores for All, RLA, and MATH. This approach starts with the assumption that each task score of 1 represents an unacceptable performance. It follows that candidates should only be allowed a small number of unacceptable performances in order to receive an overall passing score. In this score model, the scores for All, RLA, and MATH are a count of the number of 1 scores for each of the tasks. Higher scores indicate a lower level of performance. The two panels in Figure 9 summarize the scores created using this method. The left panel

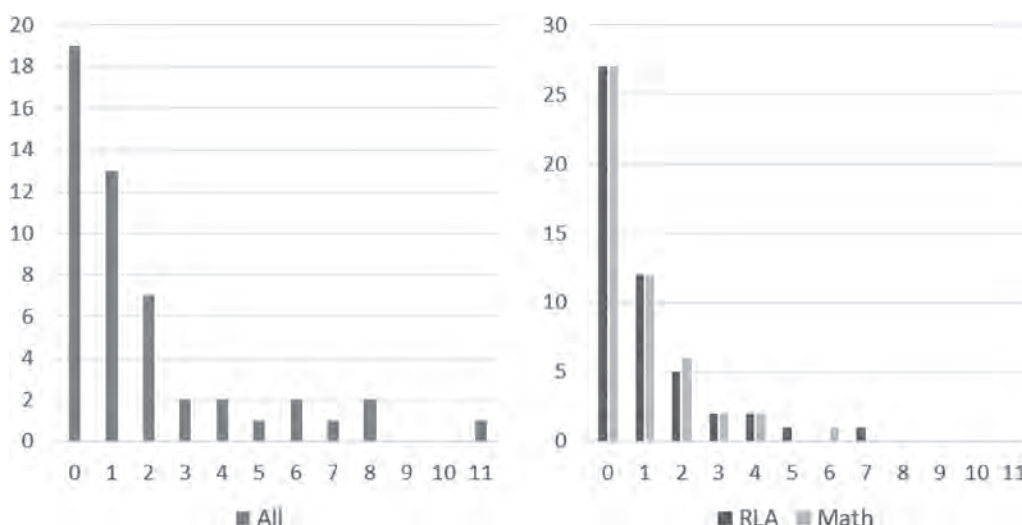


Figure 9 Number of participants receiving different counts of low or 1 scores for all, reading language arts (RLA), and elementary mathematics (MATH) tasks.

Table 18 Number of Participants Receiving Low Scores for All, Reading Language Arts (RLA), and Mathematics (MATH) Tasks

Tasks receiving a 1 (<i>low</i>) score	All		RLA		MATH	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
0–1	32	64	39	78	39	78
2	7	14	5	10	6	12
3	2	4	2	4	2	4
4–11	9	18	4	8	3	6

shows the 14 tasks that make up the measure of all tasks. The right panel breaks out the 7 RLA and 7 MATH tasks. For All, RLA, and MATH, a majority of candidates completed the tasks with no more than a single 1 score.

There is currently no standard for how many 1 scores might be acceptable for a candidate to still be awarded a teaching license. These types of determinations are typically made by standards panels made up of subject matter experts to recommend an acceptable passing standard. However, for the purposes of illustration, if we assume that a passing score can have no more than three instances of 1 or inadequate scores, then 82% of participants would pass for All. If we assumed for RLA and MATH no more than two 1 scores are allowed, then 88% and 90% would pass RLA and MATH respectively (Table 18). However, we do not have sufficient data to fully explore the potential of using this method as part of a licensure decision, either independently or in combination with a sum score. We do, however, think that this approach warrants further investigation in a larger-scale study.

Are the FACT Performance Measures Associated With Participants’ Academic and Professional Attainment?

In order to provide preliminary validity evidence, we examined score differences for groups of participants with different levels of educational and professional experience. Although we recognize that the study sample was small, and for some of the background characteristics there was relatively restricted variation (e.g., education level, professional experience), we nonetheless think it is informative to see if any differences do exist. In Table 19, we summarize the breakdown for education level, type of preparation program, and completion of student teaching, and report the mean score for each of the groups.

We ran one-way analysis of variance to determine whether the mean differences among groups were significant. There was not a significant effect of teacher preparation program (master’s degree program and undergraduate degree program) at the $p < .05$ level for All tasks, RLA tasks, or MATH tasks. Likewise, there was not a significant effect of having completed

Table 19 Mean Score for All, Reading Language Arts (RLA), and Elementary Mathematics (MATH) Tasks for Different Groups of Candidates

	<i>n</i>	All		RLA		MATH	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Highest education level							
Bachelor's degree or higher	21	32.21	3.29	16.35	2.00	15.86	2.08
Senior	18	28.94	5.10	14.87	2.96	14.07	2.86
Junior or sophomore	10	25.70	6.37	12.64	3.63	13.06	3.06
Teacher preparation program							
Master's degree program	21	31.40	4.27	15.92	2.20	15.48	2.64
Undergraduate degree program	28	28.18	5.74	14.19	3.43	13.99	2.82
Completed student teaching							
Yes	40	29.98	5.30	15.18	3.10	14.80	2.76
No	9	28.96	4.83	14.52	3.05	14.44	2.24

Table 20 Participant Perceptions of FACT Assessment Tasks and Delivery Platform (*n* = 59)

Survey question	Agree (%)	Somewhat agree (%)	Somewhat disagree (%)	Disagree (%)
The kinds of teaching knowledge, skills, and abilities required by the tasks are important for effective teaching	86	12	2	0
The kinds of teaching knowledge, skills, and abilities required by the tasks felt authentic	61	34	2	4
The kinds of teaching knowledge, skills, and abilities required by the tasks are a real focus of my teacher preparation program	53	35	12	0
The kinds of teaching knowledge, skills, and abilities required by the tasks should be assessed as part of teacher licensure	63	32	3	2
My performance on the tasks accurately reflects my actual teaching ability	12	39	4	12
Some of the mathematics tasks were difficult for me to complete successfully because I was unfamiliar with the mathematics topic	10	30	30	30
Some of the reading language arts tasks were difficult for me to complete successfully because I was unfamiliar with the reading language arts topic	14	32	32	22
The task directions provided a clear description of what I needed to do	51	41	8	0
I found the testing interface and tools clear and easy to use (even if the tasks themselves might have been difficult)	49	36	10	5

student teaching (master's degree program and undergraduate degree program) at the $p < .05$ level for All tasks, RLA tasks, or MATH tasks. However, there was a significant effect of education level on scores for All tasks at the $p < .05$ level [$F(2,47) = 4.76, p = .002$], for MATH tasks at the $p < .05$ level [$F(2,47) = 4.73, p = .013$], and for RLA tasks at the $p < .05$ level [$F(2,47) = 6.51, p = .003$]. Post hoc comparisons using the Tukey HSD test indicated that education level for junior or sophomore was significantly different from Bachelor's degree or higher for All ($p = .002$), RLA ($p = .002$), and MATH ($p = .016$), but there was no significant difference for All, RLA, or MATH for the comparison of junior or sophomore and senior, or senior and bachelor's degree or higher.

How Do Participants View the FACT Performance Tasks and the Testing Experience?

More than 90% of participants agreed or somewhat agreed with statements indicating the value and authenticity of the FACT tasks (Table 20). However, endorsement was substantially lower when participants were asked if performance on the FACT tasks accurately reflected their teaching ability. Only 51% selected agreed or somewhat agreed. A majority of participants disagreed or somewhat disagreed that the FACT tasks were challenging because they were unfamiliar with the RLA and MATH topics that were assessed. Over 90% of participants agreed or somewhat agreed that the directions for FACT were clear and 85% agreed or somewhat agreed that the testing interface and tools were clear and easy to use.

Table 21 Participant Perceptions of Time Allowed for Testing ($n = 59$)

Survey question	Too short: I needed more time (%)	Adequate: I had the right amount of time (%)	Too long: I had more time than I needed (%)
Was the right amount of time provided to practice with the tools and functionality in the orientation task?	17	79	4
Was the right amount of time provided to <i>prepare</i> for the task performances?	25	58	17
Was the right amount of time provided to complete the task performances?	44	54	2

A majority of participants reported that they had adequate time to learn to use the FACT interface, to prepare for a performance, and to carry out a performance (Table 21). However, 44% of participants reported that the amount of time provided to carry out their performance was too short.

FACT test takers also responded to the following two open ended questions about their experience completing the tasks: “Please describe your overall impression of the FACT assessment. What did you find interesting or noteworthy?” and “What do you think are benefits of a performance assessment like FACT? Please explain.” Responses to the questions were reviewed and themes identified. The responses were then coded for each of these themes and the theme descriptions modified for responses that were difficult to code. Six categories were identified, and the participant responses coded and summarized. Responses for each category are presented in Appendix C. In general, the responses echoed the survey patterns from the Likert items summarized just above. Most of the candidates offered positive impressions of the FACT tasks. Roughly half were supportive of the focus on practice and felt the tasks were authentic. Roughly 20% of candidates liked the emphasis on student thinking and how it helped them think about what they still need to learn. About 20% of candidates commented that allowing interaction with students as part of the performance would improve FACT.

Taken together the Likert responses and open-ended responses indicate that candidates were intrigued by the FACT approach and felt that these assessment tasks focused on important competencies that were both part of their professional preparation and important to effective teaching. The candidates also felt that FACT performance tasks could be used to improve on current approaches to licensure testing.

Discussion

The main purpose of this pilot study was to evaluate the measurement properties of a new generation of assessment tasks designed for use as part of initial licensure testing. This pilot study explored six research questions that focused on administration, scoring, task and test measurement properties, and participant perceptions about FACT. The main findings for each of these questions are summarized below.

Can the Tasks Be Administered as Intended?

The administration was largely successful. Out of the 59 candidates who completed the assessment, 57 (97% of the sample) had video records uploaded for scoring. For these 57 candidates, the 14 tasks selected for scoring included a total of 798 task performances. Of the score performances, 783 (98% of the sample) were scoreable. These results strongly suggest that the FACT delivery platform is functioning well and that only minor enhancements are required to ensure that there is no data loss during administration and processing.

We also analyzed the amount of time required for administration. Our goal was to create tasks that could be completed quickly to allow for administering a greater number and range of tasks in a given testing period. The amount of time allocated for preparation varied by task with 4 min allowed for 50% of tasks, 3 min for 25% of tasks, 2 min allowed for 20% of tasks, and 1 min allowed for 5% of tasks. For all tasks, the large majority of participants took substantially less time for preparation than was allocated. For the performance portion of each of the 20 tasks, 2.5 min were allocated. For about a quarter of the tasks, participants were clustered near the time limit for the performance, suggesting that some

participants may not have had sufficient time to complete their performance. On average, participants took 194 s or about 3.75 min to complete the entire task, including preparation and performance.

Participants were also asked on the postsurvey if sufficient time was provided to prepare and to carry out the task performance. Twenty-five percent of participants felt they could use additional time to prepare. Nearly half of the candidates (44%) felt that the time provided for the performances was too short. These results suggest that some tasks may either need revision to require less involved performances or warrant a longer allocated time.

Can the Tasks Be Accurately and Efficiently Scored?

The scoring procedures were developed to support quick scoring and thus reduce the costs of human rating. The average time spent rating per task was 130 s or just over 2 min. The tasks and the associated scoring rules were also developed to limit the range of responses and the complexity of the scoring procedures with the goal of increasing agreement among raters (and consequently reducing scoring error). Across all scoring occasions, raters agreed for 62% of the ratings. For 10 of these tasks, 2% or less of the ratings were discrepant. For two tasks, 3% of the ratings were discrepant. And for the final two tasks, either 4% or 5% were discrepant. The sample of scored performances was too small for a trustworthy calculation of an agreement statistic such as the QWK that adjusts the proportion of agreement for the likelihood of guessing.

Although the agreement rate was lower than what was originally hoped for, it is still in keeping with agreement rates that are often observed for human rating of performance tasks. The formative scoring study identified a number of changes to future scoring that are likely to improve rater agreement. These proposed changes include creating a content training portion to provide background on the content concepts included in each task, adding additional training cases, creating a practice set of training cases so raters can receive feedback, allowing raters to review the performance video multiple times, and providing real time feedback on validity case scoring (Weren *et al.*, 2020).

Is There a Reasonable Distribution of Scores Within and Among Tasks?

The tasks were designed to provide evidence across a range of participant abilities. One consideration was ensuring that the criteria used to identify performances were not too stringent for beginning teachers. Although there is no hard guideline for what constitutes a “safe to practice” level for entry level teachers, we set out to develop tasks that a majority of candidates could complete with an acceptable or good level of performance. We wanted to ensure that a low score of 1, which roughly corresponds to an unacceptable level of performance, never represented a sizeable proportion of the sample for any given task. Although the results indicated that there was variation across the different tasks, for the majority of tasks, fewer than 20% of participants received a score of 1.

We also examined variation in average scores and their standard deviations. There was relatively limited variation in mean score (on a scale of 1–3) among tasks with a low task mean score of 1.94 to a high task mean score of 2.26. Differences in the standard deviations among tasks are somewhat more variable ranging from a low standard deviation of .48 to a high of .78. These results suggest that there was more variation in scores within than among tasks. They also indicate that the FACT performance tasks are sensitive to a range of candidate ability.

Do the Tasks Provide Reliable Measures of the Competencies Assessed?

It is useful to create sum scores for the full set of 14 scored tasks, and also subscores for the seven RLA and seven MATH tasks. These scores provide insight into how the FACT performance tasks might perform on a full test design. The results indicate that there was substantial variation for each of these three scale scores and that each score has relatively high alpha reliability: $\alpha_{All} = .87$, $\alpha_{RLA} = .77$, $\alpha_{MATH} = .79$. However, it is important to keep in mind that the sample for this study was relatively small and that these reliabilities could change with larger and more representative samples. Nonetheless, these initial results are encouraging and suggest that the FACT performance tasks have promise for creating reliable measures of teaching performance.

Are the Measures Associated With Participants’ Academic and Professional Attainment?

Validity arguments for assessments such as FACT often evaluate the relationship among test scores and construct relevant learning opportunities. To evaluate this type of validity claim, it is useful to examine evidence that level of professional and

academic attainment is associated with FACT scores. The presurvey asked about participant's highest education level, type of teacher preparation program (i.e., bachelor's or master's degree program), and whether the participant had completed student teaching. For each of these areas, we looked to see if there were differences in FACT scores for participants with different levels of professional experience and academic attainment. We found that there was a significant difference for All, RLA, and MATH scores for participants' level of education. However, there was no significant difference for either the type of teacher preparation program or for having completed student teaching. These results should be interpreted with caution given the small numbers of participants in the various comparison groups leading to analyses that were most likely substantially underpowered for detecting significant effects. However, this preliminary analysis does indicate that FACT assessment may be sensitive to candidates' level of professional preparation as indicated by the year of attainment in their teacher education program.

How Do Participants View the Tasks and the Testing Experience?

A large majority of participants agreed that the FACT tasks are authentic, focus on valuable competencies, are a focus of their preparation, and should be included on licensure assessments. However, a smaller proportion of the candidates felt that FACT tasks accurately reflected their teaching ability (51%), which might simply reflect the reality that no one assessment can provide information on the range of competencies required for effective teaching.

Limitations

The development and research results presented in this report represent a positive step toward a new type of assessment task designed to provide evidence of teaching performance suitable for use as part of a teacher licensure assessment. However, it is important to recognize that findings from this type of preliminary study are subject to a number of limitations. One limitation is related to the tasks themselves. As a first-generation effort, these tasks almost certainly fall short of the level of quality that could be achieved after a few rounds of development and revision. Improvements to the task design, scoring material, and training could support significantly higher levels of agreement among raters and ultimately more reliable test scores. For these reasons, some of the results we have reported likely represent a lower bound on the measurement quality of the task and item statistics.

A second limitation relates to the test design process. ECD involves specifying the knowledge, skills, and abilities that are the focus of the development process. Both task development and assessment design need to provide evidence of the specified knowledge, skills, and abilities in order to develop assessments that support explicit interpretations about score meaning (e.g., Mislevy *et al.*, 2003). Typically, frameworks are developed to clearly delineate the features of the construct to be assessed and then test items are developed to provide evidence for each of the construct dimensions. A FACT test design, therefore, requires a framework that specifies critical tasks of teaching, student learning targets, and the number of assessment tasks that need to be developed to provide evidence of each. The initial development reported in this pilot did not follow a full ECD process. Instead the focus was on developing prototype performance tasks that allowed for initial inquiry into the feasibility of this type of assessment technology. This initial type of pilot work is more a proof of concept than full ECD test development. To investigate the viability of this approach for developing a FACT test, it will be important to carry out a more complete ECD process. This needs to start by specifying a test design framework followed by task development that is explicitly focused on providing evidence for each of the identified framework components (see for example, Phelps *et al.*, 2020). Without a full ECD process it is difficult to evaluate the extent to which the pool of tasks that were developed for the pilot (Table 1) represent the desired construct, and whether task performance was associated with underlying construct dimensions or differences in task quality.

A third limitation relates to the pilot study design. The pilot was conducted with a relatively small sample of participants from a small number of EPP programs. These participants are almost certainly not representative of teacher candidates across the country. Basic descriptives, like the difficulty of tasks, will change when the tasks are administered to more representative samples. The conditions of testing are also dissimilar to typical conditions for licensure testing. Because there are no stakes attached to the results, candidates may put in different levels of effort than they would if they were taking the test under high stakes conditions (i.e., a pass/fail decision on receiving a teaching license). This could influence the amount of time participants take, how hard they try, the resulting difficulty parameters for the tasks, and even summary

statistics such as test reliability. Testing conditions could also influence how candidates respond to questions about their testing experience.

Finally, the current development and study administration was conducted without attending to a number of critical components of assessment development. For example, the current tasks were not developed to address accessibility for candidates with specific disabilities (e.g., blind, legally blind or low vision; deaf or hard of hearing). Consequently, the current tasks may not meet guidelines for test fairness. However, we anticipate that steps could be taken to make these tasks accessible by using various technologies such as screen readers, tactile manipulatives, and American Sign Language captioning.

These shortcomings will need to be addressed in future development and research cycles. A useful next step would be to revise the current tasks and scoring routines and implement identified issues with test accessibility. A subsequent pilot should then be conducted with a larger and more diverse sample of participants to reevaluate measurement properties. Ideally, subsequent studies would also follow a full ECD process to develop an assessment framework with the associated knowledge and performance tasks.

Conclusion

The work presented in this report represents an important step in developing a new type of licensure test that provides evidence of both the knowledge and skill that define the competencies used in effective subject matter teaching. This new licensure design, FACT, is referred to as “foundational,” because it is designed to assess the foundational tasks of teaching that function as the critical building blocks for the work of teaching at all grade levels and for all subjects. This focus is particularly relevant for use with entry-level teachers to ensure they have opportunities to develop and demonstrate the foundational competencies needed for beginning teaching. FACT focuses on assessing teaching competencies, because safe and effective teaching requires both knowledge and performance skills. Although FACT builds on prior research and development of assessments of CKT, there is a clear recognition that CKT can only provide evidence of what teachers know, but not necessarily what they can do in respect to content teaching. The new FACT performance tasks provide a promising new measurement technology that can complement CKT to provide evidence of both the knowledge and skill that define critical competencies needed for safe and effective beginning teaching.

Note

- 1 The electronic survey form failed for a number of administrations. Participants who could not complete the electronic form completed a paper-and-pencil version with identical questions. The data collected on the electronic administration was later combined with hand entered data from the paper-and-pencil version.

References

- Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, 59(5), 389–407. <https://doi.org/10.1177/0022487108324554>
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., & Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47(1), 133–180. <https://doi.org/10.3102/0002831209345157>
- Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, 54, 106–148. <https://doi.org/10.1111/j.1467-6494.1986.tb00391.x>
- Carlisle, J. F., Correnti, R., Phelps, G., & Zeng, J. (2009). Exploration of the contribution of teachers' knowledge about reading to their students' improvement in reading. *Reading & Writing*, 22(4), 457–486. <https://doi.org/10.1348/000711005X66419>
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309–319. <https://doi.org/10.1037/1040-3590.7.3.309>
- Copur-Gencturk, Y., Tolar, T., Jacobson, E., & Fan, W. (2018). An empirical study of the dimensionality of the mathematical knowledge for teaching construct. *Journal of Teacher Education*, 70(3), 485–497. <https://doi.org/10.1177/0022487118761860>
- Correnti, R., & Phelps, G. (2010). *Investigating the relationship between teachers' knowledge, literacy practice and growth in student learning*. Paper presented at the American Educational Research Association annual meeting, Denver, CO.

- Gitomer, D. H., Phelps, G., Weren, B., Howell, H., & Croft, A. J. (2014). Evidence on the validity of content knowledge for teaching assessments. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 493–528). Jossey-Bass. <https://doi.org/10.1002/9781119210856.ch15>
- Goldschmidt, P., & Phelps, G. (2010). Does teacher professional development affect content and pedagogical knowledge: How much and for how long? *Economics of Education Review*, 29(3), 432–439. <https://doi.org/10.1016/j.econedurev.2009.10.002>
- Herbst, P., & Kosko, K. (2014). Mathematical knowledge for teaching and its specificity to high school geometry instruction. In J. Lo, K. R. Leatham, & L. R. Van Zoest (Eds.), *Research trends in mathematics teacher education* (pp. 23–45). Springer. https://doi.org/10.1007/978-3-319-02562-9_2
- Hill, H. C., & Ball, D. L. (2004). Learning mathematics for teaching: Results from California's mathematics professional development institutes. *Journal for Research in Mathematics Education*, 35(5), 330–351. <https://doi.org/10.2307/30034819>
- Hill, H. C., Ball, D. L., Blunk, M., Lewis, J., Phelps, G., Sleep, L., & Zopf, D. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26(4), 430–511. <https://doi.org/10.1080/07370000802177235>
- Hill, H. C., Dean, C., & Goffney, I. M. (2007). Assessing elemental and structural validity: Data from teachers, non-teachers, and mathematicians. *Measurement: Interdisciplinary Research and Perspectives*, 5(2–3), 81–92. <https://doi.org/10.1080/15366360701486999>
- Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, 42(2), 371–406. <https://doi.org/10.3102/00028312042002371>
- Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, 105(1), 11–30. <https://doi.org/10.1086/428763>
- Howell, H., Lai, Y., Phelps, G., & Croft, A. (2016, April 8–12). *Assessing mathematical knowledge for teaching beyond conventional mathematical knowledge: Do elementary models extend?* Paper presented at the American Educational Research Association annual meeting, Washington, DC. <https://doi.org/10.13140/RG.2.2.14058.31680>
- Iaconangelo, C., Phelps, G., & Gitomer, D. (2020). *Dimensionality and validity of the Content Knowledge for Teaching construct using cognitive diagnostic modeling and known groups comparisons* [manuscript submitted for publication].
- Kersting, N. B. (2008). Using video clips of mathematics classroom instruction as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement*, 68(5), 845–861. <https://doi.org/10.1177/0013164407313369>
- Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, 49(3), 568–589. <https://doi.org/10.3102/0002831212437853>
- Kleickmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., Krauss, S., & Baumert, J. (2013). Teachers' content knowledge and pedagogical content knowledge: The role of structural differences in teacher education. *Journal of Teacher Education*, 64, 90–106. <https://doi.org/10.1177/0022487112460398>
- Krauss, S., Baumert, J., & Blum, W. (2008). Secondary mathematics teachers' pedagogical content knowledge and content knowledge: Validation of the COACTIV constructs. *ZDM Mathematics Education*, 40(5), 873–892. <https://doi.org/10.1007/s11858-008-0141-9>
- Liu, S., & Phelps, G. (2019). Does teacher learning last? Understanding how much teachers retain their knowledge after professional development. *Journal of Teacher Education*. Advance online publication. <https://doi.org/10.1177/0022487119886290>
- Luecht, R. M. (2017). Data and scale analysis for credentialing examinations. In S. Davis-Becker & C. W. Buckendahl (Eds.), *Testing in the professions: Credentialing policies and practice* (pp. 123–152). Routledge. <https://doi.org/10.4324/9781315751672-7>
- McCrorry, R., Floden, R., Ferrini-Mundy, J., Reckase, M., & Senk, S. (2012). Knowledge of algebra for teaching: A framework of knowledge and practices. *Journal for Research in Mathematics Education*, 43(5), 584–615. <https://doi.org/10.5951/jresmetheduc.43.5584>
- Mikeska, J. N., Kurzum, C., Steinberg, J. H., & Xu, J. (2018). *Assessing elementary teachers' content knowledge for teaching science for the ETS® educator series: Pilot results* (Research Report No. RR-18-20). Educational Testing Service. <https://doi.org/10.1002/ets2.12207>
- Mikeska, J. N., Phelps, G., & Croft, A. J. (2017). *Practice-based measures of elementary science teachers' content knowledge for teaching: Initial item development and validity evidence* (Research Report No. RR-17-43). Educational Testing Service. <https://doi.org/10.1002/ets2.12168>
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02
- Mohr-Schroeder, M., Ronau, R. N., Peters, S., Lee, C. W., Bush, W. S. (2017). Predicting student achievement using measures of teachers' knowledge for teaching geometry. *Journal for Research in Mathematics Education*, 48(5), 520–566. <https://doi.org/10.5951/jresmetheduc.48.5>
- Phelps, G. (2005). *Content knowledge for teaching reading* (unpublished manuscript). University of Michigan, Ann Arbor, MI.
- Phelps, G. (2009). Just knowing how to read isn't enough! What teachers know about the content of reading. *Educational Assessment, Evaluation and Accountability*, 21(2), 137–154. <https://doi.org/10.1007/s11092-009-9070-6>

- Phelps, G., Corey, D., DeMonte, J., Harrison, D., & Ball, D. L. (2012). How much English language arts and mathematics instruction do students receive? Investigating variation in instructional time. *Educational Policy*, 26(5), 631–662. <https://doi.org/10.1177/0895904811417580>
- Phelps, G., Gitomer, D. H., Iaconangelo, C. J., Etkina, E., Seeley, L., & Vokos, S. (2020). Developing assessments of content knowledge for teaching using evidence-centered design. *Educational Assessment*, 25(2), 91–111. <https://doi.org/10.1080/10627197.2020.1756256>
- Phelps, G., Gitomer, G., Weren, B., Croft, A. (2014). *Developing content knowledge for teaching assessments for the Measures of Effective Teaching study* (Research Report No. RR-14-33). Educational Testing Service. <https://doi.org/10.1002/ets2.12031>
- Phelps, G., Howell, H., & Liu, S. (2019). Exploring differences in mathematical knowledge for teaching for prospective and practicing teachers. *ZDM*, 52, 255–268. <https://doi.org/10.1007/s11858-019-01097-x>
- Phelps, G., Kelcey, B., Jones, N., & Liu, S. (2016). Informing estimates of program effects for studies of mathematics professional development using teacher content knowledge outcomes. *Evaluation Review*, 40(5), 383–409. <https://doi.org/10.1177/0193841X16665024>
- Phelps, G., & Schilling, S. (2004). Developing measures of content knowledge for teaching reading. *Elementary School Journal*, 105(1), 31–48. <https://doi.org/10.1086/428764>
- Phelps, G., & Sykes, G. (2020). The practice of licensure, the licensure of practice. *Phi Delta Kappan*, 101(6), 19–23. <https://doi.org/10.1177/0031721720909582>
- Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal*, 50(5), 1020–1049. <https://doi.org/10.3102/0002831213477680>
- Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1–23. <https://doi.org/10.1177/003172178705700101>
- Smith, P. S., & Banilower, E. R. (2015). Assessing PCK: A new application of the uncertainty principle. In A. Berry, P. Friedrichsen, & J. Loughran (Eds.), *Re-examining pedagogical content knowledge in science education* (pp. 88–104). Routledge.
- Tatto, M. T., Schwille, J., Senk, S., Ingvarson, L., Peck, R., & Rowley, G. (2008). *Teacher education and development study in mathematics (TEDS-M)*. Teacher Education and Development International Study Center, College of Education, Michigan State University.
- Tröbst, S., Kleickmann, T., Heinze, A., Bernholt, A., Rink, R., & Kunter, M. (2018). Teacher knowledge experiment: Testing mechanisms underlying the formation of preservice elementary school teachers' pedagogical content knowledge concerning fractions and fractional arithmetic. *Journal of Educational Psychology*, 110(8), 1049–1065. <https://doi.org/10.1037/edu0000260>
- van Driel, J. H., Verloop, N., & de Vos, W. (1998). Developing science teachers' pedagogical content knowledge. *Journal of Research in Science Teaching*, 35(6), 673–695. [https://doi.org/10.1002/\(SICI\)1098-2736\(199808\)35:6<673::AID-TEA5>3.CO;2-J](https://doi.org/10.1002/(SICI)1098-2736(199808)35:6<673::AID-TEA5>3.CO;2-J)
- Weren, B., Kinsey, D., Dickerman, D., & Phelps, G. (2020). *Scoring FACT performance task responses: Procedures, results and lessons* [Manuscript in preparation].

Appendix A

Task of Teaching Framework From Measures of Effective Teaching Study (Phelps et al., 2014)

Task of teaching	Description	Mathematics	Reading Language Arts
1. Anticipating student challenges, misconceptions, partial strengths, interests, capabilities, and background knowledge	This task of teaching describes the work of integrating knowledge about content to be learned and about anticipating how students are likely to interact with the content and content practices. This work is key in planning instruction, both in advance and in the moment as adjustments are made. These anticipations can become the basis for selecting appropriate explanations, examples, or tasks as instructional strategies	<ul style="list-style-type: none"> Anticipating student challenges in reasoning about and doing mathematics due to the interplay of content demands and students' understanding Anticipating the impact of limited English language proficiency on students' comprehension of mathematics concepts and on their ability to present mathematical ideas, make mathematical arguments, and give explanations Anticipating likely misconceptions, partial conceptions, and alternate conceptions about particular mathematics content and practices Anticipating student interest and motivation around particular mathematics content and practices 	<ul style="list-style-type: none"> Anticipating student challenges in reading, writing, speaking, or listening due to the interplay of content demands and students' understanding Anticipating the impact of limited English language proficiency on students' comprehension of text and speech and on their written and spoken expression Anticipating likely misconceptions, partial conceptions, and alternate conceptions about particular ELA concepts, texts, resources, and processes Anticipating student interest and motivation around particular ELA concepts, texts, resources, and processes Anticipating how students' background knowledge, life experiences, and cultural background can interact with new ELA concepts, texts, resources, and processes
2. Evaluating student ideas evident in work, talk, actions and interactions	This task of teaching describes the work of making sense of things that students do, say, and produce. It can involve deciding if an idea is valid and/or if an explanation is adequate in a particular context, identifying evidence of partial or alternate understandings, or analyzing errors. Evaluating includes characterizing, analyzing, describing, and judging student work	<ul style="list-style-type: none"> Evaluating student work, talk, or actions in order to identify conceptions in mathematics, including incorrect or partial conceptions Evaluating student explanations or arguments for use of appropriate mathematical practices Evaluating non-standard responses for evidence of mathematical understanding and in terms of efficiency, validity, and generalizability Evaluating student representations for evidence of mathematical understanding Evaluating or comparing multiple solutions to a mathematics problem or explanations of a mathematical concept or procedure Evaluating discussion among groups of students for evidence of understanding mathematics concepts and practices 	<ul style="list-style-type: none"> Evaluating student work, talk, or actions in order to identify conceptions, including incorrect or partial conceptions, about ELA concepts, texts, and processes Evaluating student work, talk, and actions for evidence of strengths and weaknesses in reading, writing, speaking, and listening Comparing multiple responses to an ELA task Evaluating discussion among groups of students for evidence of understanding ELA concepts, texts, and processes

Appendix A Continued

Task of teaching	Description	Mathematics	Reading Language Arts
3. Explaining concepts, procedures, representations, models, examples, definitions, and hypotheses	This task of teaching describes the crafting and using of appropriate explanations at any point in time when explanations are called for, including in setting purposes for instruction, planning, giving feedback, or in responding to students in the moment. Explaining also includes activities that show, such as modeling	<ul style="list-style-type: none"> Explaining mathematical concepts, or why a mathematical idea is “true” Explaining mathematical procedures Explaining mathematical representations Explaining mathematical models Explaining mathematical examples Explaining mathematical definitions Explaining mathematical hypotheses Explaining why a particular definition, model, or representation serves a particular purpose Interpreting a particular representation in multiple ways to further understanding Explaining why a practice, process, or procedure is efficient or appropriate Explaining the context of a mathematical conjecture and why it is a conjecture 	<ul style="list-style-type: none"> Explaining literary or language concepts, using definitions, examples, and analogies when appropriate Explaining processes of reading, including why certain processes are appropriate for particular texts and/or tasks Explaining processes of writing, including why certain processes are appropriate for particular tasks Explaining processes of listening and speaking, including why certain processes are appropriate for particular topics, groups, and/or task Explaining the use of representations to support understanding or development of text Explaining hypotheses about texts
4. Creating and adapting resources for instruction (examples, models, representations, explanations, definitions, hypotheses, procedures)	This task of teaching describes the work of inventing new instructional tools or adapting available ones to fit particular purposes. While curricular resources can be a source of material for adaptation, student-generated materials can also be adapted into opportunities for learning. The creation or adaptation of instructional tools is thought of as having an instructional goal, examples of which are in the subject-specific bulleted lists. Examples are selected for a reason — to introduce new material, to showcase certain attributes of content, or to challenge students’ current conceptions	<ul style="list-style-type: none"> Creating and adapting examples to introduce a concept or illustrate an idea or to demonstrate a strategy, procedure or practice Creating and adapting examples that support particular mathematical strategies or to address particular student questions, misconceptions, or challenges with content Creating and adapting representations or models to support students’ mathematical understanding Creating and adapting multiple representations or models to support students’ mathematical understanding across representations Creating and adapting representations or models that support multiple interpretations Creating and adapting definitions to fit instructional purposes Adapting student-generated conjectures to support instructional purposes Creating and adapting procedures for working with content 	<ul style="list-style-type: none"> Creating and adapting examples or model texts to introduce a concept or to demonstrate a literary technique or a reading, writing, or speaking strategy Creating and adapting examples or model texts to develop understanding of a concept, literary technique, or literacy strategy, or to address particular student questions, misconceptions, or challenges Creating and adapting representations (e.g., graphic organizers, think-alouds) to support understanding or development of text or other products Creating and adapting definitions of ELA concepts and processes to fit instructional purposes Adapting student contributions to fit instructional purposes Creating and adapting procedures for writing, speaking, listening, or working with texts Creating and adapting analogies to support student understanding of ELA concepts, texts, and processes

Appendix A Continued

Task of teaching	Description	Mathematics	Reading Language Arts
5. Evaluating and selecting resources for instruction (examples, models, representations, explanations, definitions, hypotheses, procedures)	This task of teaching describes the work of evaluating instructional resources and selecting those most appropriate for a particular instructional purpose. This work can occur through advance planning or during instruction. Teachers may evaluate and select from curricular materials or from potential resources generated during the course of instruction, including work generated by students	<ul style="list-style-type: none"> Evaluating and selecting examples to introduce a concept or illustrate an idea or to demonstrate a strategy, procedure or practice Evaluating and selecting examples that support particular mathematical strategies or to address particular student questions, misconceptions, or challenges with content Evaluating and selecting representations or models to support students' mathematical understanding, or in terms of validity, generalizability, or fit to the concept, calculation, etc. to be represented Evaluating and selecting representations or models that support multiple interpretations Evaluating and selecting definitions to fit instructional purposes Evaluating and selecting explanations of mathematical concepts for potential to support mathematical learning or in terms of validity, generalizability, or explanatory power Evaluating and selecting procedures for working with mathematics content Evaluating and selecting student-generated content to support instructional purposes Evaluating and selecting resources for their potential to support mathematical learning 	<ul style="list-style-type: none"> Evaluating and selecting examples or model texts to introduce a concept or to demonstrate a literary technique or literacy strategy Evaluating and selecting examples to develop understanding of a concept, literary technique, or literacy strategy, or to address particular student questions, misconceptions, or challenges with content Evaluating and selecting representations (e.g., graphic organizers, think-alouds) to support understanding or development of text or other products or in terms of fit to the concept to be represented Evaluating and selecting definitions of ELA concepts and processes to fit instructional purposes Evaluating and selecting explanations of ELA concepts, texts, and processes for their potential to support learning Evaluating and selecting procedures for writing or working with text Evaluating and selecting resources to support particular instructional goals Evaluating and selecting analogies to support student understanding of ELA concepts, texts, and processes
6. Developing questions, activities, tasks, and problems to elicit student thinking	This task of teaching includes the work of developing and using questions, tasks, and problems to engage students in developing content understanding. The development of questions, tasks, and problems is used to make student understanding visible and to build deeper and more accurate understanding	<ul style="list-style-type: none"> Creating or adapting problems or questions with the potential to elicit student mathematical thinking, discussions, justifications, or explanations Creating or adapting problems that support particular mathematical strategies and practices Creating or adapting questions, activities, tasks, or problems that demonstrate desired mathematical characteristics Creating or adapting classes of problems that address the same mathematical concept or that systematically vary in difficulty and complexity Creating or adapting questions, activities or tasks to elicit evidence that students have a particular mathematical understanding or skill 	<ul style="list-style-type: none"> Creating or adapting questions with the potential to elicit discussion or analysis of a text or a literary concept Creating or adapting prompts or questions with the potential to elicit productive student writing Creating or adapting activities or tasks that support the development of a particular literary understanding or skill Creating or adapting questions, activities or tasks to elicit evidence that students have a particular literary understanding or skill

Appendix A Continued	Task of teaching	Description	Mathematics	Reading Language Arts
7. Evaluating and selecting student tasks (questions, tasks, problems) to elicit student thinking	This task of teaching includes the work of evaluating and selecting things for students to work on or interact with. The evaluation and selection of student tasks is thought of as having an implicit instructional goal and relies on anticipations about how students will understand and interact with the tasks	<ul style="list-style-type: none"> Evaluating and selecting problems or questions to elicit student mathematical thinking, justifications, or explanations Evaluating and selecting problems that support particular mathematical strategies and practices Evaluating and selecting questions, tasks, or problems that demonstrate desired characteristics Evaluating and selecting classes of problems that address the same mathematical concepts or that systematically vary in difficulty or complexity Evaluating and selecting questions, activities or tasks to elicit evidence that students have a particular mathematical understanding or skill 	<ul style="list-style-type: none"> Evaluating and selecting questions, activities or tasks to elicit discussion about a specific text or literary concept Evaluating and selecting questions to elicit productive student writing Evaluating and selecting questions, activities or tasks to support the development of a particular literary understanding or skill Evaluating and selecting questions, activities or tasks to elicit evidence that students have a particular literary understanding or skill 	
8. Doing the work of the student curriculum	This task of teaching includes doing the student level tasks that make up the intended curriculum. Doing the student level work is not uniquely demanded by teaching, but is an essential part of the content work necessary to do in the course of teaching, especially in preparation for assigning student tasks. This work is often incorporated into or prerequisite for doing other tasks such as anticipating student difficulties or selecting problems or tasks	<ul style="list-style-type: none"> Doing the work that will be demanded of the students as part of the intended curriculum 	<ul style="list-style-type: none"> Doing the work that will be demanded of the students as part of the intended curriculum 	

Appendix B

FACT Tasks

RLA 11

FACT
Question 1 of 20
1:50
👁️
▶️ NEXT

Allowed preparation time: 2 minutes


The students in your first-grade class are generally performing at grade level. You notice that many students are not regularly using end punctuation in their writing, resulting in run-on sentences. As part of your writing instruction, you have been working on improving sentence structure.

Your task is to do the following.

- Show students how to determine where to add end punctuation to the declarative sentences.
- Narrate your thinking as you punctuate the sentences.
- Explain how the revisions you made helped to improve the quality of the writing.

You must use the work space in order to get full credit for your response.

When you are ready to begin your recording, select the microphone. You may use the stylus to write on the work space during your performance. Once you select the microphone to begin, you cannot stop and restart the recording.



your work space

👉 Pointer
■ Pen color & thickness
🖍️ Highlighter
🧼 Eraser
🗑️ Erase all

I had fun with my friend we
played on the swings.

RLA 12

FACT
Question 2 of 20
2:32
👁️
▶️ NEXT

Allowed preparation time: 3 minutes


In order to informally assess third-grade students' reading fluency, you ask students to choose a short passage from a book that they are reading independently and have them read it to you aloud.

You have noticed that a student, Sofia, has struggled with comprehension of text. Listen to Sofia's reading of the passage below, from *Because of Winn-Dixie*.

After you listen to Sofia read the passage, your task is to do the following.

- Provide Sofia with targeted verbal feedback about her reading fluency.
- Begin your feedback with the phrase "I noticed . . ."
- Link your suggestions to how reading in this way can improve her comprehension.

You may use the work space in your performance if you wish.



Select the play button to start the video.
The video file can only be played once.

your work space

👉 Pointer
■ Pen color & thickness
🖍️ Highlighter
🧼 Eraser
🗑️ Erase all

The dog went running over to the manager, wagging his tail and smiling. He stood up on his hind legs. You could tell that all he wanted to do was get face to face with the manager and thank him for the good time he was having in the produce department, but somehow he ended up knocking the manager over. And the manager must have been having a bad day, because lying there on the floor, right in front of everybody, he started to cry. The dog leaned over him, real concerned, and licked his face.

"Please," said the manager. "Somebody call the pound." (p.9)

Video:

(Sofia reads rapidly without stops or pauses for punctuation and without inflection.)

RLA 13

FACT
Question 3 of 20
3:41

Allowed preparation time: 4 minutes

Students in a sixth-grade class have read the poem "Nothing Gold Can Stay" by Robert Frost. As part of a whole-class discussion on theme, the students concluded that the theme of the poem was "Nothing good lasts forever." Students noted the last line of the poem, "Nothing gold can stay," as evidence of this theme. You ask the students to work with a partner to find additional evidence in the poem that supports the theme "Nothing good lasts forever."


Included is a video of a discussion between two students as they find additional evidence to support the theme "Nothing good lasts forever."

You hear Alex and Jordan's discussion as you are monitoring the students' conversation. You hear similar ideas from other students. You decide that the class would benefit from further discussion about how authors present the theme of a work.

Your task is to do the following.

- Identify for the students their limited conception about the language the author uses to show the poem's theme.
- Use one example from the poem to help students expand their understanding of using evidence to justify a theme.
- Explain to the students how their limited conception affects their understanding of the poem's theme.

You must use the work space in order to get full credit for your response.



Nothing Gold Can Stay

Robert Frost

Nature's first green is gold,
Her hardest hue to hold.
Her early leaf's a flower;
But only so an hour.
Then leaf subsides to leaf.
So Eden sank to grief,
So dawn goes down to day.
Nothing gold can stay.

Video:

Alex: I think the only place in the poem the author shows the idea that nothing good lasts forever is right here in the last line. *(Alex looks down and underlines the final line of the poem)*

Jordan: *(Jordan looks at Alex when he is done writing.)* Why do you think that?

Alex: *(Alex and Jordan look at each other when talking.)* Because it says nothing gold can stay

Jordan: Is that the only time he says that? *(Jordan looks down at tablet.)*

Alex: *(Alex looks at tablet.)* I think so. The other lines talk about leaves and flowers and other stuff, and the last line is the only one where the author says anything about how long things stay

Jordan: *(Jordan looks up and at Alex.)* OK. I couldn't find any other lines where the author talks about things staying around.

Alex: *(Alex looks at Jordan.)* Yeah. I think we got it

Jordan: Me too.

RLA 14

FACT Question 4 of 20 2:45 **NEXT**

Allowed preparation time: 3 minutes

The students in a fourth-grade class have researched and drafted short informational pieces about the advantages and disadvantages of different energy sources. Sameer has written his first draft about wind energy.


As part of the revision process, you confer with Sameer about the organization of his writing.

Your task is to do the following.

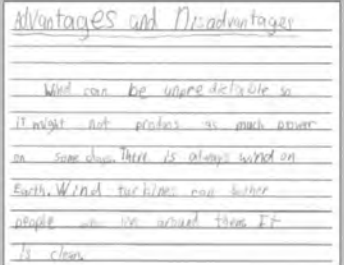
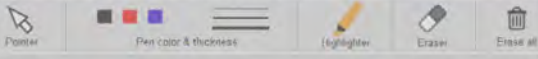
- Explain to Sameer how the organization of his current draft interferes with the reader's ability to understand which facts are advantages and which facts are disadvantages.
- Begin your feedback with the phrase "I noticed . . ."
- Use examples from Sameer's draft in order to support your feedback.

You must use the work space in order to get full credit for your response.

When you are ready to begin your recording, select the microphone. You may use the stylus to write on the work space during your performance. Once you select the microphone to begin, you cannot stop and restart the recording.



your work space



RLA 15

FACT Question 5 of 20 1:50 **NEXT**

Allowed preparation time: 2 minutes


Students in your kindergarten class are learning to identify the sounds they hear in individual words. They have been practicing segmenting words in which one letter is represented by only one sound. Today you are going to continue to develop their phonemic awareness by focusing on words that have digraphs, one sound that is represented by two letters.

Your task is to do the following.


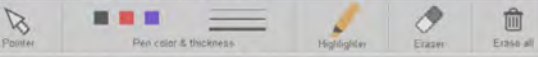
- Explain to students that they are going to practice segmenting words in which two letters make one sound.
- Model for students how to segment the words on the work space (back, chin, shop).
- Explain how hearing the sounds in words will help them as readers and writers.

You must use the work space in order to get full credit for your response.

When you are ready to begin your recording, select the microphone. You may use the stylus to write on the work space during your performance. Once you select the microphone to begin, you cannot stop and restart the recording.



your work space



RLA 21

FACT
Question 6 of 20
1:46
👁️
NEXT

Allowed preparation time: 2 minutes

When students read, they often encounter words that they do not know. One strategy that can help them to identify the meaning of unknown words is structural/word part analysis. It is helpful to identify prefixes, suffixes, and root words.

Fourth-grade students have been introduced to common prefixes and suffixes and their meanings. During a unit on ecosystems, students encounter a passage containing an unknown word in their science textbook.

To help students apply their previous knowledge of word part analysis to a new situation, you will use the work space to model for students how to use prefixes and root words to determine the meaning of the word **nonliving**.

Your task is to do the following.

- Model for students how to use a prefix and root word to determine the meaning of the word **nonliving**.
- Help students generalize the prefix non- by using these words as examples in your explanation: **nonfiction, nonstop**.
- Make sure to include how your definition of **nonliving** fits the context in which it is located.

You must use the work space in order to get full credit for your response.

When you are ready to begin your recording, select the microphone. You may use the stylus to write on the work space during your performance. Once you select the microphone to begin, you cannot stop and restart the recording.

your work space

Pointer
Pen color & thickness
Highlighter
Eraser
Erase all

*In ecosystems, many organisms support each other. For example, food sources decompose and provide nutrients to organisms. This is one way that **nonliving** organisms work with other organisms in the environment.*

RLA 22

FACT
Question 7 of 20
1:52
👁️
NEXT

Allowed preparation time: 2 minutes

The students in a second-grade class are generally performing at grade level. As part of your writing instruction, you have been working on improving sentence quality.

You notice that many students are using short, choppy sentences in their writing.

Use the sentences provided in the work space to model how to combine sentences.

Your task is to do the following.

- Show students how to combine three short sentences into one sentence.
- Narrate your thinking as you combine the sentences.
- Explain how the revisions you made helped to improve the quality of the writing.

You must use the work space in order to get full credit.

When you are ready to begin your recording, select the microphone. You may use the stylus to write on the work space during your performance. Once you select the microphone to begin, you cannot stop and restart the recording.

your work space

Pointer
Pen color & thickness
Highlighter
Eraser
Erase all

I was in the park. I saw a squirrel. It was furry and gray.

RLA 23

FACT
Question 8 of 20
3:45
👁️
NEXT

Allowed preparation time: 4 minutes

Students in a fifth-grade class often use the terms "main idea" and "theme" interchangeably. You decide to use the following fable as the basis for your instruction about how to explain the difference between main idea and theme.

You have identified the main idea and theme of a fable for the class and have written each of them out for the students to see.

Your task is to do the following.

- Use the main idea and the theme that are provided to explain the difference between the concepts of "main idea" and "theme" to the students.
- Define each term and use the main idea and theme from the fable to illustrate the difference between the two terms
- Explain to the students how this type of literary analysis supports their understanding of texts

You must use the chart provided in the work space in your explanation to get full credit for your response.

"The Fox and the grapes"

One hot summer's day a Fox was strolling through an orchard till he came to a bunch of grapes just ripening on a vine which had been trained over a lofty branch. "Just the thing to quench my thirst," quoth he. Drawing back a few paces, he took a run and a jump, and just missed the bunch. Turning round again with a "One, Two, Three!", he jumped up, but with no greater success. Again and again he tried after the tempting morsel, but at last had to give it up, and walked away with his nose in the air, saying: "I am sure they are sour."

When you are ready to begin your recording, select the microphone. You may use the stylus to write on the work space during your performance. Once you select the microphone to begin, you cannot stop and restart the recording.

your work space

Pointer
Pen color & thickness
Highlighter
Erase
Erase all

Main Idea	Theme
-----------	-------

Main Idea: Fox sees grapes high on a vine. He jumps, but he can't reach them. He then decides they are sour, so he isn't missing anything.

Theme: It's easy to dislike what you can't have.

RLA 24

FACT
Question 9 of 20
2:53
👁️
NEXT

Allowed preparation time: 3 minutes

Fourth-grade students are writing personal narratives. You are teaching the students to improve the quality of their writing by adding sensory details (sight, sound, taste, touch, smell).

You have defined the term "sensory details" for your class and drafted the following passage to use in your example.

My favorite place is a lake in New Hampshire. I go there each summer. There is so much to do. Each morning I swim in the lake. The mountains are beautiful. I hike on the trails. I also like to play games in the woods with my friends.

You are going to model for students how to add sensory details.

Your task is to do the following.

- Add sensory details to the sentence "Each morning I swim in the lake."
- Begin your recording with the statement "I am going to show you where I am going to add sensory details."
- Think aloud while you explain to the students why adding sensory details to their writing is important.

You must use the work space in order to get full credit for your response.

When you are ready to begin your recording, select the microphone. You may use the stylus to write on the work space during your performance. Once you select the microphone to begin, you cannot stop and restart the recording.

your work space

Pointer
Pen color & thickness
Highlighter
Erase
Erase all

My favorite place is a lake in New Hampshire. I go there each summer. There is so much to do. Each morning I swim in the lake. The mountains are beautiful. I hike on the trails. I also like to play games in the woods with my friends.

RLA 25

FACT Question 10 of 20 2:42 **NEXT**

Allowed preparation time: 3 minutes

The students in a first-grade class are relying on using only words that they know how to spell in their writing. You want to model for them how to use developmental spelling by using the sounds that they know to spell words.


You plan to use the Elkonin Boxes below to model for students how to put one letter in each box using the sounds that they know in the words **boat**, **seat**, and **bike**.

Your task is to do the following.

- Explain to students your expectation that they try to spell new words using the sounds that they know.
- Model for students how to use developmental spelling by segmenting the words into sounds and then writing the sounds that you hear. You should write just one letter in each box, and that letter might not represent conventional spelling.
- Explain to students how using developmental spelling will help them improve their writing and reading.

You must use the work space in order to get full credit for your response.

When you are ready to begin your recording, select the microphone. You may use the stylus to write on the work space during your performance. Once you select the microphone to begin, you cannot stop and restart the recording.



your work space

Pointer Pen color & thickness Highlighter Eraser Erase all

Boat
Seat
Bike

MATH 11

FACT Question 11 of 20 2:24 **NEXT**

Allowed preparation time: 3 minutes

The fourth-grade students in your class have been learning about characteristics of two-dimensional shapes.


You want to introduce students to the concept of lines of symmetry.

Your task is to do the following.

- Explain to the students what a line of symmetry is in language that is appropriate for students at their grade level.
- Include at least one example of a line of symmetry in your explanation.

You must use the work space in order to get full credit for your response.

When you are ready to begin your recording, select the microphone. You may use the stylus to write on the work space during your performance. Once you select the microphone to begin, you cannot stop and restart the recording.



your work space

Pointer Pen color & thickness Highlighter Eraser Erase all

MATH 12

FACT
Question 12 of 20
3:54
👁️
NEXT

Allowed preparation time: 4 minutes

The fifth-grade students in your class have been working on adding decimals. They are familiar with place-value language and the standard algorithm for adding decimals.


One student explained how he found the answer to $25.3 + 9.46$.

Review the video of Sameer explaining his work.

Your task is to do the following.

- Pose one question to Sameer about the work he did to solve the problem that will elicit more information about his understanding of place value.
- Once you have posed the question to Sameer, end your performance. Do not pose other questions, and do not provide an answer to the question.

You may use the work space in your performance if you wish.



Select the play button to start the video.
The video file can only be played once.

your work space

Pointer
Pen color & thickness
Highlighter
Eraser
Erase all

$$\begin{array}{r}
 25.3 \\
 + 9.46 \\
 \hline
 34.76
 \end{array}$$

Video:

Sameer: First I wrote the problem up and down and lined up the decimals. (*Sameer looks down at his tablet.*) I wrote down the 6 ... (*Sameer looks up.*) ... since that was by itself. (*Sameer looks down at his tablet.*) Then I added 3 and 4 and got 7, ... (*Sameer looks up.*) ... so, I wrote that down. (*Sameer looks down at his tablet.*) Then 5 plus 9 is 14, ... (*Sameer looks up.*) ... so, I wrote down 4 and put the 1 up top. (*Sameer looks down at his tablet.*) Then 1 plus 2 is 3, so I wrote down 3, ... (*Sameer looks up.*) ... and my answer was thirty-four point seven six.

MATH 13

FACT
Question 13 of 20
3:48
👁️
NEXT

Allowed preparation time: 4 minutes

The third-grade students in your class have been working on finding equivalent fractions.

You want to use an area model to explain to the students why $\frac{2}{3}$ is equivalent to $\frac{6}{9}$.

Your task is to do the following.

- Demonstrate how to represent $\frac{2}{3}$ using the area model on the work space.
- Model and explain how to partition the same area model into equal parts so that it represents $\frac{6}{9}$.
- Explain why the area model shows that $\frac{2}{3}$ is equivalent to $\frac{6}{9}$.


You must use the work space in order to get full credit for your response.

your work space



Pointer
Pen color & thickness
Highlighter
Eraser
Erase all

--	--	--

When you are ready to begin your recording, select the microphone. You may use the stylus to write on the work space during your performance. Once you select the microphone to begin, you cannot stop and restart the recording.



MATH 14

FACT Question 14 of 20 4:50  **NEXT** 

Allowed preparation time: 5 minutes

The second-grade students in your class are able to add two one-digit numbers, and they understand that the two digits of a two-digit number represent amounts of tens and ones. You recently introduced addition of two two-digit numbers to them.


Two students, Sofia and Neel, used different methods to find the answer to $48 + 25$.

Review the students' work and the video of the students explaining their methods.

Your task is to do the following.

- Explain to Neel what he did incorrectly.
- Use Sofia's method to explain to Neel how to find the correct answer to the problem.

You must use the work space in order to get full credit for your response.



Select the play button to start the video.
The video file can only be played once.

your work space

Pointer
Pen color & thickness
Highlighter
Eraser
Erase all

Sofia $\begin{array}{r} 48 \\ + 25 \\ \hline 73 \end{array}$	Neel $\begin{array}{r} 48 \\ + 25 \\ \hline 63 \end{array}$
---	--

Video:

Neel: (Sofia and Neel are looking at each other.) How did you add the numbers, Sofia?



Sofia: (Sofia looks down at her tablet.) I added 8 plus 5 in the ones place and got 13. (Neel looks down at his tablet.) The 3 goes in the ones and the ten goes over the tens. Then I added my tens.

I had 4 plus 2, plus one more, so I wrote 7 in the tens place.

Sofia: (Sofia and Neel look up at each other.) My answer is 73. How did you get 63?

Neel: (Neel looks down at his tablet.) First I added 8 and 5. I got 13, and I put down the 3. (Neel looks up at Sofia.) Next I added 4 and 2. I got 6, so the answer is 63.

MATH 15

FACT Question 15 of 20 3:26  **NEXT** 

Allowed preparation time: 4 minutes

The third-grade students in your class have been working on dividing whole numbers. They have been thinking about how to represent division problems and the two different models of division that can be used, which are described as follows.

- In one model of division, the quotient (answer) is the number of groups.
- In the other model of division, the quotient (answer) is the size of each group.

Two students, Ruby and Diego, used different methods to correctly find the answer to $6 \div 3$.


Review the video of Ruby and Diego explaining their methods.

You want to use the two students' methods to emphasize the two different models of division for the students in the class.

Your task is to do the following.

- Restate each method and elaborate on it to support other students' understanding.
- Visually represent each method in a way that reinforces the two different models of division.
- Emphasize the difference in the meaning of the quotient in the two students' answers.


You must use the work space in order to get full credit for your response.



Select the play button to start the video.
The video file can only be played once.

your work space

Pointer
Pen color & thickness
Highlighter
Eraser
Erase all



Video:

Diego: *(Diego and Ruby are looking at each other.)* What did you do to answer the problem, Ruby?

Ruby: *(Ruby looks down at her tablet.)* Well, I started with 3, then 3 more is 6. *(Ruby looks up at Diego.)* So, two threes is 6, so the answer is 2. Did you do the same thing?

Diego: Mine was a little different. *(Diego looks down at his tablet.)* I thought that if I have 3 buckets for 6 things, then I can start by putting 1 thing in each bucket. Then I can put a second thing in each bucket. *(Diego looks up at Ruby.)* Then I do not have any more things to put in the buckets. So, 6 divided by 3 is 2.

MATH 21

FACT
Question 16 of 20
3:51
👁️
NEXT ➔

Allowed preparation time: 4 minutes

The first-grade students in your class have been working on adding within 20. One student, Ben, explained how he used counters to find the answer to $8 + 7$.


The counters Ben used are represented on the work space, and additional counters are available for you to use, if you wish.

Review the video of Ben using the counters to find the sum.

Your task is to address a misconception that Ben has by doing the following.

- Explain to Ben what he did incorrectly.
- Demonstrate and explain to Ben how to correctly use the **count-on strategy** to find the answer.

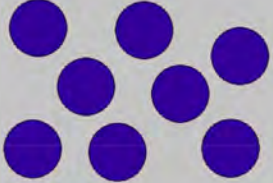
You must use the work space in order to get full credit for your response.

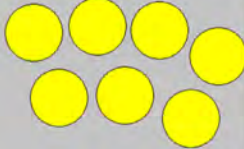


Select the play button to start the video.
The video file can only be played once.

your work space

Pointer
● ●
Pen color & thickness
Highlighter
Erase
Erase all





Video:

Ben: *(Ben looks down at the tablet and points to the blue circles.)* There are 8 over here. So, 8, *(Ben points to one of the yellow circles)* ... 9, 10, 11, 12, 13, 14. *(Ben points to one circle each time he says a number.)*

Note: Video shows the arrow moving from circle to circle while the student counts.

MATH 22

FACT Question 17 of 20 3:45 **NEXT**

Allowed preparation time: 4 minutes

The fourth-grade students in your class have been working on using number lines to solve word problems involving addition and subtraction of fractions. You want to show them how to use a number line to represent and solve a problem with a comparison structure. A problem with a comparison structure involves two distinct amounts and the difference between them.

Your task is to do the following.

- Use the problem provided on the work space.
- Represent the problem using the number line provided on the work space.
 - Demonstrate how to label the number line.
 - Represent the comparison structure of the problem.
- Explain how to use the number line to solve the problem.

You must use the work space in order to get full credit for your response.

When you are ready to begin your recording, select the microphone. You may use the stylus to write on the work space during your performance. Once you select the microphone to begin, you cannot stop and restart the recording.

your work space

Mia and Wyatt each work in a garden. Mia works in the garden for $\frac{4}{5}$ of an hour, and Wyatt works in the garden for $\frac{1}{5}$ of an hour. How much more time did Mia work in the garden than Wyatt did?

MATH 23

FACT Question 18 of 20 3:45 **NEXT**

Allowed preparation time: 4 minutes

The second-grade students in your class are able to add two one-digit numbers, and they understand that the two digits of a two-digit number represent amounts of tens and ones. You recently introduced addition of two two-digit numbers to them.

Two students, Kia and Rey, used different methods to find the answer to $29 + 34$.

Review the students' work and the video of the students explaining their methods.

You want to compare Kia's method with Rey's method to emphasize concepts of place value for the students in the class.

Your task is to do the following.

- Compare Kia's method and Rey's method. Include at least one similarity or difference between their methods in your comparison.
- Emphasize concepts of place value in your comparison.

You may use the work space in your performance if you wish.

Select the play button to start the video. The video file can only be played once.

your work space

Kia	Rey
$\begin{array}{r} 29 \\ + 34 \\ \hline 63 \end{array}$	$\begin{array}{r} 29 \\ + 34 \\ \hline 13 \\ + 50 \\ \hline 63 \end{array}$

Video:

Rey: How did you get your answer?

Kia: (*Kia looks down at her tablet.*) First I added the ones. 9 plus 4 is 13. (*Rey looks down at his tablet.*) I put down the 3 in the ones place and I wrote the 1 above the tens. Then in the tens, 1 plus 2 is 3, and 3 plus 3 is 6, so I put down 6 in the tens place. (*Kia and Rey look up at each other.*) The answer is 63. What did you do, Rey?

Rey: (*Rey looks down at his tablet.*) 9 plus 4 is 13 and 20 plus 30 is 50. (*Rey looks up at Kia.*) Then 13 more than 50 is 63, so the answer is 63.

MATH 24

FACT
Question 19 of 20
3:30
👁️
➡️ NEXT

Allowed preparation time: 4 minutes

The fourth-grade students in your class have been learning about finding factor pairs for whole numbers between 1 and 100.

You want to introduce students to the concept of composite numbers.

The following is a dictionary definition of a composite number.

composite number:

1. an integer exactly divisible by at least one positive integer other than itself or 1.
2. a number that is the product of at least two numbers other than itself and 1.

Your task is to do the following.

- Explain to the students what a composite number is in language that is appropriate for students at their grade level.
- Include at least one example of a composite number in your explanation.

You must use the work space in order to get full credit for your response.

When you are ready to begin your recording, select the microphone. You may use the stylus to write on the work space during your performance. Once you select the microphone to begin, you cannot stop and restart the recording.

your work space
🗑️ Erase

👉 Pointer
■ ■ ■ Pen color & thickness
🖍️ Highlighter
🧽 Eraser

MATH 25

FACT
Question 20 of 20
2:56
👁️
➡️ NEXT

Allowed preparation time: 4 minutes

The third-grade students in your class have been working on using visual fraction models to generate equivalent fractions and then explaining why the fractions are equivalent.

One student created the following set model to represent $\frac{3}{4}$.

If the set model represents $\frac{3}{4}$, your task is to do the following.

- Explain to the student why the set model represents $\frac{3}{4}$.
- Demonstrate to the class how to create a different set model that uses 12 counters and represents $\frac{3}{4}$.

If the set model does not represent $\frac{3}{4}$, your task is to do the following.

- Explain to the student why the set model does not represent $\frac{3}{4}$.
- Demonstrate to the class how to correctly create a set model that uses 12 counters and represents $\frac{3}{4}$.

You must use the work space in order to get full credit for your response.

When you are ready to begin your recording, select the microphone. You may use the stylus to write on the work space during your performance. Once you select the microphone to begin, you cannot stop and restart the recording.

your work space
🗑️ Erase

👉 Pointer
● ● Pen color & thickness
🖍️ Highlighter
🧽 Eraser

Appendix C

Participant Comments about FACT

Examples of participant responses to the open-ended questions about the FACT performance tasks are provided below under each of the identified categories. The number of candidates giving a response coded at that category and the corresponding percent of total number of candidates is provided in parentheses.

Assesses teaching skills: (*n* = 27, 46%)

- “I liked that I was able to get tested on how I verbally instructed students instead of whether I was aware of information being taught. I feel this method is good to incorporate on exams and it should be taught more in education courses. Some students may be aware of information on paper but need more help expressing this to students for a level of understanding.”
- “I found it interesting that the FACT assessment was actually assessing my ability to teach a child. I have not experienced another teacher performance assessment that does this, which I believe is necessary in the teacher licensure process. The skills assessed are much more realistic to daily teaching life in this assessment than in any other I have encountered.”
- “I enjoyed how this test used examples that I would find in actual classrooms from students. It is helpful seeing the videos in order to see how the students are thinking when explaining their answer. I enjoyed this test because it did not just ask me how I would teach/explain something, but I was able to actually demonstrate how I would approach this to a student.”

Authentic to teaching: (*n* = 25, 42%)

- “I think it is more realistic than a multiple-choice exam. It is much more practical and gives real-life issues that may happen in a classroom anywhere. It makes you think about your responses and how you might approach a problem in a natural way. It is also much more specific than generic one size fits all multiple-choice tests.”
- “I found that the FACT assessment was a much more authentic way for me to be able to showcase my teaching abilities. I was required to plan, evaluate, and provide instruction to students in a similar way I would in the classroom. I think that these were all real tasks/situations that would arise in the classroom.”
- “It has a great way of testing who we are as teachers, in practice, without being a burden on the classroom. This test was faster and arguably a better indicator of my actual practical skills as a teacher because of the timed nature. It presents situations that are so common that it would be the same effect if someone came into my current field placement and waited for students to ask questions to test me.”

Addresses student misconceptions: (*n* = 11, 19%)

- “I think that the benefit of this test is that one is able to see how a teacher can actually teach their students who may be struggling. It allows the person taking the test to show their skills of re-teaching when a student is struggling and also allows for the person grading the test to be able to evaluate how the test taker will handle different situations in their future classroom as they are allowed to showcase their different skills.”
- “A performance test like FACT appropriately assesses the abilities of teacher candidates to address student misconceptions, reteach topics not fully understood by students, and introduce new concepts. Since basic information on the content was provided, the assessment was based more on teaching rather than understanding or memorizing content.”
- “My overall impression was that the assessment was challenging. It addressed a wide variety of topics and really made you think about responding to students and their needs. I think the student samples and explanations were more challenging since you are not talking to real life students in a class setting. However, I found this assessment to be beneficial in preparing me for future student struggles and misconceptions.”

Thinking on your feet: ($n = 9$, 15%)

- “Teachers have to be able to explain, draw, discuss, give feedback, etc. on the spur of the moment all the time. Allowing only a couple of minutes of thinking and prep time is very similar to what actually happens in the classroom, so this feels authentic. Almost all other assessments I have taken feel like they have little to no connection to the daily classroom practice and challenges.”
- “I think the benefits of a performance assessment like FACT is to give pre-service teachers the opportunity to showcase their teaching abilities with planning, evaluating and providing instruction. It required me to quickly evaluate and make an instructional decision which occurs in the classroom every day. I think this is more beneficial than asking pre-service teachers to show what they know through multiple choice test.”
- “I think there are benefits in that it puts teachers on the spot and requires them to pull from prior experiences to be able to respond to these situations. In my opinion, this form of assessment is more beneficial than things like multiple choice tests and long analysis or reflections.”

Determining where I need to improve: ($n = 8$, 14%)

- “The benefit of having this kind of assessment is that it will guide you in determining what teacher preparation programs are really doing to prepare future teachers. It will also allow you to determine what kinds of content teachers need reinforcement in. This assessment made me realize that I need to review content in all areas because it is important to know what knowledge they should leave with as they enter the next grade level.”
- “I can determine if I am able to explain concisely and effectively because of the time constraint. I can notice what topics in language arts and math I need to improve because FACT gave a variety of elementary education questions found in a typical school curriculum that I could be asked to teach to students.”
- “I was really impressed and a bit startled by the nature of the test. Besides the initial shock it had a great way of being able to show what I was actually prepared, or not prepared to do as a teacher. It showed a variety of tasks, and by providing the definitions of some of the skills it made a clear distinction between the content and the pedagogy. It was really interesting that some of the tasks were in direct response to a student’s needs, and that those responses were almost entirely based on our assessment.”

Include student interaction: ($n = 12$, 20%)

- “To really reflect teaching performance you need student feedback, because you don’t really want teachers to teach at students. You want teachers that take student input into account and teach with them, especially in today’s focus on student-centered tasks. Normally, I would give students some time to reflect and think about tasks as well, not just explain at them.”
- “Some parts felt a bit awkward because I am used to leading a more interactive lesson where students answer my questions posed.”
- “One thing I wish was for the ability to have a conversation because my teaching style is to have more of a conversation with my students vs. me directly teaching like that.”

Suggested citation:

Phelps, G., Bridgeman, B., Yan, F., Steinberg, J., Weren, B., & Zhou, J. (2020). *Preliminary evidence on measurement characteristics for the Foundational Assessment of Competencies for Teaching performance tasks* (Research Report No. RR-20-27). Educational Testing Service. <https://doi.org/10.1002/ets2.12310>

Action Editor: Elizabeth Stone

Reviewers: Michael Kane and Caroline Wylie

ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS RESEARCHER database at <http://search.ets.org/researcher/>