Measuring the Power of Learning.™

Research Report
ETS RR–18-20

Assessing Elementary Teachers' Content Knowledge for Teaching Science for the *ETS*® Educator Series: Pilot Results

Jamie N. Mikeska

Christopher Kurzum

Jonathan H. Steinberg

Jun Xu

December 2018

ETS Research Report Series

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.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Assessing Elementary Teachers' Content Knowledge for Teaching Science for the *ETS*® Educator Series: Pilot Results

Jamie N. Mikeska, Christopher Kurzum, Jonathan H. Steinberg, & Jun Xu

Educational Testing Service, Princeton, NJ

The purpose of this report is to examine the performance of assessment items designed to measure elementary teachers' content knowledge for teaching (CKT) science as part of the *ETS*® Educator Series. The Elementary Education: CKT Science assessment is 1 component of licensure examination through the *PRAXIS*® assessments. The Elementary Education: CKT Science assessment is designed to determine whether kindergarten through 6th-grade elementary teacher candidates have the essential content knowledge needed for teaching elementary science as new teachers at the entry level. This report provides information about the development of the Elementary Education: CKT Science framework and the associated assessment items. The main part of the report focuses on the evidence gathered when piloting 104 CKT science assessment items with 417 preservice and novice elementary teachers. This evidence included the following: (a) how these new CKT science assessment items function, including their item difficulties and discrimination, and reliabilities of the pilot forms and of the classification of examinees; (b) how teachers perceive the importance and relevance of these new assessment items; (c) how teachers' performances on these items relate to their background characteristics and professional and academic preparation; and (d) how teachers' performances on the CKT science items compare to their performances on an assessment designed to measure only their science subject matter knowledge.

Keywords Science; licensure assessment; content knowledge for teaching; pedagogical content knowledge; specialized content knowledge; subject matter knowledge

doi:10.1002/ets2.12207

In this report, we examine the performance of assessment items designed to measure elementary teachers' content knowledge for teaching (CKT) science as part of the *ETS*® Educator Series. The Elementary Education: CKT Science assessment is one component of licensure examination through the *PRAXIS*® assessments. The Elementary Education: CKT Science assessment is designed to determine whether kindergarten through sixth-grade elementary teacher candidates have the essential content knowledge needed for teaching elementary science as new teachers at the entry level. In particular, the CKT assessments for elementary licensure were designed as part of collaboration between ETS and the TeachingWorks organization at the University of Michigan.

The CKT assessments focus on the knowledge that elementary teachers use as they engage in various tasks of teaching in four different content areas.[1] These tasks of teaching explicitly target the ways in which teachers use their content knowledge as they respond to critical content challenges that arise within their daily work, such as eliciting and interpreting students' ideas, selecting resources for instruction, and generating explanations and examples. As such, the CKT assessment in each subject area aims to assess both the content knowledge that novice teachers need to do the work of the student curriculum (e.g., knowledge of key science concepts and ideas targeted within particular grade levels) and the specialized knowledge that they apply when engaged in specific tasks to teach the student curriculum. For example, on the Elementary Education: CKT Science assessment, teacher candidates may be asked to determine which procedures would be most useful to address a specific scientific investigation question, to select a scientific model or representation that is well matched to a particular instructional goal, or to evaluate a student-generated scientific explanation for consistency in a student's claim, evidence, and reasoning. These aspects of CKT are critical for ensuring that novice teachers are well prepared and have the full range of content knowledge needed to enter the teaching profession.

Corresponding author: J. N. Mikeska, E-mail: jmikeska@ets.org

This research report presents the results from the piloting of 104 CKT science assessment items with 417 preservice and novice elementary teachers from across the United States. In particular, this report addresses the following research questions:

1. How do preservice and novice elementary teachers perform on assessments of CKT elementary science?
2. How do preservice and novice elementary teachers perceive the importance and the relevance of assessments of CKT elementary science?
3. Are there significant differences in teachers' performances on the Elementary Education: CKT Science pilot assessment based on background characteristics and professional and academic preparation?
4. How do scores on the Elementary Education: CKT Science pilot assessment compare to scores on the PRAXIS Elementary Education: Multiple Subjects Science licensure test?

The report begins with a brief overview describing the use of licensure assessments in teaching and the theory of CKT. In the second section, we describe the pilot study's methodology, including the blueprints and item development processes, the sample, and the data collection and analysis processes used. We then present the main findings from the pilot study, organized by the four research questions stated above, and end with a discussion of the implications of these results for assessing CKT elementary science.

Overview

Purpose and History of Licensure Assessments in Teaching

Licensure and certification examinations are one of the primary means for determining whether novices are ready to enter a profession, and these examinations have been used in a variety of fields, such as teaching, medicine, law, real estate, accounting, and architecture (Clauser, Margolis, & Case, 2006; Wang, Schnipke, & Witt, 2005). The main purpose of these examinations is to ensure that individuals who will practice in a given profession have met certain standards or requirements (Raymond & Luecht, 2013; Schmitt, 1995). These examinations ultimately serve to protect the public by identifying those who are ready to work in a profession while barring unqualified novices from entering the workforce in a particular area (Shimberg, 1981). As noted by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), these examinations target the "standards of competence needed for effective performance" and "protect the public by excluding persons who are deemed to be not qualified to do the work of the profession or occupation" (p. 175).

Across professions, a variety of assessment formats are employed for licensure and certification examinations, ranging from selected-response formats to constructed-response formats to performance tasks (Tannenbaum, Robustelli, & Baron, 2008). As Raymond and Luecht (2013) mentioned, a number of factors influence decisions about an assessment format, including "the purpose of the credential and the inferences to be made from test scores, the desired level of score reliability, administration costs, scoring logistics, and the perceived fidelity of the implementation" (p. 402). In addition, a common trend across licensure and certification examinations is the approach taken to specify the knowledge, skills, and abilities used to define the test content domain and develop the test specifications. In particular, some form of job analysis is typically used to identify the main tasks or activities in which novices need to be able to engage successfully to demonstrate beginning competency in that profession (Reese & Tannenbaum, 1999). A task inventory questionnaire is frequently used to further refine the importance of these job-specific tasks and activities and determine the content that should be assessed in the examination. This general approach has been used extensively when developing teacher licensure assessments.

Across the nation, all states mandate some form of teacher testing requirements to assess novice teachers' readiness to enter the profession (Wilson, 2016). Historically, teacher licensure assessments have focused on assessing candidates' basic academic skills, including their ability to read, write, and solve mathematical problems and their more in-depth content knowledge of specific subjects they will teach, such as their knowledge of physics or biology (Wilson, 2016). The focus on teachers' knowledge of the subject matter in licensure examinations has a strong grounding in the idea that teachers cannot teach what they do not know (Wilson, Floden, & Ferrini-Mundy, 2001). However, in addition to assessing teacher candidates' basic academic skills and more in-depth content knowledge in the subject areas, a little more than half of the states also require novices to pass licensure examinations in teaching pedagogy to receive certification (Wilson, 2016).

Licensure assessments of teachers' pedagogical knowledge frequently target the knowledge and skills required for engaging in general teaching practices across content areas, such as understanding the purpose and use of formative

Table 1 Science Examples of Tasks Designed to Assess Common Content Knowledge and Practice-Based Content Knowledge (Specialized CKT)

Science content area	Tasks that require only common content knowledge	Tasks that require mostly specialized CKT
Physical science	Explain the distinction between physical and chemical changes	Choose instructional materials that would be most appropriate for illustrating the distinction between physical and chemical changes
Life science	Identify the role of animals in plant pollination	Evaluate models that students generated showing the role of animals in plant pollination for evidence of scientific understanding
Earth and space science	Determine patterns in the locations of earthquakes and volcanoes and what causes those patterns	Select which of several pictorial representations would be most useful to display evidence of plate movement

Note. CKT = content knowledge for teaching.

and summative assessments or knowing about principles and strategies for classroom management or for addressing the needs of diverse learners. For example, the PRAXIS Principles of Teaching and Learning assessment series provides a compilation of examinations at different grades designed to assess beginning teachers' knowledge of foundational educational principles.[2] However, more recent trends in the field have moved toward assessing the knowledge and skills that teachers use as they engage in the actual work of teaching in the subject areas.

Content Knowledge for Teaching

Across disciplines, many researchers agree with the theoretical notion that there is a form of knowledge that is unique to teaching and distinctly different from what well-educated individuals know about a specific discipline (Gess-Newsome, 2015; Hill, Ball, & Schilling, 2008; Loughran, Mulhall, & Berry, 2004). In their seminal piece, Ball, Thames, and Phelps (2008) called this *content knowledge for teaching* and defined it as "knowledge 'entailed by teaching'" (p. 399), meaning the knowledge that is relevant to teaching practice in a specific discipline. In their practice-based conceptualization of CKT, this knowledge base includes two key components—subject matter knowledge and pedagogical content knowledge—which are further subdivided into various domains. Subject matter knowledge includes both common content knowledge, defined as one's understanding of core disciplinary concepts that are used in a wide variety of settings and professions, and specialized content knowledge that is "tailored to the work that teachers do with curriculum, instruction, and students" (Ball, Hill, & Bass, 2005, p. 16). Pedagogical content knowledge includes teachers' knowledge of content and students, content and teaching, and content and curriculum, all of which focus on the knowledge demands that occur at the intersection of the content, students, and teaching.

These latter parts of CKT—the specialized content knowledge and pedagogical content knowledge—focus directly on the professional knowledge that teachers use as they engage in critical tasks of teaching, such as eliciting and interpreting students' ideas and selecting resources for instruction. Despite the growing evidence for the importance of the full range of science teachers' CKT, efforts to develop measures assessing their CKT have focused mainly on one aspect of science teachers' subject matter knowledge—their common content knowledge (Minner, Martinez, & Freeman, 2012). To date, the other aspects of science teachers' CKT have been addressed to a limited extent by the current assessments available in science education, although there are a handful of notable exceptions (Jin, Shin, Johnson, Kim, & Anderson, 2015; Mikeska, Phelps, & Croft, 2017; Sadler, Sonnert, Coyle, Cook-Smith, & Miller, 2013; Smith & Taylor, 2010). Table 1 provides a few examples in three science content areas illustrating the distinction between assessment tasks that require teachers to use their common content knowledge alone and those that require teachers to draw upon other aspects of their CKT. As shown in this table, common content knowledge tasks focus on one's understanding of the disciplinary core ideas in science, while the other task examples target the ways in which science teachers use their conceptual understanding to engage in critical tasks of teaching science.

Calls for the creation of valid and reliable assessments of science teachers' CKT are at the forefront of recent appeals to develop indicator systems to track science and math teachers' learning on a national scale (National Research Council, 2013). These measures also can support better understanding of the mechanisms through which particular types of

learning opportunities are more or less effective for preservice and practicing science teachers (Wilson, 2013). Despite these calls, the field has made only limited progress in developing CKT science assessment items to test the full range of knowledge used in science teaching. The present study addresses this gap by developing and validating a set of CKT science assessment items emphasizing the knowledge that novice science teachers are required to use when they address the critical tasks of teaching elementary science.

Methods and Data Sources

In this section, we begin with an account of the development team's approach to designing the test content specifications and CKT elementary science items. These details help to set the stage for understanding how these items measure the full range of elementary science teachers' CKT, going beyond assessing their subject matter knowledge alone. We then describe the sample, data collection, and analysis methods for this pilot study.

Developing the Content Knowledge for Teaching Science Blueprints

The development team for the Elementary Education: CKT Science component of the ETS Educator Series used evidence-centered design (ECD) principles to develop the test content specifications for this new licensure assessment (Mislevy & Riconscente, 2006). ECD has been described as a "methodology for designing assessments that underscores the central role of evidentiary reasoning in assessment design" (Mislevy, Almond, & Lukas, 2003, p. 20). In particular, ECD emphasizes the importance of test specifications as a direct outgrowth of considerations regarding the test's purpose, the test-taking population, the use of the scores, what knowledge is sufficient for safe and effective entry into the profession, the assessment items needed to elicit evidence about test takers' knowledge, and how this evidence will be used to support claims that test takers possess sufficient knowledge for entry-level practice (Tannenbaum et al., 2008). As the focus for this report is on the psychometric quality of the items, a more detailed discussion of the ECD development process used is outside its scope. Instead, here we provide an overview of the process that the CKT science development team used to determine the essential content knowledge needed for teaching elementary school science at the entry level. In doing so, the development team sought to define two aspects: (a) the critical content knowledge students in the given grade range are expected to master, what we refer to as the *student-level content domain*, and (b) the science-specific teaching practices, or tasks of teaching science, that support student learning in the science content domain. It is the intersection of these two parts that defines the CKT required to support student learning in science.

The first step focused on identifying the science content that was most foundational to the elementary science curriculum and instruction and would be more likely to put students' future educational success at risk if that content was not well understood. To do so, the development team decided to draw upon the recently released Next Generation Science Standards (NGSS) (NGSS Lead States, 2013), which defined a set of student performance expectations in each of the four major science content domains: physical science; life science; earth and space science; and engineering, technology, and applications of science. Each performance expectation is a short statement identifying what a student should be able to do to show that he or she has met a particular standard. The performance expectations are unique in that each one combines one disciplinary core idea with one scientific or engineering practice and one crosscutting concept — what the NGSS refers to as "three-dimensional learning" (NGSS Lead States, 2013). In creating the NGSS, the developers were intent on including only the most critical aspects of science disciplinary content knowledge at the K – 12 level. *A Framework for K-12 Science Education* explained how a committee of science researchers, teacher educators, scientists, and teachers applied the following criteria to determine these core ideas: "(a) have broad importance across multiple sciences or engineering disciplines or be a key organizing principle of a single discipline, (b) provide a key tool for understanding or investigating more complex ideas and solving problems, (c) relate to the interests and life experiences of students or be connected to societal or personal concerns that require scientific or technological knowledge, and (d) be teachable and learnable over multiple grades at increasing levels of depth and sophistication" (National Research Council, 2012, p. 31). Each core idea had to meet at least two of these criteria, although the goal was to have each one meet three or all four criteria. Because the determination of these core ideas in science was recently conducted by a national committee using these well-specified criteria, the CKT science development team decided that these core science ideas should form the basis of the student-level content domain for this assessment. Table 2 highlights the major core ideas and component ideas composing each

Table 2 Science Content Categories and Specifications

Content domain	Core ideas	Component ideas
Physical science	Matter and its interactions	Structure and properties of matter
		Chemical reactions
	Motion and stability: Forces and interactions	Forces and motion
		Types of interactions
	Energy	Definition of energy, conservation of energy, and energy transfer
		Relationship between energy and forces
		Energy in chemical processes and everyday life
	Waves and their application in technologies for information transfer	Wave properties
		Electromagnetic radiation
		Information technologies and instrumentation
Life science	From molecules to organisms: Structures and processes	Structure and function
		Organization for matter and energy flow in organisms
	Ecosystems: Interactions, energy, and dynamics	Interdependent relationships in ecosystems
		Cycles of matter and energy transfer in ecosystems
		Ecosystem dynamics, functioning, and resilience
		Societal interactions and group behavior
	Heredity: Inheritance and variation of traits	Inheritance of traits and variation of traits
	Biological evolution: Unity and diversity	Evidence of common ancestry and diversity
		Natural selection
		Adaptation
		Biodiversity and humans
Earth and space science	Earth's place in the universe	The universe and its stars
		Earth and the solar system
		History of the planet Earth
	Earth's systems	Earth materials and systems
		Plate tectonics and large-scale system interactions
		The roles of water in Earth's surface processes
		Weather and climate
		Biogeology
	Earth and human activity	Natural resources
		Natural hazards
		Human impacts on Earth systems
Engineering, technology, and applications of science	Engineering design	Defining and delimiting an engineering problem
		Developing possible solutions
		Optimizing the design solution

Note. These science core ideas and component ideas are taken directly from the Next Generation Science Standards (NGSS Lead States, 2013) accessed at http://www.nextgenscience.org/

of the four science content domains, which was used to specify the content knowledge that novice teachers need to do the work of the student curriculum.

The second step involved specifying the most critical tasks of teaching in which elementary science teachers are required to engage from their first day on the job. We began this work by reviewing relevant standards and related documents—both those that were general in nature, such as the Interstate New Teacher Assessment and Support Consortium (InTASC) Standards (Council of Chief State School Officers, 2011) and the Council for the Accreditation of Educator Preparation Standards,[3] and those that targeted science-specific teaching practices from the National Research Council, such as the National Science Education Standards (National Research Council, 1996) and *A Framework for K-12 Science Education* (National Research Council, 2012). We also reviewed literature and empirical research that proposed a core set of instructional teaching practices in science (Kloser, 2014; Windschitl, Thompson, Braaten, & Stroupe, 2012) and in other disciplines.[4] In particular, we sought to identify the science-specific teaching practices that have been nominated or shown in the research literature to hold the most promise for positively impacting student learning. From this review of the relevant standards and research literature, our team drafted a list of the key instructional practices

in which elementary science teachers engage during their daily work, which could involve planning for instruction, interacting with parents and colleagues, engaging with their students during class time, or reflecting on their instruction or students' work products after an instructional episode. We then considered which ones seemed most critical for entry-level practice and revised the tasks-of-teaching list accordingly. The goal was to create a working draft for the committee of external practitioners to review and revise as part of the validity argument.

To focus the list on the content challenges that novice elementary science teachers face, we decided to organize the list by the instructional tools that science teachers use (e.g., scientific models, explanations, or investigations) and the instructional practices in which they engage with each of these tools. This approach is similar to a process developed in mathematics to specify the mathematical work of teaching for elementary teachers (Selling, Garcia, & Ball, 2016) and provided a way for our team to determine the instructional tools and practices that were fundamental to the work of beginning elementary science teachers for safe and effective entry into the profession. As such, the framework, which is shown in Table 3, focuses on critical tasks of teaching elementary science for novice practitioners and is organized into seven different instructional tool categories. The instructional practices noted in this table target problems that novice science teachers regularly encounter and must solve while teaching and provides the key contexts in which science teachers draw on their CKT science.

After creating this initial draft of the Elementary Education: CKT Science blueprints, the development team recruited a cadre of experienced science content experts, assessment developers, and science education researchers from Educational Testing Service (ETS) to conduct an internal review. Another round of intense and structured reviews of these test content specifications occurred with the convening of a national advisory committee, which consisted of 13 members (six current elementary science teachers and seven current or former science education faculty members) who had experience teaching science in an elementary or middle school classroom and had familiarity with the NGSS. Each of the college faculty members also had in-depth experience preparing teacher candidates to teach science in an elementary classroom. Appendix A provides details about these committee members, their affiliations, and their most recent positions. The main purpose of this external review was to support the validity argument for the CKT science assessment by confirming that the content (both the content descriptions and the tasks of teaching science) being assessed is important for safe and effective practice of beginning elementary teachers. In particular, we gathered recommendations about critical content that was missing, nonessential content that could be removed, and revisions to the language to improve clarity. We also sought their feedback on the relative importance of each of the content domains and tasks of teaching science categories for use in assembling the pilot forms.

Item Design and Development

The next step in the team's development process was to specify the nature and kinds of items that could be used to elicit evidence of novice elementary science teachers' CKT. As noted, the CKT science items were conceptualized at the intersection of the two aspects of the content framework — the student-level content domain and the tasks of teaching science. In particular, the team sought to specify how teachers use their conceptual understanding to productively engage in these tasks of teaching and how to develop assessment items that would measure that knowledge.

To develop appropriate items, we utilized a task structure for the CKT science assessment items that situates teachers within the work of teaching science. To do so, our team built upon recent work conducted in mathematics and science suggesting that embedding assessment tasks in scenarios that describe instructional situations and provide relevant information about students, curriculum, or the classroom is productive for assessing teachers' CKT (Gitomer, Phelps, Weren, Howell, & Croft, 2014; Hill, Schilling, & Ball, 2004; Mikeska et al., 2017). Therefore, each of the items designed to assess elementary teachers' CKT science included an instructional scenario to situate the knowledge being used in a specific task of teaching; this design targets a practice-based conceptualization of CKT.

For this study, the development team, which included a set of 30 outside item writers — some of which were external practitioners who had previously participated on the national advisory committee — authored a set of 104 items designed to assess elementary teachers' CKT science using two main item formats: (a) single-selection multiple-choice items and (b) technology-enhanced items, including multiple-selection multiple-choice items, tabular/grid items, drop-down selection items, and ordering items. Appendix B provides an example of a single-selection multiple-choice item that was designed to assess teachers' CKT about weather and climate. Appendix C shows an example of a technology-enhanced item, in this case a multiple-selection multiple-choice item, created to assess teachers' CKT about magnetic interactions

Table 3 Tasks of Teaching Science Organized by Instructional Tool

Instructional tool	Tasks of teaching science
Scientific instructional goals, big ideas, and topics	Selecting or sequencing age-appropriate, grade-level instructional goals or big ideas for a topic
	Identifying the big idea(s) or instructional goal(s) of an instructional activity
	Choosing which science ideas or instructional activities are most closely related to a particular instructional goal
	Linking science ideas to one another and to particular activities, models, and representations within and across lessons
Scientific investigations and demonstrations	Selecting investigations or demonstrations that facilitate understanding of disciplinary core ideas, scientific practices, or crosscutting concepts
	Evaluating investigation questions for quality (e.g., testable, empirical)
	Determining the variables, techniques, or tools that are appropriate for use by students to address a specific investigation question
	Critiquing scientific procedures, data, observations, or results for their quality, accuracy, or appropriateness
	Evaluating and selecting media for engaging students in virtual investigations not possible in firsthand situations
	Supporting students in generating questions for investigation or identifying patterns in data and observations
Scientific resources (texts, curriculum materials, journals, and other print and media-based resources)	Evaluating instructional materials and other resources for their ability to sufficiently address scientific concepts; engage students with relevant phenomena; develop and use scientific ideas; promote students' thinking about phenomena, experiences, and knowledge; provide a sense of purpose; take account of students' ideas; and assess student progress
	Choosing resources that support the selection of accurate, valid, and age-appropriate goals for science learning
Student ideas (including common misconceptions, alternate conceptions, and partial conceptions)	Analyzing student ideas for common misconceptions regarding intended scientific learning
	Selecting diagnostic items and eliciting student thinking about scientific ideas and practices to identify common student misconceptions and the basis for those misconceptions
	Developing or selecting instructional moves, approaches, or representations that provide evidence about common student misconceptions and help students move toward a better understanding of the idea, concept, or practice
	Identifying the connections between students' talk and work and scientists' talk and work
Scientific language, discourse, vocabulary, and definitions	Selecting scientific language that is precise, accurate, grade appropriate, and illustrates key scientific concepts
	Anticipating scientific language and vocabulary that may be difficult for students
	Supporting and critiquing students' participation in and use of verbal and written scientific discourse and argumentation
	Modeling the use of appropriate verbal and written scientific language in critiquing arguments or explanations, in describing observations, in using evidence to support a claim, etc.
Scientific explanations (includes claim, evidence, and reasoning)	Critiquing student-generated explanations or descriptions for their generalizability, accuracy, precision, or consistency with scientific evidence
	Selecting explanations of scientific phenomena that are accurate and accessible to students
Scientific models and representations (analogies, similes, metaphors, simulations, illustrations, diagrams, data tables, performances, videos, animations, graphs, examples)	Evaluating or selecting scientific models and representations that predict or explain scientific phenomena or address instructional goals
	Engaging students in using, modifying, creating, and critiquing scientific models and representations that are matched to an instructional goal
	Evaluating student models or representations for evidence of scientific understanding
	Generating or selecting diagnostic questions to evaluate student understanding of specific models or representations
	Evaluating student ideas about what makes for good scientific models and representations

Table 4 Content Knowledge for Teaching Science Items by Content Domain

Content domain	Core ideas	Pilot Form 1,[a] *n* (%)	Pilot Form 2,[a] *n* (%)	Test content specifications distribution goal (%)
Physical science	Matter and its interactions	6 (11.5)	6 (11.5)	—
	Motion and stability: Forces and interactions	5 (9.6)	4 (7.7)	—
	Energy	3 (5.8)	4 (7.7)	—
	Waves and their application in technologies for information transfer	3 (5.8)	5 (9.6)	—
	Total	17 (32.7)	19 (36.5)	30
Life science	From molecules to organisms: Structures and processes	5 (9.6)	6 (11.5)	—
	Ecosystems: Interactions, energy, and dynamics	2 (3.8)	3 (5.8)	—
	Heredity: Inheritance and variation of traits	4 (7.7)	3 (5.8)	—
	Biological evolution: Unity and diversity	3 (5.8)	2 (3.8)	—
	Total	14 (26.9)	14 (26.9)	30
Earth and space science	Earth's place in the universe	5 (9.6)	3 (5.8)	—
	Earth's systems	8 (15.4)	6 (11.5)	—
	Earth and human activity	2 (3.8)	4 (7.7)	—
	Total	15 (28.8)	13 (25.0)	25
Engineering, technology, and applications of science	Engineering design	6 (11.5)	6 (11.5)	—
	Total	6 (11.5)	6 (11.5)	15

Note. This table shows the distribution of the total set of content knowledge for teaching (CKT) science items in each pilot form across the four science content domains. The *n* reported within the cells of this table refers to the number of CKT science items in each pilot form that addressed each domain and the component ideas within these domains. The percentages reported in the table refer to the percentage of CKT science items within each pilot form that addressed each domain or component idea.
[a] $n = 52$.

between two objects not in contact with one another. As noted in Appendices B and C, each CKT item was categorized according to the two aspects of the test content specifications—the content categories and the tasks of teaching science categories—depending on the CKT being assessed in that particular item.

These 104 items were divided into two pilot forms (52 items per form) with the goal of representing the distributions across the four content categories and seven tasks of teaching science categories recommended by the national advisory committee. Tables 4 and 5 provide details about the CKT science items developed and used across the four content domains and the seven tasks of teaching categories for each pilot form. Overall, approximately 13% of the CKT science items were technology-enhanced items (eight items on Pilot Form 1 and six items on Pilot Form 2), with the majority of items using the single-selection multiple-choice item format.

Sample

The population of teacher candidates taking the science component of the PRAXIS Elementary Education: Multiple Subjects assessment during the 2015–16 testing year was used for recruitment purposes for this pilot study. The science subtest (referred to hereafter as PRAXIS 5005) is 50 minutes in length with 50 multiple-choice questions, for which scaled scores between 100 and 200 are generated based on the number of items each candidate answers correctly. This test measures candidates' subject matter knowledge in three science content domains—earth, life, and physical science—with each domain having approximately the same number of questions and covering material typically taught in a bachelor's degree program in elementary education.[5] As of September 2015, 19 states required the PRAXIS Elementary Education: Multiple Subjects tests for licensure.[6]

Table 5 Content Knowledge for Teaching Science Items by Tasks of Teaching Science Instructional Tools

Instructional tools categories	Pilot Form 1,[a] *n* (%)	Pilot Form 2,[a] *n* (%)	Test content specifications distribution goal (%)
Scientific instructional goals, big ideas, and topics	8 (15.4)	6 (11.5)	14
Scientific investigations and demonstrations	12 (23.1)	12 (23.1)	16
Scientific resources	4 (7.7)	7 (13.5)	8
Student ideas	11 (21.2)	8 (15.4)	18
Scientific language, discourse, vocabulary, and definitions	3 (5.8)	4 (7.7)	10
Scientific explanations	5 (9.6)	8 (15.4)	18
Scientific models and representations	9 (17.3)	7 (13.5)	18

Note. This table shows the distribution of the total set of content knowledge for teaching (CKT) science items in each pilot form across the instructional tools categories. While these specifications are organized by instructional tool, each CKT science item was designed to assess the knowledge targeted within a single task of teaching science (see Table 2). The *n* reported within the cells of this table refers to the number of CKT science items in each pilot form that addressed the tasks of teaching science within each of these instructional tools categories. The percentages reported in the table refer to the percentage of CKT science items within each pilot form that addressed the tasks of teaching science within each instructional tools category.
[a] $n = 52$.

Table 6 Eligibility Survey Classification for Recruited Participants

	% (*n*) of total sample[a]
Eligible for participation[b]	
Final year of teacher preparation program	32.9% (308)
Recent graduate of a teacher preparation program	14.3% (134)
Master's or alternative route teacher preparation program	18.4% (172)
Second to last year of teacher preparation program	9.8% (92)
Current teacher with fewer than 3 years of teaching experience	8.1% (76)
Recently started teacher preparation program	2.7% (25)
Not eligible for participation[c]	
Graduated from teacher preparation program more than 2 years ago	1.9% (18)
Current teacher with 3 or more years of teaching experience	3.3% (31)
Eligibility status unclear	8.6% (81)

[a] $N = 937$. [b] $N = 807$, 86.1% overall. [c] $N = 130$, 13.9% overall.

The recruitment efforts for this pilot study targeted the population of teacher candidates who completed the PRAXIS 5005 science subtest between September 2015 and August 2016. A staggered approach was used to recruit from this population and consisted of three groups, depending on when they completed the PRAXIS 5005 science subtest: between September 2015 and January 2016 (Recruitment Group 1), between February and May 2016 (Recruitment Group 2), and between June and August 2016 (Recruitment Group 3). The staggered approach was mapped to align with the PRAXIS 5005 administration dates and to provide our team with a more manageable outreach volume during the recruitment window.

All teacher candidates in Recruitment Groups 1 ($N = 6{,}962$) and 2 ($N = 5{,}898$) who completed the PRAXIS 5005 exam during this time were sent a recruitment letter via e-mail; those who expressed interest in participating were then sent a consent form and link to an eligibility survey to be completed online. For Recruitment Group 3, we purposefully identified candidates from underrepresented groups — in this case, all candidates who were male, were minority,[7] and/or scored in the two bottom quartiles for their PRAXIS 5005 performance — to diversify the sample pool in terms of representativeness to the extent possible and then filled in with candidates with other background characteristics to meet our recruitment target of 800 participants.[8] This decision meant that we ended up reaching out to a subset ($n = 1{,}625$) of individuals from Recruitment Group 3. Altogether, a little over 6% of the recruited participants (937 of the 14,485 candidates who received recruitment e-mails) expressed interest in participating in this pilot study, although only 807 of these candidates were deemed eligible, as shown in Table 6.

The PRAXIS 5005 test takers during the September 2015 to August 2016 testing year were used as the reference sample for pilot recruitment and sample selection. Key background information was used to determine representativeness, most

Table 7 Participant Background Characteristics

Demographic variable	PRAXIS 5005 test-taker population,[a] *n* (%)	Eligible participants from recruited sample,[b] *n* (%)	Pilot Form 1 participants,[c] *n* (%)	Pilot Form 2 participants,[d] *n* (%)
Gender				
Male	1,658 (9.2)	59 (7.3)	19 (9.0)	18 (8.7)
Female	16,358 (90.8)	746 (92.7)	191 (91.0)	189 (91.3)
Race/ethnicity[e]				
White	14,007 (77.7)	658 (81.7)	164 (78.1)	168 (81.2)
Asian	422 (2.3)	18 (2.2)	2 (1.0)	7 (3.9)
Hispanic	746 (4.1)	28 (3.5)	10 (4.8)	9 (4.3)
African American or Black	1,253 (7.0)	54 (6.7)	21 (10.0)	13 (6.3)
Native American	63 (.3)	2 (.2)	0 (.0)	1 (.5)
Two or more races	349 (1.9)	15 (1.9)	4 (1.9)	4 (1.9)
PRAXIS performance quartile				
Quartile 1 (100 – 158)	4,291 (23.8)	125 (15.5)	42 (20.0)	39 (18.8)
Quartile 2 (159 – 166)	4,212 (23.4)	173 (21.5)	52 (24.8)	59 (28.5)
Quartile 3 (167 – 176)	4,833 (26.8)	232 (28.8)	66 (31.4)	52 (25.1)
Quartile 4 (177 – 200)	4,680 (26.0)	275 (34.2)	50 (23.8)	57 (27.5)
Region[f]				
Northeast	6,756 (37.5)	297 (36.9)	74 (35.2)	67 (32.4)
Midwest	320 (1.8)	18 (2.2)	6 (2.9)	9 (4.3)
South	7,858 (43.6)	343 (42.6)	93 (44.3)	83 (40.1)
West	2,897 (16.1)	146 (18.1)	37 (17.6)	48 (23.2)
Teacher preparation program[g]				
Undergraduate education program (bachelor's)	11,322 (62.8)	508 (63.1)	135 (64.3)	145 (70.0)
Fifth-year postbaccalaureate program	578 (3.2)	30 (3.7)	2 (1.0)	10 (4.8)
Master's degree education program	3,574 (19.8)	185 (23.0)	51 (24.3)	34 (16.4)
Alternate route program	1,555 (8.6)	58 (7.2)	17 (8.1)	12 (5.8)

[a]Between September 1, 2015, and August 31, 2016. $N = 18{,}206$. [b]$N = 805$. PRAXIS 5005 data were unavailable for two recruited participants. [c]$N = 210$. [d]$N = 207$. [e]The respective counts and proportions of those with missing race/ethnicity were as follows: PRAXIS 5005 test-taker population, $n = 967$, 5.4%; eligible participants from recruited sample, $n = 25$, 3.1%; Pilot Form 1 participants, $n = 8$, 3.8%; Pilot Form 2 participants, $n = 5$, 2.4%. [f]The respective counts and proportions of those with missing or other regional classifications were as follows: PRAXIS 5005 test-taker population, $n = 185$, 1.0%; eligible participants from recruited sample, $n = 1$, .1%; Pilot Form 1 participants, $n = 0$, .0%; Pilot Form 2 participants, $n = 0$, .0%. [g]The respective counts and proportions of those with missing or other teacher preparation program status were as follows: PRAXIS 5005 test-taker population, $n = 449$, 2.5%; eligible participants from recruited sample, $n = 3$, .4%; Pilot Form 1 participants, $n = 0$, .0%; Pilot Form 2 participants, $n = 1$, .5%.

of which was self-reported by test takers at the time of registration using the standard PRAXIS questionnaire.[9] It should be noted that test-taker responses to specific questions were consolidated for purposes of analysis, namely, by race/ethnicity[10] and state of residency.[11] Our goal was to have the pilot sample represent, to the extent possible, similar distributions in the demographic characteristics and PRAXIS 5005 performance scores across the reference sample. In particular, to better examine item performance, especially item discrimination, it was important to recruit participants who would likely vary in their scores on these measures. To capture this variation in the recruited sample, we decided to rank order all test takers who took the PRAXIS 5005 during this 12-month time frame based on their PRAXIS 5005 scaled score and then divide the full group into PRAXIS performance quartiles.[12]

Based on program recommendations, we aimed for 460 examinees (230 participants per pilot form) and selected them purposefully according to gender, race/ethnicity, previous PRAXIS 5005 performance quartiles, geographic region, and teacher preparation category. Table 7 provides a comprehensive picture of teachers' background characteristics from the PRAXIS 5005 test-taker population during this 12-month period, the recruited and eligible sample, and the final Pilot Forms 1 and 2 participant pools. The PRAXIS 5005 test-taker population during this 12-month period was just over 90% female, slightly more than 75% White, largely from the Northeast and South regions, and largely engaged in undergraduate education programs. These statistics are relatively consistent with patterns found in particular with respect to gender and race/ethnicity in Elementary Education: Multiple Subjects test-taker data (Steinberg, Ling, & Delaney, 2016) and with respect to region, as noted earlier based on PRAXIS 5005 adoption patterns across states; this suggests that the

participating sample for this study was not significantly different on these characteristics from the overall PRAXIS 5005 test-taker population. Comparing the PRAXIS 5005 test-taker population during this 12-month period to the eligible sample, the group of eligible participants was slightly skewed more toward those who self-identified as female, White, and enrolled in master's degree education programs, while also having higher PRAXIS 5005 performance. The pattern for this last background characteristic is similar to previous PRAXIS performance based on an unpublished ETS study involving test takers who volunteered to take CKT English language arts or mathematics items following an operational PRAXIS Elementary Education form (Phelps, Bunde, Howell, & Steinberg, 2015). Examinee characteristics across the two forms are compared in the next section on data collection.

Data Collection

Once all 460 slots were filled, form assignments were completed to balance the composition across as many of the key demographic indicators as possible. Chi-squared tests of independence within and across gender groups, race/ethnicity groups (defined as being from a predominant minority group or not from a predominant minority group[13]), region, and PRAXIS 5005 performance quartile showed no bias in assignments toward one form or the other. As shown in Table 7, across both pilot forms, there was great similarity by gender with slight differences by race/ethnicity (higher proportion of African American examinees on Pilot Form 1, higher proportion of White examinees on Pilot Form 2), PRAXIS 5005 performance (higher proportion of examinees in the third performance quartile on Pilot Form 1, higher proportions of examinees from the second and highest performance quartiles on Pilot Form 2), and by region (higher proportion of examinees from the South on Pilot Form 1, higher proportion of examinees from the West on Pilot Form 2). These differences were somewhat more pronounced by teacher preparation program (higher proportion of examinees in master's degree programs on Pilot Form 1, higher proportion of examinees in undergraduate programs on Pilot Form 2).

The pilot window occurred during November 2 – 29, 2016, and the forms were administered online in an untimed setting using a proprietary platform developed by ETS. Participants were directed to complete the full online assessment during one 90-minute block of time in an undisturbed location of their choice, preferably at home or in their school settings, if available. Participants were allowed as much time as needed to complete the assessment. They were also instructed to answer the CKT science items without the use of any external resources, such as the Internet, teaching resources, or other colleagues. In addition to the set of 52 CKT science items, each online assessment form contained a perceptions survey and a background information questionnaire. The perceptions survey asked teachers to report on the extent to which they found these CKT science items to be challenging, how well they thought the set of CKT science items related to their learning in their teacher education programs and professional development, and the alignment between the knowledge assessed in these new items and the knowledge used in the work of teaching science in the elementary classroom. The background information questionnaire was designed to collect data on participants' background characteristics, teaching preparation and experience, certification status, and professional and academic preparation using well-known proxies, such as grade point average (GPA) and undergraduate major/minor.[14]

Test takers needed to complete 100% of the CKT science items to be included in the final analysis sample. For Pilot Form 1, there were initially 224 participants, and 14 were removed for not meeting the 100% CKT item completion requirement, leaving a total of 210 examinees. For Pilot Form 2, there were initially 219 participants, and 12 were removed for not meeting the 100% CKT item completion requirement, leaving a total of 207 examinees. Altogether, less than 6% of participants were excluded from the subsequent data analysis described for failure to achieve full completion on the pilot form.

Data Analysis

Item and Form Performance

Upon establishing the final samples for each of the forms, classical test analyses were conducted to evaluate the baseline psychometric quality of the complete forms and the associated CKT science items within them. Each item was scored as either correct or incorrect and counted as one point in the final scoring model. Descriptive statistics were generated for each item (proportion correct and point-biserial correlations) and for each form (Cronbach's alpha reliability), separately,

and items were removed based on recommendations in the field (Brennan, 2006). Regarding item difficulty as measured by proportion correct, items below .20 and above .95 were flagged for removal (Crocker & Algina, 1986). With respect to point-biserial correlations, although .30 was a more desired minimum value, .20 was an absolute accepted minimum value, and any items with point-biserial correlations less than .20 were also flagged for removal (Crocker & Algina, 1986). Distractor analyses[15] were also conducted to evaluate any potential incorrect answer keys or other instances where those participants performing well on the pilot form could have been inappropriately misled by one or more answer choices. Any such flagged items were brought to the attention of the science assessment developers, with their recommendations for retention or exclusion serving as the official decisions for the final scoring model.

In total, these analyses resulted in the removal of six items from Pilot Form 1 and seven items from Pilot Form 2 for the respective final scoring models. Two items were removed (one item per pilot form) for having low item difficulty, and eight items (three items for Pilot Form 1, five items for Pilot Form 2) were removed owing to low point-biserial correlations. Three items were removed (two items from Pilot Form 1, one item from Pilot Form 2) because both had low item difficulty and low point-biserial correlations. No items were removed owing to concerns with examinees' performance on the distractors. Reported findings for the item difficulty and discrimination and form reliabilities are based on the set of items with adequate item performance (46 items for Pilot Form 1, 45 items for Pilot Form 2).

Reliability of Classification

The reliabilities of classification accuracy and consistency were calculated for each form to evaluate the efficacy of decisions regarding whether examinees' test scores can be classified as passing relative to supplied cut-scores (Livingston & Lewis, 1995). The rate of classification *accuracy* ranges from 0% to 100% and represents the extent to which decisions made on the basis of this form of this test would agree with those made from all possible forms of the test without examinees having an opportunity to practice (i.e., an estimate of the examinees' true scores). The rate of classification *consistency* also ranges from 0% to 100% and represents the extent to which decisions made on the basis of one form of a test would agree with those made using an alternate form of the same test.

For tests with only a single cut-score, the method for determining classification *accuracy* would be based on a 2×2 table in which observed scores and "true" scores or "all forms average" (Livingston & Lewis, 1995, p. 180) are classified into passing or failing status. *Accuracy* is defined as the sum of the percentage of cases in which both the observed and "true" classifications agree. Classification *consistency* is defined similarly with the observed classifications and estimates of the alternate form classifications compared. Livingston and Lewis (1995) described the method used to estimate the true and alternate form score distributions given only the observed score distribution, the cut-score, and the effective test length. As the results in the present study were based on a pilot administration of the two forms, no cut-scores exist yet. Therefore the analyses in this section were run using a series of hypothetical cut-scores for estimating classification accuracy and classification consistency. The hypothetical cut-scores for each form were 50%, 60%, and 70% of the available raw score points. This range of cut-scores is comparable to the passing scores on other PRAXIS science tests, such as PRAXIS 5005, in which the raw passing scores range from approximately 56% to 66% of the available points.

It should be noted that the proportions of examinees accurately or consistently classified tend to increase as the differences between the mean score and the hypothetical cut-scores increase. Sample sizes of examinees at or near the hypothetical cut-scores become small when the mean scores are much larger or smaller than the hypothesized cut-scores. This can lead to lower likelihoods of inaccurate classification of examinees in the hypothesized cut-score region and therefore increase the proportions of examinees correctly and/or consistently classified. This becomes important when looking at rates for certain demographic subgroups, particularly race/ethnicity. Regarding evaluation of both sets of ratings, .80 was an absolute accepted minimum value, and .90 was a more desired minimum value.

Teachers' Perceptions of the Content Knowledge for Teaching Science Items' Importance and Relevance

At the end of each pilot form, each participant responded to a series of eight Likert-scale questions regarding his or her perceptions of the importance and relevance of the CKT science items. Means and standard deviations were calculated for

each perception survey question. We examined the overall trends in these results to determine the extent to which teachers felt that these CKT science items were challenging, targeted knowledge that was critical to the work of teaching elementary science, and linked to their learning opportunities in their teacher education programs and professional development. This part of the analysis served as a form of face validity to indicate whether the participants perceived these CKT science items to measure important and relevant knowledge for teaching elementary science.

Relationship Between Teachers' Performance and Background

To evaluate possible group differences in performance, we examined how participants' performance on these CKT science pilot forms related to selected background characteristics. The primary variables of interest were in two categories: demographic (e.g., gender and race/ethnicity) and professional/academic preparation (e.g., undergraduate GPA, undergraduate major/minor, certification status, and type of teacher preparation program). The first set of variables was considered to determine whether any group differences in performance may indicate the potential of adverse impact. On the other hand, the second set of variables provided some initial external validity evidence regarding how well teachers' performances align to other well-known proxies of their knowledge.

The first step in this part of the analysis was to construct the groups for comparison purposes. For the two demographic categories, we used the following groups: men and women for gender and African American, Asian, Hispanic, White, and two or more races for race/ethnicity. Because other research has shown differential passing rates on licensure exams in relation to demographics, most notably in terms of men passing these exams at higher rates than women (Gitomer, Brown, & Bonett, 2011; Steinberg et al., 2016) and White candidates passing these exams at higher rates than minority candidates (Gitomer, 2007; Nettles, Scatton, Steinberg, & Tyler, 2011), we anticipated similar group differences on the CKT science pilot forms.

For GPA, we divided participants into two groups—those who had a GPA of 3.5 or higher and those with a GPA less than 3.5. Similarly, we categorized participants into two groups for certification, according to whether they had (or were working toward) elementary certification with a focus in science (science certification vs. no science certification). We hypothesized that participants with higher GPAs and those who had more extensive science academic preparation would be more likely to have higher scores on the CKT science pilot forms on average, consistent with findings reported by Gitomer et al. (2011) and Steinberg et al. (2016) with respect to passing licensure exams.

In terms of their professional preparation, we examined group differences in performance based on where participants were in their teacher education programs: in or planning to enroll in a teacher education program (in program), graduated but not currently teaching (graduated), or currently teaching with fewer than 3 years of experience (teaching). We hypothesized that their program placement may be related to the extent of their teaching experience, whereby those who had graduated and were currently teaching likely had more teaching experience than those who were still in the program (albeit this experience was still limited compared to more seasoned elementary teachers). Because some research has suggested that science teachers' knowledge develops in practice (Lee, Brown, Luft, & Roehrig, 2007; Loughran, Berry, & Mulhall, 2006), we anticipated that those who had more teaching experience would have higher performance scores on average.

Finally, on the background questionnaire, participants selected which major and minor they held (or were working toward) in their undergraduate degree from a list of 13 options, such as elementary education, English or language arts, or mathematics or computer science. There was also an "other" category that allowed participants to enter their own response if they did not find it in the original list. Participants entered a variety of different majors/minors, such as business, religion, dance, autism studies, and gender studies, into the "other" category. We used this information to generate three main categories for analysis purposes: (a) education only with no minor or a minor that is not in a science discipline (education focus); (b) majored and/or minored in a science discipline, which was usually combined with a major or minor in education (science focus); and (c) no education or science undergraduate background (other focus). Similar to our prediction about certification, we anticipated that participants who had more extensive opportunities to learn about science content and/or pedagogy through a major or minor in a science discipline or in education would also tend to be higher performers on the CKT science pilot, in line with Gitomer et al. (2011). After categorizing teachers into these varied groups, we then conducted a series of *t*-tests or analyses of variance (ANOVA), as appropriate, by pilot form to identify any significant differences across subgroups.

Relationship Between Test Scores

To gather initial external validity evidence regarding the degree to which scores relate to other indicators, analyses were conducted to relate CKT performance to previous PRAXIS 5005 performance. This analysis was conducted by examining the correlations between CKT percentage correct values (range of 0% – 100% correct) and PRAXIS 5005 scale scores (range of 100 – 200) for each participant. We also created a scatterplot to visually represent the relationship between participant scores on these two assessments.

In particular, the theory of CKT suggests that strong subject matter knowledge is a necessary, but not sufficient, condition related to teachers' CKT. That is, in these CKT science items, teachers are required to use their understanding of science concepts to engage in critical tasks of teaching science, such as analyzing students' scientific thinking or evaluating scientific models or investigations. One would not expect teachers to do well on the CKT pilot form if they did not have the requisite content knowledge to apply in these tasks of teaching science. For example, it would be difficult (if not impossible) for one to evaluate whether a particular student diagram and written explanation reveal an understanding of why we see different moon phases without an understanding of the reason for this scientific phenomena.

In terms of this part of the analysis, the preceding logic suggests that there will be a general trend in these results whereby many teachers would score similarly across both assessments (e.g., by scoring on the lower side of the scale on both tests, or vice versa). However, there would also be a subset of teachers that may have comparatively higher scores on the subject matter knowledge assessment compared to the CKT science pilot form. This line of reasoning also suggests that teachers who do not have the requisite subject matter knowledge in science would not perform well on assessments of their CKT science. In terms of the analysis, we would expect to see a moderate correlation between participants' scores on these two assessments — one designed to measure teachers' science subject matter knowledge alone (PRAXIS 5005) and one designed to measure their CKT science — which would indicate that they are measuring two related, yet distinct, constructs.

Results

Teachers' Performance on the Pilot Forms

Item and Form Performance

Table 8 provides data on the CKT science item statistics by pilot form, while Figures 1 and 2 provide a graphical representation of the score distributions based on percentage correct across the full sample. On average, teachers' performances were similar across CKT science pilot forms. Findings show that the distribution of scores on each form fell across almost the full range of the scale, although no participant received a score lower than 10% correct on either pilot form. Likewise, the score distributions from both forms reveal a slightly negative skew, suggesting that the scores tend to cluster a bit more on the higher end of the score range. In particular, on both forms, approximately three-quarters of participant mean scores fell between the 40% and 90% correct range. In addition, the scale reliabilities for both forms fell between .85 and .90, which is within acceptable ranges for assessments of this type and length slated for high-stakes licensure decisions, in line with related findings from Steinberg, Brenneman, Castellano, Lin, and Miller (2014).

Table 9 shows the results of the analyses for item difficulty within each pilot form as measured by percentage correct. To describe the results from the pilot forms, difficult items were classified as those answered correctly by less than 40% of participants, whereas easier items were classified as those 80% or more of the participants answered correctly. Across both pilot forms, findings show that the majority of items were of moderate difficulty. Results indicate that only a small subset of the CKT science items were answered correctly by more than 80% of participants or by less than 40% of participants.

Table 10 shows the findings of the point-biserial correlations between each CKT science item score and the total raw score for that pilot form. This part of the analysis indicates how well the items discriminated among lower and higher performing participants. Overall findings indicate that the majority of items show at least a moderate relationship to the criterion measure (total raw score), suggesting that a large proportion of the items differentiate quite well. In fact, across both pilot forms, more than 80% of the CKT science items had biserials above .30.

As noted earlier, each CKT science item is classified according to one of the content domains and one of the instructional tool categories (see example items in Appendices B and C). Tables 11 and 12 highlight participants' mean proportion correct scores on the CKT science items in terms of the two major aspects of the framework: (a) science content domains

Table 8 Content Knowledge for Teaching Science Item Statistics by Pilot Form

	Pilot Form 1[a]	Pilot Form 2[b]
Total score points	46	45
Mean proportion correct (*SD*)	.62 (.17)	.60 (.19)
Minimum score	6	8
Maximum score	44	44
Scale reliability	.85	.89

Note. The *N* here refers to the total number of participants who reached 100% completion on each pilot form. The total score points for each pilot form only include the items that had adequate item difficulty (between .20 and .95) and point-biserial correlations (>.20).
[a]$N = 210$. [b]$N = 207$.

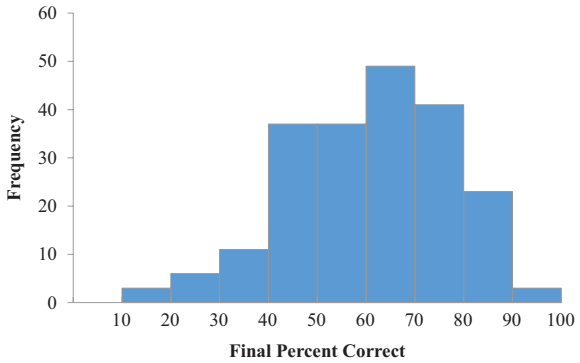


Figure 1 Score distribution for Pilot Form 1 (percentage correct).

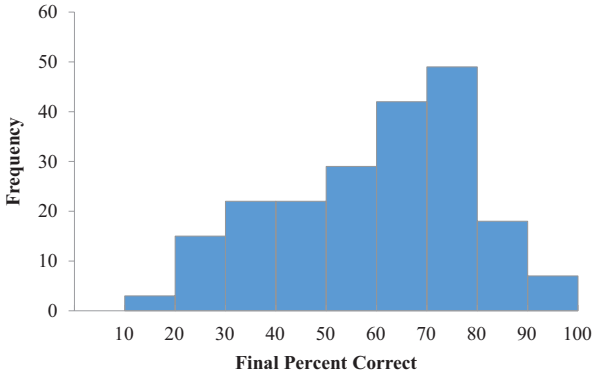


Figure 2 Score distribution for Pilot Form 2 (percentage correct).

and (b) the instructional tool categories highlighting the critical tasks of teaching science for entry-level elementary teachers. Findings indicate that participants' mean scores are fairly consistent across the four science content domains (Table 11), although this claim is better supported for Pilot Form 1, for which the average mean scores ranged from 60% to 63% in these four categories, in comparison to a 57%–71% span on Pilot Form 2. Second, in terms of the tasks of teaching science (Table 12), findings are similar in that participants' mean scores tended to be fairly consistent across categories, although this is more accurate for Pilot Form 1 than Pilot Form 2. However, in general, these findings should be taken with caution owing to the small sets of items within each of these categories.

Reliability of Classification

Table 13 displays the results of these analyses for African American examinees, White examinees, and all examinees on each form. It should be noted that results for African American examinees must be evaluated with caution, as sample sizes were quite small. Findings show that the rates of accurate and consistent classification all met the anticipated 80%

Table 9 Item Difficulty by Pilot Form

Difficulty level	Item percentage correct	Pilot Form 1,[a] % (*n*)	Pilot Form 2,[b] % (*n*)
Easier	20 – 39.9	8.7 (4)	15.6 (7)
	40 – 59.9	30.4 (14)	17.8 (8)
	60 – 79.9	50.0 (23)	57.8 (26)
Harder	80 – 94.9	10.9 (5)	8.9 (4)

Note. The *n* here refers to the number of content knowledge for teaching (CKT) science items in each form that were used to create the final participant scores. The total score points for each pilot form only include the items that had adequate item difficulty (between .20 and .95) and point-biserial correlations (>.20).
[a] $n = 46$. [b] $n = 45$.

Table 10 Item Discrimination by Pilot Form

Item discrimination	Point-biserial correlation	Pilot Form 1,[a] % (*n*)	Pilot Form 2,[b] % (*n*)
Better	.80 – .89	0.0 (0)	6.7 (3)
	.70 – .79	2.2 (1)	2.2 (1)
	.60 – .69	8.7 (4)	11.1 (5)
	.50 – .59	6.5 (3)	17.8 (8)
	.40 – .49	45.7 (21)	31.1 (14)
	.30 – .39	19.6 (9)	22.2 (10)
Poorer	.20 – .29	17.4 (8)	8.9 (4)

Note. The *n* here refers to the number of CKT science items in each form that were used to create the final participant scores. Point-biserial correlations were calculated between each CKT science item and the overall scale score for that pilot from. The total score points for each pilot form only include the items that had adequate item difficulty (between .20 and .95) and point-biserial correlations (>.20).
[a] $n = 46$. [b] $n = 45$.

threshold for being considered acceptable, although not all reached the 90% threshold for being considered desirable. In addition, these rates are similar between forms at these points and only differ at the 70% raw score point level (accuracy, Pilot Form 1 = 88.0%, Pilot Form 2 = 91.3%; consistency, Pilot Form 1 = 83.4%, Pilot Form 2 = 88.6%).

Teachers' Perceptions of the Importance and Relevance of the Content Knowledge for Teaching Science Pilot Items

As part of the overall assessment, teachers completed a survey on their perceptions of the importance and relevance of the CKT science items. As shown in Table 14, the perceptions survey consisted of eight Likert-scale items asking about their evaluation of the difficulty of these items (e.g., "I found the questions to be challenging") and how well they thought the items' content aligned with their learning opportunities in teacher education programs and professional development (e.g., "The questions covered material that is/was covered in my teacher preparation program"). Participants responded to each item on the perceptions survey on a 4-point Likert scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). Mean scores on these items ranged from 2.44 to 3.29.

Notably, the following results help support the face validity of the assessment, as results point to test takers believing the assessment covered appropriate material relevant to the work of teaching science, that it was challenging, and that people outside of their field would find the items more difficult. Participants tended to agree with the statements "I found the questions to be challenging" ($M = 3.12$, $SD = .65$) and "the questions covered material that I teach or expect to teach in the classroom" ($M = 3.12$, $SD = .73$). In the former case, while those taking Pilot Form 1 were slightly more likely to agree ($M = 3.21$, $SD = .59$) than those taking Pilot Form 2 ($M = 3.04$, $SD = .70$), the directionality of this finding ran opposite to the performance results on the individual forms, as shown in Table 8, where average performance was slightly higher on Pilot Form 1 ($M = .62$, $SD = .17$) compared to Pilot Form 2 ($M = .60$, $SD = .19$). One possible hypothesis for this finding may relate to prior preparation, as shown in Table 7, where those taking Pilot Form 1 were more likely to be enrolled in master's degree programs (24%) than those taking Pilot Form 2 (16%).

However, participants tended to slightly agree with the statement "The questions covered material that I have encountered in my professional development experiences" ($M = 2.72$, $SD = .85$). Finally, the statement that had the lowest mean

Table 11 Content Knowledge for Teaching Science Item Statistics by Science Content Domains

Content domain	Pilot Form 1[a]	Pilot Form 2[b]
Physical science		
Number of items	16	16
Mean proportion correct (*SD*)	.63 (.17)	.58 (.14)
Scale reliability	.68	.71
Life science		
Number of items	13	11
Mean proportion correct (*SD*)	.61 (.20)	.62 (.13)
Scale reliability	.59	.66
Earth and space science		
Number of items	11	12
Mean proportion correct (*SD*)	.60 (.10)	.57 (.20)
Scale reliability	.59	.67
Engineering, technology, and applications of science		
Number of items	6	6
Mean proportion correct (*SD*)	.63 (.12)	.71 (.06)
Scale reliability	.46	.64
Total		
Number of items	46	45
Mean proportion correct (*SD*)	.62 (.15)	.60 (.15)
Scale reliability	.85	.89

Note. The N here refers to the total number of participants who reached 100% completion on each pilot form.
[a] $n = 210$. [b] $n = 207$.

Table 12 Content Knowledge for Teaching Science Item Statistics by Tasks of Teaching Science Categories

Instructional tool	Pilot Form 1[a]	Pilot Form 2[b]
Scientific instructional goals, big ideas, and topics		
Number of items	8	5
Mean proportion correct (*SD*)	.62 (.12)	.65 (.07)
Scale reliability	.59	.45
Scientific investigations and demonstrations		
Number of items	11	12
Mean proportion correct (*SD*)	.66 (.15)	.66 (.10)
Scale reliability	.62	.69
Scientific resources		
Number of items	4	5
Mean proportion correct (*SD*)	.69 (.18)	.65 (.24)
Scale reliability	.23	.42
Student ideas		
Number of items	7	7
Mean proportion correct (*SD*)	.57 (.18)	.64 (.13)
Scale reliability	.50	.66
Scientific language, discourse, vocabulary, and definitions		
Number of items	2	3
Mean proportion correct (*SD*)	.72 (.04)	.40 (.13)
Scale reliability	.24	.23
Scientific explanations		
Number of items	5	8
Mean proportion correct (*SD*)	.59 (.17)	.58 (.16)
Scale reliability	.31	.56
Scientific models and representations		
Number of items	9	5
Mean proportion correct (*SD*)	.57 (.17)	.49 (.13)
Scale reliability	.46	.49
Total		
Number of items	46	45
Mean proportion correct (*SD*)	.62 (.15)	.60 (.15)
Scale reliability	.85	.89

Note. The *N* here refers to the total number of participants who reached 100% completion on each pilot form.
[a] $n = 210$. [b] $n = 207$.

Table 13 Reliability of Classification by Pilot Form

Potential qualifying scores	Subgroup	Sample size	Percentage classified correctly	Percentage classified consistently
Pilot Form 1				
50% ≥ 26	AA	15	88.3	84.5
	White	130	90.8	87.3
	Total	145	90.2	86.4
60% ≥ 32	AA	7	85.3	82.1
	White	94	87.2	82.4
	Total	101	87.0	82.0
70% ≥ 37	AA	3	92.6	89.5
	White	47	87.1	82.3
	Total	50	88.0	83.4
Pilot Form 2				
50% ≥ 26	AA	6	88.7	84.7
	White	120	91.1	87.5
	Total	126	90.5	86.6
60% ≥ 32	AA	2	92.2	89.6
	White	85	87.2	84.1
	Total	87	87.6	84.5
70% ≥ 37	AA	0	[a]	[a]
	White	37	90.4	87.3
	Total	37	91.3	88.6

Note. AA = African American.
[a]Could not be computed because there were no examinees in this subgroup who obtained a score at or above this qualifying score.

Table 14 Participants' Perceptions on the Content Knowledge for Teaching Science Items

			Pilot Form 1		Pilot Form 2	
Perception survey question	*N* (participants)	Full sample, mean (*SD*)	*N*	Mean (*SD*)	*N*	Mean (*SD*)
I found the questions to be challenging.	415	3.12 (.65)	209	3.21 (.59)	206	3.04 (.70)
Answering these questions made me think about some aspect(s) of teaching this content that I had not considered previously.	415	3.29 (.62)	209	3.28 (.61)	206	3.30 (.62)
I found it difficult to choose among answer options.	415	2.79 (.71)	209	2.81 (.72)	206	2.77 (.70)
The questions focused on material that is/was covered in my teacher preparation program.	393[a]	2.54 (.85)	201	2.50 (.86)	192	2.59 (.85)
The questions covered material that I teach or expect to teach in the classroom.	395[a]	3.12 (.73)	201	3.11 (.79)	194	3.14 (.66)
The questions covered material that I have encountered in my professional development experiences.	371[a]	2.72 (.85)	192	2.70 (.89)	179	2.75 (.81)
I think elementary school teachers should be able to answer most of these questions correctly.	415	2.96 (.70)	209	2.96 (.71)	206	2.96 (.68)
I think people in professions other than teaching should be able to answer most of these questions correctly.	415	2.44 (.72)	209	2.45 (.72)	206	2.42 (.73)

Note. Participants responded to each perception question based on a 4-point scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). The score points of each perception question range from 1 to 4.
[a]Only these three perception questions had "not applicable" as one of the potential answer choices. For these three questions, any "not applicable" responses were treated as missing, which explains the reduced number of participants reported for each of these three questions.

score was "I think people in professions other than teaching should be able to answer most of these questions correctly" ($M = 2.44$, $SD = .72$), suggesting that they were less likely to agree with this statement on average. Taken together, these results show that teachers found the assessment items to be challenging and material they expect to teach in the classroom. Interestingly, though, teachers felt that they had not consistently seen material like this in their professional development training. There were slight differences in responses to this question by teaching experience, $F(2, 361) = 2.99$, $p = .052$, where recent graduates who are not yet teaching agreed with this statement more on average ($M = 2.94$, $SD = .79$) than those with less than 3 years of teaching experience ($M = 2.60$, $SD = .86$). This result might point to the need for a broader range of material covered in professional development sessions. Finally, as this is an assessment of teachers' CKT science, it is encouraging that participants were less likely to think that people outside of their profession should be able to answer most of the questions correctly.

Relationship Between Teachers' Performance on the Content Knowledge for Teaching Science Pilot Forms and Their Background Characteristics and Professional and Academic Preparation

In our analysis, we examined the relationship between teachers' scores on the CKT science pilot forms and various indicators of their background and their professional and academic preparation. Because the results are based on raw scores from two different pilot forms, we conducted the following analyses by pilot form instead of pooling data across forms. In addition, it is important to note that there are sample size differences between many of the groups reported in this section, resulting in unbalanced groups that may exceed current statistical analysis guidelines and prevent strong conclusions being drawn from the results.

For each of the groups, we examined three assumptions before conducting analyses using either a *t*-test or ANOVA: independence of cases, homogeneity of variance, and normality. In examining the performance scores within each comparison group (e.g., gender, race/ethnicity), we found that the assumptions for independence and homogeneity of variance were met for all groups. However, the normality assumption was often violated according to the Shapiro – Wilk results from the normality tests, consistently so overall, and also among female examinees, White examinees, those in the highest undergraduate GPA category (3.5 – 4.0), those whose major/minor had a science focus, those with elementary certification in science, and those planning to enroll or currently enrolled in a teacher preparation program. However, a visual inspection of the histograms indicated that the distributions of teachers' scores were close to normal for both pilot forms. In addition, because both ANOVA and *t*-test are considered robust tests against the normality assumption, which means that they tolerate violations to the normality assumption well, we decided that these statistical tests were still appropriate to use for these analyses.

The results reported from these analyses are shown in Figure 3 (by gender), Figure 4 (by race/ethnicity), Figure 5 (by GPA and major/minor), and Figure 6 (by certification and professional preparation). Details about how we classified teachers into these groups are provided in the methods section earlier in this report. Each figure contains the mean and standard deviations for each group by pilot form.

Gender

Results of independent-samples *t*-tests found that there were significant differences in final proportion correct on the CKT science scores per form. For Pilot Form 1, the independent-samples *t*-test indicated that men scored, on average, higher ($M = .73$, $SD = .17$, $N = 19$) than women ($M = .61$, $SD = .16$, $N = 191$), $t(208) = 3.07$, $p = .002$. For Pilot Form 2, the result was similar: Men scored, on average, 14 percentage points higher ($M = .73$, $SD = .19$, $N = 18$) than women ($M = .59$, $SD = .19$, $N = 189$), $t(205) = 2.93$, $p = .004$.

Race/Ethnicity

A one-way ANOVA showed that differences on total proportion correct for CKT science scores on Pilot Form 1 were not statistically significant, $F(4, 204) = .69$, $p = .600$ (means and standard deviations in Figure 4) across race/ethnicity groups (e.g., African American, Asian, Hispanic, White, two or more races). In contrast, a one-way ANOVA for Pilot Form 2 showed that differences were statistically significant, $F(4, 200) = 2.55$, $p = .040$, across race/ethnicity groups. However, a Tukey's HSD (honest significant difference) post hoc test of the scores on Pilot Form 2 indicated that there were no

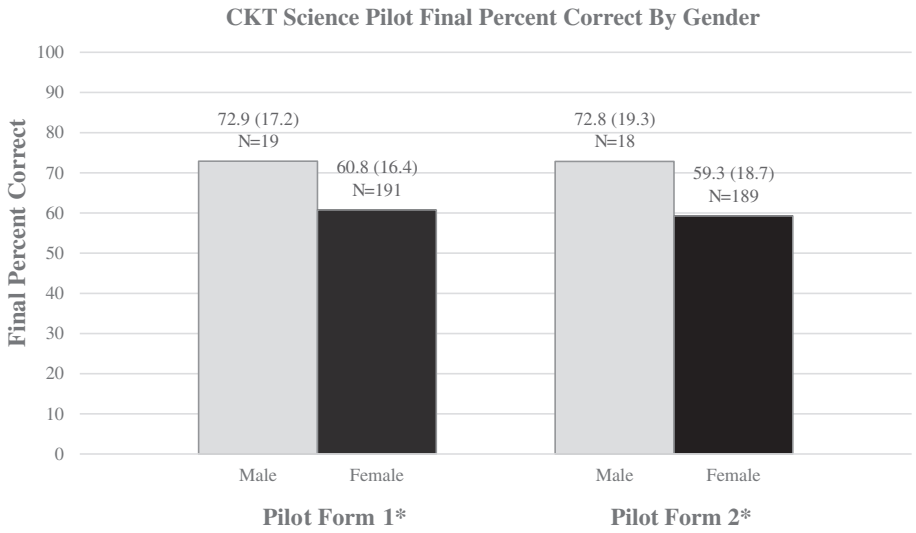


Figure 3 Content knowledge for teaching science mean scores by gender. The numbers above each bar represent the mean and standard deviation for that category. The *N* indicates the sample size of that category. **p* < .05.

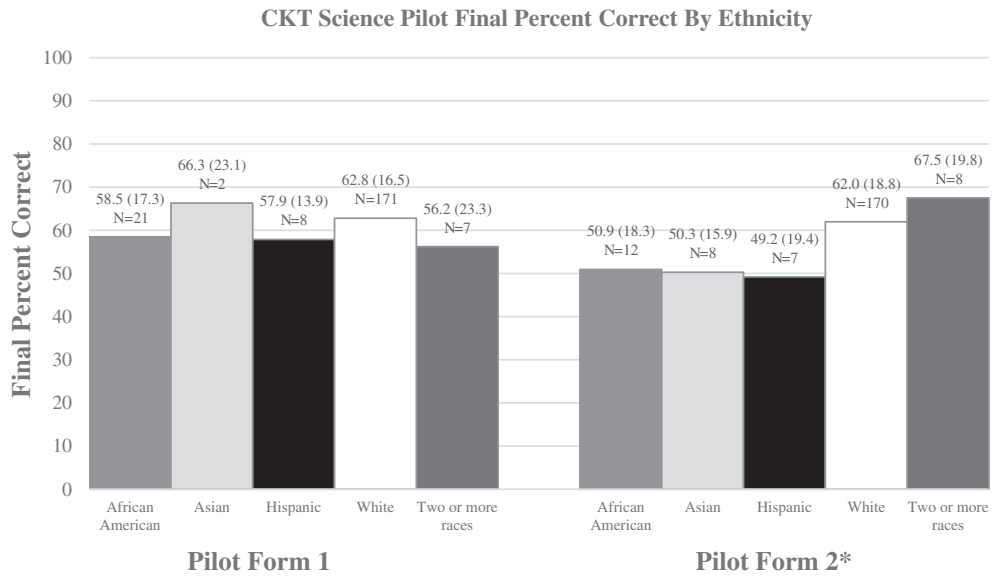


Figure 4 Content knowledge for teaching science mean scores by ethnicity. The numbers above each bar represent the mean and standard deviation for that category. The *N* indicates the sample size of that category. **p* < .05.

statistically significant differences between the mean scores of each racial/ethnic group. This result is mainly due to the pairwise test being more conservative. In addition, the fact that one of the key ANOVA assumptions—in this case, the normality assumption—was not met may have influenced these results.

Undergraduate Grade Point Average

For Pilot Form 1, an independent-samples *t*-test found that participants with higher GPAs had significantly higher total proportion correct scores ($M = .64$, $SD = .15$, $N = 130$) than participants with GPAs under 3.5 ($M = .58$, $SD = .18$, $N = 79$), $t(207) = 2.51$, $p = .013$. Pilot Form 2 had similar results, where participants with higher GPAs showed significantly higher total proportion correct scores ($M = .63$, $SD = .20$, $N = 125$) than participants with GPAs under 3.5 ($M = .57$, $SD = .18$, $N = 81$), $t(204) = 2.34$, $p = .020$.

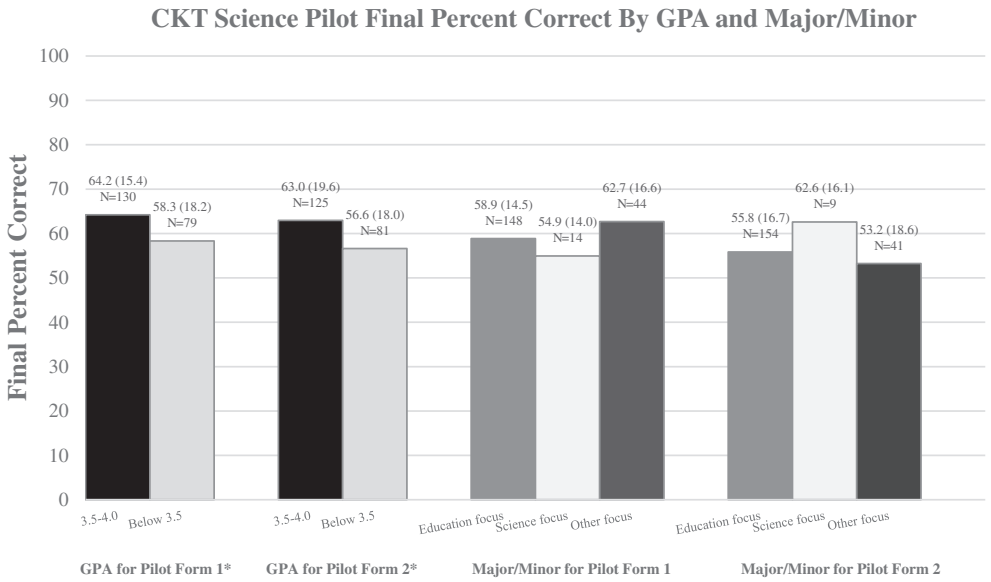


Figure 5 Content knowledge for teaching science mean scores by grade point average and major/minor. The numbers above each bar represent the mean and standard deviation for that category. The *N* indicates the sample size of that category. $^*p < .05$.

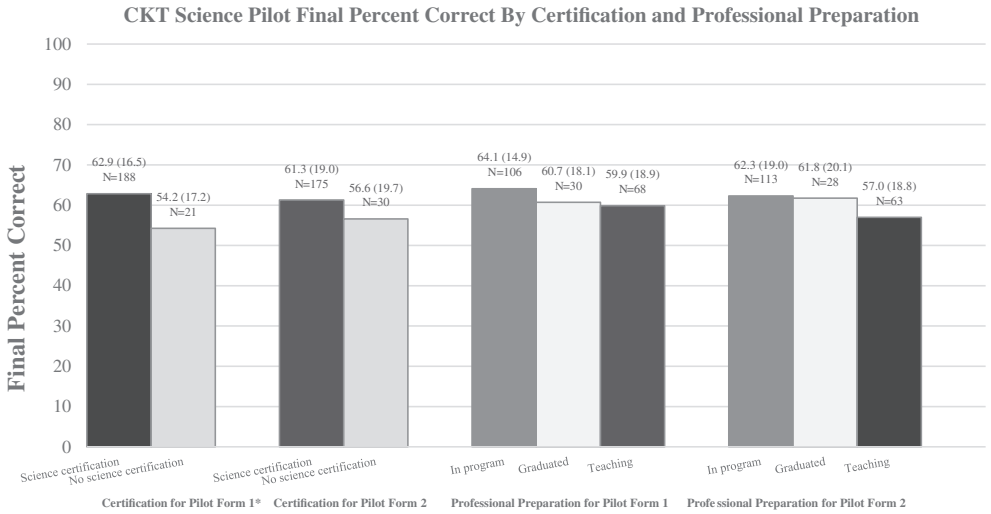


Figure 6 Content knowledge for teaching science mean scores by certification and professional preparation. The numbers above each bar represent the mean and standard deviation for that category. The *N* indicates the sample size of that category. $^*p < .05$.

Undergraduate Major and Minor

A one-way ANOVA showed that for Pilot Form 1, the relationship between undergraduate focus (i.e., education focus, science focus, or other focus) and total proportion correct scores was not significant, $F(2, 203) = 1.79$, $p = .170$ (means and standard deviations in Figure 4). In Pilot Form 2, a one-way ANOVA showed that the relationship between undergraduate focus and total proportion correct scores was not significant, $F(2, 201) = 1.17$, $p = .314$.

Teaching Certification

An independent-samples *t*-test for Pilot Form 1 indicated that on average, participants who had (or were working toward) an elementary certification with an emphasis in science ($M = .63$, $SD = .16$, $N = 188$) had significantly higher total proportion correct scores than participants who had (or were working toward) an elementary certification with no emphasis

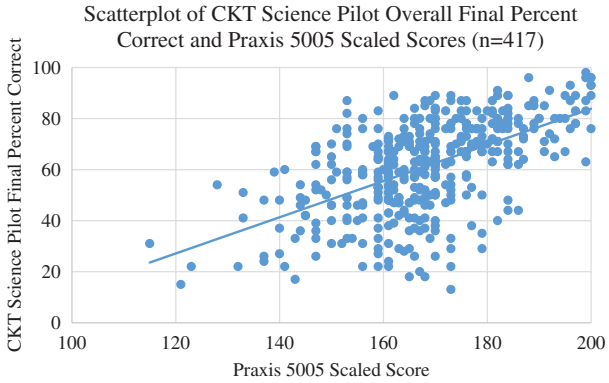


Figure 7 Comparing participant scores between content knowledge for teaching science pilot and PRAXIS 5005 assessment.

in science ($M = .54$, $SD = .17$, $N = 21$), $t(207) = 2.26$, $p = .025$. Pilot Form 2 did not have a significant difference between the two groups.

Professional Preparation

One-way ANOVAs showed that the relationship between professional preparation (i.e., planning to enroll or currently enrolled in a teacher education program, recently graduated but not teaching, and less than 3 years of teaching experience) and proportion correct scores was not significant, $F(2, 201) = 1.44$, $p = .240$ and $F(2, 201) = 1.64$, $p = .196$ (means and standard deviations in Figure 6), in Pilot Forms 1 and 2, respectively.

Relationship Between Performances on the Content Knowledge for Teaching Science Pilot Forms and the PRAXIS 5005 Science Exam

On average, findings reveal a moderate correlation, zero-order $r = .567$, disattenuated $r = .684$, $p < .001$,[16] between teachers' performance on these two science knowledge assessments. This correlation suggests that there is some consistency in terms of how well teachers performed on these two assessments, although the correlation is not high enough to suggest that these assessments are measuring the same construct. Figure 7 is a scatterplot illustrating the relationship between teachers' scores on both assessments across the full set of study participants. This scatterplot confirms the general expected trend in that a number of teachers score similarly on both assessments: on the lower end of both or on the higher end of both. However, there is also a set of teachers who appear to have higher scores on the PRAXIS 5005 subject matter knowledge assessment but do not perform as well on the CKT science pilot assessment. For example, the scatterplot shows that there is a subset of teachers who scored above 159 on the PRAXIS 5005 subject matter knowledge assessment, which is the cut-score for the majority of states using this assessment and translates to a percentage correct value well above 50%, but who failed to achieve 50% correct on the CKT science pilot assessment, reflecting the inherently increased difficulty of the CKT pilot assessment. The other important feature is that the scatterplot shows that no participants performed poorly on the PRAXIS 5005 subject matter knowledge assessment and also scored well on the CKT science pilot assessment. In general, the pattern noted in the scatterplot adheres to the expected pattern and hypothesis described earlier.

Discussion

The purpose of this study was multifaceted. In particular, we focused on examining the performance of assessment items designed to measure elementary teachers' CKT science for use in the Elementary Education: CKT Science component of the ETS Educator Series, including (a) how these new CKT science assessment items function, including their item difficulty and discrimination, form reliability, and classification of examinees; (b) how teachers perceive the importance and relevance of these new assessment items; (c) how teachers' performances on these items relate to their background characteristics and professional and academic preparation; and (d) how teachers' performances on the CKT science items compare to their performances on an examination of their science subject matter knowledge. This evidence was gathered

when piloting two CKT science assessment pilot forms (52 items per form) in an online, untimed testing environment with 417 preservice or novice elementary teachers.

Overall findings reveal that the majority of these CKT science assessment items show adequate item functioning in terms of item difficulty and item discrimination. In fact, less than 13% of the overall item pool was removed because of inadequate item functioning. Both pilot forms captured a wide distribution of scores across the scoring range, and the scale reliabilities on both forms fell within an acceptable range. In general, the patterns in teachers' performance and item functioning were fairly consistent across both pilot forms, although there was some variation in teachers' scores across the science content domains and the instructional tool categories. However, it is difficult to make any strong claims on this last point owing to the limited number of items in these categories within each pilot form. Overall, this analysis suggests that the pilot forms meet or are close to meeting the requirements for operational testing.

Findings also suggest that this sample of preservice and novice teachers recognized the value and importance of the CKT science being assessed with these new items, which provides important face validity evidence for these items. Participants tended to agree that the assessment items were tied closely to the knowledge that they use—or expect to use—as elementary science teachers. Despite this tendency to view the importance of the knowledge assessed in these items favorably, many teachers also noted that their teacher education programs and professional development experiences did not always address this material explicitly. These findings suggest that there may be a discrepancy between the CKT science that teachers are routinely called on to use in their work with students and the focus of their learning opportunities for preparing them to teach elementary science.

In terms of participant background and preparation, findings show three significant differences, although it is important to note that strong conclusions should not be drawn based on these analyses owing to the small sample size, especially within certain groups. On average, men scored significantly higher on the CKT science pilot forms than women. In addition, those with undergraduate GPAs of 3.5 or higher had significantly higher scores on the pilot forms compared to those with GPAs below 3.5. Finally, participants who were planning to or had already received their elementary certification with an emphasis in science had significantly higher scores on Pilot Form 1 than those participants whose certification emphasized a different subject area. Again, these latter relationships trend in anticipated directions (Gitomer et al., 2011; Steinberg et al., 2016), providing some initial external validity evidence for these new measures.

Lastly, findings showed moderate correlations between teachers' scores on the CKT science pilot forms and their scores on an assessment designed to measure their understanding of the science subject matter alone. This finding is promising in that the theory of CKT suggests that subject matter knowledge is necessary but not sufficient for measuring teachers' CKT. The CKT items are designed to assess the content knowledge that teachers apply in the work they do with curriculum, instruction, and students. Without an understanding of the subject matter, one would not expect teachers to perform well on the CKT science assessment. However, a strong conceptual understanding of the subject matter alone does not necessarily mean that one would perform at a high level, as these practice-based CKT items target the ways in which science teachers use their conceptual understanding to engage in the critical tasks of teaching science. This positive correlation is encouraging in that it suggests that both assessments are measuring science knowledge. Yet, the moderate nature of the correlation suggests some divergence, indicating they are likely measuring distinct aspects of science knowledge. Similar findings in terms of moderate correlations between measures of science teachers' subject matter knowledge and measures designed to assess these practice-based aspects of CKT more closely have been reported in the literature (Davidowitz & Potgieter, 2016; Grobschedl, Harms, Kleickmann, & Glowinski, 2015; Mikeska, Phelps, & Croft, 2016). Additional research, especially the use of cognitive interviews, would be useful for better understanding this relationship and the knowledge and reasoning that teachers use as they respond to these CKT science items.

Despite the mainly positive findings from this pilot study, there are a few limitations to keep in mind. First, the recruited sample for this study was not necessarily a representative sample of the population of teachers who typically take licensure assessments each year. For practical reasons, we focused our recruitment on teachers who had recently taken one particular elementary teaching licensure examination. As the results note, the recruited population was primarily located in the Northeast and South, which was largely an artifact of the states that have adopted PRAXIS 5005, and had a larger proportion of test takers who performed toward the higher end of the PRAXIS score range. To compensate, we did attempt to include all of the recruited participants at the lower end of the PRAXIS performance range, although not all of them ended up participating in this study.

Other limitations are related to the study's limited sample size and limited item pool, which precluded examining the structure, or subdomains, of CKT in science. Examining the structural validity of CKT science instruments requires both a large enough item pool to explore the extent to which the CKT items map onto the hypothesized component structure and a large enough sample to support factor analysis, neither of which was accomplished in the study reported here. Finally, it is unclear to what extent the testing environment — in this case, an untimed, online administration at a location of the participants' choice and without any stakes attached — may have impacted participants' motivation and involvement in this study. It is unclear how similar results would be for participants completing one of the pilot forms under conditions that more closely mimic the testing environment for operational licensure assessments.

Conclusion

This pilot study is one of the important steps in the development and validation process for the science component of ETS's PRAXIS Elementary Education: CKT Test licensure exam. This study gathered critical evidence for assessing how well these particular items functioned as part of test forms designed to assess novice elementary teachers' readiness to enter the profession. As noted earlier, while teachers cannot teach content they do not understand, research emphasizes that content knowledge alone is insufficient for effective teaching and that the specialized science knowledge teachers need differs from that of the scientist (Ball et al., 2008; National Academies of Sciences, Engineering, & Medicine, 2015; National Research Council, 2013; Shulman, 1986). Most importantly, research has suggested that CKT, especially the specialized and pedagogical content knowledge components, is a key mediator in science teachers' abilities to engage in critical teaching practices, such as interpreting students' ideas, constructing explanations, and selecting and modifying resources for instruction (Davis, Petish, & Smithey, 2006; Kloser, 2014; National Research Council, 2007; Windschitl et al., 2012). As such, developing licensure assessments that measure the CKT elementary science needed for beginning practitioners is critical to ensuring that only those who meet these standards enter the teaching profession.

Acknowledgments

A number of colleagues have been instrumental in supporting the development of the test content specifications and assessment items for the Elementary Education: CKT Science component of ETS's PRAXIS Elementary Education: CKT licensure exam. First, the authors extend their deep appreciation for the extensive efforts on this project from the team of science assessment development specialists at ETS, which was led by Israel Solon and included Declan Burke, Phil Falcone, Marshall Freedman, Stephen Horvath, Stephanie Lanzel, Jesse Miner, and Beth Nichols. In addition, the item development work would not have been possible without the team of outside item writers who contributed their extensive understanding of the CKT elementary science to design many of the instructional scenarios. The authors are also grateful to Joyce Marie Cromer, Kullen Day, Kathy Gill, Kristin Gunckel, Deborah Hanuscin, Stephanie Ashworth Kawamura, Maribel Magdaleno, Julianne Paul, Deborah Roberts-Harris, Kimberly Ann Robertson, Kathleen Roth, Deborah Smith, and Robert Timothy Smith for serving as members of the national advisory committee and providing critical feedback on the test blueprints. The authors also appreciate the support of a cadre of fellow ETS colleagues — Jason Bonthron, Joe Ciofalo, Andrew Croft, Nathan Lederer, Dawn Leusner, Kathy Miller, Pavan Pillarisetti, Sharon Temmer, and Georgiana Weingart — who committed their time and expertise to various components of this pilot study, such as identifying and recruiting study participants, programming the online pilot forms, and exporting and organizing the data files for analysis. Finally, the authors are grateful for the guidance and support from the National Observational Teaching Examination leadership team — in particular, Geoffrey Phelps, Karen Riedeburg, and Eric Steinhauer — in conceptualizing and conducting this pilot study.

Notes

1 To date, the ETS Educator Series has developed CKT assessments in English language arts and mathematics. The CKT Science component is slated to be operational in fall 2017, while the CKT social studies component is currently set for release in fall 2019.

2 For more information, please see https://www.ets.org/praxis/prepare/materials/5621 (Early Childhood), https://www.ets.org/praxis/prepare/materials/5622 (Grades K – 6), https://www.ets.org/praxis/prepare/materials/5623 (Grades 5 – 9), and https://www.ets.org/praxis/prepare/materials/5624 (Grades 7 – 12).

3 http://www.caepnet.org/standards/

4 http://www.teachingworks.org/work-of-teaching/high-leverage-practices
5 For more detailed information about the content test specifications for this assessment, refer to the Elementary Education: Multiple Subjects Study Companion located at https://www.ets.org/s/praxis/pdf/5001.pdf.
6 A current list of states that require this test can be found at https://www.ets.org/s/praxis/pdf/passing_scores.pdf. Since September 2015, two additional states have adopted this test.
7 These minority groups consisted of African American, Asian, Hispanic, and Native American test takers.
8 Although our goal was to have approximately 400 candidates (200 per pilot form) participate in this study, we aimed to successfully recruit approximately 800 participants, anticipating that a subset of them would likely end up withdrawing from the study.
9 A general list of questions can be found at https://www.ets.org/s/praxis/pdf/pdt_registration_form.pdf.
10 Asian included test takers describing themselves as Asian American/Asian, Southeast Asian American/Southeast Asian, or Pacific Island American/Pacific Islander. Hispanic included test takers describing themselves as Mexican, Mexican American, or Chicano; Puerto Rican; or Other Hispanic, Latino, or Latin American.
11 States were classified according to the four primary census regions: Northeast, Midwest, South, and West.
12 About 20% of the overall and recruited samples were repeat test takers. For purposes of describing the sample characteristics, highest scores for repeat test takers were used, as these are used to inform individual test-taker licensure decisions and therefore best reflect the characteristics of the CKT test takers with respect to PRAXIS performance.
13 Those not from a predominant minority group consisted of White, other, those of two or more races, and missing.
14 It was accepted that certain background characteristics, such as certification status, had changed from the time examinees completed the PRAXIS registration questionnaire.
15 Distractor analyses involved examining the percentage of test takers who selected each of the answer options and noting the average mean score of those examinees who selected each option. Typically, one would expect to see that examinees with an overall higher mean score on the assessment were also the ones who more often selected the intended answer and that examinees who selected each of the distractors had on average lower mean scores on the overall assessment.
16 This was corrected for attenuation based on using the average Cronbach's alpha reliability estimate of .79 from the four forms used during the 2015–16 testing year and then applying the standard adjustment formula (Spearman, 1904).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ball, D. L., Hill, H. C., & Bass, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? *American Educator, 29*(1), 14–22, 43–46.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*, 389–407. https://doi.org/10.1177/0022487108324554

Brennan, R. L. (Ed.). (2006). *Educational Measurement* (4th ed.). Westport, CT: ACE/Praeger.

Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 701–731). Westport, CT: Praeger.

Council of Chief State School Officers. (2011, April). *Interstate teacher assessment and support consortium (InTASC) model core teaching standards: A resource for state dialogue*. Washington, DC: Author.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart, and Winston.

Davidowitz, B., & Potgieter, M. (2016). Use of the Rasch measurement model to explore the relationship between content knowledge and topic-specific pedagogical content knowledge for organic chemistry. *International Journal of Science Education, 38*, 1483–1503.

Davis, E. A., Petish, D., & Smithey, J. (2006). Challenges new science teachers face. *Review of Educational Research, 76*, 607–651. https://doi.org/10.3102/00346543076004607

Gess-Newsome, J. (2015). A model of teacher professional knowledge and skill including PCK. In A. Berry, P. Friedrichsen, & J. Loughran (Eds.), *Re-examining pedagogical content knowledge in science education* (pp. 28–42). New York, NY: Routledge. https://doi.org/10.4324/9781315735665

Gitomer, D. H. (2007). *Teacher quality in a changing policy landscape: Improvements in the teacher pool*. Princeton, NJ: Educational Testing Service.

Gitomer, D. H., Brown, T. L., & Bonett, J. (2011). Useful signal or unnecessary obstacle? The role of basic skills tests in teacher preparation. *Journal of Teacher Education, 62*, 331–345.

Gitomer, D. H., Phelps, G., Weren, B. H., Howell, H., & Croft, A. J. (2014). Evidence on the validity of content knowledge for teaching assessments. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 493–528). San Francisco, CA: Jossey-Bass. https://doi.org/10.1002/9781119210856.ch15

Grobschedl, J., Harms, U., Kleickmann, T., & Glowinski, I. (2015). Preservice biology teachers' professional knowledge: Structure and learning opportunities. *Journal of Science Teacher Education, 26*, 291–318.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*, 372–400.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematical knowledge for teaching. *Elementary School Journal, 105*(1), 11–30. https://doi.org/10.1086/428763

Jin, H., Shin, H., Johnson, M. E., Kim, J., & Anderson, C. W. (2015). Developing learning progression-based teacher knowledge measures. *Journal of Research in Science Teaching, 52*, 1269–1295. https://doi.org/10.1002/tea.21243

Kloser, M. (2014). Identifying a core set of science teaching practices: A delphi expert panel approach. *Journal of Research in Science Teaching, 51*, 1185–1218. https://doi.org/10.1002/tea.21171

Lee, E., Brown, M. N., Luft, J. A., & Roehrig, G. H. (2007). Assessing beginning secondary science teachers' PCK: Pilot year results. *School Science and Mathematics, 107*(2), 52–60. https://doi.org/10.1111/j.1949-8594.2007.tb17768.x

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179–197.

Loughran, J., Berry, A., & Mulhall, P. (2006). *Understanding and developing science teachers' pedagogical content knowledge*. Rotterdam, Netherlands: Sense.

Loughran, J., Mulhall, P., & Berry, A. (2004). In search of pedagogical content knowledge in science: Developing ways of articulating and document professional practice. *Journal of Research in Science Teaching, 41*, 370–391. https://doi.org/10.1002/tea.20007

Mikeska, J. N., Phelps, G., & Croft, A. (2016, January). *Going beyond subject matter knowledge: Developing measures to assess content knowledge for teaching elementary science*. Paper presented at the annual meeting of the Association for Science Teacher Educators, Reno, NV.

Mikeska, J. N., Phelps, G., & Croft, A. (2017). *Developing and validating assessments of content knowledge for teaching elementary science* (Unpublished manuscript).

Minner, D., Martinez, A., & Freeman, B. (2012). *Compendium of research instruments for STEM education: Part I: Teacher practices, PCK, and content knowledge*. Washington, DC: Community for Advancing Discovery Research in Education.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report No. 03-16). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Erlbaum.

National Academies of Sciences, Engineering, and Medicine. (2015). *Science teachers learning: Enhancing opportunities, creating supportive contexts*. Washington, DC: National Academies Press.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academies Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K–8*. Washington, DC: National Academies Press.

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

National Research Council. (2013). *Monitoring progress toward successful K–12 STEM education: A nation advancing?:* Washington, DC: National Academies Press. https://doi.org/10.17226/13509

Nettles, M. T., Scatton, L. H., Steinberg, J. H., & Tyler, L. L. (2011). *Performance and passing rate differences of African American and White prospective teachers on Praxis examinations* (Research Report No. RR-11-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02244.x

Next Generation Science Standards Lead States. (2013). Next generation science standards: For states, by states. Washington, DC: National Academies Press.

Phelps, G., Bunde, H., Howell, H., & Steinberg, J. (2015, February). *Assessing content knowledge in teacher preparation*. Presentation at the 67th annual conference of the American Association of Colleges of Teacher Education, Atlanta, GA.

Raymond, M. R., & Luecht, R. M. (2013). Licensure and certification testing. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 391–414). Washington, DC: American Psychological Association.

Reese, C. M., & Tannenbaum, R. J. (1999). Gathering content-related validity evidence for the school leaders licensure assessment. *Journal of Personnel Evaluation in Education, 13*, 263–282.

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*, 1020–1049. https://doi.org/10.3102/0002831213477680

Schmitt, K. (1995). What is licensure? In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 3–32). Lincoln, NE: Buros Institute of Mental Measurements.

Selling, S. K., Garcia, N., & Ball, D. L. (2016). What does it take to develop assessments of mathematical knowledge for teaching? Unpacking the mathematical work of teaching. *Mathematics Enthusiast, 13*(1), 35–51.

Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist, 36,* 1138–1146.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14. https://doi.org/10.3102/0013189x015002004

Smith, P. S., & Taylor, M. J. (2010, March). *New tools for investigating the relationship between teacher content knowledge and student learning*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

Spearman, C. (1904). The proof of measurement and association between two things. *American Journal of Psychology, 15*, 72–101.

Steinberg, J., Brenneman, M., Castellano, K., Lin, P., & Miller, S. (2014). *A comparison of achievement gaps and test-taker characteristics on computer-delivered and paper-delivered Praxis I tests* (Research Report No. RR-14-35). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12033.

Steinberg, J., Ling, G., & Delaney, C. (2016, April). *Balancing quality and opportunity for elementary education licensure candidates within multiple frameworks*. Roundtable presentation at the annual meeting of the American Educational Research Association, Washington, DC.

Tannenbaum, R. J., Robustelli, S. L., & Baron, P. A. (2008). Evidence-centered design: A lens through which the process of job analysis may be focused to guide the development of knowledge-based test content specifications. *CLEAR Exam Review, 19*, 26–33.

Wang, N., Schnipke, D., & Witt, E. A. (2005). Use of knowledge, skill, and ability statements in developing licensure and certification examinations. *Educational Measurement: Issues and Practice, 24*(1), 15–22.

Wilson, S. M. (2013). Professional development for science teachers. *Science, 340*(6130), 310–313. https://doi.org/10.1126/science.1230725

Wilson, S. M. (2016). *Measuring the quantity and quality of the K–12 STEM teacher pipeline* (Education White Paper). Menlo Park, CA: SRI International.

Wilson, S. M., Floden, R., & Ferrini-Mundy, J. (2001). *Teacher preparation research: Current knowledge, gaps, and recommendations*. Seattle, WA: Center for the Study of Teaching and Policy, University of Washington.

Windschitl, M., Thompson, J., Braaten, M., & Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Science Education, 96*, 878–903. https://doi.org/10.1002/sce.21027

Appendix A

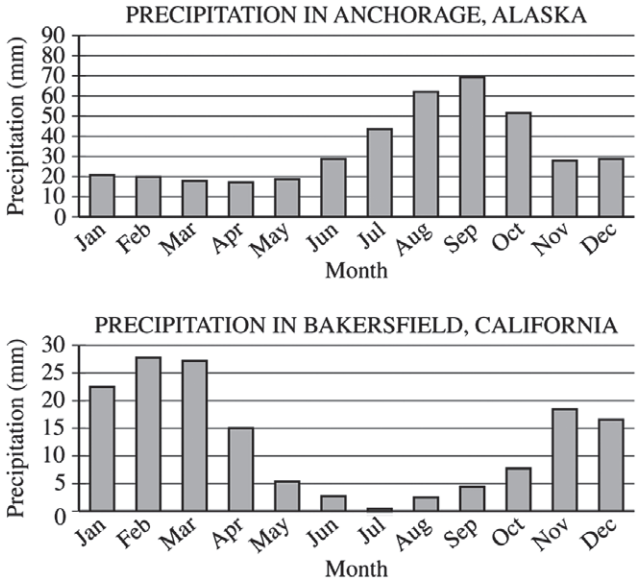
Roster for the Elementary Content Knowledge for Teaching Science National Advisory Committee

Participant name	University or school name	Location	Position
Joyce Marie Cromer	Council Traditional School	Mobile, AL	Science teacher, Grades K–5
Kullen Day	Eastern Pulaski Elementary School	Winamac, IN	Science department head
Kathy Gill	Willett Elementary School	Davis, CA	Elementary science specialist
Kristin Gunckel	University of Arizona	Tucson, AZ	Associate professor
Deborah Hanuscin	University of Missouri	Columbia, MO	Professor of science education and physics
Stephanie Ashworth Kawamura	Pine Lane Intermediate	Parker, CO	Discovery teacher, talented and gifted, Grades 3–6
Maribel Magdaleno	Estates Elementary	Naples, FL	Fifth-grade teacher
Julianne Paul	Salt Lake City School District	Salt Lake City, UT	Science coach
Deborah Roberts-Harris	University of New Mexico	Albuquerque, NM	Assistant professor
Kimberly Ann Robertson	Paul L. Dunbar Learning Center	Dallas, TX	Campus instructional coach in science, Grades PreK–5
Kathleen Roth	California State Polytechnic University	Pomona, CA	Senior research scientist
Deborah Smith	The Pennsylvania State University	University Park, PA	Assistant professor of education (retired)
Robert Timothy Smith	N/A	Bellefonte, PA	Consultant

Appendix B

Example Content Knowledge for Teaching Science Item Assessing Teachers' Practice-Based Content Knowledge in Earth and Space Science

Prior to a lesson on predicting weather outcomes, Ms. Monroe asked her students to look at the data presented in the two bar graphs showing average monthly precipitation in Anchorage, Alaska, and in Bakersfield, California.



Ms. Monroe would like to determine which students have not noticed the different scales on the two *y*-axes. Which question would best identify those students?

(A) "Which three months produce the least precipitation in each location?"
(B) "Which location has less precipitation during the summer months?"
(C) "Which location has the most precipitation during February and March?"
(D) "Which location has the most precipitation during November and December?"

		Content Topics											
		LS				PS				ESS			ETS
		1	2	3	4	1	2	3	4	1	2	3	1
Instructional Tools	I												
	II												
	III												
	IV												
	V												

Instructional Tool

VII. Scientific Models and Representations

Task of Teaching Science

d. Generating or selecting diagnostic questions to evaluate student understanding of specific models or representations

Content Topic

ES2.D: Weather and Climate

	VI											
	VII									X		

Performance Expectation

3-ES2-1. Represent data in tables and graphical displays to describe typical weather conditions expected during a particular season.

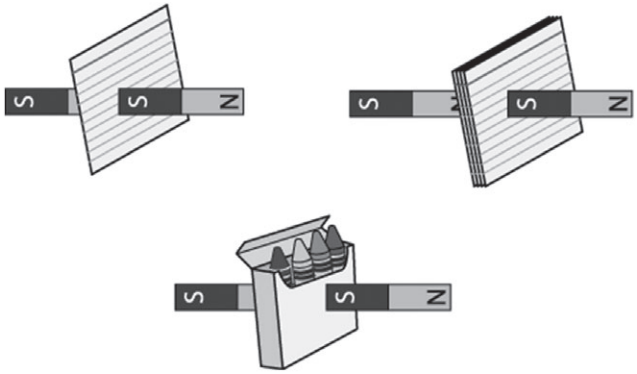
KEY: D

Rationale To answer this question, a teacher candidate needs to recognize the question that will elicit a different answer from a student who has misread the *y*-axis scales. Option A does not involve a comparison between graphs, while options B and C will produce the same response whether the student uses the *y*-axis scale or compares the height of the bars during those months. In option D, a student comparing the height of the bars would provide an incorrect response, indicating a need to point out the difference in the two scales.

Appendix C

Example Content Knowledge for Teaching Science Item Assessing Teachers' Practice-Based Content Knowledge in Physical Science

Four teams of students worked with two bar magnets each. Each team performed three trials testing one magnet's ability to attract the other through the following objects: a single index card, four index cards, and a box of four crayons as shown.



All teams observed that the magnetic attraction was weaker as the number of index cards was increased and even weaker when the box of crayons was used. The students then suggested additional experimental questions that they felt would help interpret their results.

Answering which three of the following student-generated questions would provide the most useful additional data?

(A) Would replacing the box of crayons with the same thickness of index cards make a difference?
(B) How are horseshoe magnets and bar magnets different?
(C) Did we follow the directions properly during the box of crayons trial, or did we do something wrong?
(D) How does the distance between the magnets by itself affect the attraction between the magnets?
(E) What would happen if we used two boxes of crayons instead of one?

Content Topics													
		LS				PS				ESS			ETS
		1	2	3	4	1	2	3	4	1	2	3	1
Instructional Tools	I												
	II						X						
	III												
	IV												
	V												
	VI												
	VII												

Instructional Tool

II. Scientific Investigations and Demonstrations

Task of Teaching Science

c. Determining the variables, techniques, or tools that are appropriate for students to address a specific investigation question

Content Topic

PS2. Motion and Stability: Forces and Interactions

Performance Expectation

3-PS2-3. Ask questions to determine cause and effect relationships of electric or magnetic interactions between two objects not in contact with each other.

Key: A, D, E

Rationale: To answer this question, a teacher candidate needs to identify questions that would provide data that would clarify the observations already made. The most important issue to resolve is whether using a box of crayons rather than index cards added an additional uncontrolled variable. Using two boxes of crayons instead of one would provide additional data on the effect of increasing the amount of material, and thus distance, between the two magnets. Knowing the difference between bar magnets and horseshoe magnets would not explain the student observations. As all groups reported the same observations, the prospect that the same random error, as mentioned in option C, occurred in all four groups can be ruled out.

Suggested citation:

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). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12207

Action Editor: Elizabeth Stone

Reviewers: Malcolm Bauer and Kevin Larkin