Developing and Using a Scalable Assessment to Measure Preservice Elementary Teachers' Content Knowledge for Teaching about Matter

Katherine E. Castellano and Jamie N. Mikeska

ETS

Acknowledgements

This study was supported by a grant from the National Science Foundation (Award No. 1813254). The opinions expressed herein are those of the authors and not the funding agency. We are grateful to the preservice elementary teachers who participated in this study to field test this newly developed instrument to measure content knowledge for teaching (CKT) about matter and its interactions. We are also grateful to three colleagues (Dante Cisterna, Josie Melton, and Liz Orlandi) who each lead a team of outside item writers to develop and refine the CKT items used in this study and to each of the item writers for their insights and dedication throughout the CKT item development process. Finally, our research team also included research support staff (Barbara Weren, Andrew Finnegan, and Ruopei Sun) who were instrumental in organizing the logistical aspects of the item development work, recruiting study participants, monitoring data collection, and supporting item analysis. Correspondence concerning this article should be addressed to Katherine Castellano, Educational Testing Service, 660 Rosedale Rd., Princeton, NJ 08541. E-mail: kecastellano@ets.org.

Abstract

There is strong agreement in science teacher education of the importance of teachers' content knowledge for teaching (CKT), which includes their subject matter knowledge and their pedagogical content knowledge. However, there are limited instruments that can be used efficiently and effectively on a large scale to assess and study elementary science teachers' CKT. Such measures would support strategic monitoring of large groups of science teachers' CKT and the investigation of comparative questions about science teachers' CKT longitudinally across the professional continuum or across teacher education or professional development sites. To address this gap, this study focused on designing an automatically scorable summative assessment that can be used to measure preservice elementary teachers' (PSETs') CKT in one high-leverage science content area: matter and its interactions. We conducted a field test of this CKT instrument with 822 PSETs from across the United States and used the response data to examine how this instrument functions as a potential tool for measuring PSETs' CKT in this science content area. Results suggest this instrument is reliable and can be used on large scale to support valid inferences about PSETs' CKT in this content area. In addition, the dimensionality analysis showed that all items measure a single construct of CKT about matter and its interactions, as participants did not show any differential performance by content topic or work of teaching science instructional tool categories. Implications for progressing the field's understanding of the nature of CKT and approaches to developing summative instruments to assess science teachers' CKT are discussed.

Keywords: content knowledge for teaching, matter and its interactions, preservice teachers, elementary

Developing and Using a Scalable Assessment to Measure Preservice Elementary Teachers' Content Knowledge for Teaching about Matter

Research has suggested that effective teachers not only have a deep knowledge base focused on understanding the content they need to teach and how students learn, but they also need to know how to use that knowledge to provide high-quality instruction to K-12 students within the content areas (Abell, 2013; Ball et al., 2008; National Academies, 2015; National Research Council, 2013; Shulman, 1986). For example, science teachers need to understand alternative conceptions and specific assets that students bring with them and use their content knowledge to select scientific models, representations, and investigations they can use to address these alternative conceptions and leverage student assets productively. Many scholars refer to this collective knowledge as content knowledge for teaching (CKT), as it is the knowledge that teachers are required to draw upon as they engage in the work of teaching within specific content areas (Hill et al., 2004, 2007, 2008; Kersting, 2008; Krauss et al., 2008; Phelps et al., 2014).

Despite strong agreement in science teacher education on the importance of teachers' CKT, there are limited instruments that can be used efficiently (e.g., manageably timed instrument with automatic scoring) and effectively (e.g., inferences about science CKT supported by strong validity evidence) on a large scale to assess elementary science teachers' CKT – which includes their subject matter knowledge and their pedagogical content knowledge, such as their knowledge of students and instructional strategies for teaching specific science topics (Minner et al., 2012; National Research Council, 2013; Wilson, 2016). A report titled *Monitoring Progress Toward Successful K-12 STEM Education* published by the National Research Council (2013) argued for the importance of direct measures that can be used to assess K-12 teachers' CKT. Such measures would support analysis of how well the field is progressing towards attaining the

goal of successful K-12 STEM education. These measures also would support decision making by school districts and professional development programs about how best to target the limited time they have to provide professional learning opportunities to teachers. These measures could provide information on teachers' CKT in particular STEM areas, which would support the development of more robust teacher learning experiences.

The challenge of having access to efficient and effective instruments for assessing teachers' CKT is especially acute in elementary science teacher education, an area that increasingly struggles with recruiting and maintaining effective science teachers. To date, current measures that could be used on a large scale across teacher preparation programs mainly assess science teachers' subject matter knowledge (Sadler et al., 2013; Smith, 2010; Trystad et al., 2014). The few that measure the practice-based components of science teachers' CKT, specifically their pedagogical content knowledge, typically require more extensive time and resources to administer and score, such as open-ended survey responses, interviews, observations, and use of graphic organizers (Henze & van Driel, 2015; Loughran et al., 2004; Park & Oliver, 2008; Park & Suh, 2015; Roth et al., 2011). This makes it challenging to efficiently investigate comparative questions about science teachers' CKT longitudinally across the professional continuum or across teacher education or professional development sites.

To address this gap, in our research project we focused on designing an automatically scorable assessment that can be used to measure preservice elementary teachers' (PSETs') CKT in one high-leverage, foundational science content area: matter and its interactions. If successful, our design and development process could serve as a model for future work in this area, particularly in terms of expanding the development of CKT instruments to other science areas and grade levels. In two earlier studies (Author, 2022a; 2022b), we examined validity evidence

based on content and response processes of the individual CKT matter items to ensure that they were adequately measuring the construct of interest – PSETs' CKT about matter – and examined PSETs' perceptions of the importance of the CKT being measured in each item and the items' connection to the work of teaching elementary science. Both studies showed promising findings indicating that the CKT matter items engaged the PSETs in leveraging their CKT about matter to answer each item and that the PSETs perceived these items to have a strong connection to the work of teaching elementary science. In this study, we aimed to determine if we could adequately develop an instrument that measures elementary PSETs' CKT about matter and could be administered and scored efficiently and effectively. Study questions attend to four key strands of validity evidence—validity evidence based on test content, internal structure, relations to other variables, and face validity—(see, e.g., American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014) to evaluate the adequacy of the CKT about matter instrument.

First, as part of validity evidence based on test content, we examine how well the items function, or behave, and focus on classical item statistics such as proportion correct, item-total correlations, and distractor analyses. The first research question is: To what extent do the CKT about matter items on this instrument exhibit adequate item functioning? This first part of the analysis seeks to ensure that the CKT matter items do not have any issues in terms of how they function, which would compromise the validity of inferences based on the full instrument results.

Second, we examine validity evidence based on the internal structure of the CKT about matter instrument by using multidimensional Item Response Theory (MIRT) models to answer the study's second research question: What is the nature, or structure, of the CKT used by PSETs in one science content area (matter and its interactions)? This part of the analysis will allow us to

confirm or refute the hypothesized dimensionality of this construct, as described in the theoretical framework and the methods sections. This dimensionality analysis will also allow us to determine the potential for reporting scores to support users of the CKT about matter assessment results (in this case, teacher educators) to make sense of this information.

Third, we examine the instrument's external validity or evidence based on relations to other variables. We compare performance on the CKT instrument by personal background variables, educational obtainment variables, and teacher preparation program variables to determine if certain groups of PSETs that have more relevant background and preparation experiences to the construct of interest (CKT about matter) perform better than those who do not. We also correlate our CKT matter scores with two external measures of science content knowledge—one that measures PSETs' science content knowledge across all topics and another that assesses the PSETs' subject matter knowledge about properties and changes in matter. We hypothesize that both should be related to PSETs' responses on the CKT about matter instrument and show moderate correlations, although we anticipate that the measure that is more aligned with this content area will be more highly correlated. The study's third research question is: How does PSETs' performance on the CKT matter instrument (a) compare across relevant background and preparation variables and (b) relate to their performance on other science knowledge measures?

Finally, we examine PSETs' perceptions of the face validity of the items on the CKT about matter instrument by asking: How do the PSETs perceive the importance, clarity, rigor, and relevance of the instrument for assessing their CKT about matter and its interactions? While face validity alone is not a useful indicator to determine the adequacy of a new assessment, it is an important part of ensuring that the users of specific assessments—in our case, the PSETs

themselves – view the items positively in terms of their relevance to the construct being measured and their importance to the work of elementary science teaching.

By collectively addressing these research questions, we accumulate validity evidence supporting the use of the test scores as a measure of PSETs' CKT in this content area. In addition, findings add to the science education literature about the nature of elementary science teachers' CKT and raise questions about the structure of this knowledge base. Implications for progressing the field's understanding of the nature of CKT and approaches to developing summative instruments to assess science teachers' CKT are discussed.

Theoretical Framework: CKT and the Work of Teaching Science Framework

In their seminal piece, Ball et al. (2008) define CKT as including both subject matter knowledge (SMK) and pedagogical content knowledge (PCK). SMK is typically defined as teachers' understanding of the key concepts, principles, and ideas that comprise a specific discipline, as well as the relationships between them (Ball & McDiarmid, 1989; Kennedy, 1998). For example, science teachers' SMK includes their understanding of specific science content, such as understanding conservation of matter, the particle model of matter, and changes in matter. PCK is conceptualized as the knowledge the teachers use to transform the subject matter to support student learning, which includes their understanding of common student challenges and ways to represent conceptual ideas (Shulman, 2015). Shulman's initial conceptualization of PCK as "the blending of pedagogy and content" (1987, p. 8) resulted in differing conceptualizations of this knowledge domain and sparked much scholarship, debate, and consternation in the field of science teacher education that continues to this day (Carlson & Daehler, 2019; Hashweh, 2005; Magnusson et al., 1990; Park & Oliver, 2008). Both aspects of

science teachers' CKT have been a target of science teacher preparation and professional development for many decades (Davis et al., 2006; Van Dijk et al., 2007).

Research has suggested that CKT is a key mediator in teachers' abilities to engage successfully in critical teaching practices such as interpreting students' ideas, constructing explanations, and selecting and modifying resources for instruction (Baumert et al., 2010; Hill et al., 2007; 2008; Schneider & Plasman, 2011). This is especially the case for elementary science teachers, who tend to struggle with understanding science content and knowing how to elicit and interpret students' ideas and how to use topic-specific instructional strategies and content representations to support student learning effectively (Davis et al., 2006; Kloser, 2014; National Research Council, 2007; Windschitl et al., 2012). In our study, we use a framework called the Work of Teaching Science (WOTS) framework (Mikeska et al., 2018), which was developed by a team of elementary science teachers, science teacher educators, science content experts, researchers, and assessment developers, to characterize the CKT used by individual teachers when teaching a specific science content area or topic. Instead of focusing solely on teachers' PCK or conceptualizing the knowledge base in terms of discrete knowledge components, this framework explicitly highlights the full breadth of science teachers' CKT, which includes both their SMK and their PCK, and starts by identifying the specific instructional practices science teachers engage in. By doing so, this framework targets the "knowledge in action" that is leveraged by elementary science teachers in their daily work. As such, the WOTS framework is a useful tool to guide the development of science teaching scenarios where teachers are required to leverage both their SMK and PCK to engage in a science teaching practice.

As shown in Table 1, this framework identifies specific content challenges that novice elementary science teachers face in the work of teaching and is organized by the instructional

tools that elementary science teachers use and interact with (e.g., scientific models and explanations). For each of the instructional tools identified, the authors identified a set of key instructional practices fundamental to the work of beginning elementary science teachers. Engaging in each instructional practice requires that elementary science teachers draw upon both their SMK and their PCK. A similar approach to conceptualizing science teachers' knowledge has been used in other studies such as when studying the development and use of a CKT assessment to measure secondary physics teachers' CKT about energy (Etkina et al., 2018; Phelps et al., 2020) and when examining the performance of a set of assessment items to measure PSETs' CKT across three different science content areas (Mikeska et al., 2018).

[Insert Table 1 about here.]

Background: Tools for Measuring and Analyzing Science Teachers' Knowledge

In science education, teachers' knowledge has been measured using a wide array of approaches (Chan & Hume, 2019). These varied approaches use different kinds of instruments and have resulted in a compilation of both qualitative and quantitative data. To date, most empirical research in science education has focused on examining individual science teachers' knowledge, although a few more recent studies explored science teachers' collective knowledge (Akerson et al., 2017; Nilsson, 2014). Traditionally, instruments to measure science teachers' knowledge tend to address either their SMK or their PCK, with only a few studies targeting the development and use of instruments that measure both in an integrated fashion (Etkina et al., 2018; Mikeska et al., 2018; Phelps et al., 2020). In addition, most instruments require the use of approaches that are difficult to scale.

In terms of measuring science teachers' SMK, historically studies used more easily accessible proxies, such as the number or type of content courses they took in college, their grade

point average, or their college degree (Monk & King, 1994; Wilson et al., 2001). Yet, since some research has suggested these kinds of proxies are not strongly related to student learning outcomes or engagement, more recent studies have turned to the use of more direct measures of science teachers' SMK. In some cases, this has involved self-report measures where teachers rate themselves on how knowledgeable they are about certain science topics or concepts or how well prepared they are to teach certain science topics or concepts (Banilower et al., 2007; Diamond et al., 2013; Supovitz & Turner, 2000). In other cases, studies have used instruments where teachers respond to science assessment items designed to assess their understanding of specific scientific concepts and the relationship between concepts (Menon & Sadler, 2016; Rice, 2005; Shugart & Hounshell, 1995). Sometimes these instruments are either the same ones used to assess K-12 students' understanding of the science content or adaptations of student science assessments for use with science teachers (Sadler, Coyle, et al., 2013; Sadler, Sonnert, et al., 2013). In other cases, researchers have developed their own instruments to assess teachers' understanding of scientific concepts they need to teach K-12 students (Jüttner, 2013; McConnell et al., 2017; Nixon et al., 2019). Finally, other studies engage teachers in in-depth interviews to elicit their SMK within particular science topics and conduct qualitative analysis of the interview responses to assess the quality of science teachers' SMK (Nixon et al., 2016), sometimes using specific approaches like concept profiles to probe their SMK during the interviews (Arzi & White, 2008) or tools like concept maps to examine science teachers' SMK (Nixon et al., 2017).

In terms of science teachers' PCK, a recent literature review in science education (Chan & Hume, 2019) resulted in the analysis of 99 peer-reviewed empirical research studies published from 2008 to 2018 that investigated science teachers' PCK and identified two main approaches that researchers used to determine and study science teachers' PCK. The first approach uses

teachers' self-reports to answer questions on written tests, open-ended surveys or interviews, or questionnaires where they recall their instructional practice or respond to prompts about others' practice (Großschedl et al., 2015; Kirschner et al., 2016; Luft & Chang, 2014; McNeill et al., 2016; Park et al., 2020; Sorge et al., 2019). Like measures of science teachers' SMK, some researchers used PCK tests that were scored using researcher-established correct answers or scoring rubrics. For example, Kirschner et al. (2016) developed a PCK assessment of 17 items, most of which were open-ended items, to measure secondary physics teachers' PCK and coded the responses as either incorrect, partially correct, or correct. They used the assessment to compare different groups of physics teachers' PCK (e.g., in-service physics teachers vs. preservice physics teachers).

Other researchers use open-ended surveys or interviews where teachers are asked to respond to a set of questions about their PCK either in writing or verbally. Then, the researchers use qualitative analysis to categorize the responses and examine the nature and development of science teachers' PCK. For example, Bergqvist et al. (2016) analyzed ten secondary teachers' lesson plans and conducted interviews with the teachers about their teaching to examine their PCK about students' understanding, representations, and instructional strategies when teaching about chemical bonding. Similarly, Lee et al. (2007) used a rubric to analyze beginning secondary science teachers' interview responses and classroom observations for evidence of two PCK aspects – their understanding of student learning and of instructional strategies – and understand if and how their PCK develops.

In addition to these analysis approaches, researchers have used two frameworks – Content Representations (CoRes) and Pedagogical and Professional-experience Repertoires (PaP-eRs) – across studies to investigate both preservice and in-service teachers' PCK across

various science topics (Bertram & Loughran, 2012; Carpendale & Hume, 2021; Hume & Berry, 2011; Loughran et al., 2006; Loughran et al., 2008; Nilsson & Elm, 2017; Nilsson & Karlsson, 2019) and have used PCK mapping to examine the relationships between various PCK components (e.g., knowledge of student ideas; knowledge of instructional strategies; etc.) (Chan, 2022). However, many of these approaches have not been used in studies with large sample sizes or sophisticated designs due to the resource-intensive process needed for coding the survey or interview data or constructing PCK maps.

The second approach for examining science teachers' PCK focuses on examining teachers' actual instructional performance, either using simulated performance tasks or observing them (in person or via video records) while teaching science to their students in their actual classrooms (Chan & Hume, 2019). These observations may be complemented with the use of teaching artifacts, such as copies of teachers' lesson plans or written assignments, or teachers' verbal reflections or justifications about their instructional decision making. Such data sources have been used to investigate science teachers' PCK during their lesson preparation, their actual lesson enactment, or during their reflection post-lesson. For example, Marshall et al. (2016) developed an observation protocol to measure science teachers' use of research-supported instructional practices in their teaching. In another study, Roth et al. (2011) engaged in-service elementary science teachers in a professional development program where they had opportunities to analyze videos of science instruction through two lenses – attending to student thinking and the science content storyline. Their findings indicated that teachers in the intervention group not only increased their understanding of the science content but also improved in their ability to analyze science instruction for these specific PCK aspects.

Across the 99 PCK studies that Chan and Hume (2019) reviewed, researchers primarily focused on examining the nature of science teachers' PCK, with a smaller set of studies examining the development of their PCK, the relationship between PCK and other variables, and changes to science teachers' PCK due to the use of an intervention. Most notable is that across these 99 empirical studies, only six of them addressed the development of a PCK measurement instrument and only half of those six did so in the context of preservice teacher education.

In the last few years, a handful of research teams have made progress in conceptualizing and developing CKT science assessments that target the measurement of both aspects of CKT – teachers' SMK and PCK – in an integrated fashion. This work is grounded in and builds off the previous measurement work done in other content areas, most frequently in mathematics education, where researchers aimed to assess mathematics teachers' CKT using assessment items with embedded teaching scenarios (Gitomer et al., 2015; Hill et al., 2005; Phelps & Howell, 2016). In science, two primary efforts have been undertaken to date. First, Mikeska and colleagues (Mikeska et al., 2017) developed, piloted, and collected initial validity evidence on a set of 39 CKT assessment items across three science content areas (structure and properties of matter, ecosystems, and Earth's place in the universe). They scored upper elementary teachers' responses on selected-response and constructed-response CKT items and investigated the item properties and item set scores across the full sample. Findings indicated that the CKT items captured variation in teachers' responses at different item difficulty levels and that most items showed a moderate to strong relationship to the overall scale score, suggesting that this CKT item development approach has much promise. In a follow up study, Mikeska et al. (2018) piloted two separate CKT science assessment forms (52 CKT items per form) with preservice and novice elementary teachers. These items were intended for potential use as part of a high-

stakes elementary teacher licensure assessment. This study examined how the items functioned, teachers' perceptions of the importance and relevance of the items, and how teachers' performance relates to their background and another assessment measure. Findings showed that most items exhibited adequate item functioning and that teachers perceived the items as relevant to the work of teaching elementary science.

Second, more recently, Phelps et al., 2020 used an evidence-centered design approach to develop a CKT assessment using both selected response and constructed response items for measuring secondary physics teachers' CKT about teaching energy. They note that their "approach seeks to integrate rather than isolate the types of content knowledge used in teaching" (p. 107). In this study, evidence gathered from exploratory factor analysis suggests that the CKT about the energy construct is unidimensional and that the CKT energy assessment they designed is reliable with strong measurement qualities.

Based on the prior research that has been conducted to develop, use, and refine tools and instruments for assessing science teachers' CKT, there are a few notable takeaways relevant to this study. First, while most researchers recognize that teachers' CKT is subject specific (e.g., science vs. mathematics), there are differing viewpoints regarding the grain size of science teachers' knowledge base in terms of the extent to which this knowledge base is specific to particular domains (e.g., matter and its interactions vs. force and motion) or particular content areas or topics (e.g., properties of matter vs. conservation of matter). Second, since most researchers conceptualize science teachers' PCK as a set of distinct, yet interrelated, components, they tend to develop and use tools and approaches that target individual PCK aspects, such as teachers' knowledge of students' understanding and their knowledge of instructional strategies for teaching specific science topics to K-12 students. To date, it has been

rare for researchers to develop instruments that measure these CKT in an integrated fashion, despite compelling research that suggests science teachers tend to draw upon multiple forms of knowledge simultaneously as they engage in critical teaching practices (Barnett & Hodson, 2001; Fischer et al., 2012; Gess-Newsome et al., 2018). Finally, most of the tools and approaches developed and used to date, especially for examining science teachers' PCK, require extensive amounts of time and resources for both data collection and analysis (e.g., human scoring of open-ended items). This resource drain means that it is difficult to imagine using these tools and approaches on a large scale to monitor and track changes to science teachers' CKT across time or across sites. It also suggests that the development of instruments that could be used to measure science teachers' knowledge efficiently on a large scale would be an asset to the field of science education.

Our research study was designed to directly address these gaps and challenges. In this study, we developed an automatically scorable instrument for large scale use in assessing PSETs' CKT about matter and its interactions. Unlike previous instruments that were more time-consuming to administer and score, an instrument for large-scale use allows teacher preparation programs to implement it easily and receive actionable results in real time. The instrument focuses on a single content domain—matter and its interactions—that is a high-leverage, foundational one in science education central to understanding many other scientific ideas but complex to teach and difficult to learn (National Research Council, 2013; Talanquer, 2009; Tsarpalis & Sevian, 2013). Due to shifts in how concepts of matter are introduced in the elementary years as shown in the Next Generation Science Standards, there is also a lack of content-specific teaching knowledge relevant to teaching about matter in the elementary years (Smith & Plumley, 2016).

One unique aspect of the CKT matter instrument is that each item is designed to measure both aspects of PSETs' CKT in this area, as both SMK and PCK together form the usable knowledge that science teachers must leverage when engaging in specific science teaching practices. We first describe our development approach for designing CKT matter assessment items and assembling a CKT matter assessment instrument. We then share our analysis approach and findings related to the instrument's validity and reliability, including dimensionality analysis targeting questions about the nature, or structure, of science teachers' CKT, analysis about how their CKT relates to different PSET background and preparation variables and other knowledge measures, and analysis on the PSETs' perceptions of this newly developed CKT instrument.

Methods and Data Sources

Sample

Our research team recruited PSETs nationwide who took ETS's *Praxis*® 5005: Elementary Education: Science Subtest licensure assessment between January 2018 and June 2019 to participate in this study. To meet our goal of complete data for at least 800 PSETs[1],while allowing for expected sample attrition, we oversampled by 20 percent, resulting in an initial recruitment of 960 PSETs. The recruited PSETs were selected through a stratified random sample with four stratifying variables: gender (Male, Female), geographical location (Midwest, Northeast, South, West), race/ethnicity (White vs Not-White), and Praxis elementary science test quartiles (Q1-Q4). These variables were used to create 64 (2*4*2*4) distinct cells (e.g., Male-Midwest-White-Q1) from which we sampled the same proportion of preservice teachers as in our Praxis Science test-taker population. Participants who completed all study components, including

[1] We chose this sample size to support estimation of our IRT models (e.g., Yen & Fitzpatrick, 2006).

the CKT about matter field test, test perceptions survey, background questionnaire, self-efficacy survey, and the AIM Horizon test on matter, were compensated $150 for their time.

Of the 960 recruited participants, 822 completed the CKT matter field test, which met our target sample size of at least 800. The 822 PSETs who completed the field test were representative of the *Praxis*® science test population, as seen in Table 2. From the background questionnaire, we also found that the majority of participating PSETs were enrolled in undergraduate teacher education programs (64%) and included elementary education as one of their undergraduate majors (65%).[2]

[Insert Table 2 about here]

CKT Matter Instrument Development

For this study, we developed and used 60 CKT matter assessment items. Each CKT matter item has a similar structure – with an opening scenario, question, and set of options from which to select or respond. We developed a variety of item types, including multiple choice single select (*n*=24), multiple choice multiple selection (*n*=17), grid (*n*=9; in which PSETs make selections per row of a grid), inline choice (*n*=5; in which PSETs fill in blanks in a sentence), and matching (*n*=5) items, and incorporated different stimuli within them, such as students' written work, transcripts of students' conversations, video clips, and graphics. Since our goal was to develop an instrument that could be used on a large-scale and administered and scored efficiently, all the CKT items developed are discrete, automatically-scorable items. In addition, as shown in Figure 1, each item was designed to measure teachers' CKT at the intersection of one of the five matter content topics (e.g., conservation of matter) and one of the seven WOTS

[2] Other education programs for the 822 participants included: Master's degree programs (17%), fifth-year post-baccalaureate programs leading to a Master's degree (9%), alternate-route programs designed to expedite the transition of non-teachers to a teaching career (8%), fifth-year post-baccalaureate programs not leading to a Master's degree (1%), and other types of teacher preparation programs (1%).

instructional tool categories (e.g., student ideas). For example, the sample CKT matter item in Figure 2 assesses the matter content topic of "properties of matter materials" and the WOTS instructional tool category of "scientific resources". The figure includes a rationale for the item describing the knowledge that a PSET should draw upon when responding to this item.[3]

[Place Figures 1 and 2 about here.]

Items were developed through an iterative process that included peer review by item writers, expert panel review by five teacher educators, PSET review and interaction through cognitive interviews (n≈5 PSETs/item) and pilot testing (n≈200). The pilot participants were recruited similarly as the field test participants and selected to represent the *Praxis*® test-taking population (see Table 2 for population demographics). The expert panel reviews and cognitive interviews (in which PSETs thought out loud as they independently worked through a set of items) provided content and response process validity evidence that our items were assessing CKT for science and not purely subject matter knowledge (Mikeska et al., 2021; Mikeska et al., accepted). In particular, this earlier study used cognitive interviews to examine the CKT that elementary science teachers leveraged when responding to 118 different CKT matter items and found that most of the developed CKT matter items did meet the hypothesized assessment intent and justification, which suggests the items functioned as intended and were adequately assessing the teachers' CKT in this science content area. The pilot testing with a representative sample of *Praxis*® Elementary Science (5005) test-takers provided invaluable data on item performance.[4] Results from each review were used to revise and improve items or drop them from further consideration, for example, when an item's opening instructional scenario was unclear or when

[3] To preserve the security of the test form, the given item is a sample item that we developed but did not ultimately appear on the field test.
[4] PSETs who participated in the pilot were ineligible to participate in the field test.

an item was deemed to have multiple keys, respectively. Sixty developed items were used in the field test form that was built to match the test blueprint, which specified proportions of CKT matter items by five content matter topics and the seven WOTS instructional tool categories (i.e., each item was written to a cell in the table given in Figure 1). The 60-item form was intended to take about an hour to complete, with about a minute per item.

Data Collection and Analysis

All data collection occurred via online surveys and assessments in fall 2019. In addition to our CKT matter assessment field test form; study participants completed a background and self-efficacy questionnaire (Riggs & Enochs, 1990); an 8-item perceptions survey asking participants their perceived clarity, rigor, and relevance of the CKT matter assessment items, which was developed and used on a previous study (Mikeska et al., 2018); and an external measure of their SMK on the science topic of matter and its interactions using a previously developed and validated instrument (the AIM Properties of and Changes in Matter Elementary School Teacher Assessment developed by Horizon Research, Inc. as part of the National Science Foundation grant DUE-0928177). We used both the perceptions survey and AIM test as part of our analyses.

To address our first research question about item behavior we obtained classical item statistics, (e.g., proportion correct and item-total correlations), conducted distractor analyses [e.g., flagging if any distractors correlated positively with the total score, flagging if high-scoring participants selected a distractor more often than the key (correct choice)], and examined item timing data using our 822 field test participants. We used this analysis to create our final test form, dropping any items that did not function as expected, while still adhering to our test blueprint target percentages for items by content area and WOTS.

To address our second research question about the internal structure of the final form of the assessment, we used the results from the field test of PSETs' responses to these CKT matter items to conduct a dimensionality analysis examining the nature of their CKT and, ultimately, to make scaling decisions about what scores the test best support. We fit various Item Response Theory (IRT) Models to assess the internal structure, which are shown in Figure 3. We first fit a uni-dimensional (1D) IRT model (Figure 3a) that assumes all the items measure a single construct as well as a set of MIRT models that correspond to the theorized content and WOTS structures.[5]

Given low numbers of items by some of the content and WOTS, we grouped like matter content topics together into two groupings and likewise grouped the seven WOTS into four groups. The two dimensions by content are a dimension for materials and properties of matter (21 items) and one for model of matter, changes in matter, and conservation of matter (31 items). We grouped the materials and properties of matter topics together, as they collectively focus on teaching about descriptive properties that characterize matter and how properties can be used to explain the behavior of different types of matter and the suitability of specific materials for intended purposes. We collapsed the three other topics into one category due to their focus on understanding and explaining physical and chemical changes in matter.

The four dimensions by WOTS are: WOTS 1-3: Scientific goals, resources, and models (20 items), WOTS 4-5: Student ideas and scientific language (14 items), WOTS 6: Scientific explanations (9 items), and WOTS 7: Scientific investigations (9 items). We grouped the first three WOTS categories (scientific goals, resources, and models) together because the science teaching practices in these categories involve science teachers in making decisions about,

[5] We fit all (M)IRT models in R (R Core Team, 2020) using the "mirt" package (Chalmers, 2012).

evaluating, and critiquing the types and nature of instructional materials and activities, including ways to represent phenomena, that they can use to address their instructional goals and promote student learning. We consolidated WOTS category 4 (student ideas) and WOTS category 5 (scientific language) into one group because the science teaching practices in these two categories focus on the work that science teachers do to consider and anticipate students' previous experiences and understandings, including their use of scientific language, and figure out how to address specific difficulties or gaps they identify.

To assess the extent that PSETs integrated their content knowledge with teaching practices differently by content sub-area, we fit a 2-dimensional (2D) correlated factor MIRT model (or "between-item" MIRT model; e.g., Adams, Wilson, and Wang, 1997) model with separate dimensions for the two content subdomains (see Figure 3b). If PSETs were integrating content knowledge with teaching practices differently for materials and properties of matter items than items assessing models of matter, changes in matter, and conservation of matter, we would find that this model fits better than the simple 1D model.

To assess the extent that PSETs integrated their content knowledge with teaching practices differently by different teaching practices (i.e., WOTS categories), we fit a 4D correlated factor MIRT model with dimensions by sets of instructional tools (see Figure 3c). This 4D model allows us to explore the extent that PSETs are similarly proficient at integrating content knowledge with each of the 4 sets of instructional tools. For instance, a moderate correlation between the scientific explanations and scientific investigations dimensions would indicate that not all PSETs who are proficient at integrating science content knowledge with scientific explanations are proficient at such integration with scientific investigations. In contrast, estimated latent correlations near 1 among the dimensions would indicate that PSETs are rank

ordered similarly in terms of their proficiency at integrating content knowledge with each of the four sets of instructional tools, and provide evidence in favor of the simpler 1D model.

Finally, to assess the extent that PSETs perform distinctly on any given content *or* instructional tool category over and above their general integration of content knowledge and teaching practice, we fit a 7D model (see Figure 3d). In this model, all items load on an overall dimension that represents the integrated construct of CKT as well as one of the two content dimensions and one of the four instructional tool dimensions (i.e., each item loads on a total of 3 dimensions). To identify this model, all dimensions are constrained to be independent (i.e., correlations between dimensions equal 0). If PSETs' performance cannot be explained primarily by the overall CKT dimension, then this 7D model will exhibit better model fit than the simple 1D model. If PSETs are not always integrating content knowledge and teaching practices but instead, for instance, are performing distinctly on materials and properties of matter items regardless of the WOTS assessed by those items and likewise are performing distinctly on scientific explanations items regardless of content assessed, then this 7D model would likely fit better than the 1D model (and we would find high loadings on these auxiliary dimensions compared to the overall dimension); we would have evidence that PSETs had distinct content and teaching practice proficiencies rather than an overall integrated proficiency of CKT for matter.

[Insert Figure 3 about here.]

To assess model fit, we used typical model fit criterion: Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC), where smaller values indicate better model fit. For the cases where the MIRT models are nested within the 1D model, we also performed a likelihood ratio test to test if these models fit significantly better than the 1D model. In addition,

we inspected the estimated latent correlations among the dimensions of the 2D content and 4D instructional tools models. Estimated latent correlations near 1 would be evidence against the more complicated MIRT models and in favor of the simpler 1D model.

After determining which model fit best, we checked item fit for that model using the root mean square deviation (RMSD) item fit statistic, which quantifies the deviation between the observed and model-based item characteristic curves. Thresholds for good item fit range from 0.10 to 0.15 (e.g., Yamamoto, Khorramadel, & von Davier, 2016; OECD, 2014; Oliveri & von Davier, 2011). Depending on the identified best-fitting model, we considered other checks of the model adequacy, such as testing for local dependencies if the 1D model was found to fit best. We then scaled the test accordingly and computed the marginal IRT reliability (Thissen & Wainer, 2001).

To address the third research question about relationships with external variables, we first compared performance on the CKT matter test with personal background variables (gender and race/ethnicity), educational variables (undergraduate grade point average (GPA), undergraduate major/minor, and highest educational level obtained), and teacher preparation program (TPP) variables (type of teacher preparation program -- Bachelor's program, Master's program, alternate-route program), and emphasis of the program on K-5 science education, the five matter and its interactions content areas, and the 7 WOTS instructional tools). All variables were taken from the background questionnaire. See Supplemental Online Appendix A for how the variables were created from the background questionnaire questions and response options. In some cases, substantively similar response options were collapsed into a single category, and, in others, response options were dropped if they were distinct and had very small sample sizes ($n \leq 5$). We used one-way ANOVAs (or, equivalently, two-sample t-tests if only two groups) to test for

differences in mean performance between the groups for each variable. If the overall ANOVA was significant, we followed up with Tukey-adjusted pairwise comparisons.

The first set of variables—personal background variables—allow us to investigate potential adverse impact. The assessment items went through several cycles of review to ensure they were fair to all PSETs. Ideally, there would be no performance differences by gender identity or race/ethnicity. However, there has been previous evidence of differential performance by gender with males outperforming females, and by race/ethnicity, with White preservice teachers outperforming non-White preservice teachers on teacher licensure assessments (e.g., Gitomer et al., 2011; Steinberg et al., 2016; Gitomer, 2007; Nettles et al., 2011). Our hope is that these historical differences by gender and race/ethnicities differences would not be evident in our CKT matter assessment for PSETs, although they may persist.

The educational and TPP variables serve as proxies of PST preparation, training, and knowledge related to the assessed construct: CKT of matter and its interactions. For instance, we hypothesize that PSTs with higher undergraduate GPAs, undergraduate major/minors more related to education and science, and TPPs that emphasize the content of the CKT matter assessment would perform better.

We also examined the relationship between our CKT matter and two external measures of SMK in science (*Praxis*® Science and AIM). These two direct measures of SMK are also continuous test scores and thus we took the Pearson correlation between the CKT matter scores and each of the *Praxis*® Science and AIM test scores. Competency in SMK is a required foundation for strong CKT, and thus we expected find moderate correlations between our CKT matter assessment and each of the science SMK assessments. Because AIM is also focused on

matter and its interactions, whereas *Praxis*® Science covers science topics more broadly, we expected a higher correlation with AIM.

Lastly, to address our fourth research question regarding face validity, we reviewed the PSETs' perceptions of the CKT matter assessment. At the end of their test sessions, participants were asked to rate their level of agreement (strongly disagree, disagree, agree, or strongly agree) with eight statements about the CKT matter assessment items they had just completed. These questions largely focused on participants' perceived understanding, rigor, and relevance of the items. This brief survey provides a quick snapshot of PSETs' perceptions. The extent that the assessment clearly assessed CKT about matter and its interactions, or content validity evidence, was established through cognitive interviews and expert reviews of all items (Cisterna et al., 2022; Mikeska et al., 2022). However, the brief end-of-assessment perceptions survey also provides some content validity evidence. PSTs' perception of the items helps us discern if performance was contaminated by confusion regarding the content or utility of the test items. For ease of interpretation, we compared the percentage of PSTs who expressed any level of disagreement (strongly disagree or disagree) with the percentage who expressed any level of agreement (strongly agree or agree) across the 8 perceptions items.

Results

Research Question 1: Validity Evidence Based on Test Content

Findings from the item analysis showed that the CKT matter items generally functioned well. Table 3 indicates the number of CKT matter items flagged by item type using common flagging criteria (e.g., California Department of Education [CDE], 2020; Castellano & McCaffrey, 2021). Only eight of the 60 items were flagged due to aberrant item statistics (e.g., too difficult or low item discrimination), an additional two were flagged for review because of

concerns identified in their empirical item characteristic curves (ICCs) (e.g., curves for probability of selecting a distractor increasing instead of decreasing with increasing ability), and two more items were flagged for item timing (median item time > 90 seconds), resulting in a total of 12 items for review. No items were flagged as too easy (proportion correct ≥ .95), and two items were flagged as too difficult (proportion-correct ≤ .2). Six items were flagged for having low item-total correlations (≤ .2), or not discriminating between low and high performers, with no items having negative item-total correlations. In any case that an item was flagged for an issue with a distractor (e.g., positive distractor-total correlation, high-scoring test takers selecting a distractor more often than key, or mean total test score higher for those selecting a distractor than the key), it was also flagged for either being too difficult or not discriminating. Overall, on average, PSETs responded correctly to about 56% of the items, and no test-takers earned more than 59 out of the total 60 possible score points. Cronbach's alpha for the overall test was 0.914.

After a careful review of all the flagged items, the final form was constructed using 52 of the 60 field tested items, dropping six of the items flagged due to poor item statistics and the two items identified by the ICC review. The other two items flagged due to poor item statistics were grid items and it was obvious the issues arose due to a specific row in the items. We dropped those rows and rescored the items to address the identified issues. For the final 52-item form, the mean proportion correct was 0.61 (min=.28, max=.92), mean item-total correlation was 0.40 (min=.23, max=.55), and Cronbach's alpha was 0.918. We also verified that the 52-item final form still met our test blueprint in terms of proportions of items by each of the five content areas and seven WOTS instructional tools. This final form was used in the dimensionality and scaling analyses.

[Insert Table 3 about here.]

Research Question 2: Internal Structure Validity Evidence

To first provide a sense of the performance by the two matter content topics and four WOTS instructional tool categories, Table 4 provides some descriptive statistics for their simple summed scores. Cronbach's alpha reliability is reasonable (.75 to .86) for the scores with 14 or more items but for the two WOTS subscores based on only 9 items, the reliability is low (0.6 and .69). The percentage correct scores for each area averaged over all PSETs ranged from 58% to 64% for the two topic areas and 56% to 65% for the four WOTS categories. That is, performance is generally moderate regardless of content or teaching practice assessed.

[Insert Table 4 about here.]

Turning to the dimensionality analyses to examine more systematically the internal structure of the CKT matter assessment, findings showed that none of the MIRT models fit better than the simple 1D model. As seen in Table 5, the AIC for the 1D and 2D model by content area are very similar and the smallest values across the models. In terms of BIC, the 1D model has the smallest value, indicating better model fit. Moreover, the likelihood ratio tests comparing the 2D content and 4D WOTS models to the 1D model were each not significant at the $\alpha = 0.05$ significance level, indicating that they did not fit significantly differently (better) than the 1D model. The loadings on the auxiliary content and WOTS dimensions of the 7D model were generally much smaller than the loading on the overall dimension, which coupled with the higher AIC and BIC for this model, suggest that it is not a better fit or more descriptive of the data than the simpler 1D model.

[Insert Table 5 about here.]

In addition, the latent correlations (estimated by the MIRT models) among the content and WOTS dimensions in the 2D and 4D models, as shown in the "latent correlation estimate"

column of Table 6, were all estimated between .97 and .99, indicating very strong correspondence of performance among the different subareas. That is, PSETs who performed well (or poorly) in one area tended to perform well (or poorly) in another. As another estimate of latent correlation, the last column of Table 6 shows the correlations between the raw sub-scores (i.e., simple number-correct scores for each content or WOTS category) disattenuated for measurement error. They are all also essentially 1. Accordingly, both sets of pairwise correlation estimates between latent dimensions by content and WOTS categories are nearly perfectly correlated, suggesting that performance is not differentiated by content area or instructional tool category.

[Insert Table 6 about here.]

The strong correspondence between performance on the content categories may not be surprising given the two content categories are both sub-areas of the focused science topic of matter and its interactions. Their near perfect correlation indicates that PSETs are integrating content knowledge for each content area with teaching practices similarly well. This is not to say that all PSETs are good at integrating content knowledge and teaching practices, but rather that PSETs are rank ordered in terms of their proficiency with integrating content knowledge and teaching practices just as well for each of the two matter content topics.

Exploratory factor analysis also revealed that a 1D solution was appropriate as seen in the scree plot in Figure 4, with only one eigenvalue greater than 1 and a large drop off after the first eigenvalue at which point the eigenvalues begin to level off (e.g., Kline, 2004). The model fit, latent correlation estimates, and scree plot all provide strong evidence in favor of a

unidimensional solution.[6] That is, the items measure a common construct of CKT for matter and its interactions, and PSETs are not showing any differential performance by content area or instructional tools.

Accordingly, we scaled the test with a 1D two-parameter-logistic (2PL) IRT model (see e.g., Ryan & Brockmann, 2018 for a more detailed description of a 2PL IRT model), which includes item parameters for item difficulty and item discrimination (or the extent the item discriminates/distinguishes between high- and low-performing examinees). The item difficulty estimates ranged from -2.3 to +1.45 (mean = -.53), indicating a range of difficulties from easy to moderately difficult, and the item discriminations ranged from 0.52 to 2.26 (mean = 1.12). All items fit well with this model as indicated by RMSD values all less than or equal to the threshold of 0.10 (values ranged from 0.03 to 0.10).

As one more check on the adequacy of the identified best-fitting model, we tested for violations of local independence or the extent that PSET scores on one item depended on scores on another item beyond that of the general CKT matter ability. Such violations would indicate that the simple 1D model was not capturing all the variation in PSET performance. We used Fisher's z-transformed Yen's (1984) Q3 statistic and adjusted for multiple comparisons (controlling the false discovery rate) with the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). We would be concerned if several item pairs were flagged with all of them items having a feature in common such as item type, type of stimuli used, or content assessed. Out of the 1,326 possible item pairs of the 52 items, only one item pair was flagged (at the

[6] Reliabilities of the subscores by content or WOTs categories after controlling for variance due to the overall general factor of CKT of matter and its interactions (i.e., ω_{HS} defined in Rodriguez, Reise, & Haviland, 2016) were also all close to 0, ranging from 0.00 to 0.03, which provides further evidence that reporting subscores by content or WOTS categories would not provide any more distinct or reliable information than is in the total score. Moreover, the coefficient omegaH was .95, indicating that 95 percent of the variance in the total score can be attributed to the single overall CKT dimension.

alpha=.05 level). Accordingly, modeling PSET performance with a single general CKT matter dimension in the IRT framework was further supported as sufficient.

We estimated PSETs' latent abilities using the inverted-TCC (test characteristic curve) method, which transforms a PSET's number correct score to a theta estimate (e.g., Kolen & Tong, 2010). The marginal IRT reliability was 0.91. We then scaled the test so that it had mean 300, standard deviation 10, and average conditional standard error of measurement of 3. The scale ranges from 265 to 335. We primarily wanted to avoid using a scale similar to *Praxis*® (with a score range of 100 to 200) so the test scores were not misinterpreted. The distribution of the PSETs CKT matter test scores is given in Figure 5. It shows that the majority of scores were low to mid-ranging with few high performers.

Research Question 3: Validity Evidence Based on Relations with Other Variables

The results of the statistical tests investigating the relationships between the CKT matter scores and PSET variables are provided in Table 7. In terms of relationships between CKT matter scores and personal background variables, there was no significant difference by gender identity. However, PSETs who identified as White performed significantly better (by about 7 score points on average) than those who identified as Asian or Asian American and Black or African American. In addition, PSETs who identified as two or more races scored significantly higher by 6.5 points on average than those who identified as Black or African American. There were no other significant differences among race/ethnicities. We discuss this finding further in the Discussion.

[Insert Table 7 about here.]

In terms of relationships between CKT matter scores and education variables, PSETs with high undergraduate GPAs scored significantly higher by 4 points on average than those with

low GPAs. That is, consistent with our hypothesis, PSETs with better undergraduate records performed better on the CKT matter assessment. There was no significant difference by the focus of PSETs' undergraduate major or minor, but surprisingly, those with Master's degrees (about 5.5% of the sample) scored significantly lower than those who were pursuing a Bachelor's (by 3.8 points on average) or had already obtained a Bachelor's degree (by 4.7 points on average). We probe this result further in the discussion.

In terms of relationships between CKT matter scores and TPP variables, there was no significant difference in scores by TPP type (Bachelor's program, Master's program, alternate-route program). PSETs in TPPs that had a stronger emphasis on K-5 science education and those in programs that more strongly emphasized the WOTS scored significantly higher (by 2 and 3 points, respectively) than those in programs with limited or no emphasis in each of these areas, which is consistent with our hypotheses. However, inconsistent with our expectations, PSETs in programs that had limited or no emphasis on the five matter and its interactions topics scored significantly higher by 1.6 points on average than those in programs that reported emphasis in these areas. See the Discussion for further exploration of this result.

In terms of the relationships between the CKT matter scores and related external measures, we found that PSETs' scale scores on the CKT matter test were correlated 0.53 (disattenuated $r = 0.63$) with *Praxis*® Science and 0.66 with the Horizon AIM test (disattenuated $r = 0.72$). Given that the *Praxis*® Science test broadly measures science content knowledge, whereas the AIM specifically assesses content knowledge for matter and its interactions—the same content area as our CKT matter assessment—the stronger correlation with the AIM is consistent with our expectations. Since neither external measure is a measure of CKT, the moderate size of the correlations is supportive of the conjecture that our CKT matter assessment

is assessing a construct distinct from, but related to, pure subject matter knowledge. Moreover, these moderate correlations are similar to those reported in the literature between measures of science teachers' SMK and measures designed to assess these practice-based aspects of CKT more closely (Davidowitz & Potgieter, 2016; Grobschedl et al., 2015; Mikeska et al., 2017; Mikeska et al., 2018).

Research Question 4: Face Validity Evidence Based on PSETs' Perceptions

The participating PSETs' perceptions of the CKT matter assessment are summarized in Table 8. The majority of PSETs' (77%) found the items clear. Although the items were not confusing, the majority of PSETs' (78%) found them challenging and were evenly split as to whether it was difficult to choose among the answer options. The items may have been challenging in part because they stretched beyond aspects of teaching PSETs had previously considered (93%) or material covered (55%) in their teacher preparation programs (by time of participating in our study). However, the majority of PSETs agreed that the assessed material is relevant to what they will teach in an elementary school classroom (85%) and that elementary school teachers should be able to answer the items correctly (82%). PSETs were roughly split (47% disagree vs 53% agree) on whether professionals in other fields than teaching should also be able to answer most questions correctly, suggesting that about half of these PSETs see the CKT assessed in these items as primarily within elementary teachers' purview.

[Insert Table 8 about here.]

Discussion

Across the last couple decades, there has been a tremendous amount of effort and attention in science education to examining and studying the two main domains of science teachers' CKT – their SMK and PCK. These efforts have resulted in a plethora of varied

approaches and data sources for investigating science teachers' CKT. Most of the approaches tend to target either science teachers' SMK or one or more aspects of their PCK. As noted earlier, only limited research and development efforts have focused on developing tools or instruments that measure these CKT domains in an integrated fashion. However, research has suggested that as elementary teachers engage in the work of teaching science, they frequently draw upon multiple forms of knowledge simultaneously in an integrated manner. As such, this research study makes three important contributions to science teacher education and the measurement of science teachers' CKT.

First, study results suggest that the instrument our team designed and field tested with a national sample of PSETs is efficient and effective and supports valid inferences about PSETs' CKT about matter and its interactions. One direct implication of these findings is that this instrument has the potential to be used on a large scale across groups of PSETs within and across teacher education programs. Although our assessment instrument is limited to a specific science content area, we specifically chose matter and its interactions because it is a high-leverage content area that is fundamental to the K-12 science curriculum. The general consensus by PSET participants that the assessed material is indeed material they expect to be able to teach in an elementary science classroom along with reviews by content experts provides validating evidence for our choice. Similar assessment development efforts could be undertaken to make CKT science tests for other content areas or topics, as earlier research has suggested that science teachers' knowledge, especially their PCK, is domain and topic specific (Azam, 2019; Mavhunga, 2020). Further research could investigate the extent that CKT assessment in multiple science content areas are needed. If PSETs' performance across multiple such assessments is highly correlated, then one or two such assessments may be sufficient. Similarly, future research

could extend our test blueprint to the broader science domain across content areas to determine if it is feasible to have a single, general science CKT assessment within a reasonable test time.

Second, our results suggest that our CKT matter assessment is assessing a construct distinct from pure SMK. The magnitude of the correlations between the CKT matter assessment scores with the *Praxis*® Science and AIM Horizon tests were consistent with expectations for a related, but distinct construct. Although the field test PSETs did not have any personal stake in their performance, their timing data suggests that for the most part, they attended to the items and did not simply skip through them (average time PSETs spent per item was 57 seconds which is close to the intended 60 seconds per item). PSETs were also compensated for their time so had some external motivation to complete all the administered surveys and assessments with a good-faith effort. As with any new assessment program, further research would need to confirm that our results held in more high-stakes settings such as when PSET grades or licensure depended on the CKT matter scores.

Third, performance on the CKT matter instrument varies by PSET variables in ways that were generally consistent with expectations: PSETs with higher undergraduate GPAs and those in TPPs that emphasize K-5 science education and emphasize the WOTS performed better on the CKT matter assessment. However, there were some instances in which findings ran counter to our a priori hypotheses: PSET performance differed across some races/ethnicities, PSETs with Master's degrees performed worse than those pursuing or already obtained a Bachelor's degree and those in TPPs that emphasized the five matter content topics performed worse (though by about 1.5 points) than those in programs that had little or no emphasis on these topics. Despite efforts during item development to check items for fairness concerns, there is potential for adverse impact by race/ethnicity. Further construct and item review by additional experts of

diverse backgrounds and race/ethnicities and additional data collection with oversampling (than is represented in the teacher population) by each non-White race/ethnicity to obtain more robust sample sizes (e.g., currently N=20 for Asian or Asian American) to support differential item functioning analyses by race/ethnicities would help further probe the potential for adverse impact.

Upon further inspection, we also found similar relationships with the external measures of subject matter knowledge—both for the observed race/ethnicity differences and those by education level obtained and program focus. For instance, the participating Black or African American PSETs also scored significantly lower on Praxis Science than several different race/ethnicity groups—Asian or Asian American, Hispanic or Latino, White, and two or more races—and had significantly lower undergraduate GPAs than White PSETs (proportion of Black/African American with GPAs of 3.5 or above was 30% versus 68% for White PSETs, p-value=0). Similarly, those with Master's degrees as their highest education level obtained also had significantly lower AIM scores on average and lower (but not significant) average Praxis Science scores. The relatively small portion of the sample with Master's degrees may be lower performing in general. Another possibility is that PSETs with Master's degrees may have had their disciplinary science courses several years ago as compared to the PSETs pursuing or having obtained a Bachelor's degree, who likely had such critical and relevant courses (for the assessment) more recently. Those in TPPs that emphasized the matter topics also scored lower on average on Praxis Science than those in TPPs that had little or no emphasis on these topics (although not statistically significant; p-value =0.12). Further targeted research could help investigate these findings further with additional statistical modeling (e.g., that includes controlling for multiple variables at once and allows for interactions among variables), follow up

with a sample of PSETs to obtain further information about their Master's degrees and TPP programs, and/or with further data collection designed to collect representative samples by these variables.

Finally, our results provide empirical evidence suggesting that there are strong connections between various aspects of PSETs' CKT in one content area. That is, the nature of PSETs' CKT within a topic area may be more integrated and less siloed than previous research has suggested. The unidimensional structure of the CKT about matter and its interactions construct indicates that the CKT matter items are not measuring separate dimensions of this construct. Although the dimensionality analyses did not reveal a test structure indicated by the test design, this result: (1) is not surprising given that tests are often found to be unidimensional even if built to assess the integration of content and (some type of) practice as with the California Science Test (CDE, 2020), (2) does not suggest that the design of the test should be narrowed as the five matter content topics and seven WOTS instructional tools are needed to cover the construct, and (3) does not suggest that there is little differentiation in PSETs' CKT matter ability. Rather, PSETs perform at varying levels, but PSETs that can implement teaching practices well with one content matter topic tend to also be able to do so in other matter content topics, and similarly, PSETs that can implement a particular teaching practice well across matter content topics tend to also be able to do so with other teaching practices. Although we did not find psychometric evidence to support reporting separate scores by any of the matter content topics or the WOTS instructional tools, that is not to say that reporting such results at an aggregate level, such as for a classroom or teacher preparation program would not be meaningful. Further analyses (with more data per institution) are needed to evaluate the value added of reporting out along these dimensions for classrooms or entire preparation programs.

Choi and Papageorgiou (2020), for example, found evidence against reporting TOEFL subscores for individuals but for reporting them for institutions as long as they had more than 50 examinees. Finer-grain reporting at the aggregate level would allow teacher educators to diagnose gaps in their PSETs' preparation and identify specific areas in which the program could strength its emphasis on CKT.

Based on these results, we see a few possibilities for using such an assessment on a large scale. One such use would be as part of a national indicator system tracking science teachers' quality and professional growth across their career. Another use case could focus on identifying patterns of strengths across teacher groups using group-level summaries of performance by content topics and/or WOTS and identify areas of focus for teacher preparation and professional development support. A final use could be to incorporate as part of future research studies to examine and monitor science teacher learning and the relation of science teachers' CKT to various contextual factors, such as their beliefs about students or their previous teaching experiences, and student outcomes or to compare teacher learning across content areas.

It is still the case that few districts, states, teacher education programs, or professional development programs have approaches to efficiently collecting data on science teachers' CKT. The results from this study suggest that developing automatically scorable CKT science assessments that target science teachers' knowledge in the work of teaching science is one viable solution to this widescale challenge and CKT assessment response data can provide useful information about science teachers' CKT. The hope is that such information could be used for monitoring or making decisions about teachers' professional development – although future research will need to examine this claim empirically.

References

Abell, S. K. (2013). Research on science teacher knowledge. In *Handbook of research on science education* (pp. 1119-1164). Routledge.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *The Standards for Educational and Psychological Testing,* American Educational Research Association: Washington, D.C.

Arzi, H. J., & White, R. T. (2008). Change in teachers' knowledge of subject matter: A 17-year longitudinal study. *Science Education*, *92*(2), 221-251.

Azam, S. (2019). Distinguishing topic-specific professional knowledge from topic-specific PCK: A conceptual framework. *International Journal of Environmental & Science Education*, *14*(5), 281-296.

Ball, D. L., & McDiarmid, G. W. (1989). The subject-matter preparation of teachers. In Houston, W. R. (Ed.), Handbook of research on teacher education (pp. 437–465). New York: Macmillan.

Banilower, E. R., Heck, D. J., & Weiss, I. R. (2007). Can professional development make the vision of the standards a reality? the impact of the national science foundation's local systemic change through teacher enhancement initiative. *Journal of Research in Science Teaching, 44*(3), 375-395.

Barnett, J., & Hodson, D. (2001). Pedagogical context knowledge: Toward a fuller understanding of what good science teachers know. *Science Education*, *85*(4), 426-453.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B, 57*, 289–300.

Bergqvist, A., Drechsler, M., & Chang Rundgren, S. N. (2016). Upper secondary teachers'

knowledge for teaching chemical bonding models. *International Journal of Science Education*, *38*(2), 298-318.

Bertram, A., & Loughran, J. (2012). Science teachers' views on CoRes and PaP-eRs as a framework for articulating and developing pedagogical content knowledge. *Research in Science Education*, *42*(6), 1027-1047.

California Department of Education. (2020). *California Science Test 2018-2019 Technical Report.* (Contract #CN150012). Sacramento, CA: California Department of Education. Retrieved from https://www.cde.ca.gov/ta/tg/ca/documents/cast19techrpt.pdf

Carpendale, J., & Hume, A. (2019). Investigating practising science teachers' pPCK and ePCK development as a result of collaborative CoRe design. In *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 225-252). Springer, Singapore.

Castellano, K. E., & McCaffrey, D. F. (2021). Student test characteristics. In OECD (Eds.). *Global Teaching InSights: Technical Report. (Ch. 17).* OECD: Paris.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1 – 29. doi: 10.18637/jss.v048.i06.

Chan, K. K. H. (2022). A critical review of studies using the pedagogical content knowledge map approach. *International Journal of Science Education*, 1-27.

Chan, K. K. H., & Hume, A. (2019). Towards a consensus model: Literature review of how science teachers' pedagogical content knowledge is investigated in empirical studies. *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science*, 3-76.

Choi. I., & Papageorgiou, S. (2020). Evaluating subscore uses across multiple levels: A case of reading and listening subscores for young EFL learners. *Language Testing, 37*(2), 254 – 279.

Cisterna, D., Bookbinder, A. K., Mikeska, J.N., & Lakhani, H. R. (2022). Elementary preservice teachers' perceptions of assessment tasks that measure content knowledge for teaching about matter. *Journal of Science Teacher Education*, 1-28. https://doi.org/10.1080/1046560X.2021.2015831

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.

Diamond, B. S., Maerten-Rivera, J., Rohrer, R., & Lee, O. (2013). Elementary teachers' science content knowledge: Relationships among multiple measures. *Florida Journal of Educational Research*, *51*(1), 1-20.

Etkina, E., Gitomer, D., Iaconangelo, C., Phelps, G., Seeley, L., & Vokos, S. (2018). Design of an assessment to probe teachers' content knowledge for teaching: An example from energy in high school physics. *Physical Review Physics Education Research*, *14*(1), 010127.

Fischer, H. E., Borowski, A., & Tepner, O. (2012). Professional knowledge of science teachers. In *Second international handbook of science education* (pp. 435-448). Springer, Dordrecht.

Gess-Newsome, J., Taylor, J. A., Carlson, J., Gardner, A. L., Wilson, C. D., & Stuhlsatz, M. A.

(2019). Teacher pedagogical content knowledge, practice, and student achievement. *International Journal of Science Education*, *41*(7), 944-963.

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. (2015). Evidence on the validity of content knowledge for teaching assessments. *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*, 493-528.

Großschedl, J., Harms, U., Kleickmann, T., & Glowinski, I. (2015). Preservice biology teachers' professional knowledge: Structure and learning opportunities. *Journal of Science Teacher Education*, *26*(3), 291-318.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371-406.

Hume, A., & Berry, A. (2011). Constructing CoRes—a strategy for building PCK in pre-service science teacher education. *Research in Science Education*, *41*(3), 341-355.

Jüttner, M., Boone, W., Park, S., & Neuhaus, B. J. (2013). Development and use of a test instrument to measure biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK). *Educational Assessment, Evaluation and Accountability*, *25*(1), 45-67.

Kennedy, M. M. (1998). Education reform and subject matter knowledge. *Journal of Research in*

Science Teaching, *35*(3), 249-263.

Kirschner, S., Borowski, A., Fischer, H. E., Gess-Newsome, J., & von Aufschnaiter, C. (2016). Developing and evaluating a paper-and-pencil test to assess components of physics teachers' pedagogical content knowledge. *International Journal of Science Education*, *38*(8), 1343-1372.

Kolen, M. J. & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice, 29*(3), 8-14.

Kline, R. B. (2004). *Principles and Practice of Structural Equation Modeling, Second Edition.* Guilford Publications: New York, NY.

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.

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

Loughran, J., Mulhall, P., & Berry, A. (2008). Exploring pedagogical content knowledge in science teacher education. *International Journal of Science Education, 30*(10), 1301–1320.

Marshall, J. C., Smart, J., & Alston, D. M. (2016). Development and validation of Teacher Intentionality of Practice Scale (TIPS): A measure to evaluate and scaffold teacher effectiveness. *Teaching and Teacher Education*, *59*, 159-168.

Mavhunga, E. (2020). Revealing the structural complexity of component interactions of topic-specific PCK when planning to teach. *Research in Science Education*, *50*(3), 965-986.

McConnell, T. J., Parker, J. M., & Eberhardt, J. (2013). Assessing teachers' science content

knowledge: A strategy for assessing depth of understanding. *Journal of Science Teacher Education*, *24*(4), 717-743.

Menon, D., & Sadler, T. D. (2016). Preservice elementary teachers' science self-efficacy beliefs and science content knowledge. *Journal of Science Teacher Education*, *27*(6), 649-673.

Mikeska, J.N., Kurzum, C., Steinberg, J., & Xu, J. (2018). Assessing elementary science teachers' content knowledge for teaching science for the ETS Educator Series: Pilot results. *ETS Research Report Series* (Research Report No. RR-18-20). Princeton, NJ: ETS. https://onlinelibrary.wiley.com/doi/full/10.1002/ets2.12207

Mikeska, J.N., Cisterna, D., Lakhani, H., Bookbinder, A.K., Myers, D.L., & Vaval, L. (accepted). Knowledge in use: Examining elementary science teachers' responses to assessment tasks designed to measure their content knowledge for teaching about matter and its interactions. *Science Education.*

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. *ETS Research Report Series*, *2018*(1), 1-30.

Mikeska, J. N., Phelps, G., & Croft, A. J. (2017). Practice-based measures of elementary science teachers' content knowledge for teaching: Initial item development and validity evidence. *ETS Research Report Series*, *2017*(1), 1-72.

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.

Monk, D. H., & King, J. (1994). Multi-level teacher resource effects on pupil performance in

secondary mathematics and science. In Ehrenberg, R. G. (Ed.), Contemporary policy issues: Choices and consequences in education (pp. 29–58). Ithaca, NY: ILR Press.

National Research Council. (2013). *Monitoring progress toward successful K-12 STEM education: A nation advancing?* Washington, DC: The National Academies Press. doi: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

Nilsson, P., & Elm, A. (2017). Capturing and developing early childhood teachers' science pedagogical content knowledge through CoRes. *Journal of Science Teacher Education*, *28*(5), 406-424.

Nilsson, P., & Karlsson, G. (2019). Capturing student teachers' pedagogical content knowledge (PCK) using CoRes and digital technology. *International Journal of Science Education*, *41*(4), 419-447.

Nixon, R. S., Campbell, B. K., & Luft, J. A. (2016). Effects of subject-area degree and classroom experience on new chemistry teachers' subject matter knowledge. *International Journal of Science Education*, *38*(10), 1636-1654.

Nixon, R. S., Hill, K. M., & Luft, J. A. (2017). Secondary science teachers' subject matter knowledge development across the first 5 years. *Journal of Science Teacher Education*, *28*(7), 574-589.

Nixon, R. S., Toerien, R., & Luft, J. A. (2019). Knowing more than their students:

Characterizing secondary science teachers' subject matter knowledge. *School Science and Mathematics*, *119*(3), 150-160.

Organisation for Economic Cooperation and Development (OECD). (2014). Scaling PISA Data. *PISA 2015 Technical Report* (Ch. 9). Paris: OECD. Retrieved from 09_Chapter_09_PISA2015.pdf (oecd.org)

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*(3), 315 – 333.

Phelps, G., Gitomer, D. H., Iaconangelo, C. J., Etkina, E., Seeley, L., & Vokos, S. (2020). Developing assessments of content knowledge for teaching using evidence-centered design. *Educational Assessment*, *25*(2), 91-111.

Phelps, G., & Howell, H. (2016). Assessing mathematical knowledge for teaching: The role of teaching context. *The Mathematics Enthusiast*, *13*(1), 52-70.

R Core Team. (2020). R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria.

Rice, D. C. (2005). I didn't know oxygen could boil! What preservice and inservice elementary teachers' answers to 'simple' science questions reveals about their subject matter knowledge. *International Journal of Science Education*, *27*(9), 1059-1082.

Riggs, I., & Knochs, L. (1990). Towards the development of an elementary teacher's science teaching efficacy belief instrument. *Science Education, 74*(6), 625-637. doi:10.1002/sce.3730740605

Rodriguez, A. Reise, S., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137 – 150.

Roth, K. J., Garnier, H. E., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. (2011). Videobased lesson analysis: Effective science PD for teacher and student learning. *Journal of Research in Science Teaching*, *48*(2), 117-148.

Ryan, J., & Brockmann, F. (2019). *A practitioner's introduction to equating with primers on classical test theory and Item Response Theory.* Council of Chief State School Officers, Washington, D.C.

Sadler, P. M., Coyle, H., Smith, N. C., Miller, J., Mintzes, J., Tanner, K., & Murray, J. (2013). Assessing the life science knowledge of students and teachers represented by the K–8 national science standards. *CBE—Life Sciences Education*, *12*(3), 553-575.

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal*, *50*(5), 1020-1049.

Shugart, S. S., & Hounshell, P. B. (1995). Subject matter competence and the recruitment and retention of secondary science teachers. *Journal of Research in Science Teaching*, *32*(1), 63-70.

Shulman, L. S. (2015). PCK: Its genesis and exodus. In *Re-examining pedagogical content knowledge in science education* (pp. 13-23). Routledge.

Smith, P. S., & Plumley, C. L. (2016). *A review of the research literature on teaching about the small particle model of matter to elementary students.* Chapel Hill, NC: Horizon Research, Inc.

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.

Supovitz, J. A., & Turner, H. M. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching, 37*(9), 963-980.

Talanquer, V. (2009). On cognitive constraints and learning progressions: The case of "structure of matter." *International Journal of Science Education*, *31*(15), 2123–2136. http://doi.org/10.1080/09500690802578025

Tsaparlis, G., & Sevian, H. (2013). *Concepts of matter in science education*. Dordrecht: Springer. http://doi.org/10.1007/978-94-007-5914-5

Van Dijk, E. M., & Kattmann, U. (2007). A research model for the study of science teachers' PCK and improving teacher education. *Teaching and Teacher Education*, *23*(6), 885-897.

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

Wilson, S. M., Floden, R. E., & Ferrini-Mundy, J. (2001). Teacher preparation research: Current knowledge, gaps, and recommendations. East Lansing: Michigan State University.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2016). Scaling PIAAC cognitive data. In OECD (Eds.) *Technical Report of the Survey of Adult Skills (PIAAC; 2nd Edition).* Paris: OECD. Retrieved from PIAAC_Technical_Report_2nd_Edition_Full_Report.pdf (oecd.org)

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–45

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: Praeger.

Table 1. Work of Teaching Science (WOTS) framework

Instructional Tools	Examples of Instructional Practices
1. Scientific Instructional Goals, Big Ideas, and Topics	Choosing which science ideas or instructional activities are most closely related to a particular instructional goal
2. Scientific Resources (texts, curriculum materials, etc.)	Evaluating instructional materials for their ability to address scientific concepts; engage students with relevant phenomena; promote students' scientific thinking; and assess student progress
3. Scientific Models and Representations	Evaluating or selecting scientific models and representations that predict or explain scientific phenomena or address instructional goals
4. Student Ideas	Analyzing student ideas for common misconceptions regarding intended scientific learning
5. Scientific Language, Discourse, and Vocabulary	Anticipating scientific language and vocabulary that may be difficult for students
6. Scientific Explanations	Critiquing student-generated explanations or descriptions for their accuracy, precision, or consistency with scientific evidence
7. Scientific Investigations and Demonstrations	Selecting investigations or demonstrations that facilitate understanding of disciplinary core ideas, scientific practice, or cross-cutting concepts

Table 2. Comparing field test sample with *Praxis*® population

Characteristic	Field Test Sample	*Praxis*® Population
Female	93%	92%
White	80%	80%
Midwest	4%	4%
Northeast	22%	22%
South	46%	49%
West	28%	25%
Praxis Q1 (lower scoring)	22%	26%
Praxis Q2	27%	27%
Praxis Q3	24%	23%
Praxis Q4 (higher scoring)	27%	24%

Note: Praxis Q1 to Q4 represent the four quartiles of the *Praxis*® Elementary Science test scores in our population (those who took this test from January 2018 to June 2019).

Table 3. Numbers of items flagged by item type (includes flags by item statistics, visual inspection of empirical item characteristic curves, and item timing)

Item Type	Total Number of Items	Items with 1 Flag n (%)	Items with 2 or More Flags n (%)	Items Removed n (%)
Multiple Choice Single Selection	24	2 (8%)	1 (4%)	2 (8%)
Multiple Choice Multiple Selection	17	2 (12%)	2 (12%)	3 (18%)
Grid Multiple Selection	9	0 (0%)	4 (44%)	2* (22%)
Inline Choice Single/Multiple Selection	5	1 (20%)	0 (0%)	1 (20%)
Match Multiple Selection	5	0 (0%)	0 (0%)	0 (0%)

*Note: For the 2 other flagged grid items, it was obvious that the poor item statistics were due to a particular row in the grid so we dropped that row and rescored the items.

Table 4. Descriptive statistics for the two matter content topics and four teaching practice sum scores.

Type	Subarea	Number of Items	Reliability (Cronbach's alpha)	Mean (Percent Correct)	SD (Percent Correct)
Matter Topic	Materials & Properties	21	0.84	64%	23%
Matter Topic	Model, Change, and Conservation of Matter	31	0.86	58%	20%
Teaching Practice	WOTS 1-3: Goals, Resources, & Models	20	0.82	56%	23%
Teaching Practice	WOTS 4-5: Student Ideas and Language	14	0.75	65%	22%
Teaching Practice	WOTS 6: Explanations	9	0.69	64%	24%
Teaching Practice	WOTS 7: Investigations	9	0.60	60%	22%

Table 5. Model fit statistics

Model	AIC[1]	BIC[1]	logLik	Model fit test compared to 1D model for nested models
1D	47011.2	47501.2	-23401.6	
2D by Content[2]	47010.6	47505.4	-23400.3	$P(\chi^2(1) > 2.59) = 0.11$
4D by WOTS[3]	47029.6	47547.9	-23404.8	$P(\chi^2(6) > -6.39) = 1.0$
7D model with overall dimension and dimensions by Content and WOTS[4]	47032.1	48012.1	-23308.0	

Notes:

1. Models with smaller AIC and BIC show better model fit. The 1D model has the smallest BIC and about the same (though slightly larger) AIC as the 2D model by content areas. The last column provides the p-value for the chi-squared test of model fit compared to the 1D model. Not significant results indicate that the model does not fit significantly better than the 1D model.
2. The two dimensions by content are a dimension for materials and properties of matter (21 items) and one for model of matter, changes in matter, and conservation of matter (31 items).
3. The four dimensions by WOTS are: WOTS 1-3: Scientific goals, resources, and models (20 items), WOTS 4-5: Student ideas and scientific language (14 items), WOTS 6: Scientific explanations (9 items), and WOTS 7: Scientific investigations (9 items).
4. The 7D model includes an overall dimension and then 6 dimensions by Content and WOTS, which are the four dimensions for the WOTS followed by the two for Content. Each item loads on the overall dimension as well as one of the content dimensions and one of the WOTS dimensions. All dimensions are constrained to be independent.

Table 6. Correlations between dimensions and simple (raw) sub-scores

Model	Dimensions	Latent correlation estimate from MIRT model	Correlation between simple (raw) sub-scores	Disattenuated correlation between simple (raw) sub-scores
2D Content	1 & 2	0.98	0.84	0.99
4D WOTS	1 & 2	0.99	0.80	1.02
4D WOTS	1 & 3	0.98	0.75	1.00
4D WOTS	1 & 4	0.98	0.72	1.03
4D WOTS	2 & 3	0.97	0.71	0.99
4D WOTS	2 & 4	0.98	0.70	1.04
4D WOTS	3 & 4	0.97	0.66	1.02

Note: The dimensions for each model are numbered in the order they appear in the footnote for Table 5. For example, for the 4D model by WOTS, dimension 1 = Scientific goals, resources, and models, and dimension 2 = student ideas and scientific language. Disattenuated correlation estimates can be greater than 1 but should be treated as 1. Raw sub-scores refer to the number of correctly responded items for each content or WOTS category.

Table 7. Summary of statistical tests examining the relationship between CKT matter scores and PST variables.

Type	Group	Category	N	Score Mean	Score SD	Test Statistic[1]	df1	df2	p-value	Significant Post-hoc Pairwise Comparisons
Personal Background	Gender	Female	764	299.86	9.79	-1.72	62.06		0.09	N/A
		Male	57	302.58	11.64					
	Race/Ethnicity	Asian or Asian American	20	293.85	8.65	10.02	4	807	0.00	White > Asian or Asian American (p=.01)
		Black or African American	63	294.32	9.42					White > Black or African American (p=0)
		Hispanic/Latino	27	296.52	10.44					
		White	669	301.00	9.69					
		Two or More Races	33	300.79	11.04					Two or More Races > Black or African American (p=.02)
Education	Undergrad GPA	Low (2.0 to 2.99)	54	296.50	11.14	4.52	2	819	0.01	High > Low (p=.01)
		Moderate (3.0 to 3.49)	233	299.64	9.72					
		High (3.5 to 4.0)	535	300.61	9.86					
	Major/Minor	Education Focus	636	299.93	9.78	0.29	2	819	0.75	N/A
		Science Focus	31	301.03	8.78					
		Other Focus	155	300.41	10.85					
	Degree Obtained	Pursuing Bachelor's	525	299.99	9.71	4.50	2	819	0.01	Pursuing Bachelor's > Master's (p=.03)
		Bachelor's Obtained	251	300.94	10.11					Bachelor's Obtained > Master's (p=.01)
		Master's Obtained	46	296.20	10.94					
Teacher Preparation Program (TPP)	TPP Type	BA/BS Program	541	300.04	9.64	0.88	2	817	0.41	N/A
		Master's Program	215	300.58	10.74					
		Alternate-route Program	64	298.70	9.92					
	TPP K-5 Emphasis	Emphasis	614	300.57	10.17	2.65	394.79		0.01	N/A
		Limited/No Emphasis	208	298.57	9.12					
	TPP Content Emphasis	Emphasis	392	299.25	9.84	-2.24	815.15		0.03	N/A
		Limited/No Emphasis	430	300.80	10.00					
	TPP WOTS Emphasis	Emphasis	650	300.71	10.05	3.82	288.75		0.00	N/A
		Limited/No Emphasis	172	297.63	9.19					

1.If the grouping variable has two groups, a Welch’s two-sample t-test was used. If the grouping variable has more than two groups, a one-way ANOVA was used. If significant, Tukey’s post-hoc pairwise comparisons were conducted.

Table 8. Summary of CKT matter assessment perceptions survey results

Perception	Strongly Disagree/ Disagree	Strongly Agree/ Agree
Clarity of CKT Matter Assessment Items		
I found the questions to be unclear or confusing.	77%	23%
Rigor of CKT Matter Assessment Items		
I found the questions to be challenging.	22%	78%
I found it difficult to choose among answer options.	50%	50%
Coverage of CKT Matter Assessment Items in Elementary Teacher Preparation		
Answering these questions made me think about some aspect(s) of teaching this content that I had not considered previously.	7%	93%
The questions focused on material that is/was covered in my teacher preparation program.	55%	45%
Relevance of CKT Matter Assessment Items to Elementary Teaching		
The questions covered material that I teach or expect to teach in the elementary classroom.	15%	85%
I think elementary school teachers should be able to answer most of these questions correctly.	18%	82%
I think people in professions other than teaching should be able to answer most of these questions correctly.	47%	53%

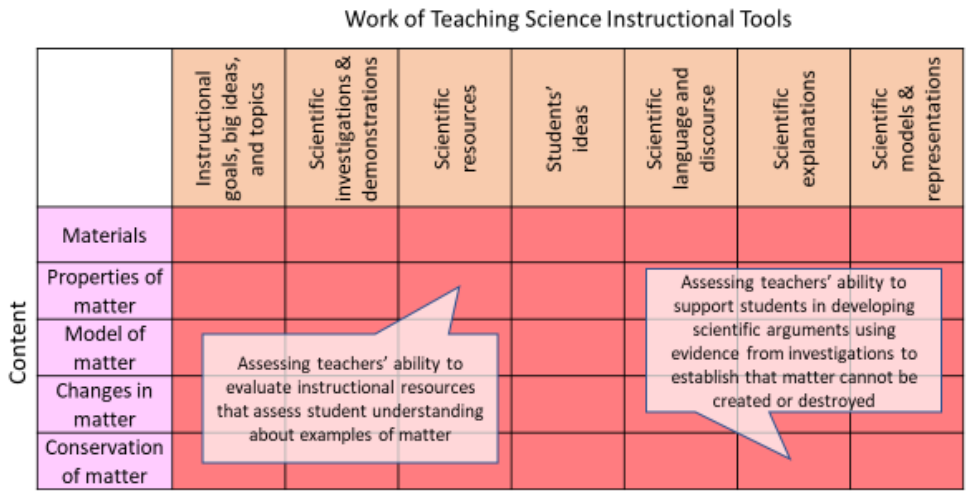
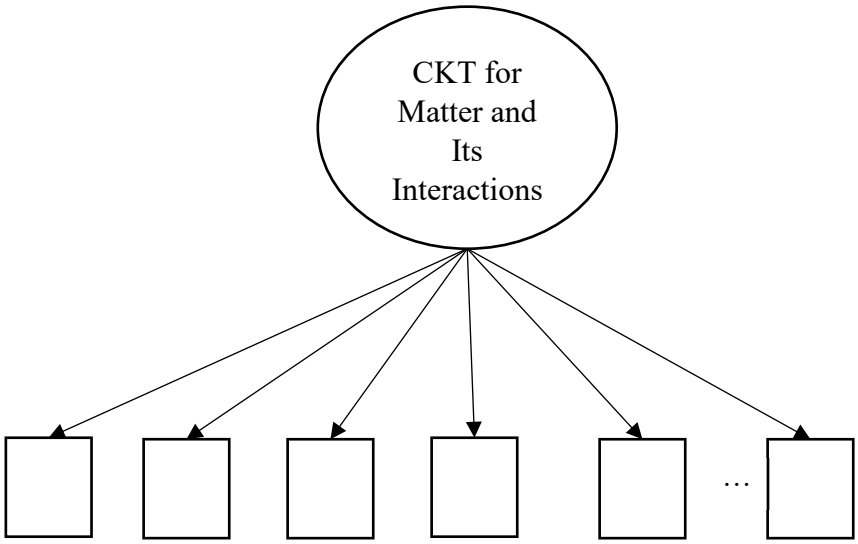


Figure 1. CKT matter item matrix that shows that each item was developed to align to both a content sub-area and a Work of Teaching Science instructional tool category

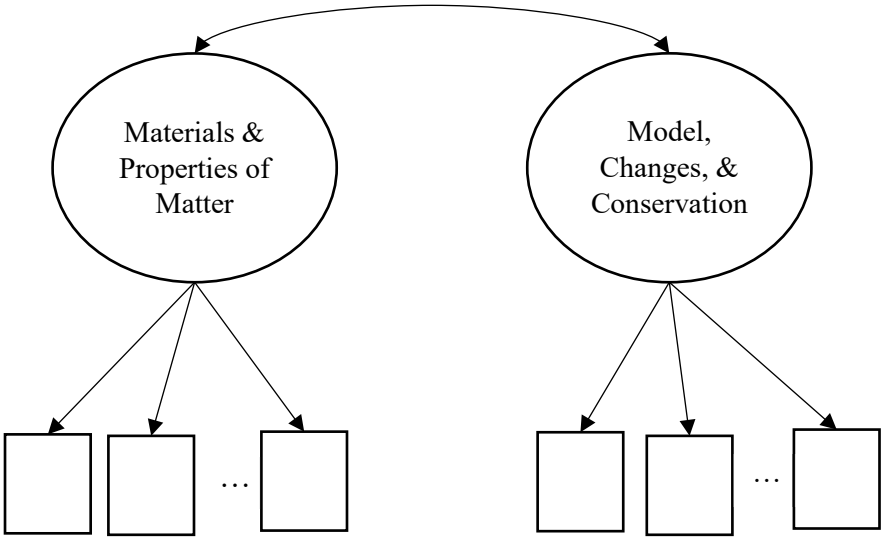
In Ms. Quintana's second-grade class, students explore the properties of different solids and liquids. Based on the exploration findings, students create definitions for solids and liquids. While completing the definition for liquids, one student makes the claim that "all substances that look like they take the shape of their containers are liquids." Ms. Quintana is planning to include a follow-up activity for students to collect more data and refine their ideas.
Which TWO of the following materials will best challenge the claim and help the student improve his or her definition?
A) Maple syrup B) Ice block C) Salt D) Milk E) Rice
Rationale for Item: *Salt and rice are solids; yet since they are granular they appear to take the shape of their container when poured into something larger than the individual units. Ms. Quintana should select both materials to help students understand that small solids also appear to take the shape of the container—as liquids do, but individual shapes of the grains can be observed and do not change.*
Item Metadata: • Work of Teaching Science Instructional Tool Category: Scientific resources • Work of Teaching Science Instructional Practice: 2.2 Choosing resources that support the selection of accurate, valid, and age-appropriate goals for science learning • Matter Content Topic: Properties of matter and materials

Figure 2. CKT matter item example with accompanying rationale for item

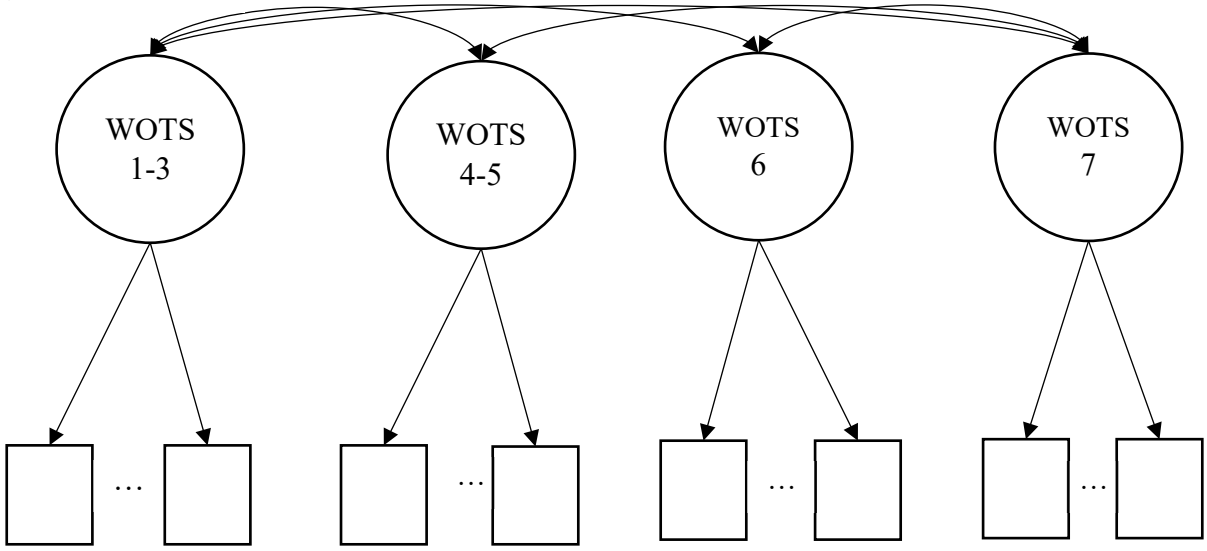
(a) Unidimensional (1D) Model



(b) 2D Content Model



(c) 4D WOTS Model



(d) 7D Model

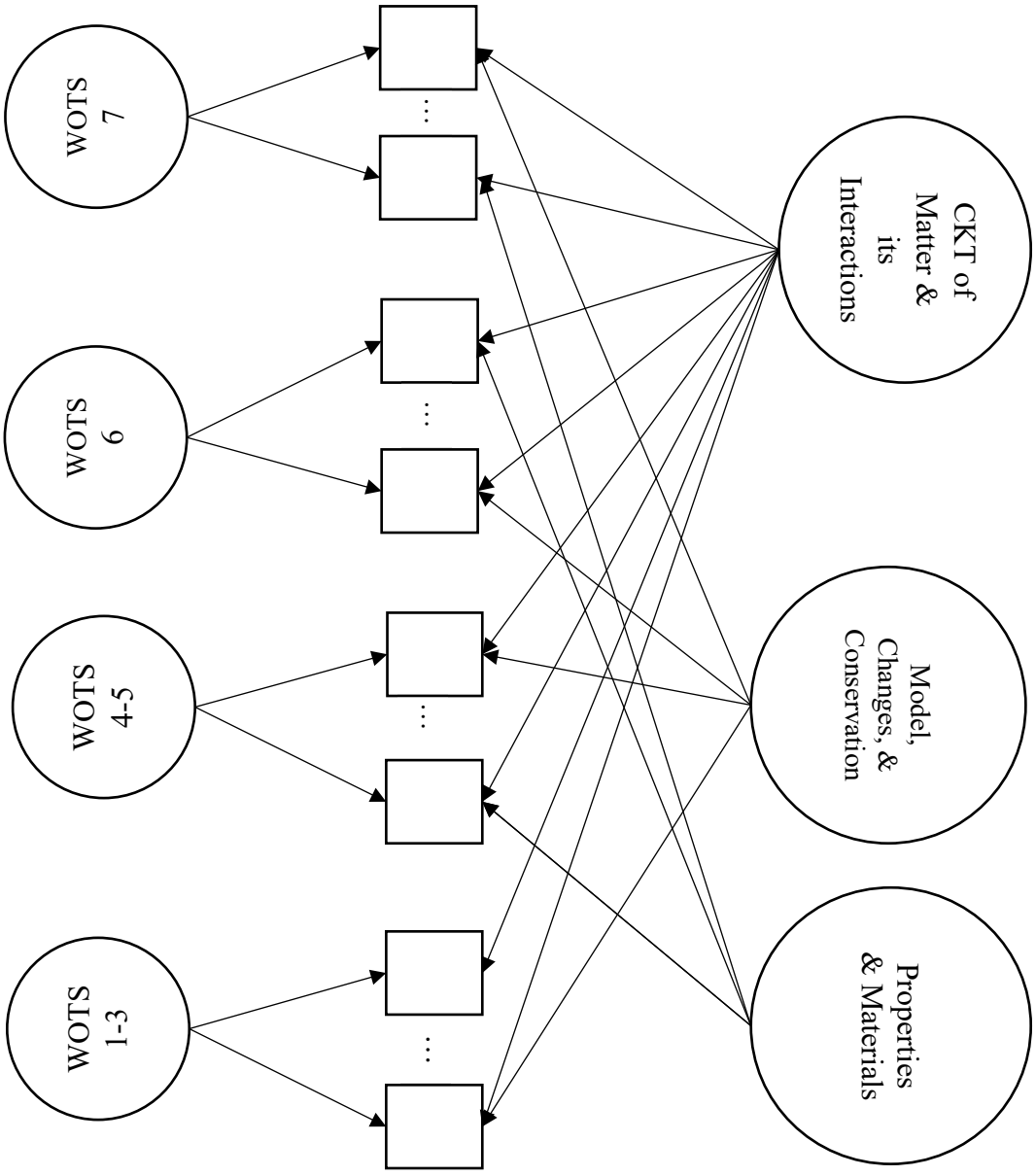


Figure 3. MIRT model diagrams

Note: Circles represent latent constructs and boxes represent items.

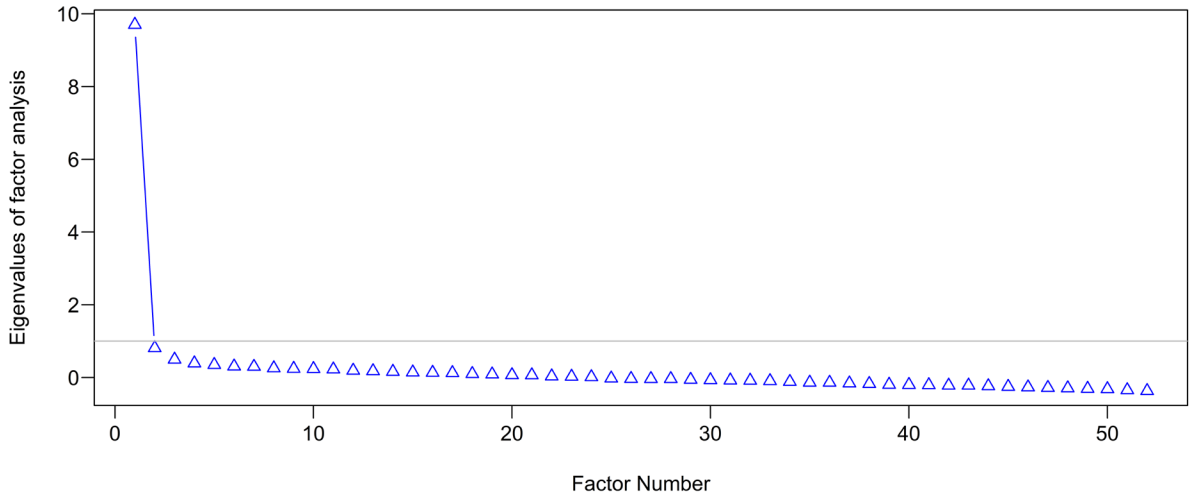


Figure 4. Scree plot of field test item data.

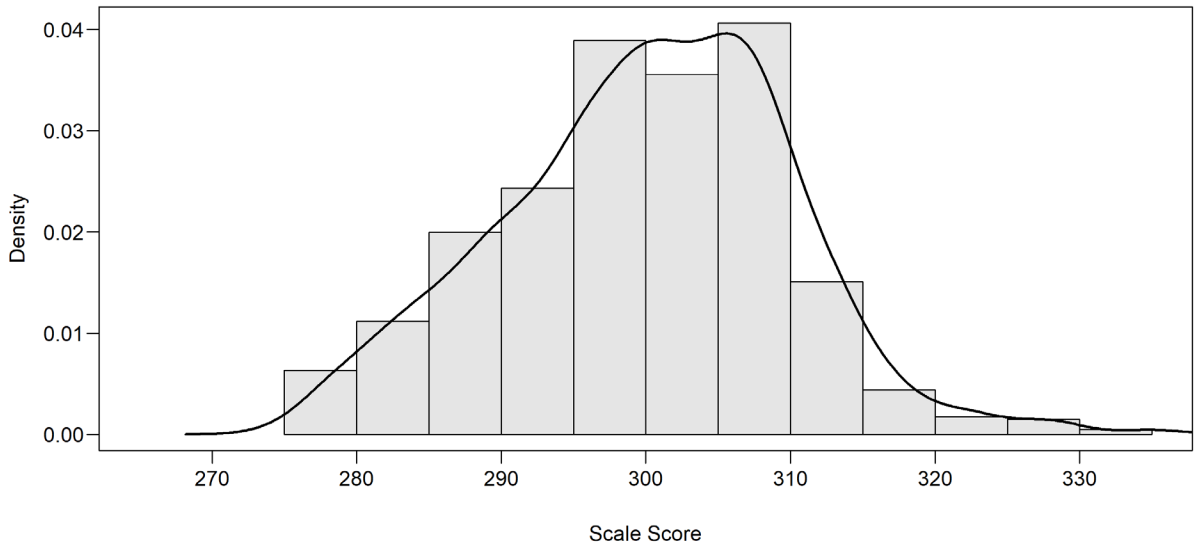


Figure 5. Distribution of CKT matter test scores.

Supplemental Online Appendices

Appendix A: Definition of the Variables Used in the Study Analysis Based on the Background Questionnaire

Table A1. Translation of background questionnaire response options to categories used to create variables in the study

Variable	Question Asked	Response Options	Frequency*	Category used in Analysis
Gender	How do you identify yourself?	Female	764	Female
		Male	57	Male
		Other _____	1	*Not used*
		Prefer not to answer	0	*Not used*
Race/ethnicity	Which of the following best describes you? (Select all that apply.)	American Indian or Alaskan Native	2	*Not used*
		Asian or Asian American	20	Asian or Asian American
		Black or African American	63	Black or African American
		Hispanic/Latino	27	Hispanic/Latino
		Native Hawaiian or other Pacific Islander	1	*Not used*
		White	669	White
		Other _____	2	*Not used*
		Prefer not to answer	5	*Not used*
		*Selected 2 or more options***	33	Two or more races
Undergrad GPA	What is your cumulative undergraduate grade point average to date (based on a system in which 4.0 = A)?	3.5 to 4.0	535	High
		3.0 to 3.49	233	Moderate
		2.5 to 2.99	47	Low
		2.0 to 2.49	7	
		1.5 to 1.99	0	*Not used*
		Below 1.5	0	*Not used*
Major/Minor focus	What is or was your undergraduate major? (Select all that apply.) AND What is or was your undergraduate	Early childhood education	636	Education focus (If selected any of these options for either major or minor question AND did not select any science major or minor.)
		Elementary education		
		Secondary education		
		Other education (e.g., administration, counseling and guidance, school psychology)		
		Health education		
		Career and technical education		

	minor? (Select all that apply.)	Other _____ [If write-in entry corresponded to education field, such as "Childhood Education (K-6)", "Middle school education"]		
		Natural sciences (e.g., biology, chemistry, Earth, physics)	31	Science focus (If selected any of these options for either the major or minor question.)
		Other _____ [If write-in entry corresponded to a science field, such as "Biomedical sciences", "Health science".]		
		Art or music	155	Other focus (If selected any of these options for either major or minor AND did not select any of the education or science options for either major or minor.)
		English or language arts		
		Foreign language		
		Mathematics or computer science		
		Social sciences (e.g., social studies, history, geography, psychology)		
		Other _____ [If write-in entry did NOT correspond to an education or science field, such as "Accounting", "Business".]		
Education Obtained	What is the highest education level you have attained?	College Freshman (first year)	0	Pursuing Bachelor's
		College Sophomore (second year)	15	
		College Junior (third year)	80	
		College Senior (fourth or final year)	430	
		Bachelor's degree	82	Bachelor's
		Bachelor's degree plus additional credits	169	
		Master's degree	30	Master's
		Master's degree plus additional credits	16	
		Doctoral degree	0	*Not used*

Teacher Preparation Program	Which of the following best describes your teacher preparation program?	Undergraduate teacher education program (BA or BS)	530	Bachelor's
		Fifth-year post-baccalaureate program (not leading to a master's degree)	10	
		Other _____ [If write-in entry indicates Bachelor's program.]	1	
		Fifth-year post-baccalaureate program (leading to a master's degree)	71	Master's
		Master's degree education program (MA, MS, EdM, MAT)	141	
		Other _____ [If write-in entry indicates Master's program.]	3	
		Alternate-route program designed to expedite the transition of non-teachers to a teaching career	64	Alternate-route program
		Other _____ [If write-in entry did not indicate one of the other program types.]	2	*Not used*

*Note: Frequency for the major/minor focus variable is collapsed over response options because PSTs could select all that apply. For instance, PSTs could select "early childhood education" and "Elementary education".

**Note: The most common selections for multiple race/ethnicities were: Hispanic/Latino & White (n=7), Asian or Asian American & White (n=4), American Indian or Alaskan Native & White (n=4), Black or African American and Hispanic/Latino (n=3), Black or African American & White (n=2), American Indian or Alaskan Native, Hispanic/Latino & White (n=2). All other selections had only 1 record.

Creating the TPP Emphasis Variables

To produce the three teacher preparation program (TPP) emphasis variables, the following procedure was used to create two categories per variable: "Emphasis" and "Limited/No Emphasis".

1. Recode the response options to numerical ratings from 0 for "No emphasis" to 3 for "Strong emphasis."

2. For each PST, take the average of their ratings over the survey questions corresponding to the variable of interest.
    a. TPP: K-5 Science Emphasis: Q17a and Q17b
    b. TPP: Content Emphasis: Q17c-Q17g
    c. TPP: WOTS Emphasis: Q18a-Q18g

    Note that we reproduce these questions as asked below.

3. Classify PSTs who had an average rating of less than 1.5 to the "Limited/No Emphasis" category and those who had an average rating greater than or equal to 1.5 to the "Emphasis" category.

17. How much emphasis does your teacher education program place on each of the following **when learning about how to teach elementary science**?

	No Emphasis	**Little Emphasis**	**Moderate Emphasis**	**Strong Emphasis**
a) Teaching science to students in grades K through 2				
b) Teaching science to students in grades 3 through 5				
c) Teaching about properties of matter and measurements This subarea focuses on descriptive properties that characterize matter, such as texture and hardness, as well as on quantitative properties, such as weight and volume. Behaviors of different types of matter can be explained by their properties. This area also addresses the description of solids, liquids, and gases based on properties of matter. Measurements of weight and volume of different materials				

are included in this characterization. (Boundary: In grades K through 5, mass and weight are not distinguished.)				
d) Teaching about materials This subarea focuses on describing materials based on their properties. Based on a property, different materials can be related to specific purposes; for example, to explain the uses of a material based on evidence or design a solution.				
e) Teaching about the model of matter This subarea focuses on developing a particle model of matter and using this model to explain some properties of solid, liquids, and gases. The model is developed from the observation and description of macroscopic matter properties (for example, the particle model can be used to explain the behaviors of gases).				
f) Teaching about changes in matter This subarea focuses on physical and chemical changes in matter. During a physical change, a substance changes form but does not change chemical composition. During a chemical change, a new substance with a different composition and different properties is formed. Matter can change when heated or cooled. Chemical changes also may occur when two or more substances are mixed. Some changes in matter are reversible, while others are irreversible. Observations of the quantitative properties (e.g., weight, volume) and qualitative properties (e.g.,				

state of matter, color, texture, odor) of the substances are used to determine what type of change occurred.				
g) Teaching about conservation of matter This subarea focuses on the conservation of matter during physical and chemical changes. Measurements of weight provide evidence that regardless of the type of change, the total amount of matter does not change. (Boundary: In grades K through 5, mass and weight are not distinguished.)				

18. How much emphasis does your teacher preparation program place on your ability to carry out each of the following **when learning about how to teach elementary science**?

	No Emphasis	Little Emphasis	Moderate Emphasis	Strong Emphasis
a) Selecting instructional goals, big ideas, and topics This part of the work of teaching science refers to selecting, organizing, and aligning scientific ideas, activities, and representations for teaching a topic. These elements should be aligned with and in support of specific learning goals, and the learning goals should be organized so that they build to bigger scientific ideas.				
b) Scientific resources This part of the work of teaching science refers to selecting, evaluating, and using instructional materials and resources that are aligned with the learning goals. Instructional materials should promote and guide student thinking and engage students with scientific phenomena. This part of the work of teaching science also considers working with and				

assessing student ideas.				
c) Scientific models and representations This part of the work of teaching science refers to selecting and evaluating models and representations in science. Models and representations should be aligned with learning goals and provide evidence of relevant scientific phenomena. It also involves engaging students in model-based practices such as creating, using, revising, and evaluating scientific models. Moreover, it includes evaluating student understanding of models and understanding of scientific phenomena through the use of models.				
d) Student ideas This part of the work of teaching science refers to eliciting and exploring students' scientific ideas. Student misconceptions are identified and explained. It also involves selecting and developing instructional approaches to address				

students' specific scientific ideas.				
e) Scientific language, discourse, vocabulary and definitions This part of the work of teaching science refers to using scientific language and developing discursive practices in science. It includes selecting appropriate scientific language and employing strategies to improve access to academic language. It also encompasses supporting students in developing their own scientific language and practices, such as argumentation.				
f) Scientific explanations This part of the work of teaching science is focused on the use of scientific explanations. It includes selecting and evaluating scientific explanations so that accurate and accessible explanations are presented to students.				
g) Scientific investigations and demonstrations This part of the work of teaching science				

refers to selecting and evaluating science investigations or demonstrations for a particular purpose that is aligned with the Next Generation Science Standards (NGSS). It also involves evaluating procedures to design investigations and collecting and analyzing data, including those in virtual investigations. This tool includes supporting students in particular elements of the inquiry process.				