*Research Article*

# Development and Validation of a Measure of Elementary Teachers' Science Content Knowledge in Two Multiyear Teacher Professional Development Intervention Projects

Jaime Lynn Maerten-Rivera,[1] Anne Corinne Huggins-Manley,[3] Karen Adamson,[2]
Okhee Lee,[4] and Lorena Llosa[4]

*[1]Psychology Department, State University of New York at Buffalo, Buffalo, New York*
*[2]Teaching and Learning, University of Miami, Miami, Florida*
*[3]Research and Evaluation Methodology, University of Florida, Gainesville, Florida*
*[4]Teaching and Learning, New York University, New York City, New York*

*Received 7 August 2014; Accepted 22 November 2014*

Abstract: Using data collected from two multiyear teacher professional development projects employing randomized control trials, this study describes the development and validation of a paper-based test of elementary teachers' science content knowledge (SCK). Evidence of construct validity is presented, including evidence on internal structural features using Rasch measurement models. Results from 183 treatment group and 176 control group teachers from Project 1 demonstrated that the SCK test had acceptable person reliability at baseline; at later time points the test was easy for the teachers and person reliability was below acceptable. Results from Project 1 informed changes made to the test for use in Project 2, including an increase in the difficulty level and the development of two equated forms. Results from the 148 treatment and 139 control teachers from Project 2 demonstrated that the test had acceptable reliability across two time points and was a better match to teachers' SCK. © 2015 Wiley Periodicals, Inc. J Res Sci Teach 52: 371–396, 2015

**Keywords:** teachers' knowledge; science content; measurement; validity evidence; Rasch measurement; professional development

For more than two decades, science educators have met with an abundance of criticism about the state of science education leading to calls for reform (American Association for the Advancement of Science [AAAS], (AAAS, 1989,1993)), National Research Council [NRC], (NRC, 1996, 2000). The current call for science education reform is aimed at college and career readiness of K-12 students through the document *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (NRC, 2012) followed by the Next Generation Science Standards (NGSS Lead States, 2013). In addition to the changes prompted by science education reform, teachers have been subjected to increased expectations for high academic achievement for *all* students, as mandated in the *No Child Left Behind Act (NCLB)* of 2001 (PL 107–110). This Act defines the responsibilities of teachers, as accountability systems place a great deal of pressure on them to implement well-articulated curriculum, instruction, and assessment systems that foster the academic achievement of the increasingly diverse student population in the nation.

*Correspondence to*: Okhee Lee; E-mail: olee@nyu.edu
DOI 10.1002/tea.21198
Published online in Wiley Online Library (wileyonlinelibrary.com).

Recognizing the potential importance of professional development (PD) in improving teachers' knowledge, and practice, and student outcomes, scholars have identified elements of effective PD including both core and structural features (Garet, Porter, Desimone, Birman, & Yoon, 2001; Penuel, Fishman, Yamaguchi, & Gallagher, 2007; Wayne, Yoon, Zhu, Cronen, & Garet, 2008). A causal model for evaluating PD programs has been proposed (Desimone, 2009), which tests a theory of teacher change (i.e., PD alters teachers' knowledge, beliefs, or practice) and a theory of instruction (i.e., the change in practice influences student achievement). To build a stronger knowledge base about links among PD, teacher knowledge and practice, and student achievement, researchers have called for more rigorous study designs (Borko, 2004; Desimone, 2009; Wayne et al., 2008), including the use of randomized experiments (Wayne et al., 2008).

One obstacle to conducting rigorous studies is the lack of adequate measures, particularly in regard to teacher learning and change, though more recent research has begun to explore measuring teachers' science content knowledge (SCK; Desimone, 2009). There are few standardized measures of teachers' SCK and classroom practice (Liu, 2009, 2012). Additionally, educational studies have been criticized for not meeting current standards of evidence in assessing validity of measures (Desimone, 2009). Validity is a judgment of the degree to which evidence and theoretical rationale support the appropriateness of interpretations and consequences on the basis of scores (Messick, 1989). Validity can be viewed as a unified concept that integrates various aspects for which evidence is collected (Messick, 1995). One aspect of construct validity is content validity that includes evidence collected on the relevance and representativeness of the measure and is generally assessed by examining the items of the measure. A second aspect of construct validity is structural validity that includes evidence collected through an examination of the internal structure of the measure and can employ different statistical methods, including Rasch modeling. A third aspect of construct validity is external validity that includes evidence of the extent to which scores on the measure are related to other measures thought to be related. A fourth aspect of construct validity is generalizability that includes evidence collected on expected performance differences over time, across groups and settings, and in response to experimental treatments (Messick, 1995).

Furthermore, educational studies have been criticized for not utilizing modern measurement models, such as Rasch models, to establish validity and reliability evidence, instead following classical test theory (CTT; Liu, 2009, 2012). CTT is considered a traditional theory of measurement in which summated test scores are used as proxies of constructs (Novick, 1966). For example, under CTT, a test measuring teachers' SCK would be scored by summing the points the teachers received on each of the test items. Under CTT, the Cronbach's alpha estimate of reliability is typically reported and reliabilities of $\widehat{\rho} = .70$ or above are considered generally acceptable for instruments with low stakes (Nunnally, 1978). In contrast, modern theories of measurement, such as Rasch (1960) and Item Response Theory (IRT; Lord & Novick, 1968), estimate a latent variable from the item responses and then the scores on the latent variable are used as a proxy of the construct being measured. If teachers' SCK is measured by a test, the test items would be indicators of a latent variable representing that construct. Rasch and IRT models differ in the number of parameters estimated and in how the latent variables are scaled, but otherwise they are very similar statistical models. As Rasch models have fewer parameters estimated, they tend to have lower sample size requirements than IRT models, which make Rasch models very useful for measurement in research projects that typically do not have thousands of participants. Rasch models provide additional ways to estimate reliability (Linacre, 1997; Schumacker & Smith, 2007) and explore internal structure validity evidence, such as the use of fit statistics (Andrich, 1988) or the use of person-item maps (Bond & Fox, 2007), as compared to CTT models.

Using data collected from two large multiyear intervention projects, Promoting Science Among English Language Learners (P-SELL), this study describes the development and validation of a measure of elementary teachers' SCK. The measure was a paper-based test, which was developed using mainly public release items from the *National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study* (TIMSS), and selective state science assessments. Evidence of construct validity (Messick, 1989, 1995) is presented, including information on (a) the decision process for the content of items and the experts involved in the development; (b) internal structural features of the measure, including factor analyses, reliability estimates, and item analyses using Rasch measurement models; (c) external structural evidence of convergent relationships with other measures; (d) ability to differentiate between groups (treatment and control) and to detect change over time; and (e) generalizability across two projects with different sample characteristics.

Both projects took place in the same state and implemented a curricular and PD intervention at fifth grade, the grade at which the state science test counted toward school accountability. The P-SELL intervention involved a stand-alone, year-long, fifth grade science curricular and professional development. It is aimed at improving science achievement of all students with a focus on English language learners (ELLs). The intervention focused on three areas: state science standards, hands-on science inquiry, and language development for all students and ELLs in particular. The intervention was comprised of (a) curriculum materials including a student book, teachers' guide, science supplies, and online supplementary materials; (b) teacher workshops during the summer and throughout the school year; and (c) school site support for curriculum implementation. The intervention components were designed to complement and reinforce one another for the improvement of teachers' SCK and instructional practices (for a full description of the intervention, see Diamond, Maerten-Rivera, Rohrer, & Lee, 2014 and Maerten-Rivera, Ahn, Lanier, Diaz, & Lee, in press). Control schools were comparable to the intervention schools in terms of student demographics, academic achievement from previous years, and school size. The teachers in the control schools did not receive the intervention and implemented science instruction as directed by their respective district using the district-adopted curriculum. The intervention had been implemented previously with success in improving teachers' use of reform-oriented practices (Lee, LeRoy, et al., 2008; Lee & Maerten-Rivera, 2012), teachers' self-reported SCK (Lee & Maerten-Rivera, 2012), and student achievement in science and writing (Lee, Maerten-Rivera, Penfield, LeRoy, & Secada, 2008; Lee, Mahotiere, Salinas, Penfield, & Maerten-Rivera, 2009; Santau, Maerten-Rivera, & Huggins, 2011).

For each project, the intervention lasted 3 years with fifth grade teachers and used a cluster randomized controlled trial (RCT) design to evaluate the impact. Project 1, P-SELL Efficacy Study, involved treatment ($n = 183$) and control ($n = 176$) group teachers from one large, urban, and culturally and linguistically diverse school district located in the southeastern US where the intervention was implemented for 3 years from the 2010–2011 school year to the 2012–2013 school year. Project 2, P-SELL Effectiveness Study, involved treatment ($n = 148$) and control ($n = 139$) group teachers from three school districts located in the same southeastern state as Project 1 with varying teacher and student characteristics for 3 years from the 2012–2013 school year to the 2014–2015 school year. In Project 2, the intervention was implemented for 1 year out of the 3-year planned intervention. The results from the 3 years of data collection are presented for Project 1, while the results from the first year of data collection are presented for Project 2 as this project is currently being implemented. Results from Project 1 informed changes made to the test for use in Project 2.

Project 1 is characterized as an efficacy study, and Project 2 an effectiveness study. According to *Common Guidelines for Education Research and Development* (US Department of Education and National Science Foundation, 2013), the distinction between an efficacy study and an effectiveness study is made in terms of two main criteria for implementation of an intervention or strategy: (a) involvement of the developer in the implementation of the intervention and (b) under "ideal" conditions or under conditions of routine practice. In Project 1, the research team facilitated professional development workshops and provided support for teachers at the school sites. In Project 2, in collaboration with the research team, the school district personnel facilitated professional development workshops and provided support at the school sites.

## Literature Review

We begin by presenting the literature on evaluating PD in order to build the argument that better measures of teachers' SCK are needed. We then introduce other measures that have been used to assess teachers' SCK.

### *Evaluating Effectiveness of Teacher PD*

Recognizing the potential importance of PD in improving teachers' knowledge and practices and student outcomes, scholars have identified elements of effective PD (Garet et al., 2001; Wayne et al., 2008) and proposed models for evaluating PD (Desimone, 2009; Garet et al., 2001; Loucks-Horsley, Hewson, Love, & Stiles, 1998). Desimone (2009) proposed a model with the following steps: (a) teachers experience the PD; (b) the PD increases teachers' knowledge and skills and/or enhances positive attitudes and beliefs; (c) teachers use their new knowledge, skills, attitudes, and/or beliefs to improve the content of their instruction, their approach to pedagogy, or both; and (d) the instructional change fosters increased student learning, and thus increased student achievement. This model tests a theory of teacher change (i.e., PD alters teachers' knowledge, beliefs, or practices) and a theory of instruction (i.e., the change in practices influences student achievement), and both are necessary to understand how PD works (Wayne et al., 2008). If only student achievement is measured and no impact is found, then without teacher outcome measures, it is not possible to determine the point where, or why, the causal model failed.

A challenge to examining the influence of teachers on students, particularly in the causal PD model, is determining how to measure the various components with teachers' knowledge being one of the most difficult to measure (see step b in the causal model proposed by Desimone, 2009). If valid measures are not developed, there is difficulty in collecting evidence to evaluate the effectiveness of a PD model. In addition, the timing of the measures is indicative of how and when change in student achievement will occur following change in teachers' knowledge and practices, which may be of interest to PD implementers and policymakers. In many studies, the validity evidence for the measures does not stand up to current standards of evidence (Desimone, 2009). Thus, development of the next generation of measurement instruments in science education needs to involve modern measurement models such as Rasch (Liu, 2009, 2012).

### *Measuring Teachers' SCK*

Teachers' knowledge of a subject is referred to as content knowledge, whereas pedagogical content knowledge refers to knowledge of how to teach a specific topic or concept in ways that enable students to understand it (Abell, 2007; Shulman, 1986). Research on the relationship between teachers' content knowledge and instructional practices has shown inconsistent results (Krauss, Baumert, & Blum, 2008; Larson & Smith 2013; Wayne & Youngs, 2003; Weaver & Dick, 2009). One difficulty in comparing and synthesizing past studies on teachers' content knowledge is that varied methodologies have been used to measure it. Some researchers used distal information such as the number of courses completed in a content area (Baumert et al., 2010), while others have used surveys to collect teachers' self-assessment of their knowledge (Lee & Maerten-Rivera, 2012; Jacobs, Martin, & Otieno, 2008). More recently, researchers have begun to develop direct tests of teachers' content knowledge. Although most of this research has been conducted in mathematics education (Baumert et al., 2010; Hill, Ball, & Schilling, 2008), research is beginning to emerge in the field of science education (Diamond et al., 2014; Heller, Daehler, Wong, Shinohara, & Miratrix, 2012; Jüttner, Boone, Park, & Neuhaus, 2013).

Measures of teachers' SCK have been developed to assess practicing teachers' SCK (Jüttner et al., 2013; Nowicki, Sullivan-Watts, Shim, Young, & Pockalny, 2013) and to examine the effect of PD on increasing teachers' SCK (Heller et al., 2012).

Jüttner et al. (2013) outline a theoretical model that guides test development and provides steps to develop and validate a measure of biology teachers' SCK and pedagogical content knowledge on the topics of neurobiology, vertebrates, plants, and cytology. The instrument used multiple-choice and constructed response items of different cognitive complexity and was administered to a sample of 158 fifth to 12th grade German biology teachers. The Rasch partial credit model was used for the analysis of the test data. The 20 items measuring SCK had an item reliability of $\widehat{\rho} = .98$ and a person reliability of $\widehat{\rho} = .76$. The 24 items measuring pedagogical content knowledge had an item reliability of $\widehat{\rho} = .97$ and a person reliability of $\widehat{\rho} = .58$. A statistically significant low correlation ($r = .22$, $p = .006$) was found between SCK and pedagogical content knowledge. This finding is different from other studies conducted in math education that have found high correlations between the two knowledge areas (Hill, Schilling, & Loewenberg Ball, 2004; Krauss et al., 2008). The items on the SCK portion were well-matched to the teachers' estimated abilities, whereas the items on the pedagogical content knowledge portion were too difficult for the typical respondent. The measure was administered to the teachers at only one time and was not used to measure the effects of PD.

Nowicki et al. (2013) administered a 52-item science test with five subscales (i.e., life science, earth science, physical science, electricity and magnetism, and nature of science) to 27 preservice and 27 inservice grade 1 through 5 elementary teachers. The test was originally designed by Horizon Research, Inc., for grade 4 through 6 elementary students. The use of the test, which was originally designed for elementary students, with elementary teachers allowed a comparison between the student and teacher groups. The reliability estimates for the subscales of the test were reported for the pilot test data conducted with 3,000 elementary students, and $\widehat{\rho}$ ranged from .63 to .67, which is below the threshold for reliability of $\widehat{\rho} = .70$, and it is unclear which estimate of reliability was used. Additionally, information regarding the reliability of the measure for the study sample of teachers was not provided. The test was administered at one time to measure the teachers' SCK in relation to what their students were expected to know and to examine the teachers' SCK in relation to classroom instruction. For the preservice teachers, the average percentage of correct responses on the test was 79% and for the inservice teachers, it was 82%. Furthermore, the researchers observed the accuracy of science content in classroom instruction for both the preservice and inservice teachers and found that teachers' SCK test scores, science courses taken in college, comfort level with science content, and years of teaching experience (for inservice teachers only) were not significant predictors of the accuracy of science content in classroom instruction. The predictors that were significant in the model were access to kit-based resources, grade level, and a preference for teaching science.

Heller et al. (2012) examined the effects of PD using a paper-based test of teachers' SCK. The researchers conducted a randomized experiment utilizing a pre and post administration to compare three related, but systematically different, teacher science PD interventions and a control group. The study administered a teachers' SCK test on electric circuits consisting of 20 selected response items, nine yes/no items, and four constructed response items to 270 fourth grade teachers in six states. The Cronbach's alpha for the test scores across all teachers were reported as $\widehat{\rho} = .90$, but the reliability results were not presented for each administration (i.e., pre, post).

Results indicated that all three intervention groups had gains well beyond the control group with no significant differences among the three intervention groups. A 1 year follow-up indicated that the gains of the treatment groups were maintained. It is important to note that the test was on a specific content topic rather than the science content covered throughout the year.

While the studies above are notable for examining teachers' SCK, there are several limitations. First, one study is unclear about which estimate of reliability was used and reports the reliability for samples other than that used in the study (Nowicki et al., 2013). Another study reports only the Cronbach's alpha estimate of reliability under the CTT framework (Heller et al., 2012), which has been criticized as it represents only one possible source of inconsistency in scores. Furthermore, the Cronbach's alpha estimate of reliability is frequently misused since it is often reported without testing the strict assumptions that are required to be a good estimate of reliability (Raykov, 1997, 2001). Second, studies address reliability using the Cronbach's alpha estimate without providing evidence of validity. Evidence of construct validity has been discussed in the literature (Messick, 1989, 1995), yet the studies focus on outcomes without addressing the validity of the measures. An exception is Jüttner et al. (2013) who outline a theoretical model for test development and use a Rasch partial credit model for the analysis. Third, many of the measures were not developed to be used to evaluate PD that took place over multiple years, and some measures address only one specific science topic. Yet, standardized measures intended to evaluate longitudinal PD efforts are needed in order to answer the call for randomized experiments incorporating standardized measurements in longitudinal designs (NRC, 2002).

Our research, reported here, addresses the limitations found in the literature, as described above, regarding the development of teachers' SCK measures. First, we provide information on how our test was designed, how items were selected, and how our measure using a Rasch analysis was examined. The results of our research may be useful for other researchers looking for existing measures or following the steps in the development and validation of new measures. Second, the development and analysis of our measure addresses multiple pieces of validity evidence, including content validity, internal structure validity, and external structure validity. Third, our measure was comprehensive in that it was used over time in a multiyear PD to detect change on science content covered at the elementary level with a focus on the fifth grade content, not just one specific science topic.

Overall, our research contributes to the literature by taking steps in developing an instrument that has the potential to be widely used to measure elementary teachers' SCK and to compare the effectiveness of various PD interventions at increasing teachers' SCK. Ball, Hill, and Bass, 2005 wrote: "Developing rigorous measures, and having a significant number of professional developers use them, will help to build generalizable knowledge about teachers' learning. . . .[M] any studies are required in order to make sense of how differences in program content might affect teachers, teaching, and student achievement" (p. 45).

## Project 1: P-SELL Efficacy Study

### *Sample*

Project 1 took place in one large urban school district in the southeastern US with diverse student and teacher populations. During the first year (2010–2011) of the project, the K-12 student demographic composition was 24% Black, 65% Hispanic, 9% White non-Hispanic, and 2% Other; 72% received free or reduced price lunch (FRL); and 19% were designated as limited English language proficient (LEP, the federal term) or ELLs.

A RCT was conducted. At the time when schools were randomly selected to participate, there were 238 elementary schools in the district. Initially, 23 schools were removed from the pool due to participation in alternate district interventions, and nine schools were removed because they had participated in a previous version of our study. This resulted in a final pool of 206 eligible schools. From this pool, 64 schools were randomly selected to participate in the study. The 64 schools were then randomly assigned, 32 to the treatment group and 32 to the control group. All fifth grade teachers in the selected schools participated in the study.

### *Data Collection*

All teachers were asked to complete a brief survey of their background information including gender, ethnicity, native language(s) spoken, educational background (i.e., highest degree), science background (i.e., number of science methods and science content courses taken in college), and teaching experience (i.e., number of years teaching). Supplementary Table S1 presents sample descriptive statistics in terms of the demographic information and professional backgrounds as reported by the Project 1 teachers. The SCK test was administered to teachers prior to the beginning of the intervention and at the end of each school year. At each data collection, most teachers completed the test with less than 6% not completing due to a variety of reasons (e.g., refusal, teacher absent or on leave).

Time was coded as baseline (T0) when a teacher completed the test prior to beginning the intervention, Time 1 (T1) at the end of the first year, Time 2 (T2) at the end of the second year, and Time 3 (T3) at the end of the third year. A teacher could start participation during any time of the 3-year intervention. If a teacher started teaching at a school during Year 3 of the intervention, when the teacher completed the test at the beginning of the year, the time would be coded as T0, and at the end of the year, the time would be coded as T1 as it was his/her first year of participating in the intervention. Table 1 displays the number of teachers who completed the test at each time point for Project 1. The maximum number of time points that a teacher could have for Project 1 is four (Group 1), in which case the teacher completed a baseline test, participated in 3 years of the intervention, and completed the test at the end of each year.

Classroom observations were conducted with one teacher randomly selected from each school three times throughout the school year (for more information on the classroom

Table 1
*Project 1 and project 2 patterns of teacher test data collection*

| | | Treatment (*n* = 183) | | | | | Control (*n* = 176) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Project 1 (*N* =  359) | *n* | T0 | T1 | T2 | T3 | *n* | T0 | T1 | T2 | T3 |
| Group 1 | 51 | X | X | X | X | 38 | X | X | X | X |
| Group 2 | 32 | X | X | X | | 33 | X | X | X | |
| Group 3 | 49 | X | X | | | 66 | X | X | | |
| Group 4 | 10 | X | | | | 12 | X | | | |
| Group 5 | 0 | | X | X | X | 3 | | X | X | X |
| Group 6 | 14 | | X | X | | 5 | | X | X | |
| Group 7 | 26 | | X | | | 19 | | X | | |
| Group 8 | 1 | | | X | | 0 | | | X | |
| Total | 183 | 142 | 172 | 97 | 51 | 176 | 149 | 164 | 79 | 41 |

| | | Treatment (*n* = 148) | | | Control (*n* = 139) | |
|---|---|---|---|---|---|---|
| Project 2 (*N* = 287) | *n* | T0 | T1 | *n* | T0 | T1 |
| Group 1 | 126 | X | X | 134 | X | X |
| Group 2 | 3 | X | | 1 | X | |
| Group 3 | 19 | | X | 4 | | X |
| Total | 148 | 129 | 145 | 139 | 135 | 138 |

observations and scales, see Lee & Maerten-Rivera, 2012; for more information on the SCK scale, see Diamond, Maerten-Rivera, Rohrer, & Lee, 2013). Three members of the research team were trained to conduct classroom observations. The observers were trained together during a one week, four hour per day, training program led by experienced former observers. The training objective was to ensure consistent scoring by learning how the observation scales were meant to be interpreted, practicing the coding of observation notes from our previous research, and practicing observations using videotaped lessons. Throughout the period of classroom observations, 10% of the observations were conducted with pairs of observers and scores were compared to ensure observer agreements. Scores that were within one point were considered agreement. Based on this criterion, there was 95% agreement of scores between observers, and any discrepancies were discussed to make the scoring more similar in the future.

For each observation, SCK was rated using a scale of 1–5 with higher scores representing more knowledge demonstrated. If the randomly selected teacher remained in the study for multiple years, that teacher was observed throughout. If a teacher left, a new teacher from the school was randomly selected for observations. The time for the observation is coded as Year 1 if it was the teacher's first year in the study, Year 2 if it was the teacher's second year, and Year 3 if it was the teacher's third year. This was similar to the coding of time points for the test, except there was no baseline collected for classroom observations.

### *Instrument Construction and Content Validity*

The SCK test for Project 1 was aligned with the fifth grade science content standards of the state in which the research took place, at the time of developing the measure. The topics included nature of matter, energy, force and motion, processes that shape the earth, earth and space, processes of life, living things interacting with the environment, and nature of science.

Two researchers took the lead in searching for test items that mapped onto these topics from two main sources: (a) publicly released items at fourth and eighth grades in NAEP 2000 and 2005 (https://nces.ed.gov/nationsreportcard/itmrlsx/landing.aspx) and (b) publicly released items at fourth and eighth grades in TIMSS 1995, 1999, and 2003 (https://nces.ed.gov/timss/educators.asp). In addition, previously developed project items were included in the pool of possible items; these items had been developed and used on a student test in a previous version of the intervention. NAEP reports the difficulty level of each item as easy, medium, or hard, along with the percentage of national student respondents answering each item correctly. TIMSS reports the cognitive domain as factual knowledge, conceptual knowledge, or reasoning and analysis, along with the percentage of national and international student respondents answering each item correctly. The two researchers reviewed items along with the information provided about each item and ranked each item as easy, medium, or hard difficulty for the fifth grade student level. Most items were of medium or hard difficulty, with fewer items of easy difficulty.

A pilot test was developed with 34 items, including five short and extended response items. The test was piloted with a sample of 311 respondents, which included 144 K-6 teachers at five elementary schools, 137 middle school students, and 30 college students (mainly majoring in elementary education). These samples were selected because we had access to these respondents and because we expected a range of SCK across these groups. The psychometric properties of the test and the items were examined for the pilot test, and a comparison of the difficulty of the items with the ability of the sample was conducted. The Cronbach's alpha for the scores on the pilot test was $\widehat{\rho} = .84$.

A panel of researchers, district personnel, and classroom teachers reviewed the pilot test information and chose 30 items that mapped onto the science topics assessed at fifth grade for the final version of the test, which included 24 multiple-choice and six constructed response items.

Table 2
*Project 1 and project 2 test specifications*

| Strand | Easy difficulty | Medium difficulty | Hard difficulty | Total |
|---|---|---|---|---|
| Project 1 Test | | | | |
| Nature of matter | $8^{t*}$, $12^{D*}$ | $7^{t}$, $15^{n}$ | $18^{t}$ | 5 |
| Energy | $21^{N}$ | $26^{t}$, $29^{T}$ | $6^{n*}$, $10^{N}$ | 5 |
| Force and motion | | $13^{D}$, $25^{D}$, $24^{D}$, $9^{T}$ | $2^{T}$, $28^{N}$ | 6 |
| Processes that shape the Earth | | $17^{N}$ | $16^{T*}$ | 2 |
| Earth and space | | $11^{t}$, $23^{T}$ | $30^{T}$ | 3 |
| Processes of life | | $4^{t*}$ | | 1 |
| Living things interacting with environment | (21), $20^{n}$ | $14^{N}$ | $22^{N}$, $1^{D}$ | 4 |
| Nature of science | $(8^{*})$ | $3^{n}$, $5^{t}$, (17), (25) $27^{D*}$ | $(6^{*})$, $19^{n}$ | 4 |
| Total | 4 | 16 | 10 | 30 |
| Project 2 Form 1 | | | | |
| Earth and space | $18^{F}$, $14^{X}$ | $6^{T}$, $10^{N}$, $20^{TL}$, $24^{C}$, $27^{T}$ | $17^{NL*}$,$28^{P}$ | 9 |
| Life science | $31^{C}$, $32^{R}$ | $3^{T}$, $8^{X}$, $11^{T}$, $22^{T}$, $25^{TL}$ | $7^{Y}$, $13^{NL}$ | 9 |
| Physical science | $1^{T}$, $5^{X}$ | $4^{NL*}$, $15^{TL}$, $16^{M}$, $19^{X}$, $23^{M}$, $30^{NL}$ | $26^{NL}$ | 9 |
| Nature of science | $2^{T}$, $21^{NP}$ | $12^{P}$, $33^{PL}$, $29^{T}$ | $9^{n*}$ | 6 |
| Total | 8 | 19 | 6 | 33 |
| Project 2 Form 2 | | | | |
| Earth and space | $21^{T}$, $24^{S}$ | $10^{N}$, $5^{N}$, $16^{P}$, $20^{FL}$, $29^{X}$ | $6^{T}$, $17^{NL*}$ | 9 |
| Life science | $3^{F}$, $14^{T}$ | $11^{T}$, $12^{P}$, $25^{TL}$, $28^{F}$, $31^{T}$ | $7^{T}$, $13^{NL}$ | 9 |
| Physical science | $2^{M}$, $27^{F}$ | $4^{NL*}$, $15^{TL}$, $18^{T}$, $19^{T}$, $30^{NL}$, $32^{X}$ | $26^{NL}$ | 9 |
| Nature of science | $1^{T}$, $23^{n}$ | $8^{T}$, $22^{T}$, $33^{PL}$ | $9^{n*}$ | 6 |
| Total | 8 | 19 | 6 | 33 |

()Nature of science is embedded in these items; $^{*}$Short response; $^{C}$California fifth grade; $^{F}$Florida item eighth grade; $^{M}$Maine item eighth grade; $^{n}$NAEP item fourth grade; $^{N}$NAEP item eighth grade; $^{P}$NAEP item 12th grade; $^{R}$North Carolina fifth grade; $^{S}$Massachusetts eighth grade; $^{t}$TIMSS item third/fourth grade; $^{T}$TIMSS item seventh/eighth grade; $^{X}$Texas item 10th grade; $^{Y}$New York eighth grade; $^{L}$Linking item.

Only 30 items were selected because we wanted the test to take about 30 minutes to complete. Table 2 displays the test specifications, along with the item difficulty level assigned by the two project researchers who reviewed the items and source information.

The test with the correct answer for multiple-choice items denoted, and the scoring rubric for constructed response items are available as supplementary material accompanying the online article. The test was worth a total of 38 points. Each multiple-choice item was worth 1 point, one constructed response item was worth 1 point, two were worth 2 points each, and three were worth 3 points each (see the rubric for details). An entire team of raters participated in a one-hour training session prior to scoring. Then, a subgroup of 2–4 raters received training on each item and scored a subset of 10% of the tests. The agreement between raters was greater than 90%, and disagreements were resolved by group consensus, if needed.

### *Data Analysis*

This section outlines the data analysis used to examine the final version of the teachers' SCK test used in Project 1. First, the unidimensionality of the measure was examined. Second, the Rasch analysis was conducted. There were 291 respondents at T0, 336 at T1, 176 at T2, and 92 at T3 (see Table 1). The unidimensionality and Rasch analysis were conducted for each time point, and all respondents at that time point were included in the analyses. Finally, convergent validity was assessed by examining relationships between the test and other measures. All respondents with valid measures were included in these analyses.

*Unidimensionality.* The test was assessed for unidimensionality by conducting a confirmatory factor analysis (CFA) using the Mplus software (Muthén & Muthén, 2012) with the variance of the latent ability variable set to $\sigma^2 = 1$ for model identification purposes. To account for discrete or ordinal item scoring, weighted least squares estimation with adjusted means and variances was used. Model fit was evaluated using the root mean square error of approximation (RMSEA) as it is considered appropriate for this type of estimation, whereas other fit indices (e.g., CFI, TLI) are not appropriate (Muthén & Muthén, 2012) with the criteria of good model fit as RMSEA $<.06$. The 90% confidence interval for RMSEA was presented with the criteria for good model fit being a lower bound no higher than RMSEA $= .05$ and an upper bound no higher than RMSEA $= .08$ (Hu & Bentler, 1999). In addition, the standardized factor loading ($\lambda$) of each item onto the latent trait was assessed for magnitude and statistical significance; $\lambda \geq .30$ is considered an acceptable magnitude (Crocker & Algina, 1986), and for this study, $p < .05$ was considered statistically significant.

*Rasch Model.* After conducting the dimensionality analysis, the test was calibrated using the unidimensional Rasch model for dichotomous (i.e., multiple choice) items (Rasch, 1960) and the partial credit model (PCM) for polytomous (i.e., constructed response) items (Masters, 1982). The calibrations were completed separately for each time point. The probability of correctly responding to an item (i.e., obtaining a score of 1) in the Rasch model is defined as

$$P_{i(1)} = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)} \tag{1}$$

where $i$ represents an item, $b$ is the difficulty parameter of item $i$, and $\theta$ is the unidimensional latent trait measured by the set of items. The PCM defines the probability of selecting a particular category response as

$$P_{i(g)} = \frac{e^{\sum_{X_i=0}^{g} (\theta - b_{ix})}}{\sum_{X_i=0}^{m} e^{\sum_{X_i=0}^{x} (\theta - b_{ix})}} \tag{2}$$

where response options are defined as $X = 0,\ 1,\ \ldots,\ g,\ \ldots m,\ \ b_{ix}$ represents the difficulty associated with making the step from $x$–1 to $x$ on item $i$, and not making a step on item $i$ (i.e., $g = 0$) has its term set to 0, or

$$\sum_{\mathrm{X}_i=0}^{g=0} (\theta - b_{ix}) = 0 \tag{3}$$

For calculations, the Rasch software Winsteps (Linacre, 2012) was used, which uses joint maximum likelihood estimation techniques. To maintain a constant scale across the time points, the item difficulty parameter estimates ($b_i$) were anchored to the baseline parameter estimates during calibrations of T1, T2, and T3.

Both item reliability and person reliability estimates from the Rasch model were examined. Reliability estimates are reported on a 0–1 scale and, in general, reliabilities of $\widehat{\rho} = .70$ or above are considered acceptable for instruments with low stakes (Nunnally, 1978), such as those intended to answer research and evaluation questions. Item reliability is an estimate of the ability to confirm the item difficulty hierarchy of the measure and is dependent on item difficulty variance (i.e., a wide range of difficulty) and person sample size (i.e., larger samples lead to higher reliability). Low item reliability is indicative that the sample is not large enough to precisely locate the items on the latent variable. Person reliability in the Rasch model is comparable to the traditional "test" reliability but does not suffer from the same psychometric assumption problems as seen with Cronbach's alpha; low person reliability implies that the measure may not be sensitive enough to distinguish between high and low performers. Low person reliabilities may result from several sources: (a) a narrow ability range in the sample, (b) the length of the test (i.e., a longer test leads to increase in person reliability), and (c) items that are mismatched to the ability level of the population being measured.

The infit mean-square (infit) and outfit mean-square (outfit) were examined to assess the level of productive measurement provided by each of the test items. The infit is sensitive to unexpected patterns of observations by persons on items that are roughly targeted toward their ability level (i.e., persons with abilities more closely matching the item difficulty are not performing as expected). The outfit is sensitive to unexpected observations by persons on items that are relatively easy or hard for them (i.e., persons with abilities quite different from the item difficulty are not performing as expected). Infit and outfit estimates below 0.5 and above 1.5 indicate problematic item fit, yet high estimates are a greater threat to validity than low estimates. Interpretation of parameter-level mean-square fit statistics is >2.0 distorts or degrades the measurement system; 1.5–2.0 is unproductive for construction of measurement, but not degrading; 0.5–1.4 is productive for measurement; and <0.5 is less productive for measurement, but not degrading (Linacre, 2014).

Next, construct validity of the test was evaluated through the use of a person-item map for each time point. The person-item map plots the relative difficulty of the test items (on the right-hand side) to the ability of the persons measured (on the left-hand side). Persons at the base of the map are those with the lowest ability with regard to SCK, while those at the top are those with the highest ability. Similarly, items at the base of the map are easier for respondents, while items at the top are more difficult for respondents. In addition, the mean ability on the test and the mean item difficulty are noted by an "M" on each side of the map. An "S" is used to denote one standard deviation from the mean, and "T" denotes two standard deviations from the mean. This provides information regarding the difficulty of the test in relation to the ability of the respondents.

*Convergent Validity.* As part of the Rasch analysis, each person is given a person ability estimate based on her/his responses to the items. This person measure is reported in standard deviation units with 0 representing a teacher with SCK equivalent to the sample average, +1 representing a teacher one standard deviation above the average of the sample, and so forth. Evidence of validity is collected by examining the associations between the ability estimates and other constructs to which ability (or SCK) are hypothesized to relate (convergent validity).

The person measure ability estimates were used to examine relationships with other variables. First, Pearson correlation coefficients were used to examine relationships with other continuous variables, including the ability estimates from other time points, number of college science methods courses taken, and number of college science content courses taken. Second, point biserial correlations were used to examine relationships with the dichotomous variable of group (control group coded as 0 and treatment group coded as 1). In addition to the correlation coefficient, the mean and standard deviation for each time point are presented for comparison.

Third, evidence of convergent validity was investigated by relating SCK ability estimates to SCK classroom observation scores. Generalizability theory was utilized to analyze observation data ( Brennan, 2001; Cronbach, Nageswari, & Gleser, 1963), as the theory provides methods for determining if it is psychometrically appropriate to combine multiple measurements of the same construct (i.e., teachers' SCK) across facets (i.e., observations over time within a year). Crocker and Algina (1986) provide multiple methods for producing this *G*-coefficient, which can be interpreted as the percentage of variance in the average of observation scores that is due to teacher differences. Similar to reliability coefficients, it is desirable to have a *G*-coefficient $\geq$.70. For all 3 years, the *G*-coefficient for teachers' SCK through observations was above .70. For each observed teacher, the scores from the three observations were averaged to calculate an SCK observation score. For the subsample of observed teachers, the observation score was correlated with the test ability estimate.

### *Results*

*Dimensionality Results.* The dimensionality of the SCK test was assessed for all four time points. The model fit the data well at T0 (RMSEA = .036, $CI_{90\%}$ = .028, .043), T1 (RMSEA = .032, $CI_{90\%}$ = .024, .039), and T2 (RMSEA = .050, $CI_{90\%}$ = .041, .059), indicating that the items were all measuring a single construct. However, the model did not fit the data well at T3 (RMSEA = .092, $CI_{90\%}$ = .081, .103). At T3, the sample size ($n = 92$) was considerably smaller than at the other time points due to teacher attrition over time, which may have contributed to the lack of model fit. Most of the factor loadings for T3 were similar to the other time points. However, all of the constructed response items had low, and sometimes negative, factor loadings, which may have been due to the small sample size.

*Rasch Modeling Results.* The person and item reliability estimates for the teachers' SCK test at each of the time points, along with the infit and outfit for each item, is displayed in Table 3. The item reliability estimates were acceptable at T0 ($\widehat{\rho}$ = .97), T1 ($\widehat{\rho}$ = .96), T2 ($\widehat{\rho}$ = .90), and T3 ($\widehat{\rho}$ = .80). These item reliability estimates suggest that the item difficulty hierarchy of the test was confirmed in the sample at each time. The person reliability estimate was acceptable at T0 ($\widehat{\rho}$ = .72), but was low at T1 ($\widehat{\rho}$ = .65), and continued to drop at T2 ($\widehat{\rho}$ = .54) and T3 ($\widehat{\rho}$ = .58). This finding suggests that the test was not sensitive enough to distinguish between teachers with different levels of SCK, particularly at T2 and T3.

In Table 3, infit and outfit statistics below 0.5 and above 1.5 are in bold to indicate problematic item fit. The model fit the data well for each item at the T0 administration with the exception of one low outfit statistic for item 20. At T1 and T2, a couple of items had infit or outfit statistics outside of the range, yet only item 9 at T1 was above 2.0 (outfit = 2.18). At T3, there were again some items outside of the range. Of great concern is item 8 that had poor infit (9.9) and poor outfit (8.07), which indicates that persons with abilities closely matching the difficulty of this item did not answer as expected (poor infit), and persons with abilities different from the difficulty of the item did not answer as expected (poor outfit).

Figure 1 displays the person-item map for each of the times. The item difficulty parameter estimates ($b_i$) were anchored when calibrating T1, T2, and T3. Therefore, the scale is constant across the maps. At T0, the difficulty of the items tended to be lower than the ability of the teachers. The map for T0 indicates that many of the items captured the most information on teachers with ability levels 1 or 2 SDs below the average of the item difficulties. However, most teachers had a higher ability estimate, usually between 0 and 3 SDs above the item difficulty average, though a

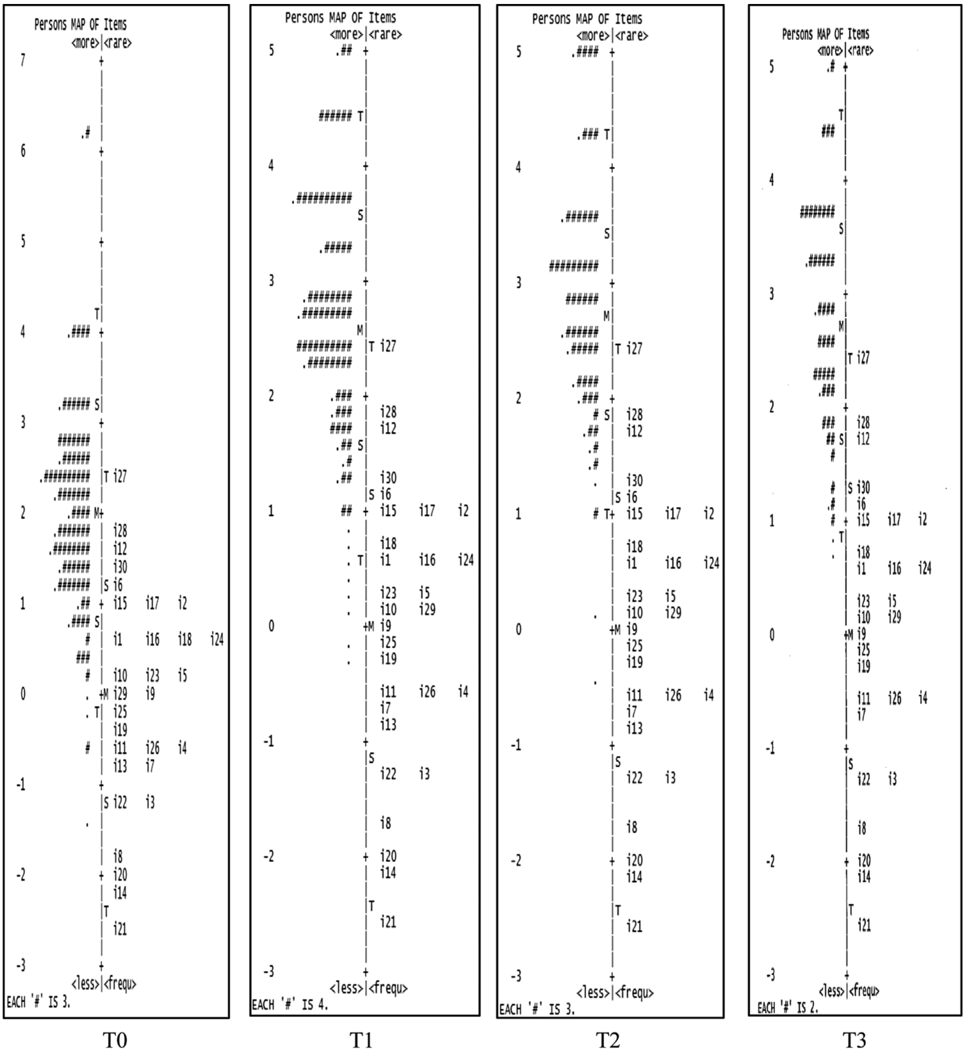
Table 3
*Project 1 Rasch modeling results*

| | | T0 | | T1 | | T2 | | T3 | |
|---|---|---|---|---|---|---|---|---|---|
| Person reliability | | .72 | | .65 | | .54 | | .58 | |
| Item reliability | | .97 | | .96 | | .90 | | .80 | |
| | Item | Infit | Outfit | Infit | Outfit | Infit | Outfit | Infit | Outfit |
| | 1 | 1.03 | 0.96 | 1.25 | 1.18 | 1.24 | 1.08 | 1.22 | 1.11 |
| | 2 | 1.00 | 1.03 | 0.93 | 0.85 | 0.85 | 0.85 | 0.95 | 1.04 |
| | 3 | 1.02 | 0.82 | 1.27 | 0.98 | 1.01 | **0.36** | 1.13 | **0.45** |
| | 4 | 1.05 | 1.23 | **1.79** | **1.57** | 1.10 | 1.18 | 1.22 | 0.87 |
| | 5 | 1.08 | 1.25 | 1.04 | 1.12 | 1.34 | 1.38 | 1.32 | 2.40 |
| | 6 | 1.00 | 0.98 | 1.10 | 1.14 | 1.17 | 1.26 | 1.03 | 0.96 |
| | 7 | 0.95 | 0.68 | 0.61 | **0.43** | 0.59 | **0.41** | **0.35** | **0.48** |
| | 8 | 0.92 | 0.73 | **1.76** | **1.92** | **1.71** | 0.83 | **9.9** | **8.07** |
| | 9 | 1.12 | 1.30 | **1.97** | **2.18** | **1.53** | 1.31 | **2.05** | **2.36** |
| | 10 | 1.10 | 1.08 | 1.04 | 1.13 | 0.76 | 0.74 | 0.76 | 0.78 |
| | 11 | 1.05 | 0.99 | 0.98 | 0.87 | 0.72 | 0.80 | 0.60 | **0.29** |
| | 12 | 1.25 | 1.25 | 1.24 | 1.26 | **1.73** | **1.89** | **1.54** | **1.54** |
| | 13 | 1.00 | 0.75 | 0.67 | **0.36** | **0.26** | **0.24** | 1.00 | 1.00 |
| Item fit | 14 | 1.02 | 0.61 | **1.75** | 0.97 | **1.68** | 1.46 | **1.93** | 1.08 |
| | 15 | 1.02 | 1.00 | 1.04 | 0.94 | 0.95 | 0.79 | 0.70 | 0.50 |
| | 16 | 0.95 | 0.85 | 1.26 | 1.32 | 1.48 | 1.33 | 1.40 | 1.43 |
| | 17 | 0.95 | 0.92 | 1.15 | 1.18 | 0.77 | 0.68 | 1.01 | 0.82 |
| | 18 | 0.95 | 0.87 | 1.10 | 1.02 | 1.16 | 1.20 | 1.06 | 0.74 |
| | 19 | 0.85 | 0.59 | 0.88 | 0.58 | 1.01 | 0.93 | 0.59 | **0.48** |
| | 20 | 0.86 | **0.32** | 0.71 | **0.25** | **0.44** | **0.07** | 0.89 | 0.89 |
| | 21 | 0.89 | 0.68 | 0.97 | 0.73 | 0.73 | **0.10** | 1.39 | 0.65 |
| | 22 | 0.90 | 0.56 | 1.16 | 0.61 | 0.52 | **0.12** | **0.44** | **0.24** |
| | 23 | 0.88 | 0.69 | 0.91 | 0.81 | 0.64 | **0.47** | 0.50 | **0.27** |
| | 24 | 1.02 | 1.09 | 1.32 | **1.99** | 1.27 | **1.56** | **1.54** | **2.33** |
| | 25 | 0.87 | 0.85 | 1.50 | **1.58** | 0.89 | 0.64 | 0.84 | 1.10 |
| | 26 | 0.89 | 0.62 | 0.96 | 0.68 | 0.62 | **0.34** | 0.60 | **0.31** |
| | 27 | 1.15 | 1.15 | **1.81** | **1.86** | 1.44 | 1.42 | 1.41 | 1.42 |
| | 28 | 1.19 | 1.33 | 1.24 | 1.32 | 1.26 | 1.27 | 1.20 | 1.22 |
| | 29 | 0.97 | 0.79 | 0.87 | 0.78 | 0.77 | 0.84 | **0.42** | **0.24** |
| | 30 | 0.92 | 0.87 | 0.91 | 0.80 | 0.98 | 0.94 | 0.72 | 0.60 |

Infit and outfit statistics below 0.5 and above 1.5 are in bold to indicate problematic item fit.

few teachers were as high as 6 SDs above the item difficulty average. At T1, the ability level of the teachers did increase, as more teachers were in the range of 3–5 SDs above average item difficulty level. However, this makes the mismatch of items to the ability level worse as the test became extremely easy. This continued at T2 and T3, as it seems the test was unable to distinguish between teachers at higher ability levels, thus contributing to the low person reliability estimate. The percentage of respondents answering items correctly also demonstrates that many of the items were not difficult for teachers. At T0, 9 multiple-choice items were answered correctly by 90% or more of the sample; at T1,12 items were answered correctly by 90% or more; and at T2 and T3, 14 items were answered correctly by 90% or more.

*Convergent Validity.* Table 4 displays the intercorrelations of the variables examined. First, correlations between the test ability estimates for the different time points were examined. The Pearson correlation coefficients were all statistically significant and large in magnitude. In addition, the general pattern is that measurements closer together were more highly correlated; for



M = mean, S = one standard deviation, T = two standard deviations

*Figure 1.* Project 1 person-item maps.

example, the T3 estimates had a higher correlation with T2 ($r = .72$, $p < .001$) than T0 ($r = .57$, $p < .001$) or T1 ($r = .68$, $p < .001$).

Number of college science methods courses and number of college science content courses were examined for their relationships with the test ability estimates at T0 and T1 only, as the person reliability estimates were higher for the T0 and T1 test. Neither of these two variables was significantly related to the test ability estimates at T0 or T1.

At T0, the point biserial correlation for the group variable with the test ability estimates was not statistically significant ($r = .02$, $p = .704$). However, at T1, group had a statistically significant correlation to the test ability estimates ($r = .15$, $p = .005$). The mean and standard deviation for the test ability estimates at each time point are displayed in Supplementary Table S2. At T0, the mean

Table 4
*Project 1 intercorrelations*

| | T0 ($n$) | T1 ($n$) | T2 ($n$) | T3 ($n$) |
|---|---|---|---|---|
| T0 | — | .66$^{**}$ (259) | .64$^{**}$ (146) | .57$^{**}$ (89) |
| T1 | — | — | .72$^{**}$ (172) | .68$^{**}$ (93) |
| T2 | — | — | — | .72$^{**}$ (91) |
| Years teaching | .07 (236) | .07 (316) | — | — |
| Science methods courses | .05 (265) | .06 (304) | — | — |
| Science content courses | .08 (269) | .06 (312) | — | — |
| Science knowledge observation scores Year 1 | .37$^{**}$ (94) | .38$^{**}$ (102) | — | — |
| Science knowledge observation scores Year 2 | — | .31$^{*}$ (46) | — | — |
| Group | .02 (281) | .15$^{**}$ (332) | — | — |

Values in parentheses represent the number of teachers for each correlation.
$^{*}p < .05$; $^{**}p < .01$.

of the treatment group ($M = 2.03$, $SD = 1.12$) was similar to that of the control group ($M = 1.97$, $SD = 1.10$). However, at T1, the treatment group mean ($M = 2.81$, $SD = 0.94$) was higher than that of the control group ($M = 2.48$, $SD = 1.20$). Although the magnitude of the effect for the correlation was small, it does indicate that the test was able to detect some change based on participating in a PD to increase teachers' SCK.

Finally, the SCK observation scores were examined for their relationships with the SCK test ability estimates. Again, the test ability estimates at T2 and T3 were not included in these analyses since the person reliability estimates were low. The Year 1 observation score was significantly correlated with both the T0 estimates ($r = .37$, $p < .001$) and the T1 estimates ($r = .38$, $p < .001$). The Year 2 observation score was significantly correlated with T1 estimates ($r = .31$, $p = .034$). The magnitudes of these correlations were moderate in size. This finding suggests that the SCK test is related to the SCK observations.

## Project 2: P-SELL Effectiveness Study

### *Sample*

Project 2 took place in three school districts with diverse student and teacher populations within the same state as Project 1. The demographic information for each of the districts pertains to the first year (2012–2013) of the project. District A was located in the northeastern part of the state with a K-12 student demographic composition of 45% Black, 8% Hispanic, 40% White non-Hispanic, and 7% Other; 52% received FRL; and 3% were designated as LEP. District B was located in the southwestern part of the state with a K-12 student demographic composition of 28% Black, 15% Hispanic, 51% White non-Hispanic, and 6% Other; 52% received FRL; and 8% were designated as LEP. District C was located in the central part of the state with a K-12 student demographic composition of 30% Black, 34% Hispanic, 28% White non-Hispanic, and 8% Other; 60% received FRL; and 14% were designated as LEP or ELLs.

An RCT was conducted. During the 2012–2013 school year, District A had 103 elementary schools, District B had 44 elementary schools, and District C had 125 elementary schools. Within each of the three school districts, 22 schools were randomly selected to participate, yielding a total of 66 participating schools. To ensure that the selected schools would have a LEP population representative of the district as a whole, half of the schools were randomly selected from schools in the district with more than the median percent LEP and the other half from schools with fewer than the median percent LEP. Within each district, half of the selected schools were randomly assigned to the treatment group and half to the control group, yielding a total of 33 schools in the treatment group and 33 schools in the control group across the three districts. All fifth grade teachers in the selected schools participated in the study.

### *Data Collection*

The data collection and coding of time points was the same as those in Project 1. The exception was that for Project 2, the maximum number of time points that a teacher could have was two (Group 1), in which case the teacher completed a baseline test, participated in one year of the study, and completed the test at the end of the year. This is because at the time of the study, only 1 year of the 3-year intervention was completed. However, we were able to examine the internal and external structure of the test along with evidence of convergent validity and assess whether we had addressed some of the weaknesses of the Project 1 test. Supplementary Table S1 presents sample descriptive statistics, and Table 1 displays the number of teachers who completed the test at each time point. At each data collection, most teachers completed the test with less than 3% not completing due to a variety of reasons (e.g., refusal, teacher absent or on leave).

### *Instrument Construction and Content Validity*

The Project 2 test differed from the Project 1 test in three main ways. First, the state adopted new science standards with 18 "big ideas" in 4 strands: the practice of science, earth and space science, life science, and physical science. Thus, the Project 2 test was developed around these strands, and the content differed somewhat from the science topics covered in the Project 1 test. Second, results of the Project 1 test suggested that the test was too easy for the teacher sample, as it was developed to measure SCK at the fifth grade level. For the Project 2 test, the overall difficulty level was increased. Third, in Project 1, the test might have been too easy for teachers over time as they were taking the same test repeatedly. For Project 2, two equated forms of the test were developed with approximately 10% of the items being linking items (i.e., appearing on both forms to link the scores from the two forms). A schedule was set up such that teachers who participated in the full 3 years of the study took Form A at T0, Form B at T1, Form A again at T2, and form B again at T3. Although they took each form twice, it was nearly 2 years in between answering the same form. In addition, this schedule of test administration should have reduced memory effects, where teachers retaking the same form might have recalled their answers to the previous form or have discussed answers with others prior to retaking the test.

Again, the same two researchers who worked for Project 1 took the lead in searching for items that mapped onto the new science standards in the state from NAEP and TIMSS public release items. They focused on more difficult items that were typically administered at the middle and high school levels. This added some challenge to finding appropriate items because it was hard to find more difficult items that covered the more basic content areas at the elementary school level. On the Project 2 test, we did not include any project-developed items; rather, if an item on a topic covered by the standards was not found in NAEP or TIMSS, we searched public release items from other states' assessments. The items considered for the pool were rated as being of easy, medium, or hard difficulty for a fifth grade teacher (as opposed to at the fifth grade level in Project 1) with consideration of information from the original sources.

Two forms were pilot tested with 33 total items on each form, of which three were constructed response items. The forms were piloted as an online test with a time limit in order to deter teachers from looking up the correct answers. Initially, fifth grade teachers in the state in which the research was conducted were recruited, excluding the three school districts participating in the research, but not enough fifth grade teachers responded. Therefore, K-5 grade teachers in another state were also recruited because we had access to these respondents. Form 1 was completed by 25 in-state fifth grade teachers and 76 out-of-state K-5 grade teachers; Form 2 was completed by 25 in-state K-5 grade teachers and 50 out-of-state K-5 grade teachers. An analysis of variance (ANOVA) demonstrated that on both Form 1 and Form 2, in-state K-5 grade teachers scored significantly higher than out-of-state K-5 grade teachers, which was probably because the content was based on the state science standards for fifth grade at which science was tested and counted toward school accountability.

The psychometric properties of the test and the items were examined for the pilot tests, including the percentage of respondents answering items correctly and a comparison of the difficulty of the items with the ability of the sample. More weight was given to information from the in-state sample when making decisions. The Cronbach's alpha for the scores on the pilot tests of Form 1 was $\widehat{\rho} = .84$, and Form 2 was $\widehat{\rho} = .78$.

Two final forms of the test were developed. For each form, 33 items were chosen that mapped onto the topics assessed at fifth grade, and each included 30 multiple-choice and three constructed response items. There were nine linking items, including all three of the constructed response items. One item from Project 1 was included on Form A, and one item from Project 1 was included on Form B. Table 2 displays the test specifications, along with the item difficulty level and source information.

The Form A test with the correct answer for multiple-choice items denoted and scoring rubric for the constructed response items are available as supplementary material accompanying the online article. Form B is not available as this form is currently being used. Both forms of the test were worth a total of 40 points. Each multiple-choice item was worth 1 point, one constructed response item was worth 2 points, and two constructed response items were worth 4 points each (see the rubric for details). A team of raters participated in a 2 hour training session prior to scoring responses to each item. All the tests were independently scored by two raters. Disagreements were resolved by a third round of scoring and group consensus, if needed. The inter-rater agreement for all three items on the pre and posttest was excellent (weighted Kappa above 0.75).

### *Data Analysis*

The psychometric analysis of internal and external structure was examined by assessing the unidimensionality and Rasch model in the same way as Project 1. There were 264 respondents at T0 and 283 at T1 (see Table 1). The unidimensionality and Rasch analyses were conducted for each time point, and all respondents at that time point were included in the analyses. Evidence of convergent validity was collected by examining the associations between the test ability estimates and the other variables in the same way as Project 1 with the exception of SCK observation scores. All respondents with valid measures were included in these analyses. Classroom observations were conducted with one teacher randomly selected from each school two times throughout the school year, and the same SCK classroom observation scale was used as in Project 1. However, in Project 2, the *G*-coefficient for teachers' SCK classroom observations was below .70, which was not acceptable. This was probably because only two observations were conducted, whereas in Project 1, three observations were conducted.

### *Results*

*Dimensionality Results.* The dimensionality of the test was assessed for both time points (i.e., T0 and T1). The model fit the data well at T0 ($RMSEA = .018$, $CI_{90\%} = .000, .028$) and T1

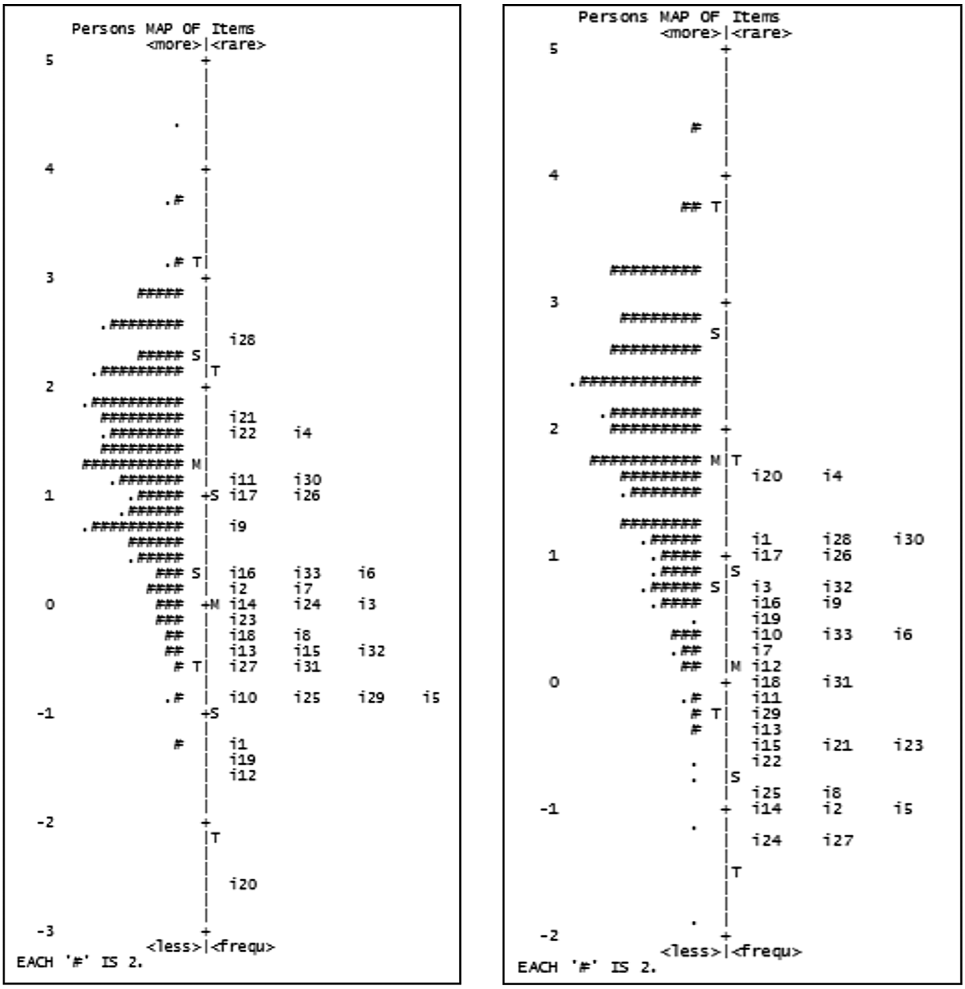
($RMSEA = .017$, $\mathrm{CI}_{90\%} = .000$, .027), indicating that the items were all measuring a single construct.

*Rasch Modeling Results.* The person and item reliability estimates for the test at each of the time points, along with the infit and outfit for each item, are displayed in Table 5. The item reliability estimates were acceptable at T0 ($\widehat{\rho} = .97$) and T1 ($\widehat{\rho} = .95$), which suggest that the item difficulty hierarchy of the test was confirmed in the sample at each time. The person reliability estimate was acceptable at T0 ($\widehat{\rho} = .77$) and at T1 ($\widehat{\rho} = .76$), which suggests that the test was sensitive enough to distinguish between teachers with different SCK at both T0 and T1. Project 2 was an improvement over Project 1 where the test had acceptable person reliability only at T0. The improved person reliability estimates were probably due to the higher difficulty level of the test, such that even at T1 when teachers' scores increased the test was able to detect differences.

Table 5
*Project 2 Rasch modeling results*

| | | T0 | | T1 | |
|---|---|---|---|---|---|
| Person reliability | | .77 | | .76 | |
| Item reliability | | .97 | | .95 | |
| | Item | Infit | Outfit | Infit | Outfit |
| Item fit | 1 | 0.97 | 0.83 | 1.06 | 1.10 |
| | 2 | 1.08 | 1.16 | 1.11 | 1.07 |
| | 3 | 0.95 | 0.85 | 1.10 | 1.08 |
| | 4 | 1.07 | 1.08 | 0.94 | 0.90 |
| | 5 | 1.06 | 0.97 | 0.90 | 0.61 |
| | 6 | 0.91 | 0.83 | 0.97 | 0.93 |
| | 7 | 1.22 | **1.59** | 1.05 | 1.04 |
| | 8 | 0.93 | 0.82 | 0.88 | 0.58 |
| | 9 | 1.02 | 1.02 | 1.14 | 1.17 |
| | 10 | 0.94 | 0.91 | 1.07 | 1.11 |
| | 11 | 1.07 | 1.09 | 1.09 | 1.15 |
| | 12 | 1.01 | 0.97 | 0.96 | 1.00 |
| | 13 | 1.02 | 0.99 | 1.19 | 1.19 |
| | 14 | 0.94 | 0.83 | 0.83 | **0.43** |
| | 15 | 0.91 | 0.79 | 1.04 | 0.89 |
| | 16 | 1.06 | 1.03 | 1.16 | 1.18 |
| | 17 | 0.99 | 0.98 | 1.00 | 1.01 |
| | 18 | 1.14 | 1.35 | 0.87 | 0.76 |
| | 19 | 1.03 | 1.06 | 1.20 | 1.29 |
| | 20 | 0.94 | 0.51 | 0.93 | 0.92 |
| | 21 | 1.05 | 1.15 | 0.93 | 0.70 |
| | 22 | 1.02 | 1.01 | 0.86 | 0.63 |
| | 23 | 0.95 | 0.87 | 0.99 | 0.98 |
| | 24 | 0.87 | 0.74 | 0.89 | 0.81 |
| | 25 | 0.93 | 0.79 | 0.84 | 0.85 |
| | 26 | 0.97 | 0.96 | 1.04 | 1.05 |
| | 27 | 1.02 | 1.10 | 1.05 | 1.16 |
| | 28 | 0.96 | 0.94 | 1.06 | 1.03 |
| | 29 | 0.96 | 0.82 | 0.90 | 0.85 |
| | 30 | 1.06 | 1.09 | 1.10 | 1.08 |
| | 31 | 0.86 | 0.75 | 0.88 | 0.82 |
| | 32 | 0.99 | 0.93 | 1.00 | 0.94 |
| | 33 | 1.03 | 1.23 | 1.06 | 1.03 |

Infit and outfit statistics below 0.5 and above 1.5 are in bold to indicate problematic item fit.



```
     Persons MAP OF Items
             <more>|<rare>
 5                 +
                   |
                   |
                   |
                .  |
                   |
                   |
 4                 +
                   |
               .#  |
                   |
                   |
                   |
               .# T|
 3                 +
            #####  |
                   |
        .########  |
                   |  i28
            ##### S|
       .#########  |T
 2                 +
      .##########  |
        #########  |  i21
        .########  |  i22    i4
        #########  |
      ########### M|
         .#######  |  i11    i30
 1         .#####  +S i17    i26
          .######  |
      .##########  |  i9
           ######  |
           .#####  |
              ### S|  i16    i33    i6
             ####  |  i2     i7
 0            ###  +M i14    i24    i3
              ###  |  i23
               ##  |  i18    i8
               ##  |  i13    i15    i32
                # T|  i27    i31
                   |
               .#  |  i10    i25    i29    i5
-1                 +S
                   |
                #  |  i1
                   |  i19
                   |  i12
                   |
                   |
-2                 +
                   |T
                   |
                   |
                   |  i20
                   |
                   |
-3                 +
             <less>|<frequ>
EACH '#' IS 2.
```

T0

```
     Persons MAP OF Items
             <more>|<rare>
 5                 +
                   |
                   |
                   |
                   |
                #  |
                   |
                   |
 4                 +
                   |
               ## T|
                   |
                   |
                   |
        #########  |
                   |
 3                 +
         ########  |
                  S|
        #########  |
                   |
    .############  |
                   |
       .#########  |
 2      #########  +
                   |
      ########### M|T
         ########  |  i20    i4
         .#######  |
                   |
         ########  |
           .#####  |  i1     i28    i30
 1          .####  +  i17    i26
            .####  |S
           .##### S|  i3     i32
            .####  |  i16    i9
                .  |  i19
              ###  |  i10    i33    i6
              .##  |  i7
               ##  |M i12
 0                 +  i18    i31
               .#  |  i11
                # T|  i29
                #  |  i13
                   |  i15    i21    i23
                .  |  i22
                .  |S
                   |  i25    i8
-1                 +  i14    i2     i5
                .  |
                   |  i24    i27
                   |
                   |T
                   |
                   |
                .  |
-2                 +
             <less>|<frequ>
EACH '#' IS 2.
```

T1

M = mean, S = one standard deviation, T = two standard deviations

*Figure 2.* Project 2 person-item maps.

In Table 5, infit and outfit statistics below 0.5 and above 1.5 are in bold to indicate problematic item fit. At T0, the outfit of item 7 was slightly above 1.5 (outfit = 1.59), and at T1 the outfit of item 14 was slightly below 0.5 (outfit = 0.43). However, neither of these items is of great concern as they are not far from the cutoff points.

Figure 2 displays the person-item map for each of the times. The item difficulty parameter estimates ($b_i$) were anchored when calibrating T0 and T1; therefore, the scale is constant across the maps. It should also be noted that the forms were equated and certain items remained the same from T0 to T1 (see Table 2 for linking items). However, most items were not the same on both forms; therefore, an item may be easy at T0 and difficult at T1 as it is not the same item. At T0, the difficulty of the items tended to match the ability of the teachers quite well. At T1, the ability level of the teachers did increase, as more teachers were in the range of 2–3 SDs above the average item

Table 6
*Project 2 intercorrelations*

| | T0 (*n*) | T1 (*n*) |
|---|---|---|
| T0 | — | $.77^{**}$ (260) |
| Years teaching | −.03 (280) | −.07 (260) |
| Science methods courses | $-.12^{*}$ (283) | $-.12^{*}$ (263) |
| Science content courses | .11 (283) | $.20^{**}$ (263) |
| Group | .03 (283) | $.13^{*}$ (265) |

Values in parentheses represent the number of teachers for each correlation.
$^{*}p < .05$; $^{**}p < .01$.

difficulty level. The mismatch is not extreme, and the person reliability estimates are acceptable. This finding suggests that in Project 2, the increased difficulty of the test improved its psychometric properties. Furthermore,the test was not too difficult. If the test were too difficult, it would have failed to accurately measure the intended construct of teachers' SCK.

*Convergent Validity.* Table 6 displays the intercorrelations of the variables examined. First, the correlation between the test ability estimates for T0 and T1 were examined. The Pearson correlation coefficient was statistically significant and large in magnitude ($r = .77$, $p < .001$), suggesting that the same construct was being measured over time.

Number of college science methods courses and number of college science content courses were examined for their relationships with the test ability estimates for both time points. The number of science methods courses had a statistically significant negative correlation with the test ability estimates at both T0 ($r = -.12$, $p = .046$) and T1 ($r = -.12$, $p = .044$). The number of college science content courses had a small positive correlation with the test ability estimates at T0, which approached statistical significance ($r = .11, p = .065$); at T1, the correlation was again small and positive but was statistically significant ($r = .20, p = .001$).

At T0, the point biserial correlation for the group variable with the test ability estimates was not statistically significant ($r = .03, p = .592$). However, at T1, group had a statistically significant correlation to the test ability estimates ($r = .13, p = .031$). The mean and standard deviation for the test ability estimates for each time point are displayed in Supplementary Table S3. At T0, the mean test ability of the treatment group ($M = 1.31$, $SD = 0.97$) was similar to that of the control group ($M = 1.26$, $SD = 0.92$). However, at T1, the treatment group mean ($M = 1.88$, $SD = 1.02$) was higher than that of the control group ($M = 1.61$, $SD = 0.96$). Although the magnitude of the effect for the correlation was small, the finding does indicate that the test was able to detect some change based on participating in the intervention. A one-way between groups ANOVA model was examined to determine if change in SCK as measured by the difference between the pretest and posttest ability estimates differed by group. Teachers participating in the intervention demonstrated a statistically significant larger increase in SCK ($F_{(1,258)} = 5.95$, $p = .015$). The partial eta squared measure of effect size (${\eta_p}^2 = .03$) suggested that 3% of the variance in difference scores was explained by group, which is small but meaningful.

## Discussion and Implications

The development of the SCK test in the study was motivated by the need for standardized measures that can be used across various studies. Specifically, this study examined (a) evidence on the internal structural features, (b) external structural evidence of convergent validity, (c) ability to differentiate between groups and detect change over time, and (d) generalizability across groups.

### *Discussion*

*Internal Structural Features.* The SCK test developed for Project 1 had acceptable person reliability estimates at T0, but the estimates were below the threshold considered acceptable at T1, T2, and T3. Furthermore, the item fit statistics suggested that some of the items on the Project 1 test were not productive for the measurement of teachers' SCK, particularly at the later time points. The person-item maps suggested that at T0 the test matched the ability level of respondents fairly well, but at T1, T2, and T3, the test became easy for respondents.

To improve the psychometric properties of the SCK test for Project 2, the overall item difficulty level was increased and two equated forms were developed with linking items so that teachers would not respond to the same items repeatedly. The results from Project 2 indicated that the person reliability was acceptable at both times, which was an improvement over Project 1. The item fit statistics also indicated that all items were productive in measuring teachers' SCK. Finally, the person-item maps indicated that the test ability estimates for teachers were higher than the items, yet ability level and items were fairly well matched. Project 2 created a balance in that the test had stronger psychometric properties, yet it was not so difficult that it was not able to accurately measure teachers' SCK.

*External Structural Evidence of Convergent Validity.* In both Project 1 and Project 2, the test ability estimates across time points were related to each other. In Project 1, there was a pattern that measurements closer together in time were more highly correlated, which provides evidence of convergent validity. In Project 1, the test ability estimates were not related to the number of science methods courses taken or the number of science content courses taken. The lack of a correlation with content courses is surprising, but other research has found that number of college science courses was not related to elementary teachers' SCK (Nowicki et al., 2013). This finding may be because the knowledge needed at the elementary level is not the same as that covered in college-level science courses. In Project 2, the correlation between SCK ability estimates and the number of science content courses taken approached significance at T0, and the correlation was significant at T1; at both times, the magnitude of the effect was small. In contrast, the correlation between SCK ability estimates and the number of science methods courses taken was significantly negative at T0 and T1; at both times, the magnitude of the effect was small. This finding suggests that the more difficult test from Project 2 was related to more SCK developed through college-level science courses, evidence of convergent relationships, whereas science methods courses may not be related to SCK.

The most compelling evidence of convergent relationships was that the test ability estimates were moderately related to the SCK observation scores in the subsample of teachers in Project 1. This finding suggests that the test was able to measure teachers' SCK needed in classroom instruction.

*Ability to Differentiate Between Groups and Detect Change Over Time.* In Project 1, the test ability estimates at T0 were not related to group, but they were at T1. This finding suggests that the test ability estimates by group were changing over time based on the effect of the intervention. This is stated with the acknowledgement that for Project 1 the person reliability estimates at T1 were not quite at the acceptable threshold. Results from Project 2 were more promising, as again the test ability estimates were not related to group at T0, but they were at T1. ANOVA results suggested small differences between the groups in the change found between the two measurement time points. Thus, the test was able to detect some change over time (between T0 and T1) in the teachers and some difference between the treatment and control group at T1. This finding suggests that a measure like our test may be useful in evaluating the effects of PD on teachers' SCK. Although the effect size was small, according to Lipsey et al. (2012); a small effect size is practically meaningful in educational interventions. Furthermore, more advanced analyses (i.e., repeated measures ANOVA, multilevel model, latent growth model) can control for other variables (e.g., baseline SCK, number of years teaching, number of science courses taken), and further analyses can determine if the effect is larger, especially in regards to particular teachers (e.g., teachers' with low baseline science knowledge).

*Generalizability Across Groups.* This research used two different projects with different sample characteristics. Results from the two projects, as discussed in the sections on Project 1 and Project 2, are similar with regard to internal structural features, external structural evidence, and the ability to differentiate between groups and detect change over time.

### *Implications*

*Contributions.* Science education researchers have consistently called for the development of more rigorous measures using more advanced methodologies, like Rasch models, to establish validity and reliability (Liu, 2009, 2012) and for these measures to be used across studies in order to build generalizable knowledge (Ball et al., 2005). Similarly, teacher PD researchers have called for the development of standardized measures of teachers' knowledge and practices (Desimone, 2009). This study provides a description of one measure used to evaluate teachers' SCK in two multiyear PD intervention projects. The study makes an important contribution to the literature on teachers' SCK, PD, and measurement. We have learned that developing a measure of teachers' SCK can be difficult but is possible. Other researchers may be able to use our tests in their own research or gain insight from the processes used and lessons learned, as discussed below.

First, previous research on teachers' SCK measures typically used the CTT framework and reported only the Cronbach's alpha estimate of reliability. We applied the Rasch measurement model, which allowed us to estimate person reliability and item reliability and to conduct a detailed item analysis. Though results from Project 1 were promising, using the information from the analyses, we were able to improve the psychometric properties of the test for Project 2 by increasing the item difficulty to more closely match the knowledge of the teachers. The challenge of matching item difficulty to ability estimates has been noted in the literature with some researchers finding that test items were, on average, too easy (Hill et al., 2004) and other researchers finding that test items were difficult for typical respondents (Jüttner et al., 2013). It is apparent that there needs to be more difficult items on a test of SCK with teachers, yet the content of the items must be consistent with the material covered in PD to reasonably expect change on the content tested. In addition to SCK, researchers have included items measuring teachers' pedagogical content knowledge (Hill et al., 2004; Hill, 2010; Jüttner et al., 2013; Krauss et al., 2008). However, adding items related to pedagogical content knowledge may increase the difficulty level of the test, as these items require specialized knowledge of content and pedagogy combined.

Second, most studies do not provide evidence of validity. Our study provides evidence of construct validity, including detailed descriptions of the decision process for selecting items and presenting test specifications. An examination of the internal structure demonstrated that the SCK test for Project 2 had acceptable reliabilities and item fit and matched the ability of respondents well. Additionally, an examination of the relationships between the test ability estimates and other measures of SCK provided evidence of convergent validity as expected.

Finally, our PD intervention is longitudinal in design, as the same teachers participate over multiple years. Our SCK test was able to detect some change over time in the treatment group compared to the control group. While other measures of teachers' SCK have been developed, most have been administered at only one point in time (Jüttner et al., 2013; Krauss et al., 2008; Nowicki et al., 2013). In addition, our intervention aims to improve elementary teachers' knowledge of comprehensive science topics, while others have covered a specific science topic (Heller et al., 2012). Developing a teachers' SCK test that is sensitive enough to detect change over time on an array of science topics proves a challenge. Many PD studies, using a number of measures, have found a pattern described as short-term growth and long-term stability (Lee & Maerten-Rivera, 2012; Supovitz, Mayer, & Kahle, 2000), where there is an initial jump in teachers' knowledge or practice, followed by sustainability. This pattern may not be what is occurring in reality. There may be continued growth by the teachers, yet the measures were not sensitive enough to detect this additional change.

*Future Research.* The results of the study point to directions for future research on measures of teachers' SCK, particularly for studies of PD. Our study had several methodological strengths. First, a randomized control design was used for the PD intervention, which has been called for in the literature as a more rigorous methodological design (Wayne et al., 2008). Second, the study used samples from two intervention projects with different teacher and student demographics, yet the results remained consistent. Even so, to evaluate the validity of the SCK test, the measure should continue to be used with different samples under different conditions to further examine the generalizability. Third, the study involved all fifth grade science teachers in the participating schools, rather than a self-selected group of volunteer teachers. Thus, the validity results are more likely to apply to the teaching population.

We also note limitations to our study that could be investigated in future research. First, we note that the test was developed to cover the fifth grade science content in a specific state, and thus the test may not be a valid measure of science content in other grades or other states. For example, when the pilot test was administered to samples for Project 2, the out-of-state sample performed differently. However, the processes used in the development and validation of our test are relevant to other researchers.

Second, the sample size in Project 1 dropped at T2 and particularly at T3 due to teacher attrition for various reasons beyond the control of the study. This impacted the results of the study, as a larger sample size would have provided better estimates in regards to the SCK test.

Third, the study examined data from the full implementation of Project 1. However, as Project 2 is currently being conducted, we were able to use data collected only from two time points over 1 year of implementation out of the 3-year planned intervention. Additional waves of data will allow us to examine the ability of the test to detect change over time.

Finally, a logical next step for further research involves testing a change model of how a PD intervention impacts teachers' SCK and how teacher change, in turn, impacts student outcomes (Desimone, 2009). The change model also provides information in regards to the influence of teacher knowledge on student outcomes, which is an area that requires more empirical evidence (Bartos & Lederman, 2014; Diamond et al., 2014; Liu, Lee, & Lin, 2010). Examining such a change model could address the ability of the test to detect teacher change and the magnitude of change in teachers' SCK needed to impact student outcomes, which would provide additional evidence regarding the validity of the SCK test.

## References

Abell, S. (2007). Research on science teachers' knowledge. In S. Abell & N. Lederman (Eds.), Handbook of research on science education (pp. 1105–1149). Mahwah, NJ: Erlbaum.

American Association for the Advancement of Science. (1989). Science for all Americans. New York, NY: Oxford University.

American Association for the Advancement of Science. (1993) Benchmarks for science literacy. New York, NY: Oxford University.

Andrich, D. (1988). Rasch models for measurement. Newbury, CA: Sage.

Ball, D. L., Hill, H., & Bass, C. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? American Educator, 29(1), 14–17, 20–22, 43–46.

Bartos, S. H., & Lederman, N. G. (2014). Teachers' knowledge structures for nature of science and scientific inquiry: Conceptions and classroom practice. Journal of Research in Science Teaching, 51(9), 1150–1184.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. American Educational Research Journal, 47(1), 133–180.

Bond, T., & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.), Mahwah, NJ: Erlbaum.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. Educational Researcher, 33(8), 3–15.

Brennan, R. L. (2001). Generalizability theory. New York, NY: Springer.

Crocker, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. Belmont, CA: Wadsworth Group.

Cronbach, L. J., Nageswari, R., & Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. The British Journal of Statistical Psychology, 16, 137–163.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. Educational Researcher, 38, 181–199.

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. Retereived from http://feraonline.org/fjer/publications.htm.

Diamond, B. S., Maerten-Rivera, J., Rohrer, R., & Lee, O. (2014). Effectiveness of a curricular and professional development intervention at improving elementary teachers' science content knowledge and student achievement: Year 1 results. Journal of Research in Science Teaching, 51(5), 635–658.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. American Educational Research Journal, 38(3), 915–945.

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. Journal of Research in Science Teaching, 49(3), 333–362.

Hill, H. C. (2010). The nature and predictors of elementary teachers' mathematical knowledge for teaching. Journal for Research in Mathematics Education, 41(5), 513–545.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. Journal for Research in Mathematics Education, 39(4), 372–400.

Hill, H. C., Schilling, S. G., & Loewenberg, B. (2004). Developing measures of teachers' mathematics knowledge for teaching. The Elementary School Journal, 105(1), 11–30.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1–55.

Jacobs, C. L., Martin, S. N., & Otieno, T. C. (2008). A science lesson plan analysis instrument for formative and summative program evaluation of a teacher education program. Science Education, 92(6), 1096–1126.

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, 25(1), 45–67.

Krauss, S., Baumert, J., & Blum, W. (2008). Secondary mathematics teachers' pedagogical content knowledge and content knowledge: Validation of the COACTIV constructs. ZDM: The International Journal on Mathematics Education, 40(5), 873–892.

Larson, M. R., & Smith, W. M. (2013). Distributed leadership: Key to improving primary students' mathematical knowledge. Journal of Mathematics Education at Teachers College, 4(1), 26–33.

Lee, O., LeRoy, K., Thornton, C., Adamson, K., Maerten-Rivera, J., & Lewis, S. (2008). Teachers' perspectives on a professional development intervention to improve science instruction among English language learners. Journal of Science Teacher Education, 19(1), 41–67.

Lee, O., & Maerten-Rivera, J. (2012). Teacher change in elementary science instruction with English language learners: Results of a multiyear professional development intervention across multiple grades. Teachers College Record, 114(8), 1–42.

Lee, O., Maerten-Rivera, J., Penfield, R. D., LeRoy, K., & Secada, W. G. (2008). Science achievement of English language learners in urban elementary schools: Results of a first-year professional development intervention. Journal of Research in Science Teaching, 45(1), 31–52.

Lee, O., Mahotiere, M., Salinas, A., Penfield, R. D., & Maerten-Rivera, J. (2009). Science writing achievement among English language learners: Results of three-year intervention in urban elementary schools. Bilingual Research Journal, 32(2), 153–167.

Linacre, J. M. (1997). KR-20/Cronbach alpha or Rasch person reliability: Which tells the truth? Rasch Measurement Transactions, 11, 580–581.

Linacre, J. M. (2012). Winsteps® Rasch measurement computer program. Beaverton, OR: Winsteps.

Linacre, J. M. (2014). A user's guide to Winsteps® ministep Rasch-model computer programs. Retrieved from http://www.winsteps.com/winman/index.htm.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, … Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. Washington, DC: National Center for Special Education Research, Institute of Education Sciences, US Department of Education. Retrieved from http://ies.ed.gov/ncser/ .

Liu, O. L., Lee, H., & Linn, M. C. (2010). An investigation of teacher impact on student inquiry science performance using a hierarchical linear model. Journal of Research in Science Teaching, 47(7), 807–819.

Liu, X. (2009). Standardized measurement instruments in science education. In W. Roth & K. Tobin (Eds.), The world of science education: Handbook of research in North America (pp. 649–677). Rotterdam, The Netherlands: Sense.

Liu, X. (2012). Developing measurement instruments for science education research. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), Second international handbook of science education (pp. 651–665). New York, NY: Springer.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Welsley.

Loucks-Horsley, S., Hewson, P. W., Love, N., & Stiles, K. (1998). Designing professional development for teachers of science and mathematics. Thousand Oaks, CA: Corwin.

Maerten-Rivera, J., Ahn, A., Lanier, K., Diaz, J., & Lee, O. (in press). Effect of a multi-year intervention on science achievement of all students including English language learners. The Elementary School Journal.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149–174.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). New York, NY: MacMillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. Educational Measurement: Issues and Practice, 14(4), 5–8.

Muthén, L. K., & Muthén, B. O. (2012). Mplus user's guide (7th ed.). Los Angeles, CA: Authors.

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

National Research Council. (2000). Inquiry and the national science education standards: A guide for teaching and learning. Washington, DC: National Academy Press.

National Research Council. (2002). Scientific research in education. Washington, DC: National Academies Press.

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

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

Novick, M. R. (1966). The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 3, 1–18.

Nowicki, B. L., Sullivan-Watts, B., Shim, M. K., Young, B., & Pockalny, R. (2013). Factors influencing science content accuracy in elementary inquiry science lessons. Research in Science Education, 43(3), 1135–1154.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York, NY: McGraw Hill.

Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. American Educational Research Journal, 44(4), 921–958.

Public Law no. 107–110. (2002). 115 Stat. 1425.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. Multivariate Behavioral Research, 32, 329–353.

Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. British Journal of Mathematical and Statistical Psychology, 54, 315–323.

Santau, A. O., Maerten-Rivera, J., & Huggins, A. C. (2011). Science achievement of English language learners in urban elementary schools: Fourth-grade student achievement results from a professional development intervention. Science Education, 95(5), 771–793.

Schumacker, R. E., & Smith, E. V. (2007). A Rasch perspective. Educational and Psychological Measurement, 67, 394–409.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 15(2), 4–14.

Supovitz, J. A., Mayer, D. P., & Kahle, J. B. (2000). Promoting inquiry-based instructional practice: The longitudinal impact of professional development in the context of systemic reform. Educational Policy, 14(3), 331–356.

US Department of Education and National Science Foundation. (2013). Common guidelines for education research and development: A report from the Institute of Education Sciences, US Department of Education and the National Science Foundation. Washington, DC: Author.

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motive and methods. Educational Researcher, 37(8), 469–479.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. Review of Educational Research, 73(1), 89–122.

Weaver, D., & Dick, T. (2009). Oregon mathematics leadership institute project: Evaluation results on teacher content knowledge, implementation fidelity, and student achievement. The Journal of Mathematics and Science, 11, 57–84.

## Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web-site.