Studies in Educational Evaluation 45 (2015) 27–36

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

SCIIENCE: The creation and pilot implementation of an NGSS-based instrument to evaluate early childhood science teaching

Joan N. Kaderavek, Tamala North, Regina Rotshtein, Hoangha Dao, Nicholas Liber, Geoff Milewski, Scott C. Molitor *, Charlene M. Czerniak

University of Toledo, United States

ARTICLE INFO

Article history:
Received 17 September 2014
Received in revised form 9 March 2015
Accepted 10 March 2015
Available online

Keywords:
Discourse analysis
Teacher assessment
Language of science in classrooms
Validity/reliability

ABSTRACT

This paper describes the development, testing and implementation of the **S***ystematic* **C***haracterization of* **I***nquiry* **I***nstruction in* **E***arly Lear***N***ing* **C***lassroom* **E***nvironments* (SCIIENCE). The SCIIENCE instrument was designed to capture best practices outlined in the National Research Council's *Framework for K-12 Science Education* as they occur within a science lesson. The goals of the SCIIENCE instrument are to (a) assess the quality of science instruction in PK-3 classrooms, (b) capture teacher behaviors and instructional practices that engage students in the lesson, promote scientific studies, encourage higher-level thinking, and (c) provide a feedback mechanism for guiding professional development of PK-3 teachers. Science educators can apply this instrument to teacher behaviors and use the data to improve classroom inquiry instructional methodology.

© 2015 Published by Elsevier Ltd.

Introduction

The Next Generation Science Standards (NGSS) and the Framework for K-12 Science Education [Framework] indicates that K-12 classroom instruction should focus on the intersection of scientific and engineering practices, disciplinary core ideas, and crosscutting concepts (National Research Council, 2012). High-quality science instruction should focus on teaching "how we come to know what we know" instead of only teaching "just what we know."

Preschool and early childhood science are overlooked as the necessary foundation for eventually achieving high quality instruction (Pratt, 2007); and yet, science especially lends itself to inquiry, exploration, and curiosity essential for establishing young children's positive attitudes towards school in general as well as towards reading, mathematics, and of course science. There is an unwritten expectation that students will naturally develop an interest in science when it is introduced in middle school or even later in junior high (Keeley, 2009). Furthermore, there is a need for early childhood science if our nation expects to improve science education at subsequent grade levels (McCormack, 2010). Eventual achievement levels in science begin in kindergarten and first grade (Chapin, 2006). Yet, many early childhood teachers are intimidated by science and not well prepared to teach science in early grades (Wenner, 1993).

Thus, to achieve the goal of having US children perform at the highest level in scientific inquiry and knowledge, a number of activities must occur that are focused on early childhood science. First, teacher trainers tasked with developing professional development [PD] in line with the current science Framework must explicitly describe what high quality science instruction looks like in early grades. Without this consensus, and an instrument to measure a teacher's level of achievement in implementing the targeted science instructional practices, researchers will not be able to determine whether a teacher will change his or her science instruction as a result of PD, document whether teaching behaviors will stay consistent over time, or determine what specific instructional practices contribute to an increase in child knowledge or skill level.

This paper describes the development, testing and application of an instrument known as the **S***ystematic* **C***haracterization of* **I***nquiry* **I***nstruction in* **E***arly Lear***N***ing* **C***lassroom* **E***nvironments* (SCIIENCE). The SCIIENCE instrument was designed to objectively capture the presence and frequency of specific best practices outlined in the Framework as they occur within a science lesson and focuses exclusively on teacher behaviors. The goals of the SCIIENCE system are to (a) provide a standardized instrument for

* Corresponding author at: University of Toledo, 2801 W. Bancroft Street, MS 311, Toledo, OH 43606-3390, United States. Tel.: +1 419 530 8040.
E-mail address: scott.molitor@utoledo.edu (S.C. Molitor).

http://dx.doi.org/10.1016/j.stueduc.2015.03.003
0191-491X/© 2015 Published by Elsevier Ltd.

assessing the quality of science instruction in a classroom setting for children grades PK-3, (b) capture the instructional practices that engage students in the lesson, promote scientific studies, encourage higher-level thinking, and (c) provide feedback for guiding professional development of PK-3 teachers.

Beyond this focus, our intent is to create an instrument that is standardized, comprehensive, geared toward early childhood classrooms and psychometrically sound. A standardized instrument is defined as an assessment that offers consistent procedures and uniform application and has the potential to compile and compare findings across teachers and different science lessons. A comprehensive instrument is needed because researchers and educators need to capture a wide range of adult behaviors as they occur within science instruction. Furthermore, the most frequently used observational instruments – Inside the Classroom Observation and Analytic Protocol (ITC COP; Horizon Research, Inc., 2000) and the *Reform Teaching Observation Protocol* (RTOP; Sawada et al., 2000) have limitations.

The Horizon rating system provides a global view of various aspects of science instruction but does not permit a fine-grained analysis of specific teacher practice (Henry, Murray, Hogrebe, & Daab, 2009; Henry, Murray, & Phillips, 2007). The RTOP also has limitation in that it uses a Likert scale to assess the quality of classroom instruction. Teachers can be challenged to understand what specifically they need to do to improve their overall instructional quality rating (i.e., to move from a level 4 to a level 5) when a Likert scale is used to evaluate the quality of teaching. The SCIIENCE instrument was developed to surmount these limitations.

The following sections in this paper elaborate the development, testing, and application of the SCIIENCE instrument at the project-level. The authors will describe: (a) the theory and prior research that informed the development of an instrument for evaluating teachers' instructional practice; (b) the theory and background that informed the development of the SCIIENCE instrument; (c) the individual codes and the way in which the SCIIENCE instrument is used to evaluate the quality of science inquiry instruction; (d) project-level data documenting the reliability and validity of the SCIIENCE instrument and the comprehensive plan developed to further document the instrument's validity; (e) an application of this instrument to improve teacher professional development; and (f) a discussion of the limitations of the instrument and future research to be completed using the SCIIENCE instrument.

Theory and background of instruments for evaluating classroom quality and instructional methods

There is a scarcity of literature that examines early childhood science education. But we can look at how teachers approach science inquiry with the hope of fostering good science inquiry technique and the benefit of making teachers more comfortable with teaching science.

Teachers play a vital role in encouraging young children to engage in sophisticated behaviors and verbal interactions (De Kruif, McWilliam, Ridley, & Wakely, 2000). The Measures of Effective Teaching project found that teachers identified as more effective caused students to learn more and teachers identified as less effective caused students to learn less (Cantrell & Kane, 2013). We believe that investing in practices and policies that support effective teaching will result in improved science inquiry. However, we also believe that teaching and learning is too complex for a single measure of performance. To identify best practices, multiple measures are required.

Current instruments that assess quality in early childhood settings include the Early Childhood Environment Rating Scale (ECERS; Harms & Clifford, 1980), now the Early Childhood Rating Scale – Revised (ECERS-R; Harms, Clifford, & Cryer, 1998). This instrument, which consists of 43 items arranged in 7 subscales, has been widely used in the early education field to assess the quality of the preschool and kindergarten learning environments in childcare centers. The ECERS-R scale consists of 43 items arranged in seven subscales. The subscales are assigned a score from 1 to 7 and an overall scale score is calculated by averaging the subscale scores. A score of 1 indicates inadequate quality, 3 indicates minimal quality, 5 indicates good quality and 7 indicates excellent quality. The ECERS-R assesses the quality of classroom routines, the quality of the activities, the availability of materials, provisions for parents and staff, and the interaction between teachers and children. In addition to assessment, the scale is used in lesson planning and professional development (Tout et al., 2010). Though the ECERS-R has a proven track record of success and is commonly used in informal child care settings, it was not designed to be used in classroom environments with a more academic focus. The ECERS-R only dedicates one item out of 43 to science.

A similar scale to the ECERS-R is the School-Age Care Environment Rating Scale (SACERS; Harms, Jacobs, & White, 1996). The SACERS scale consists of 49 items arranged in seven subscales. Like the ECERS-R, the subscales are assigned a score from 1 to 7 and an overall scale score is calculated by averaging the subscale scores. A score of 1 indicates inadequate quality, 3 indicates minimal quality, 5 indicates good quality and 7 indicates excellent quality. Though the SACERS includes environments serving children from age 5 to 12, like the ECERS-R, it was designed to assess informal child care environments and only dedicates one item to science.

Like the ECERS-R and SACERS, the Classroom Assessment Scoring System (CLASS; Pianta, Karen, Paro, & Hambre, 2008) is a commonly used instrument to observe teacher practices in early childhood classrooms. The CLASS is comprised of three domains and ten dimensions. The dimensions are scored on a scale from 1 to 7. A score of one signifies that all, or almost all, indicators in the low range are present. A score of 7 signifies that all, or almost all, indicators in the high range are present. The dimensions scores are averaged accordingly for an overall domain score in each of the three domains. Unlike the ECERS-R and the SACERS, the CLASS is appropriate for assessing informal childcare settings as well as formal classrooms by measuring teacher performance in infant settings through secondary grades. The focus of the CLASS is to rate the teacher–student interaction in the domains of emotional support, classroom organization, and instructional support. Though these domains play an important role in quality science inquiry, the CLASS does not examine a science-specific dimension.

The Early Language and Literacy Classroom Observation (ELLCO; Smith & Dickinson, 2002) is an assessment instrument used in preschool through third grade classrooms. The ELLCO is comprised of three parts: a literacy environment checklist, a general classroom observation and teacher interview and a literacy activities rating scale. The ELLCO has a proven track record of increasing literacy behaviors in children through the modification of the environment and teacher mediation (Wayne, DiCarlo, Burts, & Benedict, 2007).

Assessing the learning environment is imperative, but equally important is the assessment of the instruction provided to students (Koehler-Hak, 2008). We propose that a science instrument has potential for improving science instructional practices because other instruments such as CLASS, ELLCO, ECERS-R and SACERS have been used effectively to highlight issues in instructional domains such as language and literacy. For example, the CLASS has been used to identify how teacher instructional practices impact children's academic outcomes; specifically noting the positive impact of teachers' use of high-level inferential language (Howes et al., 2008).

In the past, there were global concerns over literacy instruction in the United States. ELLCO was one example of a response to this concern and it appears to be effective not only at evaluating language and literacy skills, but it is also having a global impact on the state of language and literacy instruction. Likewise, many outcome studies of U.S. students indicate that currently we are not doing an adequate job in the domain of science instruction (Carnegie Foundation, 2009; Committee on Science, Engineering and Public Policy, 2007). We believe that our SCIIENCE instrument can have an impact on science instruction because it enables practitioners and researchers to evaluate science teaching.

Basis of SCIIENCE instrument

There are several important principles that underlie the Frameworks and subsequently guided the development of the SCIIENCE instrument. First, the Framework indicates that high-quality K-12 science education should reflect the interconnected nature of science as it is practiced and experienced in the real world. Consequently, Framework-guided teacher assessment should consider those teacher instructional practices (i.e., pedagogical behaviors) that foster students' science and engineering knowledge across three interrelated dimensions. The three interrelated dimensions are children's knowledge and use of disciplinary core ideas (DCIs), crosscutting concepts (CCCs), and scientific and engineering practices (SEPs).

Teachers foster learning of DCIs when they focus on specific content and inquiry practices in three different science domains (life science, physical science, earth and space science) or in engineering design. Teachers vary their discourse, inquiry practices, use of expository text and other instructional behaviors in relation to domain-specific requirements. As an example, the SCIIENCE instrument has separate codes that capture the use of domain-specific vocabulary and the use of expository texts as they are incorporated into the inquiry lesson.

Teachers foster children's understanding of CCCs when they help children recognize basic principles that transcend domains and grade levels. CCCs include: patterns, cause/effect, scale/proportion/quantity, systems/system models, energy/matter, structure/function, and stability/change. As with the DCIs, a CCC is coded by the SCIIENCE coder as it is referenced during the science or engineering lesson.

In order to facilitate children's SEPs, teachers should engage children in meaningful science or engineering investigations that engage children as if they were real scientists or engineers. During the investigations, teachers' behaviors should include (for example) asking children to form hypotheses, carry out experiments, and formulate conclusions. In keeping with the Framework-guided approach to K-12 science learning, the SCIIENCE documents SEPs via codes that capture instructional practices such as (a) asking children to form a hypothesis, (b) requiring children to use data to document experimental results, or (c) asking children to support statements with evidence. In sum, the SCIIENCE was created to function as a standardized instrument for the assessment of the quality of science and engineering lessons by comprehensively capturing teacher's instructional behaviors as they occur within science/engineering instruction across all three dimensions.

SCIIENCE instrument development

With the framework described above, an initial set of teaching behaviors was identified from the *A Framework for K-12 Science Education* (National Research Council, 2007a, 2007b, 2012). These behaviors were assigned code titles, given definitions, and organized into the eight categories defined by the Framework. These codes were established to be frequency counts of the selected teacher behaviors used throughout the lesson.

Once a preliminary coding manual and instrument were developed, the research team selected video clips of early childhood science teaching to test the instrument. Four coders were assigned to individually code a video clip using the instrument. Once completed, the results of each individual's coding were compiled to determine where disagreements among coders occurred. The research team discussed the disagreements until a consensus decision was reached. Codes with ambiguous definitions were redefined. If relevant examples or counter-examples for specific codes were noted in the video clip, they were added to the manual to aid coders in correctly identifying those codes in subsequent videos.

The research team met weekly to repeat this process with new video clips. Issues regarding unclear definitions, problematic codes, or proposed new codes to add to the instrument were often raised and discussed until a consensus decision was reached, and the coding manual was adjusted accordingly. Over time, several new codes were added, problematic codes were removed, and highly similar codes were collapsed into a single code. The version of the SCIIENCE instrument used for this study contains 33 frequency codes (Table 1).

Using the SCIIENCE instrument

Each video is coded in 30-second increments. The teacher receives credit for a code if the corresponding behavior occurred at least once at any time during the 30-second segment. The coder marks all codes observed during the 30-second segment, then moves to the next segment and repeats this process. Not all utterances or activities are coded, and a single utterance or activity may receive multiple codes.

The videotaped science lessons had durations ranging from 20 minutes to over an hour. Each video was edited to provide a 20-minute format that captured the best possible observation of a teacher's science lesson and to standardize the amount of time that a teacher was observed. After a member of the research team videotaped an entire science lesson, two members of an editing team viewed the entire lesson, and edited the lesson to a 20-minute video segment using the following process if the lesson was longer than 20 minutes.

To edit the lesson into a 20 minute duration, two viewers watched the full science lesson and individually divided the video into sections indicating which portion of the science lesson consisted of prior knowledge set-up activities (i.e., the *before* segment), which part of the lesson contained the core science inquiry activities (i.e., the *during* segment), and which part of the lesson focused on discussion of the results of the inquiry (i.e., the *after* segment). It is important to capture *before*, *during*, and *after* segments because certain SCIIENCE codes are more likely to occur in some segments than others. For example, it is more likely that the code *elicit hypothesis* will occur in the *before* segment, while the code *quantitative conclusion* will typically occur when students are analyzing their data in the *after* segment of the science inquiry process.

Based on the proportions of the *before*, *during*, and *after* segments, each viewer independently selected specific points of the lesson within each of the segments. The goals of this process included (a) identifying the video sections with the highest code density and variety to obtain a teacher's best example of instructional practice, (b) maintain the proportion of the *before*, *during*, and *after* to reflect the entire unedited lesson and (c) maintain the continuity of the lesson so that segments were never less than two minutes in length.

Table 1
Description of science practices, codes and code descriptions.

Practice 1: Asking Questions and Defining Problems	
Questions are the driving force of inquiry and investigation in science. Scientific questions are often inspired by curiosity about the world and natural phenomena. In engineering, questions seek to identify problems and needs within society, which lead to the designing of technological solutions. Students should be encouraged to ask questions in science that reflect their curiosity and gaps in their knowledge of the content under investigation. These questions can be used to develop inquiry investigations and guide students' thinking during their explorations.	
1.a Prior Knowledge	Teacher asks students to recall previous knowledge from outside the current classroom. Students are asked to think about their prior learning or past experiences and this knowledge is then connected to the lesson.
1.b Elicit Hypothesis	Teacher asks students to predict the outcome of a situation, either in preparation for an experiment or as part of a hypothetical discussion.
1.c Student Idea	Teacher uses student ideas or suggestions to shape an activity or discussion. When students propose an idea for an experiment that was not planned or ask an off-topic question, the teacher accommodates and incorporates the idea into the lesson.
1.d Misconception	Teacher does not immediately correct student's misconceptions or otherwise incorrect answers. Instead of emphasizing right and wrong, emphasis is placed on using inquiry or questioning to help students recognize their misconceptions on their own.
Practice 2: Developing and Using Models	
Scientific models are visual representations of objects or phenomena. They are often used to help visualize objects or processes that cannot otherwise be directly observed and to help explain those processes. They enable scientists and engineers to examine a system or parts of a system and to communicate their explanations of these systems to others. Students should become familiar with the function of scientific models, be able to analyze and interpret those models, and learn to construct their own models of scientific phenomena.	
2.a Student Model	Teacher asks students to create models to represent scientific concepts or processes. Models can be two-dimensional (e.g., drawings) or three-dimensional (e.g., constructed from physical materials).
2.b Model Discourse	Students are engaged in discourse about an existing model. This discourse may include interpreting the model, explaining the scientific concept demonstrated within the model, or using the model to make predictions or draw conclusions.
Practice 3: Planning and Carrying Out Investigations	
Scientific investigations have two primary purposes: to systematically describe the natural world and to test theories and explanations of how the world works. The first requires the use of careful observation to obtain information and often identifies questions that must be further explored. The second requires controlled experiments that seek to isolate specific variables and obtain data to support or contradict a hypothesis. In engineering, investigations are used to test and evaluate the quality of a design. Students should gain experience in carrying out scientific investigations as a means of obtaining information and answering their scientific questions. They should become familiar with the scientific processes underlying the inquiry process.	
3.a Information Gathering	Teacher designs activities in which students are focused on obtaining data for a specific purpose. The purpose could be to test a hypothesis or design or to answer a specific question.
3.b Test Hypothesis	Teacher designs activities in which previously generated hypotheses are tested. The data obtained during the investigation will be used to support or reject the hypothesis.
3.c Equipment	Teacher provides task-specific equipment or tools for students to use to aid in inquiry activities or information gathering.
3.d Test Solution	Teacher designs activities in which previously generated solutions are tested. The solutions will be evaluated for the quality of their performance.
3.e Teacher Demonstration	Teacher demonstrates an inquiry activity for students. This can either be done as a preview for students before they perform an activity themselves or the entire activity is conducted by the teacher.
3.f Student Inquiry	Students are engaged in an inquiry activity that explores a scientific concept using a hands-on approach.
3.g Observation	Teacher encourages students to closely observe an object, a phenomenon, or their surroundings. Observation can include any of the five senses. This can occur through a statement directing a student's attention or through a question that asks students to describe what they are currently observing.
Practice 4: Analyzing and Interpreting Data	
In order to gain meaning from their investigations and experiments, scientists must consolidate and interpret the data they obtain. This allows scientists to identify patterns, examine potential relationships among variables, and determine whether or not the data supports a hypothesis. Engineers analyze experimental data in order to evaluate the quality of a design. Students should become familiar with the process of interpreting the results of their investigations in order to derive deeper meaning and greater understanding of the scientific concepts being examined. They should learn to make connections between multiple pieces of information and begin to understand the broader concepts and theories underlying a given topic.	
4.a Analysis/Interpretation	Teacher leads students to consolidate and interpret the results of their investigations. This includes recognizing patterns, comparing and contrasting objects, and determining whether data supports a hypothesis.
4.b Overarching Relationships	Teacher encourages students to recognize relationships among concepts to obtain a "big picture" view of the underlying principles.
4.c Move Past Misconceptions	Teacher uses strategies to help students move past their misconceptions. Discussions or results of experiments are used to get students to recognize their mistakes and resolve them.
Practice 5: Using Mathematics and Computational Thinking	
Mathematical concepts are integral to the fields of science and engineering. Scientists and engineers often collect measurements, perform statistical analyses of quantitative data, develop formulas to explain phenomena, and use those formulas to make predictions or construct designs. Students should become familiar with obtaining and analyzing numerical data and working with mathematical formulas, when applicable. Students can use numbers to identify and describe patterns and express relationships.	
5.a Numerical Summary	Teacher or students obtain numerical data and consolidate, organize, and/or analyze this data. This can also include statistical analyses of the data, such as computing an average.
5.b Graphical Summary	Teacher or students create a graph of data collected for interpretation and analysis. This includes bar graphs, pie charts, scatter plots, etc.
5.c Quantitative Conclusion	Teacher guides students to draw conclusions from numerical or graphical summaries of data.

Table 1 (*Continued*)

Practice 6: Constructing Explanations and Designing Solutions	
Scientists develop theories to explain the functions of the natural world. These explanations are based upon numerous sets of data and are often evaluated against further data to either support the explanation or suggest the need to revise it. Engineers use this scientific knowledge to design solutions to address various problems or needs. Students should practice developing their own explanations for observed phenomena based on scientific knowledge and evidence and to identify flaws in their explanations when they are inconsistent with evidence. Students should be able to apply their knowledge to develop designs and solve potential problems.	
6.a New Situation	Teacher helps students relate previously-learned concepts to new content. Students are asked to apply their knowledge from previous lessons to new situations, such as solving hypothetical problems.
6.b Explanation	Teacher asks students to generate their own explanations for observed or hypothetical phenomena. These are likely to be "how" or "why" questions that seek to obtain an explanation of a scientific process.
6.c Design Solution	Teacher provides students with a situation or problem and asks students to generate potential solutions. This could be a hypothetical discussion or a prelude to an activity in which students will test their solutions.
6.d Evaluate Understanding	Teacher encourages students to use metacognitive strategies to evaluate their own understanding of a concept. Students may be asked to evaluate how well they understand a concept, recognize misconceptions or flaws in their thinking, or judge their level of success or failure in an activity.
Practice 7: Engaging in Argument From Evidence	
Argumentation is common in scientific circles. Scientists may develop competing explanations and theories, especially when theories are new or when there is little information available or it is difficult to obtain. Scientific claims require support from evidence and reasoning, and arguments must be evaluated for validity and consistency. When students disagree in science, they should be encouraged to support their arguments with evidence, prior knowledge, or logical reasoning.	
7.a Disagreement	Teacher encourages and accepts multiple conflicting answers, ideas, or explanations from students. Student answers are not judged as right or wrong and emphasis is placed on obtaining multiple viewpoints.
7.b Evidence	Teacher asks students to support statements or conclusions with evidence, knowledge, or reasoning. Students are encouraged to provide support for their own thoughts and ideas.
Practice 8: Obtaining, Evaluating, and Communicating Information	
Because science consists of vast amounts of information that is continually being expanded, scientific knowledge must be recorded, communicated, and read by any who would seek to learn about or contribute to the field. Since scientific literature can be complex and highly technical, students should be exposed to it early and often in order to become familiar with its style and to develop their scientific vocabulary. Students should also learn what resources are available to obtain scientific information, how to use those resources and which of those resources are valid and unbiased sources. Students should also practice communicating their own scientific knowledge – orally and in writing – clearly, concisely, and comprehensively.	
8.a Documentation	Teacher or students record information generated during the lesson on paper, a chalkboard, or some other medium. Relevant information must be student-generated, such as students' ideas or discussions, or data obtained from an experiment, such as measurements or observations. This information can be in the form of verbal writing, numerical data, or a drawing.
8.b Vocabulary	Teacher uses appropriate science vocabulary in context during a lesson, rather than simply defining the word or asking students for a definition. The teacher reinforces the vocabulary word for students by using it in context throughout the lesson.
8.c Open-ended Question	Teacher asks questions that encourage students' own thoughts and ideas. Questions that ask students to choose from a set of pre-selected options are NOT open-ended.
8.d Sequenced Questions	Teacher uses multiple questions on a particular topic to lead students to higher levels of thinking or to move from general to specific concepts. There must be a student response between questions, and follow-up questions build on students' answers to previous questions.
8.e Clarification	Teacher asks the student to restate an unclear response or to elaborate on an incomplete response with questions such as "What do you mean?" or "Can you say more about that?" Clarification questions do not ask specifically for new information, but ask students to clarify or elaborate on their own.
8.f Expository Text	Teacher incorporates expository text into the science lesson. The expository text may be print or electronic.
8.g Technology	Teacher has students use electronic devices, such as computers or iPads, during the science lesson, or the teacher uses the technology with student involvement.
8.h Assessment	Teacher uses formative assessments during the lesson to evaluate student understanding, either through questioning or by asking students to submit a product for evaluation.

As an example of this editing process, if the teacher's entire 40-minute lesson consisted of 15 minutes of *before* activity, 20 minutes of *during*, and 5 minutes of *after* activity, the 20 minute edited version would consist of 7.5, 10, and 2.5 minutes from each respective activity. To retain continuity, additional video was selected, if needed, to provide necessary context and/or to reach the minimum of a two-minute duration for the *before*, *during* and *after* phases of the inquiry process. After independent selection of segments, the two viewers came together and discussed their decisions and reached a consensus for what segments should be included in the 20-minute video.

It is important to note that the majority of codes within the SCIIENCE instrument are focused exclusively on teacher behavior, not on student behavior or responses. For example, if a student offers a prediction (e.g., student says, "I think the ice cube will float in water"), this would not be coded as *elicit hypothesis*; the code would only be assigned if the teacher explicitly asks students to make a prediction (e.g., teacher says, "What do you think will happen when I drop this ice cube in the water?"). This ensures that a teacher only receives credit for his/her own behaviors. Conversely, a teacher will receive credit for asking a good question, even if students do not respond.

Participants

Participants in this study were eight female teachers who participated in a two-week National Science Foundation funded summer PD program for improving early childhood science teaching. The teachers taught preschool (2), kindergarten (1), 1st grade (2), 2nd grade (1), and 3rd grade (2). All six grade-school teachers taught in an urban public school system while both preschool teachers taught at a preschool that was affiliated with an

urban university. On average, the teachers had 15 years of teaching experience (SD = 8). One of the participants was a support teacher for a program designed in collaboration with the local urban university and her school district to improve teaching and learning in K-6 science education. Two of the participants were involved in their district's review and alternative compensation system, which was intended to promote teacher quality while improving the academic performance of students. One of those participants was also a peer math coach while the other was a peer literacy coach. Each participant was videotaped in her classroom in the spring before the PD and again in the fall following the PD to measure the impact of the PD on teaching quality. Inter-rater reliability and validity measures presented in the following sections are obtained from coding videos from six of the eight teachers from the fall after attending the PD sessions. One of the remaining two videos could not be used due to failure to obtain parental permission for capturing children on video, and the second was used to verify our process for editing videos down to a 20-minute length.

Table 2
SCIIENCE inter-rater reliability statistics.

Segment	MT1	MT3	MT4	MT6	MT7	MT8
Cohen's k[a]	.905	.966	.573	.830	.389	.792
PABAK[b]	.962	.985	.811	.947	.712	.909
P_o[c]	.981	.992	.905	.973	.856	.955
P_e[d]	.801	.776	.778	.844	.764	.781
P_{pos}[e]	.915	.971	.627	.844	.457	.818
P_{neg}[f]	.989	.996	.946	.986	.917	.974
PI[g]	.777	.742	.746	.830	.735	.750
BI[h]	.019	.008	.011	.019	.106	.008

[a] Cohen's kappa.
[b] Prevalence-adjusted bias-adjusted kappa.
[c] Observed proportion of agreement.
[d] Expected proportion of agreement.
[e] Proportion of positive agreement.
[f] Proportion of negative agreement.
[g] Prevalence index.
[h] Bias index.

Inter-rater reliability

The inter-rater reliability of the SCIIENCE was assessed by having two independent coders double-code a minimum of 20% of videotaped science lesson segments from six teachers in PK-3 classrooms. Cohen's kappa coefficients were calculated for each of the video segments to compare the ratings of the two coders. Cohen's kappa is a robust measure of inter-rater agreement because Cohen's kappa accounts for agreement by chance. Agreement by chance refers to the random probability that two coders will assign the same code during a 30-second segment.

However, the interpretation of Cohen's kappa is impacted by a significant prevalence trend in the data. Prevalence is the difference in the number of positive agreements when the behavior is observed by both coders and negative agreements when the behavior is not observed by both coders in relation to the total number of coding opportunities during the videotaped lesson. Specifically, the data were skewed in that it was 12 times more likely to have a negative agreement between the two coders as compared to a positive agreement. This was likely due to the observation of at most one to three codes during a 30-second increment out of the total number of SCIIENCE codes that could potentially be observed during any given segment.

To compensate for the significant prevalence in these data, prevalence-adjusted bias-adjusted kappa (PABAK) coefficients were also calculated. PABAK equally distributes the positive and negative agreements (yes/yes and no/no) and disagreements (yes/no and no/yes) between coders so that the agreement expected by chance becomes 0.50 (Byrt, Bishop, & Carlin, 1993). Cohen and PABAK kappa values over 0.80 were interpreted as almost perfect, 0.61–0.80 as substantial, 0.41–0.60 as moderate, 0.21–0.40 as fair, 0.00–0.20 as slight, and below 0.00 as poor (Landis & Koch, 1977). Following the suggestion of Cunningham (2009) we also documented, for each video segment, the observed proportion of agreement (P_o), the expected proportion of agreement (P_e), the proportion of positive agreement (P_{pos}), the proportion of negative agreement (P_{neg}), the prevalence index (PI), and the bias index (BI).

The inter-rater reliability statistics can be found in Table 2. Using the selected guidelines, three of the six video segments had Cohen's kappa values in the "almost perfect" range while all but one of the video segments had PABAK values in the "almost perfect" range. With the exception of one video segment, the calculated bias values are less than 0.05, suggesting that disagreements were random and coders did not preferentially code for presence or absence of a code when in disagreement. As expected, all video segments had high prevalence values given the sparse occurrence of many codes. These high prevalence values did create a substantial drop in Cohen's kappa values on video segments with PABAK values below 0.9. Higher levels of prevalence are known to produce low Cohen's kappa values when some disagreement between coders is observed (Byrt et al., 1993). Therefore these results suggest that SCIIENCE scores are reliable across different coders at high levels of agreement (Landis & Koch, 1977) when prevalence is accounted for.

Validity

The convergent validity of the SCIIENCE is a measure of how well SCIIENCE relates to other instruments that are used to observe a similar construct. In this case, the validity of SCIIENCE was assessed by analyzing its relationship with the Classroom Assessment Scoring System (CLASS) and the Horizon Local Systemic Change Classroom Observational Protocol (Horizon). The CLASS provides a framework for observing dimensions of classroom processes such as emotional and instructional support that contribute to the quality of PK-3 classroom settings (Paro, Pianta, & Stuhlman, 2004). The CLASS scales are rated on a 7-point Likert-type scale (1–2 = low levels of observed construct; 3–5 = moderate levels; 6–7 = high levels).

The Horizon instrument was developed to observe K-12 science or mathematics classrooms and measure the quality of the lesson design and implementation, mathematics and science content, classroom culture, and the likely impact of instruction on student understanding (Horizon Research Inc., 2000). We used the capsule rating in the Horizon instrument that is a global 5-point Likert rating of teacher effectiveness during science instruction (1–2 = low rating, 3 = moderate rating, 4–5 = high rating).

To establish the convergent validity of the SCIIENCE instrument, correlations were computed using the following data: (1) the average number of behaviors observed with SCIIENCE per 30-second video segment; (2) the total number of different behaviors observed with SCIIENCE over the course of a 20-minute video; (3) the Instructional Support domain scores from the CLASS, which include the dimensions of Concept Development, Quality of Feedback, and Language Modeling; (4) the average of the Instructional Support domain and Instructional Learning Formats dimension scores from the CLASS; (5) the capsule description scores from Horizon; and (6) the weighted capsule description scores from Horizon, with "low 3," "solid 3," and "high 3" teacher ratings being given values of 3.00, 3.33, and 3.67 respectively. The correlations can be found in Table 3. Overall, the correlation between the number of SCIIENCE codes observed and the relevant measures of CLASS and Horizon were moderate to strong, showing

Table 3
SCIIENCE criterion validity correlations.

	1	2	3	4	5	6
1. Average SCIIENCE codes[a]	–					
2. Different SCIIENCE codes[b]	.901*	–				
3. CLASS IS score[c]	.574	.741	–			
4. CLASS IS + ILS score[d]	.584	.714	.995***	–		
5. Horizon score[e]	.527	.603	.280	.273	–	
6. Weighted Horizon score[f]	.663	.698	.531	.546	.933**	–

[a] Average number of behaviors observed with SCIIENCE per 30-second video segment.
[b] Total number of different behaviors observed with SCIIENCE per 30-minute video.
[c] CLASS Instructional Support domain score.
[d] Average of CLASS Instructional Support and Instructional Learning Formats scores.
[e] Horizon capsule description score.
[f] Weighted Horizon capsule description score.
* $p < .05$, two tailed.
** $p < .01$, two tailed.
*** $p < .001$, two tailed.

that SCIIENCE is capturing similar aspects of the classroom instruction as CLASS and Horizon.

The information above describes the validity processes completed with the SCIIENCE at the current pilot level. In the next stage of development, validity measures will be undertaken with (a) more teachers from the current urban cohort, (b) teacher cohorts from geographically diverse urban communities, and (c) teacher cohorts from rural and suburban communities.

In the educational assessment community, there is debate regarding assessment validity (Moss, Girard, & Haniford, 2006). Some experts fear that a prescriptive approach to validity documentation limits the generative context and flexibility required for meaningful scientific investigation. Accordingly, Moss (2007) argues that researchers should develop an overall plan for validity work that uses a more comprehensive and flexible approach to guide decision-making regarding what evidence to pursue. Moss's position draws upon the seminal work of Cronbach (1988, 1989), Shepard (1993), and Kane (1992, 2006); she suggests that validity work should be built across multiple sources of evidence. Following this principle, we propose a number of validity projects to guide our research as we systematically move from pilot projects to research scale-up. Table 4 lists a range of validity measures that will be pursued. This thorough and systematic approach to establishing validity will allow the research team to develop, analyze, and integrate multiple types of evidence at varying levels of scale (Moss et al., 2006).

Application to professional development

Our original motivation for the development of SCIIENCE instrument was to assess the efficacy of the PD we developed for the teaching of science in PK-3 classrooms based on the Framework. After confirming the reliability and validity of the

Table 4
Validity measures for the SCIIENCE instrument.

	Research plan	Rationale	Interpretative arguments/hypotheses
1.	Complete factor analysis of 300+ teachers pre–post-PD inquiry lessons as scored by the SCIIENCE instrument.	To investigate patterns of content validity through examination of factorial loading patterns.	Factor loading should align with pre-existing theoretical notions of inquiry behaviors as described in NGSS/Framework. Factor loading will vary in relation to predictable patterns. Evaluate data in response to hypotheses such as (a) factorial loading patterns will be different across grade bands in that teachers introduce different inquiry behaviors in response to students' developmental levels and (b) patterns of SCIIENCE codes will vary in response to science/engineering content in theoretically predictable ways.
2.	Calculate statistical correlations between SCIIENCE coding tool and other measures of instructional quality at a scale-up level (300+ teachers).	To investigate convergent validity by examining relationships between SCIIENCE instruments and other accepted measures of teacher instructional quality (e.g., Horizon, RTOP, CLASS [CLASS items that rate teacher strategies that foster content knowledge and facilitate student's use of complex verbal communication skills])	There should be significant (but not perfect) correlations between SCIIENCE and other assessments of teacher instruction. Since SCIIENCE is unique from other instruments, we hypothesize it captures specific aspects of teacher inquiry instructional practice not captured in other tools.
3	Calculate statistical relationships between SCIIENCE with other school district measures of teacher quality.	To investigate predictive validity by comparing the SCIIENCE results with school-district teacher quality measures (e.g., STAR[a])	Teachers who show significantly more inquiry codes will be rated as more effective teachers via school-district teacher assessments.
4	Ask expert science educators to view teacher videos and rate teachers on a scale from "highly effective" to "ineffective."	To investigate convergent validity by determining if SCIIENCE assessment differentiates groups of teachers in alignment with opinion of experts who will be nationally recruited (K-12 science educators or science teacher educators).	SCIIENCE codes will differentiate groups of teachers who are more versus less effective.
5	Calculate statistical relationships between SCIIENCE with another teacher assessment measure that is not theoretically aligned with SCIIENCE coding tool (e.g., a measure of teachers' classroom behavior management).	To investigate discriminative validity by comparing a measure of teachers' behavioral management (e.g., CLASS items related to behavioral management) with results from SCIIENCE assessment.	SCIIENCE coding instrument captures unique inquiry-related teacher instructional patterns that differ from a different teacher instructional practice domain such as behavior management.

[a] *STAR; Renaissance Learning* (Renaissance Learning, 2009) consists of three different computer adaptive assessments. The **STAR-Early Literacy** assesses PK-3 students along three domains: word facility and skills, comprehension strategies and constructing meaning, and numbers and operations. The **STAR-Reading** is designed for independent readers through grade 12 along five domains: word knowledge and skills, comprehension strategies and constructing meaning, analyzing literary text, understanding author's craft, and analyzing argument and evaluating text. The **STAR-Mathematics** is designed for students in grades 1–12 along four domains: numbers and operations, algebra, geometry and measurement, and data analysis, statistics and probability.

SCIIENCE instrument, we used SCIIENCE data to compare instructional practices of teachers' science lessons delivered during Spring prior to our initial Summer 2-week PD institute (Summer, 2012) and during the subsequent Fall following completion of the Summer PD. Pre- and post-PD inquiry behaviors were compared using a paired *T*-test across six teachers. Results using the SCIIENCE instrument shown in Table 5 indicated that our teachers did not show the anticipated degree of change in inquiry practices following PD. Teachers did demonstrate a statistically significant increased use of open-ended questions, asking children to support their statements or conclusions with evidence, knowledge or reasoning, and designing activities to test previously designed solutions. However, teachers were less likely to obtain information from expository texts, incorporate student's ideas into the inquiry process, elicit and test hypotheses, and engage students in hands-on inquiry lessons.

These results guided the redesign of subsequent years of PD to more positively impact teacher instructional practices. Specifically, in the first pilot year of the project described (2012), teachers were exposed to inquiry practices via exposure to various disciplinary core ideas in life science, physical science, earth and space science. Much of the focus during the first year PD was on discourse practices such as open-ended questions and sequenced questions. After reviewing the 2012 results of the pre-post SCIIENCE data, we reflected that moving across content domains may not have afforded teachers to deeply explore a particular content area which may have limited their exposure to the higher-level inquiry practices such as (a) analyzing and synthesizing ideas, (b) compiling and analyzing data, and (c) supporting claims with evidence.

Table 5
Comparison of SCIIENCE code frequencies before and after PD.

SCIIENCE code	Pre-PD videos	Post-PD videos
1.a Prior knowledge	2.5	4.6
1.b Elicit hypothesis	3.8**	0.0
1.c Student idea	3.3**	0.0
1.d Misconception	6.3	5.4
2.a Student model	3.3	2.1
2.b Model discourse	0.0	0.0
3.a Information gathering	10.4	9.6
3.b Test hypothesis	5.4**	0.0
3.c Equipment	4.6	5.0
3.d Test solution	1.3	5.4*
3.e Teacher demonstration	2.9	2.9
3.f Student inquiry	40.0***	18.3
3.g Observation	16.7	16.3
4.a Analysis/interpretation	3.8*	0.8
4.b Overarching relationships	0.8	0.0
4.c Move past misconception	1.3	1.7
5.a Numerical summary	0.0	0.0
5.b Graphical summary	0.0	0.0
5.c Quantitative conclusion	0.0	0.0
6.a New situation	0.0	0.0
6.b Explanation	7.1	11.7
6.c Design solution	1.3	3.3
6.d Evaluate understanding	0.0	0.0
7.a Disagreement	2.1	4.6
7.b Evidence	1.3	6.7**
8.a Documentation	32.9	27.5
8.b Vocabulary	32.5	28.3
8.c Open-ended question	49.6	68.3***
8.d Sequenced questions	13.8	18.3
8.e Clarification	2.1	1.3
8.f Expository text	4.2**	0.0
8.g Technology	0.0	0.0
8.h Assessment	0.0	0.0

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

Accordingly, second year of our summer PD was redesigned ($N = 40$ teachers). This revised PD was designed to have teachers explore one science content domain across the entire two-week PD. During this deep exposure, our redesigned PD emphasized the full trajectory of inquiry practice including analysis and synthesis of information at the end of the science and/or engineering inquiry lesson. Importantly, we took care to expose teachers to both discourse and non discourse-based inquiry practices. For example, we implicitly exposed teachers and explicitly asked teachers to foster hypothesis formation and argumentation, use models, and evaluate and analyze data, etc. We are currently in the process of evaluating the pre- and post-SCIIENCE data results of the 40 teachers who participated in this second summer PD session. Second year teacher change in instructional practices will be compared to first year data. We believe the use of the SCIIENCE instrument and the pre-post comparison of teacher instructional practices can continue to guide development of subsequent high-quality PD.

Results and discussion

Given the urgency to improve student proficiency in science inquiry, it is very important to develop reliable observational systems for documenting the use of high-quality instructional practices by teachers, especially teachers of our youngest students. Furthermore, such observational systems can provide insight into which educational practices lead to improved outcomes for students in science classrooms. Experts have underscored the word observational systems (Hill, Charalambos, & Kraft, 2012); observational systems should include a number of elements provided in the SCIIENCE instrument. Critical elements of SCIIENCE system include the development of a theoretically strong and reliable instrument to observe teacher behavior in a specific domain, a specified rater training program, and a scoring system that allows clear interpretation of observed instructional behaviors.

The SCIIENCE instrument was developed in keeping with the specific instructional behaviors that are identified in the Framework. Therefore, the SCIIENCE instrument should be useful for research and teacher development. Since there has been limited research regarding early childhood instruction, and to our knowledge no observational instruments focusing on preschool science instruction, it is imperative that well-designed instruments document the level at which inquiry teaching behaviors occur in the early childhood classroom.

The reliability and validity data presented in this paper demonstrate the psychometric properties of the SCIIENCE instrument. We investigated the reliability of the SCIIENCE instrument by computing inter-rater reliability. Data analyses revealed high inter-rater reliability and demonstrated that independent coders identified similar instructional behaviors. This is notable since experts have found that reliability within other observational measures of science instruction was poor (Henry et al., 2009). In contrast, the data from this study shows that if the SCIIENCE instrument is used by trained observers, this instrument will accurately document the ability of teachers to implement specific inquiry practices in their classrooms.

The validity of the SCIIENCE instrument was evaluated by comparing an individual teacher's frequency of SCIIENCE codes with his or her score on two other measures of teacher instructional effectiveness. The CLASS – Pre-K was designed to measure instructional effectiveness of early childhood educator's instructional quality specifically in the realm of language and literacy. Importantly, there is no direct correspondence between instructional skills required to effectively improve a child's language and literacy ability and the instructional practices that

occur in high-quality science instruction as prescribed by the Framework. For example, the teacher is rated on a scale from 1 to 7 for the language development domain in the CLASS scoring system with regard to how well he or she uses "self-talk or parallel talk". This is defined as the degree to which the teacher maps his or her own actions or the student's actions with language and description. We would not predict that this behavior would necessarily be highly correlated with the SCIIENCE inquiry behaviors. On the other hand, we would predict that some instructional behaviors identified in the CLASS would be highly correlated with the instructional practices identified on the SCIIENCE instrument. An example of this type of instruction behavior includes "prompting thought processes," in which a teacher asks students to explain their thinking.

Therefore, the CLASS is an imperfect match for determining the validity of the SCIIENCE instrument, but it is one of the few high reliability observational instruments developed for use in early childhood classrooms. Importantly, also, we can hypothesize that a teacher who is highly rated for instructional effectiveness during language and literacy teaching is likely to be more proficient when conducting science inquiry as compared to a low-performing CLASS-rated teacher. With this in mind, these data suggest that there is an underlying construct of teaching effectiveness that is demonstrated by the moderate-level correlations between the CLASS and the SCIIENCE observation systems.

The Horizon quantitative score used in this analysis was a global capsule rating on a 5-point scale; the Horizon has been used to evaluate quality of science teaching in kindergarten to high-school level classrooms. Horizon differs from the SCIIENCE instrument in that (a) it is more subjective and (b) the SCIIENCE instrument used in this study quantifies specific inquiry behaviors instead of providing a global quality rating. Again, although an imperfect match for the SCIIENCE instrument, the moderate-level correlations are promising and in the predicted direction.

The SCIIENCE instrument and this study have limitations. First, this study was conducted with 6 teachers. As we hypothesize that training in science pedagogy results in changes in teacher rating with the SCIIENCE instrument, studies of more teachers with a range of experience and training will allow us to determine if the SCIIENCE instrument captures differences in teacher practice in predicted ways. Second, use of this instrument takes training and coders must have demonstrated a sufficient level of reliability prior to using this assessment instrument.

Third, as mentioned in our reliability data, the data are skewed in that there are many more negative agreements than positive agreements. Although this imbalance can be accommodated with specific statistical procedures, the issue should be acknowledged and does affect the interpretation of our reliability measures. Finally, and most importantly, the codes were developed because the Framework and the currently accepted best practices suggest that certain behaviors (i.e., codes) should occur during science inquiry activities. However, in the sample we have considered, and in our examination of other publically available science lessons such as FOSS videos, only a small range of codes are observed. Therefore, future research is planned to monitor changes in teacher behavior prior to and following professional development in implementing high quality science instruction. This work will facilitate our exploration of if and when teachers are able to demonstrate the full range of instructional practices identified by the SCIIENCE instrument.

Given the results of this study, we are currently augmenting the SCIIENCE instrument for future use. As described above, we developed a research plan to evaluate the validity of the instrument beyond the project-level. We anticipate that many of the frequency codes listed in Table 1 tested for reliability and validity in the present study will remain in future versions. However, we converted some frequency codes into binary codes to indicate a teacher behavior has occurred anywhere throughout the lesson, rather than recording at which time during the lesson the behavior occurred. Examples of this include 3.b Test Hypothesis, 3.c. Equipment, 5.1 Numerical Summary and 5.2 Graphical Summary. This change reflected the time-spanning nature of these activities that typically extended across segments during high quality lessons.

Although our results show good agreement with existing instruments such as CLASS and Horizon, we are also developing six global codes that will be used to rate the following overall qualities of the science lesson on a four-point scale. These ratings are designed to encompass features from other instruments that support practices highlighted in the Framework, and include Elicit Student Thinking, Balance of Teacher/Student Talk, Encourage Participation from All Students, Quality of Questions, Quality of Inquiry Activities, and Use of Discourse Techniques. We hypothesize that the correlations between the SCIIENCE global ratings and the CLASS and Horizon instruments will be higher than the current level of correlation between these systems and our current set of frequency codes. We also anticipate that as we validate the instrument beyond the project level, we will continue to improve and fine-tune this instrument.

This pilot indicates the SCIIENCE instrument is an important first step to facilitate implementation of science education principles identified by the Framework. The strengths of this instrument include objective descriptions of each of the coded SCIIENCE behavior, a rich data source that results from coded SCIIENCE observations, the ability of the SCIIENCE instrument to contribute to both professional development and teacher evaluation, and the utility of the SCIIENCE instrument for future research. We also believe that the process by which we developed this instrument can serve as a model for the development of future systems, and can serve to provide an opportunity for the science education community to provide input as to the best practices that should be measured in science classrooms.

The SCIIENCE instrument is different than many other teacher evaluation systems. Global rating is useful for general teacher evaluation, but global rating systems do not allow the teacher or the instructional coach to consider a micro-examination of teaching behaviors. An objective description of each instructional practice identified by the SCIIENCE instrument help the teacher and the instructional coach know exactly what is and is not taking place during each lesson. Furthermore, the analysis of the lesson can include not only WHAT occurs, but allows the teacher to investigate WHEN the behavior occurs and in what CONTEXT. For example, a teacher can begin to think about how quickly he or she moves into fostering high level inquiry during the lesson, what combinations of instructional behaviors are likely to occur in sequence (e.g., does a series of open-ended questions lead to an opportunity to support student argumentation?), and to help the teacher monitor changes in patterns of instructional practice over time. This information also can be used for purposes of teacher evaluation.

The level of detail and the number of instructional practices in the SCIIENCE instrument that are identified during an instructional period result in a rich data source. The frequency counts obtained in the SCIIENCE system are more amenable to statistical analysis as compared to ordinal ranking data. Also, we hypothesize that the SCIIENCE frequency codes will be more sensitive to small changes in teacher instructional practice. In contrast, documenting changes with an ordinal ranking system can be challenging. The frequency codes also will permit a fine-grained analysis of how varying inquiry practices may be more or less likely to occur within explorations of different science content. For example, we could explore whether documentation occurs more frequently in the

exploration of physical science as compared to life science. We will also be able to explore the differences in code occurrences across age groups. For example, we could investigate the frequency of modeling occurrences in a preschool classroom as compared to a 3rd grade classroom.

The objective nature of the codes and the data richness in the SCIIENCE instrument make it an important instrument for monitoring change in teacher behavior in relation to professional development and research programs investigating the effects of specific inquiry practices. In our current grant program, we intend to determine if specific PD experiences result in changes in code frequency or variety. This information will be helpful for documenting the effectiveness of our PD activities. Finally, and most importantly, we intend to use the SCIIENCE instrument to evaluate the effectiveness of specific teaching behaviors on child learning outcomes. We hypothesize that student learning outcomes will be enhanced as teachers implement a higher frequency and variety of inquiry practices.

Teacher preparation programs are now focusing on preparing teacher candidates to teach within the Framework and NGSS. Additionally, significant professional development efforts will ensue over the next few years to guide teachers to meet the expectations of the Framework and NGSS. A valid and reliable instrument is needed to help determine how and when teachers use specific patterns of instructional practices during inquiry instruction. As we have demonstrated in our analysis of the effects of our previous summer PD, the SCIIENCE instrument facilitates a granular analysis of teacher's instructional practices promoting a systemic approach to analyzing the effectiveness of our summer institutes and of our academic year coaching activities. Using the SCIIENCE instrument, we can determine exactly which instructional practices do – or do not – change as a result of PD. While some practices appear to be more readily adopted by teachers, other science practices appear to be more difficult to learn and integrate into classrooms. With systematic analysis, we believe we can fine-tune our PD to achieve optimal outcomes in teacher practice.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.stueduc.2015.03.003.

References

Byrt, T., Bishop, J., & Carlin, J. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*, 423–429.

Cantrell, S., & Kane, T. K. (2013). *MET project: Ensuring fair and reliable measures of effective teaching*. Seattle: Bill & Melinda Gates Foundation.

Carnegie Foundation (2009). *The opportunity equation: Transforming mathematics and science education for citizenship and the global economy*. New York: Institute for Advanced Study.

Committee on Science, Engineering, and Public Policy (2007). *Rising above the gathering storm: Energizing and empowering America for brighter economic future*. Washington, DC: The National Academies Press.

Chapin, J. R. (2006). The achievement gap in social studies and science starts early: Evidence from the early childhood longitudinal study. *Social Studies, 97*, 231–238.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measure, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.

Cunningham, M. (2009). More than just the kappa coefficient: A program to fully characterize inter-rater reliability between two raters. *SAS global forum 2009 conference*. Cary, NC: SAS Institute Inc Paper 242-2009.

De Kruif, R. E., McWilliam, R., Ridley, S. M., & Wakely, M. B. (2000). Classification of teachers' interaction behaviors in early childhood education. *Early Childhood Research Quarterly, 15*, 247–268.

Harms, T., & Clifford, R. (1980). *Early childhood environment rating scale*. New York: Teachers College Press.

Harms, T., Clifford, R., & Cryer, D. (1998). *Early childhood environment rating scale – Revised*. New York: Teachers College Press.

Harms, T., Jacobs, E. V., & White, D. R. (1996). *School age care environment rating scale*. New York: Teachers College Press.

Henry, M. A., Murray, K. S., Hogrebe, M., & Daab, M. (2009). *Quantitative analysis of indicators on the RTOP and ITC observation instruments*. Retrieved from http://www.mahenryconsulting.com/pdf/MA%20Henry%20Quantitative%20Analysis%20RTOP%20%20ITC%20112509%20FINAL.pdf

Henry, M. A., Murray, K. S., & Phillips, K. (2007). *Meeting the challenge of STEM classroom observation in evaluating teacher development projects: A comparison of two widely used instruments*. Retrieved from http://hub.mspnet.org/index.cfm/14975

Hill, H. C., Charalambos, Y. C., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*, 56–64.

Horizon Research, Inc. (2000a). *Validity and reliability information for the LSC classroom observation protocol*. Chapel Hill, NC. Retrieved from http://www.horizon-research.com/LSC/news/cop_validity_2000.pdf

Horizon Research, Inc. (2000b). *Inside the classroom interview and analytic protocol*. Retrieved from http://www.horizon-research.com/instruments/clas/cop.php

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly, 23*(1), 27–50.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Educaton/Praeger.

Keeley, P. (2009). *Elementary science education in the K-12 system*. NSTA Webnews Digest, April 22. Retrieved from http://www.nsta.org/publications/news/story.aspx?id=55954

Koehler-Hak, K. M. (2008). Functional assessment of academic problems: A paradigm shift for improving student outcomes. *The School Psychologist*, 50–54.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

McCormack, A. (2010, December). *It's time for more early childhood science*. NSTA Reports. Retrieved from http://www.nsta.org/publications/news/story.aspx?id=58029

Moss, P. A. (2007). Reconstructing validity. *Educational Researcher, 36*(8), 470–476.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109–162.

National Research Council (2007a). *Ready, Set, SCIENCE!: Putting research to work in K-8 science classrooms*. Washington, DC: The National Academies Press.

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

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

Paro, K., Pianta, R., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten years. *The Elementary School Journal, 104*, 409–426.

Pianta, R., Karen, M., Paro, L., & Hambre, B. (2008). *Classroom assessment scoring system (CLASS) manual, pre-k*. Baltimore: Paul H Brookes Publishing Company.

Pratt, H. (2007). *Science education's 'overlooked ingredient': Why the path to global competitiveness begins in elementary school*. NSTA Express, October 10. Retrieved from http://science.nsta.org/nstaexpress/nstaexpress_2007_10_29.htm

Renaissance Learning (2009a). *STAR Early Literacy: Technical manual*. Wisconsin Rapids, WI: Author Available from Renaissance Learning by request to research@renlearn.com

Renaissance Learning (2009b). *STAR Math: Technical manual*. Wisconsin Rapids, WI: Author Available from Renaissance Learning by request to research@renlearn.com

Renaissance Learning (2009c). *STAR Reading: Technical manual*. Wisconsin Rapids, WI: Author Available from Renaissance Learning by request to research@renlearn.com

Sawada, D., Piburn, M., Falconer, K., Turley, J., Benford, R., & Bloom, I. (2000). *Reformed teaching observation protocol: Technical Report No. IN00-1*. Tempe, AZ: Arizona State University.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450.

Smith, M. W., & Dickinson, D. K. (2002). *Early language & literacy classroom observation (ELLCO) toolkit, research edition*. Baltimore: Brookes Publishing Company.

Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., Boller, K., et al. (2010). *The child care quality rating system (QRS) assessment: Compendium of quality rating systems and evaluations*. Washington, DC: Child Trends DataBank.

Wayne, A., DiCarlo, C., Burts, D., & Benedict, J. (2007). Increasing the literacy behaviors of preschool children through environment modification and teacher mediation. *Journal of Research in Childhood Education, 22*, 5–16.

Wenner, G. (1993). Relationship between science knowledge levels and beliefs toward science instruction held by preservice elementary teachers. *Journal of Science Education and Technology, 2*, 461–468.