4-5-2019

Semi-quantitative Characterization of Secondary Science Teachers’ Use of Three-Dimensional Instruction

Senetta F. Bancroft
Southern Illinois University Carbondale

Deborah Herrington
Grand Valley State University, herringd@gvsu.edu

Roxana Dumitrache
Grand Valley State University

JOURNAL OF SCIENCE TEACHER EDUCATION
2019, VOL. 30, NO. 4, 379–408
https://doi.org/10.1080/1046560X.2019.1574512

Routledge
Taylor & Francis Group

Semi-quantitative Characterization of Secondary Science Teachers' Use of Three-Dimensional Instruction

Senetta F. Bancroft[a], Deborah G. Herrington[b], and Roxana Dumitrache[b]

[a]Department of Curriculum and Instruction, Department of Chemistry and Biochemistry, Southern Illinois University-Carbondale, Carbondale, Illinois, USA; [b]Department of Chemistry, Grand Valley State University, Allendale, Michigan, USA

ABSTRACT

This quasi-experimental study evaluated middle- and high-school science teachers' implementation of three-dimensional (3D) instruction as defined by the *Next Generation Science Standards*. Teachers participated in a long-term professional development (PD) program designed to increase their use of inquiry-based science instruction. We describe our semi-quantitative adaptation of the Educators Evaluating the Quality of Instructional Products: Science rubric version 2 (SQ-EQuIP) to facilitate the longitudinal evaluation of teacher practices with 3D instruction. SQ-EQuIP evaluations revealed that after two years, 80% of PD teachers implemented lessons where students were explicitly and coherently engaged in 3D learning, compared with 22% of comparison teachers. Further, in several cases lesson materials that should support student engagement in 3D learning were not implemented with fidelity. This discrepancy implies that PD developers must use the EQuIP not only to assess lesson or unit plans as intended by its creators, but to also evaluate the implementation of these materials from students' perspective. The small sample size restricts claims of significance. However, observed trends between teacher groups indicate long-established best practices designed to increase teacher use of inquiry-based practices may also positively impact teacher use of 3D instructional practices.

KEYWORDS

Next Generation Science Standards; professional development; three-dimensional instruction; EQuIP

Introduction

The *Next Generation Science Standards* (NGSS) (NGSS Lead States, 2013) are structured around a three-dimensional (3D) framework of teaching and learning. The three dimensions include science and engineering practices (SEPs), disciplinary core ideas (DCIs), and crosscutting concepts (CCCs). When integrated, these dimensions reflect what scientists and engineers do and how they think (NGSS Lead States, 2013), see Figure 1.

The SEPs encapsulate the skills and knowledge students need to investigate, model, and explain the natural world and to design solutions to problems. The DCIs identify core concepts underlying the science and engineering disciplines. They are teachable and learnable across multiple grade levels with increasing complexity (National Research Council [NRC], 2012).

CONTACT Senetta F. Bancroft senetta.bancroft@siu.edu Department of Curriculum and Instruction, Department of Chemistry and Biochemistry, Southern Illinois University-Carbondale, 625 Wham Dr., Mail Code 4610, Carbondale, IL

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uste.

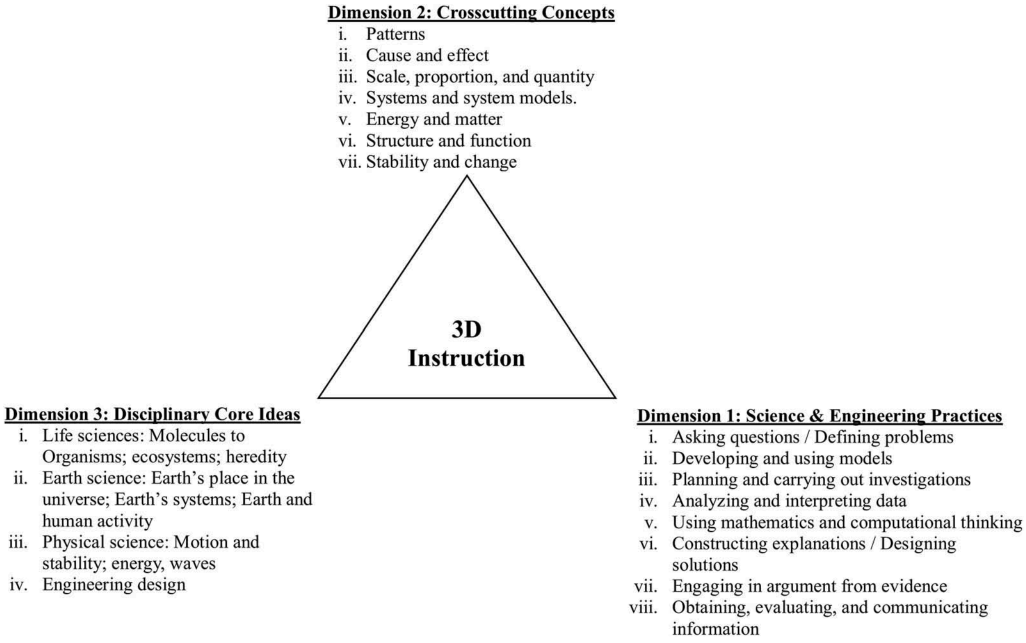

Dimension 2: Crosscutting Concepts
i. Patterns
ii. Cause and effect
iii. Scale, proportion, and quantity
iv. Systems and system models.
v. Energy and matter
vi. Structure and function
vii. Stability and change
3D Instruction
Dimension 3: Disciplinary Core Ideas
i. Life sciences: Molecules to Organisms; ecosystems; heredity
ii. Earth science: Earth's place in the universe; Earth's systems; Earth and human activity
iii. Physical science: Motion and stability; energy, waves
iv. Engineering design
Dimension 1: Science & Engineering Practices
i. Asking questions / Defining problems
ii. Developing and using models
iii. Planning and carrying out investigations
iv. Analyzing and interpreting data
v. Using mathematics and computational thinking
vi. Constructing explanations / Designing solutions
vii. Engaging in argument from evidence
viii. Obtaining, evaluating, and communicating information

Figure 1. Components of the three dimensions of science and engineering instruction as outlined by the NGSS.

The CCCs serve as a unifying link between the SEPs and DCIs through application across all disciplines to enhance the use of the SEPs (NGSS Lead States, 2013). Early versions of CCCs were present in past national science standard policy documents (American Association for the Advancement of Science [AAAS], 1989, 1993) and included systems, models, constancy, patterns of change, and scale. These common themes evolved into the unifying concepts and processes described in the *National Science Education Standards* (NSES) (NRC, 1996) and included systems, order, and organization; evidence, models and explanations; change, constancy and measurement; evolution and equilibrium; and form and function. However, CCCs in their earlier forms (AAAS, 1989, 1993; NRC 1996) were typically ignored by science teachers (NGSS Lead States, 2013). The inclusion of CCCs in the performance expectations signals the need to explicitly support student use of these ideas across concepts within a discipline as well as across multiple disciplines. Ultimately, 3D instruction has the explicit purposes of developing students' appreciation for the nature of scientific knowledge, construction of a deep understanding of science concepts, and application of a diverse set of tools to investigate and solve problems related to natural and human systems. To achieve these purposes, the NGSS address both the traditional exclusion of the CCCs in science instruction and engagement in the SEPs independent of science content (Krajcik, Codere, Dahsah, Bayer, & Mun, 2014).

The NGSS tackle this latter issue by explicitly choosing SEPs over the term "inquiry" because despite decades of policy documents and research, there remained no consensus on the meaning of inquiry (Hayes, Lee, DiStefano, O'Connor, & Seitz, 2016; NRC, 2012). SEPs, in contrast, are the specific actions that scientists and engineers engage in and can be considered as the disaggregated components of scientific investigations. However, the tenets of inquiry-based instruction outlined in the earlier NSES (NRC, 1996) (e.g., engaging in activities to answer scientific questions or supporting arguments with evidence) are preserved in the NGSS within the SEPs (NRC, 2012; NGSS Lead States, 2013). Therefore, the NSES (NRC, 1996), *Project 2061* (AAAS, 1989), *Benchmarks for Scientific Literacy* (1993), and now the NGSS all stipulate that science knowing should not be separate from the skills and tools scientists use. Thus, three-dimensional instruction — although now made more explicit in the NGSS — is not a new concept. Yet, within the 3D framework, many educators are now challenged to rethink the inquiry approach to teaching science content (Krajcik, 2015). This challenge stems from the coherency and increased complexity the 3D framework demands. Naturally, the 3D framework also challenges the criteria by which science teacher instruction are evaluated (Hayes et al., 2016). Therefore, as the science education community increasingly adopts the NGSS, there remains a gap in the evaluation of whether professional development (PD) programs shown to be effective in promoting reformed or inquiry-based science instruction as outlined in the NSES also facilitate 3D instruction (Hayes et al., 2016). In this study, teachers participated in a PD program designed in alignment with the NSES to increase their use of inquiry-based science instruction. Inquiry-based science instruction included science content, inquiry, and unifying concepts and processes as defined by the NSES. Resultantly, we hypothesized that the PD program would support participating teachers' use of 3D instruction, although the PD model was designed in alignment with the NSES rather than the NGSS. This quasi-experimental study had two purposes. The first was to adapt the Educators Evaluating the Quality of Instructional Products: Science (EQuIP) rubric version 2 (Achieve, 2014) to evaluate middle- and high-school science teacher implementation and student interactions with lesson materials. The EQuIP was designed to evaluate written lesson and unit materials. The second was to test the adapted EQuIP's ability to track changes in PD teacher's instruction over time versus a comparison teacher group and to previously completed Reformed Teaching Observation Protocol (RTOP; Sawada et al., 2002) evaluations.

Literature review

Decades of K-12 science education policy (AAAS, 1989, 1993; NRC, 1996, 2000, 2012) have called for teachers to model the SEPs for students through reformed instruction. These calls have been based on research evidence that reformed instruction improves student knowledge, reasoning, and argumentation (Blanchard et al., 2010; Wilson, Taylor, Kowalski, & Carlson, 2010). Employing instruction where students construct knowledge of core concepts in the discipline while engaging in scientific practices requires teachers to possess well-developed knowledge of content and pedagogy (Gess-Newsome, 1999; NRC, 1996; Wong & Luft, 2015). As this form of instruction often necessitates a considerable shift in pedagogical practices, changes in teachers' content knowledge, beliefs and attitudes, and pedagogical knowledge are also needed (Akkus, Gunel, & Hand, 2007; Borko & Putnam, 1995; Crawford, 2007; Enderle et al., 2014; Loucks-Horsley & Stiegelbauer, 1991; Phelps & Lee, 2003; Shumba & Glass, 1994; Wong & Luft, 2015). Fortunately, effective PD incorporating best practices such as long duration, research experiences, and pedagogy promote inquiry-based instruction (Supovitz & Turner, 2000).

Numerous instruments document teacher use of inquiry-based teaching practices in the science classroom (Heath, Lakshmanan, Perlmutter, & Davis, 2010). Several of the more commonly used instruments are summarized in Table 1.

Unsurprisingly, most of these instruments were based on reformed science instruction as called for in the NSES (NRC, 1996) and the *Benchmarks for Science Education* (AAAS, 1993). Accordingly, they focus on teacher and student actions and elements of classroom culture (e.g., students making predictions, students directing their own investigation, or teachers facilitating student discussions) called for in these reform documents. Though many also incorporate some assessment of content, most focus on what content is addressed and the accuracy of the content. Only one of the instruments, the Electronic Quality of Inquiry Protocol (Marshall, Smart, & Horton, 2009), attempts to evaluate the integration of content and actions and this is only one of 19 indicators. Further, none of the previously developed instruments explicitly assesses the integration of common themes (AAAS, 1993) or unifying concepts and processes in the NSES. Therefore, as PD program developers now support science teachers' effective use of 3D instruction, it is important that they have a tool that can document shifts in participating teachers' instructional practice in alignment with the NGSS (Hayes et al., 2016).

In this study, middle- and high-school science teachers participated in Target Inquiry, a 2.5-year PD program, designed around best practices in PD such as extended duration, research experiences, and pedagogy (Supovitz & Turner, 2000) — see Figure 2.

In a previous study (Yezierski & Herrington, 2011), the PD model was effective in increasing high-school chemistry teachers' inquiry instruction as measured by the RTOP (Sawada et al., 2002). However, with the increasing adoption of the NGSS within the United States, we wanted to evaluate whether the PD program designed around the NSES increased participating teachers' ability to use 3D instruction. Toward this goal, we used the EQuIP (Achieve, 2014). The EQuIP rubric is a qualitative tool with criteria within three categories that address alignment and overall quality of a unit or lesson with respect to the NGSS. The three categories are *Alignment to the NGSS, Instructional Supports*, and *Monitoring Student Progress*. Alignment to the NGSS category criteria relate to the coherent integration of grade-appropriate SEPs, DCIs, and CCCs in a lesson or unit (Achieve, 2014). Instructional Supports category criteria relate to using grade appropriate SEPs to make sense of phenomena and/or solving real-world problems, using the 3Ds to identify and build on students' prior knowledge, teacher facilitation of accurate facilitation of the 3Ds, scientific discourse, and differentiated instruction. Monitoring Student Progress category criteria relate to using formative and summative assessments to evaluate student learning across the 3Ds.

As reported in a series of case studies (Roseman, Fortus, Krajcik, & Reiser, 2015) conducted using the EQuIP rubric (Achieve, 2014) to critically examine curriculum materials, we also found that the EQuIP rubric provided a framework that yielded rich qualitative descriptions of the strengths and weaknesses of lessons, but did not easily facilitate tracking changes over time. The inability to easily track change hindered our ability to create a snapshot or summary of the evaluation that would easily facilitate longitudinal comparison of teacher instruction in terms of 3D integration. A comparison of teacher instruction is an important feature in documenting teacher change over the course of a long-term PD program. A second limitation of the EQuIP was its design to evaluate lesson and unit plan materials in isolation from teacher implementation of these

Table 1. Summary of science instruction observational protocol.

Instrument	Construct Focus/Instrument Design	Standards base for Instrument
Reformed Teaching Observation Protocol (RTOP; Sawada et al., 2002)	● Designed to measure degree to which instruction is aligned with science and math reforms ● Contains 25 items, divided into three categories (Lesson design and implementation; Content and process knowledge; Classroom culture) ● Each item rated on a 5-point Likert scale (0 = never present; 5 = very descriptive of lesson)	● Alignment with *National Science Education Standards* (NSES) and AAAS Benchmarks
Inquiry into Science Instruction Observation Protocol (ISIOP; Miner & DeLisi, 2012)	● Focus is on teacher instructional practices ● Uses a set of measurable class activity (teacher and student verbal and physical activities) class organization, and student disengagement codes to document different lesson events, and teacher verbal practices codes to document teacher instruction during the lesson ● Post-observation rubrics used to indicate portion of lesson spent on student- or teacher-centered activities, instructional leadership practices, and focus of content addressed	● Content standards are based on NSES ● Teaching indicators derived from the literature on inquiry-based instruction as actions that have been either theorized or demonstrated to be associated with student learning
Science Teacher Inquiry Rubric (STIR; Bodzin & Beerer, 2003)	● Rubric provides descriptors of six levels for each essential feature of inquiry ranging from teacher-centered (no evidence observed) to learner-centered ● Initially designed as both an observational and self-assessment rubric, but correlation between observation and instructor self-assessment was weak	● Based on the NSES five essential features of inquiry
Electronic Quality of Inquiry Protocol (EQuIP; Marshall et al., 2009)	● Rubric designed to assess quantity and quality of inquiry instruction ● Rubric indicators establish four levels of inquiry [pre-inquiry (level 1), developing (level 2), proficient (level 3), and exemplary (level 4)] within four constructs [Instruction, Curriculum, Discourse, and Assessment] as well as an overall assessment of the lesson	● Based on definition of inquiry in NSES
Quality of Science (QST) observation instrument (Schultz & Pecheone, 2014)	● Rubric that evaluates six domains of science teaching (Teachers' knowledge of content and pedagogy, Engaging students in learning science, Facilitating scientific discourse and reasoning, Promoting laboratory-based inquiry, Providing opportunities for applications of science, and Monitoring student learning) with a total of 18 indicators of quality science instruction	● Domains and indicators informed by/aligned with NSES, but key authors of the K-12 Framework were members of the QST advisory committee so authors promote clear overlaps between QST indicators and science and engineering practices in the Framework

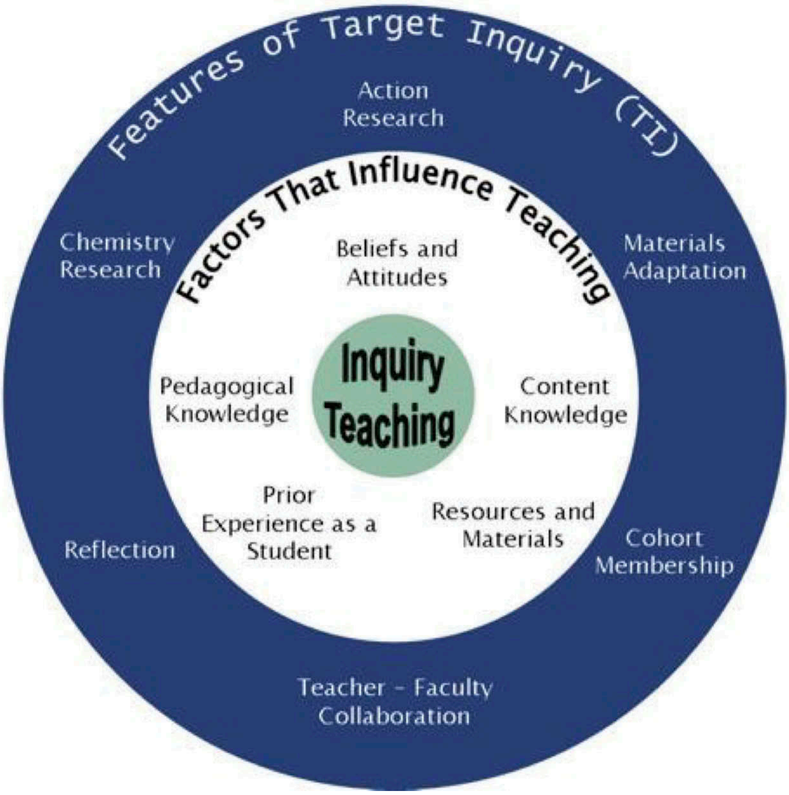


Figure 2. Target Inquiry PD model.

materials. Though the final version of the EQuIP (Achieve, 2016) attempts to quantify the evaluation by including a Likert-type rating scales for each of the three categories and the overall unit or lesson, it still focuses solely on the evaluation of lesson and unit plan materials. The EQuIP rubrics (Achieve, 2014, 2016) do not take into consideration teachers' implementation of these materials and fail to account for the coherency of the three dimensions from the *student* point of view (Community for Advancing Discovery Research in Education [CADRE], 2016; Roseman et al., 2015). Accounting for whether the students are explicitly engaged with the three dimensions needs to be an important consideration in evaluating science instructional practices, as this coherency for students has been a common weakness in K-12 instructional materials (CADRE, 2016; Roseman et al., 2015). Despite its limitations, because the EQuIP (2014) was designed by the authors of the NGSS it offers science PD program developers a common conceptual framework to evaluate PD program impact. The use of a common conceptual framework for PD program evaluation is an important goal for the science teacher PD developer community, as a common framework would maximize the benefit of such programs for teachers and students (Desimone, 2009; Heath et al., 2010).

To overcome the limitations outlined above and to facilitate the creation of a snapshot of teacher implementation of 3D instruction, we developed a semi-quantitative version of the EQuIP (SQ-EQuIP). The SQ-EQuIP facilitated a visual representation of teacher use of

3D instruction, enabling longitudinal comparison. This article describes the development and use of the SQ-EQuIP as a method for capturing teachers' use of 3D instruction as well as a means for documenting changes in teachers' use of 3D instruction over the course of a long-term PD program. The RTOP (Sawada et al., 2002) was used to triangulate the performance of the SQ-EQuIP. If the SQ-EQuIP was appropriately adapted and used by raters, we expected high RTOP scores to track with ideal or near-ideal SQ-EQuIP placements, with one exception. Based on the EQuIP rubric criteria (Achieve, 2014) an ideal lesson would be one in which students explicitly, coherently, and equally engaged with all three dimensions in an observed lesson. We define a near-ideal lesson as one in which students explicitly, coherently, but unequally engaged with the three dimensions in an observed lesson. A non-ideal lesson would be one in which students engage with one to two dimensions. Findings from Marshall, Smart, Lotter, and Sirbu (2011) indicated that the RTOP is "better suited for looking more globally at constructivist teaching practices" (p. 306) rather than inquiry-based learning. However, given the EQuIP's emphasis on the integration of all three dimensions, we would expect a lesson focusing on the science practices (e.g., the scientific method) while not meaningfully integrating science concepts to have a relatively high RTOP score but not track with an ideal or near-ideal 3D lesson. The research questions framing this study were: (a) How does PD teachers' instruction change with respect to grade appropriate 3D instruction and in relation to comparison teachers as characterized by the SQ-EQuIP? (b) How does the change in teachers' instruction as characterized by the SQ-EQuIP compare with changes tracked by the RTOP?

Methods

Context of study

This longitudinal quasi-experimental study included data from 19 teachers (see Tables 2 and 3), with 10 teachers in the intervention (PD) group (middle school [MS] = 6; high school [HS] = 4) and 9 in the comparison group (MS = 5; HS = 4). Teachers self-selected into the PD or comparison group. Schools were in suburban and rural regions of western Michigan. Overall, the two teacher groups were fairly well matched. The most notable difference between the teacher groups was years of teaching experience. On average, PD teachers had approximately 3.5 years' less teaching than comparison teachers at the start of the PD, with 12.5 and 16.3 average years of teaching experience, respectively. It is possible that the years of teaching experience contributed to whether teachers selected to be part of the intervention or comparison group, as it had been documented that years of teaching experience decreases the likelihood of teachers' willingness to change their beliefs (Pajares, 1992). Further, given the intensive nature of this PD program (a 2.5-year program with substantial summer commitments), it is reasonable to assume that even when provided compensation in terms to tuition and stipends that only more motivated teachers would opt to engage this PD experience. Though these potential differences in the treatment and comparison groups imposes some limitations on the findings from this study, given the match between the school demographics, grade level, and content between teachers in both groups, and that some teachers in the comparison group started in the treatment group but had to switch for personal reasons, asked to participate in the

Table 2. PD teacher contexts.

	Teacher Data				2012 Teacher's Schools Demographics						
Teacher #	Teaching Experience (Years)	Gender	School #	Subject	White (%)	Hispanic (%)	Black (%)	Asian (%)	American Indian (%)	Free/Red. Lunch (%)	Student Science Proficiency (%)
State Average					71	5	20	3	<1	42.5	78
Teacher 1	11	F	1	Physics, Chemistry	89	4	5	2	<1	26	61
Teacher 2	23	M	2	8th Grade Physics[a]	92	6	<1	1	1	27	89
Teacher 3	3	F	3	8th Grade Earth Science[a]	75	12	11	1	1	53	83
Teacher 4	23	M	4	6th Grade Science	58	30	3	9	1	46	85
Teacher 5	18	F	5	Biology	86	6	4	4	<1	18	67
Teacher 6	5	F	6	6th Grade Science	93	3	2	1	1	31	86
Teacher 7	2	M	7	6th –7th Grade Science	83	11	1	5	0	27	83
Teacher 8	13	M	2	8th Grade Physics[a]	92	6	<1	1	1	27	89
Teacher 9	18	F	8	Biology, Chemistry Forensics	83	7	8	1	<1	31	65
Teacher 10[b]	9	F	9	Agricultural Science	-	-	-	-	-	-	-

[a] Middle school setting, but science content taught aligns to state high school standards.
[b] School demographic data not available.

Table 3. Comparison teacher contexts.

	Teacher Data				2012 Teacher's Schools Demographics						
Teacher #	Teaching Experience (Years)	Gender	School #	Subject	White (%)	Hispanic (%)	Black (%)	Asian (%)	American Indian (%)	Free/Red. Lunch (%)	Student Science Proficiency (%)
State Average					71	5	20	3	<1	42.5	78
Teacher 11	27	F	6	6th Grade Science	93	3	2	1	1	31	86
Teacher 12	12	F	5	Biology, Anatomy	86	6	4	4	<1	18	67
Teacher 13	19	F	8	Physical Science, Chemistry	83	7	8	1	<1	31	65
Teacher 14[b]	18	F	9	Health Science	-	-	-	-	-	-	-
Teacher 15	13	F	3	Chemistry	89	4	5	2	<1	26	61
Teacher 16	20	F	10	8th Grade Earth Science	95	4	<1	<1	<1	62	81
Teacher 17	17	F	11	6th Grade Science	96	1	1	2	<1	14	92
Teacher 18	16	M	12	8th Grade Earth Science[a]	90	6	1	1	2	53	80
Teacher 19	5	M	12	6th Grade Science	90	6	1	1	2	53	80

Note. Five of the comparison teachers included in this study were in the same building as at least one PD teacher. During annual interviews (interview data not included in this study), PD and comparison teachers were explicitly asked about their collaboration efforts with teachers in their school. Only when both the PD and comparison teachers working in the same building described no meaningful collaborations to design, implement, or evaluate lessons was the comparison teacher included in this study. Three comparison teachers were excluded from this study due to reports of meaningful and sustained collaborations between PD and comparison teachers. See Herrington, Bancroft, Edwards, and Schairer (2016) for PD teachers' report in annual interviews of collaborations, or lack thereof, within their schools.

[a] Middle school setting, but science content taught aligns to state high school standards.

[b] School demographic data not available.

Table 4. Main events of three core PD experiences.

	Core Experience		
	RET Year 1	MA Year 2	AR Year 3
Main Events	• Read literature related to scientific research • Work six weeks with researcher collecting and analyzing data • Learn about and reflect on process of scientific inquiry • Present research at a regional conference • Make small modifications to two existing classroom activities to reflect process of scientific inquiry	• Read science education research and research methods literature • Design AR • Design or modify a lesson or series of lesson for AR • Implement modified lesson and collect student data for AR	• Analyze student data • Write up AR for submission to peer reviewed journal or master's thesis

next round of PD, or were engaged in other PD opportunities focused on inquiry-based science instruction within their districts, we suspect the differences in teacher motivation were minimal.

The features of the PD model are designed to encourage and improve inquiry instruction by impacting teachers' beliefs and attitudes, and content and pedagogical knowledge, as well as providing adequate resources and materials; see Figure 2. The PD included the central characteristics of high-quality PD programs (duration, cohort participation, active learning, coherence, and content-focus) (Garet, Porter, Desimone, Birman, & Yoon, 2001) in alignment with the NSES (NRC, 1996). These central characteristics were integrated in the model via the three core experiences (research experiences for teachers [RET], materials adaptation [MA], and action research [AR]); see Table 4. During the six-week summer RET, the process of scientific inquiry was emphasized. In the RET, teachers engaged in authentic scientific research and reflected on the process of scientific inquiry as they engaged in this research. They subsequently adapted materials from two classroom activities to better reflect the processes of science they experienced in the RET for implementation in their classroom. Teachers further modified a lesson or unit of lessons during the MA core experience, which they would later implement for an AR project. Post-AR data for comparison teachers were not collected, so post-AR is excluded from this study. For more detail on the design and implementation of these core experiences, see Yezierski & Herrington (2011).

Data collection

Annual classroom observations began the year prior to the start of the PD program (pre-RET), one year after the RET (post-RET), and one year after the MA (post-MA). Once within each of these three time points, teachers invited researchers to video record a lesson they perceived to be inquiry based or student centered. Video recordings captured teacher lectures, students working with materials, screen captures of student work, phenomena students were observing, and teacher–student, student–student, and whole-group interactions. Lesson-related materials such as lesson plans or handouts were also collected.

Data analysis

RTOP analysis

Video recordings of 57 lessons (30 from PD teachers, 27 from comparison teachers) were analyzed. All videos were initially analyzed using the RTOP, which has 25 items across the three scales of lesson design and implementation, content, and classroom culture (Sawada et al., 2002). Each item is evaluated on a 5-point scale (0 = never occurred to 4 = very descriptive). The lesson design and implementation scale contains five items to measure teacher ability to design and sequence activities that activate and use student ideas within the lesson. The content scale contains ten items, with five items each relating to propositional and procedural pedagogical content knowledge respectively. The classroom culture scale contains ten items, with five items each relating to communicative interactions and student–teacher interactions, respectively. Therefore, each lesson can be given a maximum score of 100 where a high score indicates high alignment with reformed practices. Internal consistency tests yielded Cronbach's alpha reliability coefficient of at least .80 for each scale (Sawada et al., 2002). All videos and any supporting lesson materials were analyzed independently by three raters using the RTOP.

Each new rater used training materials obtained from Piburn et al. (2000). Each new rater's training included independently watching each of three training videos, rating the video using the RTOP, and comparing their scores versus the provided expert scores. If needed, a new rater discussed score discrepancies with a more experienced rater on the PD team. Each video was watched, evaluated, and scores compared before the rater moved on to the next of the three training videos. After completion of this training, new raters engaged in re-rating lessons from the baseline round of data collection in this study. New raters then participated in mock RTOP negotiations in an effort to "calibrate" their views of the lesson to ensure that they were looking for the teacher and student behaviors that allowed consistent RTOP scoring across raters. New raters engaged in these mock negotiations until expert team members determined they were ready to begin rating newly recorded classroom observations. This process has been used to evaluate recorded classroom observations of several prior teacher cohorts and has yielded inter-rater reliability of over 80% (Yezierski & Herrington, 2011).

All recorded classroom observations in this study were first independently evaluated by three raters. The three independently assigned subscale and total scores were compared, and if total scores differed by more than 5 points the raters negotiated individual items that differed by more than 1 point to decide consensus scores. The level of a 5-point difference was chosen as an absolute difference of 5% (5 out of a possible 100) was considered to be a level above which it was thought that differences between the reformed natures of lessons could be detected. The idea was that with a difference of 1–2 points on a subscale would not be enough to identify one lesson as being "more reformed" than the other. In the negotiation of individual items, each rater presented specific examples from videos and/or lesson materials to justify the points they assigned. These examples were discussed in relation to the item statement until scores differed by no more than 1 point. After individual item negotiations were complete, subscores and total scores were recalculated by each rater. Final consensus was considered to be found when all three total scores were within 5 points of each other after negotiation. The average of these three consensus scores was assigned to each observed lesson and are reported in this study.

Given the small sample size, we did not seek to make claims of significance. Instead, we sought to identify trends in mean scores on each of the three RTOP scales. These trends were used to compare changes in each teacher group across time and changes between PD and comparison teacher groups.

Development of SQ-EQuIP and SQ-EQuIP analysis

Qualitative analysis with the EQuIP rubric (Achieve, 2014) was done using the same videos and lesson materials used for RTOP analysis. The NGSS — and accompanying Appendices F and G (NGSS Lead States, 2013), which describe in detail the DCIs, SEPs, and CCCs, respectively — were used to interpret the criteria statements pertaining to alignment of the lesson to the NGSS, instructional supports, and strategies teachers used to monitor student progress on the EQuIP rubric. Teacher implementation of instructional supports and monitoring student progress criteria from the latter two categories of the EQuIP rubric were evaluated during the analysis. However, the analysis presented in this study is limited to the extent the observed lessons aligned to criteria in the Alignment to the NGSS category only of the EQuIP rubric. Both teacher implementation of lesson materials and student engagement with these materials was a major focus of the PD program. Therefore, it was decided that a dimension was explicitly present in the lesson when student work or conversations, not just teacher discussion or lesson materials, included evidence of criteria related to the dimension as laid out in the NGSS. So, if the teacher attempted to incorporate a dimension in the observed lesson, but there was no evidence to support that *students* engaged with the dimension, the dimension was evaluated as implicitly present. An ideal lesson is characterized by students explicitly, coherently, and equally engaging with all three dimensions in the observed lesson.

Three videos that were a sequence of lessons from one teacher that was previously judged via RTOP evaluations as being representative of lessons that would likely have both high and low integration of science content, practices, and crosscutting concepts were purposely selected for a first round of analysis with the EQuIP rubric (Achieve, 2014). These lessons would provide the raters with a broad spectrum of lesson design and instructional practices that would yield productive discussions around the criteria on the EQuIP rubric. The first and second authors completed the qualitative analysis for the first lesson in the sequence using the EQuIP rubric. In response to each criterion on all three categories of Alignment to the NGSS, Instructional Supports, and Monitoring Student Progress on the rubric each rater described how, if at all, the criterion was demonstrated in the lesson (what was done and how it was done by teacher and students and where the criterion was included in the lesson materials if applicable). Additionally, each rater described how the demonstrated actions/content aligned to the NGSS description of the dimension. Based on this qualitative analysis, each rater independently placed a circle on a triangle with each of the three vertices of the equilateral triangle representing a dimension to visually and semi-quantitatively represent their qualitative analysis. An equilateral triangle was chosen because of its sides of equal length, equal angles, and one central point. It signals within this analysis that the three dimensions are of equal value and their integration is the central or ideal goal of an implemented science lesson or unit. A star was placed at the center of the triangle to serve denote an ideal placement on the SQ-EQuIP. Our translation of this qualitative analysis to a placement on the equilateral triangle allowed us to visually track changes in teachers' use of 3D instruction over time that the strictly qualitative description could not easily accommodate. Given this visual adaptation, we have termed it a semi-quantitative adaptation of the EQuIP. In our independent analysis of the three purposefully selected lessons, the qualitative descriptions revealed no notable differences and placements on the triangle were in agreement.

Two more selected videos (not included in this study) were analyzed using the above procedure and there was again complete agreement in the qualitative and semi-quantitative analyses. Based on discussions between raters for their rationale of placement within the triangle in relation to their qualitative analysis, criteria were created for the semi-quantitative analysis. Although we did analyze whether the dimensions present in a lesson were grade appropriate, this appropriateness was not visually represented on the SQ-EQuIP, but was noted in the narrative below the visual when a dimension was below grade level. The criteria for visually representing a lesson's alignment to the Alignment to the NGSS category were refined as analysis of the remaining videos continued. The final visual representation and set of criteria for placement are shown in Figure 3.

The final version of the visual representation does not include any points at or near the CCC vertex, since there were no lessons in this study for which this placement was appropriate. However, we recognize that it is possible that such a lesson exists beyond this study and therefore the visual representation can be revised to reflect such a lesson. Additionally, evaluated lessons include a narrative concerning what dimensions were included in the lesson and how each dimension was incorporated in the lesson.

Twenty-one of the remaining 54 videos were analyzed only by the first author. However, to ensure that there was no drift by this primary rater over time, 33 videos were analyzed by an additional rater (the third author) across the entire data analysis period (Kimberlin & Winetrstein, 2008). The third author was trained on the SQ-EQuIP

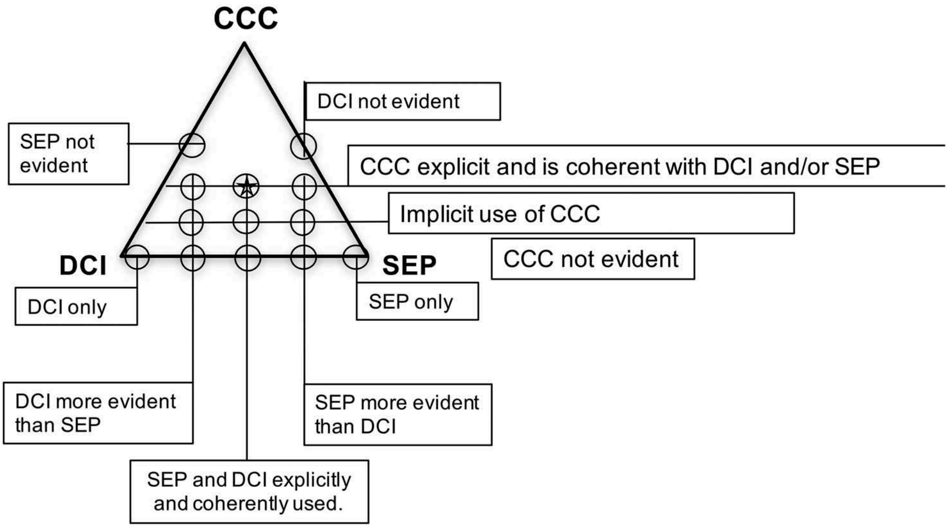

CCC
DCI not evident
SEP not evident
CCC explicit and is coherent with DCI and/or SEP
Implicit use of CCC
CCC not evident
DCI
SEP
DCI only
SEP only
DCI more evident than SEP
SEP more evident than DCI
SEP and DCI explicitly and coherently used.

Figure 3. Matrix of criteria used to evaluate video recorded lessons and accompanying materials for the semi-quantitative adaptation of EQuIP. The larger circle used to represent lesson alignment with 3D instruction is made smaller when SEP are present in the lesson, but its use is out of alignment with NGSS descriptions in Appendix F. ★ indicates a lesson where students are equally, explicitly, and coherently engaged with all three dimensions.

by evaluating three of the five videos the first two authors used to develop the SQ-EQuIP. The new rater compared and discussed her evaluations in relation to the first two authors' consensus evaluation of these videos. When it was determined that the new rater was "calibrated" on the SQ-EQuIP, she began to evaluate videos for the study.

Neither the first author nor the third author played a role in the RTOP evaluations of the pre-RET and post-RET observed lessons. Therefore, these RTOP scores had no influence on their SQ-EQuIP evaluations of these lessons. Additionally, there was consensus among raters that enough time had passed (approximately one calendar year) since their RTOP analyses of post-MA lessons that RTOP scores were not influencing SQ-EQuIP placements of the post-MA lessons. When two raters evaluated a lesson, videos and lesson materials were independently analyzed by raters. Raters then compared their qualitative evaluations of the lesson in relation to each criterion on the EQuIP rubric. Similar to the RTOP evaluations, when discrepancies were present raters discussed specific evidence from the video recording or lesson materials to support their evaluations to come to consensus. Raters then shared their placement on the SQ-EQuIP, and if needed, came to consensus on the placement based on their qualitative analyses and SQ-EQuIP placement criteria. Finally, the primary rater returned to each SQ-EQuIP analysis to ensure that the rating criteria were consistently applied to all 57 videos analyzed.

To determine SQ-EQuIP inter-rater agreement, assigned DCIs, SEPs, CCCs, the grade level at which each dimension was implemented, and SQ-EQuIP placements assigned by rater by lesson were compared. The total number of assigned DCIs, SEPs, CCCs on which raters agreed in proportion to the total number of all identified DCIs, SEPs, CCC across both raters was used to calculate an inter-rater agreement percentage. Inter-rater agreements were 82%, 63%, and 91% for DCIs, SEPs, and CCCs, respectively. Grade-level appropriateness and placement were similarly calculated. There was a 91% inter-rater agreement for both grade-level determination of the dimensions and SQ-EQuIP placement. When differences in placements on the SQ-EQuIP occurred between raters, they were typically one placement off from each other. When raters' overall assignment of the three dimensions, the grade level at which each dimension was implemented, and the raters' placement on the SQ-EQuIP were considered, there was an overall 84% inter-rater agreement.

Findings

RTOP results

PD teachers entered the PD program with higher mean pre-RET scores for all three RTOP scales relative to comparison teachers (see Figure 4).

The PD teacher group had consistent positive gains in mean scores across all three RTOP scales post-RET and post-MA. In contrast, the comparison group had a decrease in their mean scores for design and implementation and classroom culture scales post-RET. Their mean score for the content scale remained relative unchanged pre-RET to post-RET. However, the comparison group had positive gains in mean scores across all three scales from pre-RET to post-MA. Overall, the PD teacher group showed a steady gain in their mean total RTOP scores after each PD year with a total 9.1% gain in their mean total

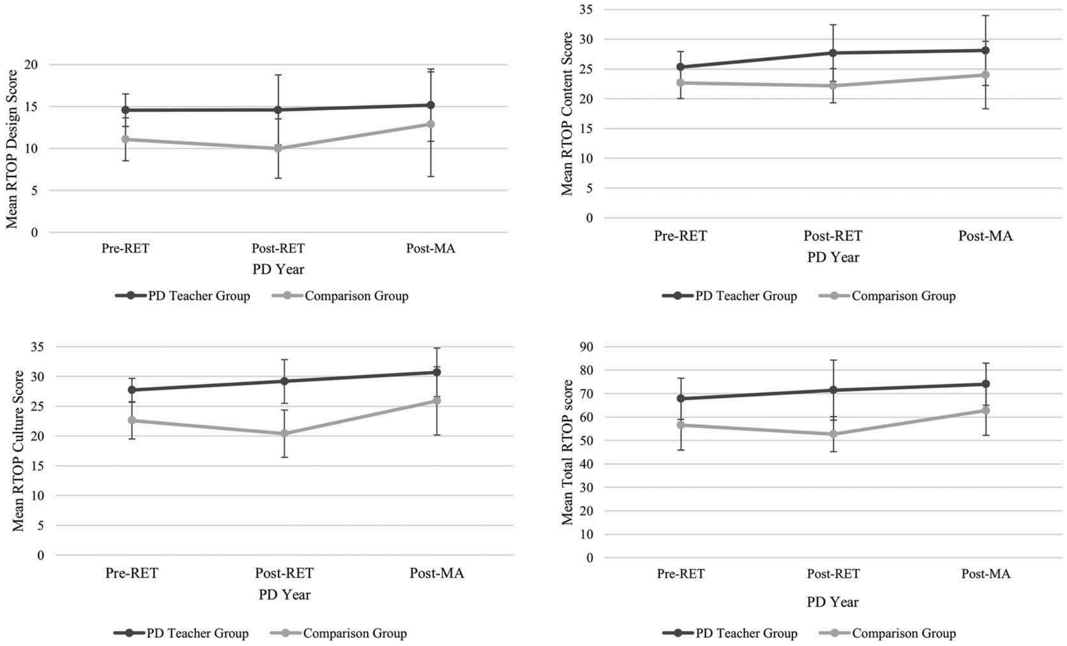


Figure 4. Mean scores for RTOP subscale and mean total RTOP scores by teacher group pre-RET to post-MA.

RTOP score from pre-RET to post-MA. The comparison teacher group had an overall 11.2% gain pre-RET to post-MA.

SQ-EQuIP results

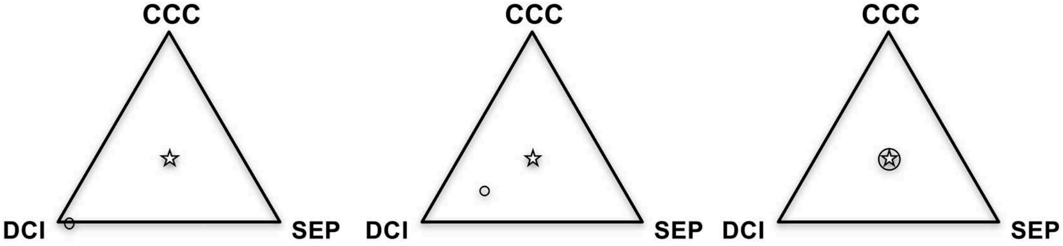
PD teachers

Overall, PD teachers' instruction became more aligned to ideal use of 3D instruction over the two years of PD (post-MA). Upon entry (pre-RET) to the PD program, two of the ten participating teachers implemented lessons that were placed at the ideal on the SQ-EQuIP and both teachers maintained this ideal placement two years after PD participation. Pre-RET, no PD teacher was placed adjacent or near to the ideal on the SQ-EQuIP. After one year in the PD program (post-RET), five of ten participating teachers were placed at or near the ideal. After two years (post-MA) in the PD program, eight of the ten participating teachers implemented lessons that were near or at the ideal on the SQ-EQuIP. Two of these eight lessons contained SEPs and CCCs found in NGSS grade bands below the grade level of the students observed in the lessons. There were three distinct pathways to this increased alignment. The first was characterized by a gradual progression in PD teachers' (n = 3) ability to implement a lesson in which the students explicitly and coherently engaged with all three dimensions. Teacher 2's SQ-EQuIP placements are used as an example of a gradual progression in Figure 5 where ideal or near-ideal placement occurred only after MA.

For two of the three teachers in this gradual progression pathway a concurrent increase in RTOP scores was tracked as shown in the example in Figure 5. However, it should be noted that these were also the only two teachers in the PD teacher group whose post-MA lessons contained SEPs or CCCs that were below grade level (both were high school teachers, but used SEPs or CCCs that were more appropriate for grades 3–5 or 6–8). The third teacher placed in this category, although achieving an ideal placement in her post-MA year, had RTOP scores that did not show concurrent increase across the time points. With an average total pre-RET RTOP score of 58.2, these teachers entered the PD program scoring the lowest on the RTOP compared with the teachers in the second and third pathways.

The second pathway was characterized by a rapid progression in the PD teachers' (n = 3) ability to implement a lesson in which the students were explicitly and coherently engaged with all three dimensions. Teacher 7's SQ-EQuIP placements are used as example of a rapid progression, shown in Figure 6, where ideal or near-ideal placement on the SQ-EQuIP occurred post-RET and was maintained post-MA. The discrepancy between the pre-RET RTOP scores and the related SQ-EQuIP placement shown in Figure 6 highlights the weakness of the RTOP to capture the lack of integration of teacher presentation of science content and practices when a lesson tended to focus on students planning and carrying out an investigation in a lesson. This specific weakness was observed in three other lessons.

For two of the three teachers in the rapid progression pathway, a concurrent increase in RTOP scores was seen from pre-RET to post-RET. With an average total pre-RET RTOP score of 72.8, these teachers entered the PD scoring relatively high on the RTOP compared with the teachers in the first pathway and slightly lower than the teachers in the third pathway.



Pre-RET (RTOP score: 56.7)
RTOP Design: 13.7
RTOP Content: 22.7
RTOP Culture: 20.3

DCI: Energy. MS-PS3-1. Collect data to find shape of position/time graph of a car moving down track. No discussion of motion in terms of energy.

SEP: 3; 4. Investigation determined completely by teacher. Teacher sets up plot for students. Students have no input. Teacher gives away shape of graph and hypothesis during warm up activity preceding investigation.

CCC: None.

Post-RET (RTOP score: 58.3)
RTOP Design: 12.0
RTOP Content: 24.3
RTOP Culture: 22.0

DCI: Energy. MS-PS3-1. Describe what distance v. time graphs looks like for various speeds and for increasing speed by collecting, plotting, and interpreting data for distance/time graphs.

SEP: 3; 4. Students begin graphing before teacher, but teacher gives away graph construction and some interpretations before students have an opportunity to finish construction and interpretation.

CCC: Patterns. Collection of data related to 3 different types of motion which was then plotted on one graph facilitated a natural comparison of patterns in rate of change by type of motion. However, unable to assess if students picked up on patterns as they were still plotting data at end of class and teacher gives away some patterns at end. No explicit discussion of a rate of change.

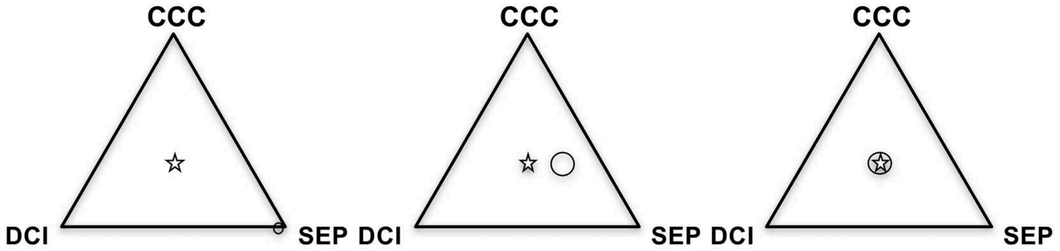
Post-MA (RTOP score: 78.7)
RTOP Design: 14.3
RTOP Content: 32.3
RTOP Culture: 32.0

DCI: Forces and interactions. MS-PS2-1. Facilitation of various experiences with force pairs, so students "discover" Newton's 3rd law.

SEPs: 3; 4. Students follow directions for lab, but very strong emphasis on students collecting, organizing, analyzing, and interpretation data.

CCC: Patterns. Student conversations indicate they are perceiving pattern that two interacting objects will experience forces of equal magnitude and opposite directions as they analyze tabulated data.

Figure 5. Teacher 2's gradual progression (explicit and coherent student engagement with 3D occurs only post-MA). Key features of lesson used to place lesson on SQ-EQuIP are underlined.



Pre-RET (RTOP score: 71.3)
RTOP Design: 17.3
RTOP Content: 24.3
RTOP Culture: 29.7

DCI: None. Students design an investigation of their choice based on their construction of a testable question method. Eg. How much whipped cream can a donut hold?
SEP: 3. Students write question, prediction, and a procedure. All unrelated to scientific question or real world phenomenon.
CCC: None.

Post-RET (RTOP score: 72.7)
RTOP Design: 16.0
RTOP Content: 25.7
RTOP Culture: 31.0

DCI: Wave properties. HS-PS4-3. Explore teacher question: How does one wave's crest or trough affect another wave's crest or trough? This is a sub-question that will help students answer bigger question of How do humans use waves? However, no discussion of interference in lesson.
SEP: 3. Unscramble procedure, student derived IV, DV, and CVs, make prediction with an explanation, in response to teacher question, and create data table. Teacher emphasizes that their IV and DV must reflect variables in the question.
CCC: Cause and effect. Students designing investigation to collect data that will describe how a change in wave amplitude is caused when wave crests and troughs overlap.

Post-MA (RTOP score: 78.7)
RTOP Design: 15.7
RTOP Content: 26.7
RTOP Culture: 29.3

DCI: Wave properties. MS-PS4-2. Explore teacher question: How do different media density affect a wave? using PhET simulation and given dataset on media and sound wave speed.
SEP: 2; 4; 6. Students identify IV and DV based on data provided on medium type and wave speed and plot related graph. Used evidence from text to support explanation of relationships observed between material type and wave speed and angle.
CCC: Patterns. Students used data analyses to identify pattern among state of matter of medium and wave speed and angle.

Figure 6. Teacher 7's rapid progression (explicit and coherent student engagement with 3D occurs only post-RET and maintained post-MA). Some key features of lesson used to place lesson on SQ-EQuIP are underlined.

The third pathway (n = 2) was characterized by the PD teachers' ability to implement a lesson in which the students were explicitly and coherently engaged with all three dimensions pre-RET and post-MA. Teacher 5's SQ-EQuIP placements are used as a typical example of this consistent ideal 3D instruction, as shown in Figure 7.

For teachers in this constant, ideal pathway high RTOP scores were typically maintained across time. With an average total RTOP score of 76.8, Teachers 5 and 6 entered the PD program with the two highest RTOP scores.

Two PD teachers (Teachers 3 and 4), both middle school teachers, did not fit any of these three pathways. Teacher 3, though showing the promise of a rapid progression pre-RET to post-RET and a concurrent increase in total RTOP scores, was unable to maintain ideal placement post-MA. The second teacher, although able to progress to authentically use SEPs from pre-RET to post-RET, was never able to explicitly and coherently engage his students with all three dimensions in any of the observed lessons.

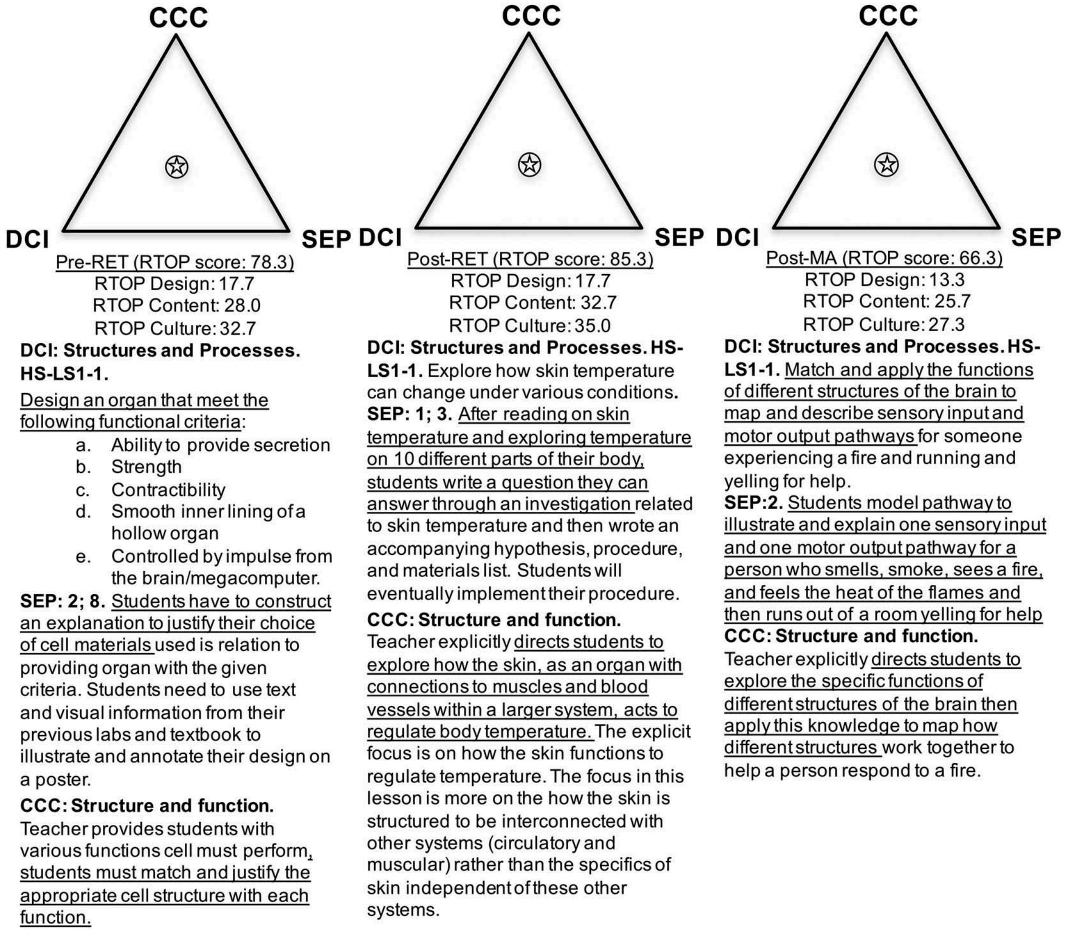
Comparison teachers

An overall increased alignment to 3D instruction was not observed for comparison teachers. None of the nine comparison teachers were at or near the ideal pre-RET. After one year, two of nine lessons were at or near the ideal. After two years, two of the nine lessons were at or near the ideal where both lessons contained SEPs and one contained a CCC found in grade bands that were below the grade level of the students observed in the lessons. Teacher 18's SQ-EQuIP placements are used as an example of a comparison teacher's lack of progression; see Figure 8.

Further (and although an extreme case), the SQ-EQuIP evaluation of one comparison teacher revealed that while her lesson materials were designed with elements that supported authentic use of SEPs such as students making predictions and using models to gather evidence to support a claim, her implementation of the lesson materials did not reflect the intended lesson design as she skipped over students making predictions as prescribed by the handout in her pre-RET lesson and consistently gave away conclusions and observations before students had any opportunity to engage with lesson activities (Figure 9). Overall, 80.0% of PD teachers' lessons and 22.2% of comparison teachers' lessons had ideal or near-ideal placement on the SQ-EQuIP after two years (post-MA).

RTOP versus SQ-EQuIP

Overall, for both groups of teachers RTOP scores were generally aligned with SQ-EQuIP evaluations. That is, high RTOP scores typically corresponded with ideal or near-ideal placements on the SQ-EQuIP. Additionally, PD teachers' higher lesson design and implementation and classroom culture RTOP scores at pre-RET align with the findings that none of the comparison teachers placed at or near ideal on the SQ-EQuIP. However, in four of the 57 lessons evaluated we found the SQ-EQuIP captured aspects of teacher instruction that the RTOP did not. Two pre-RET lessons scored high on the RTOP, suggesting these lessons were highly aligned with reformed instruction, although the SQ-EQuIP revealed the lessons to be largely devoid of science content (e.g., pre-RET lessons in Figure 6). RTOP scores in these lessons were likely higher because teachers used pedagogical strategies that are recognized as supportive of student learning such as students making predictions, exploring before teacher explanations, and teachers acting as facilitators. A second, but related, discrepancy between RTOP and SQ-EQuIP was



Pre-RET (RTOP score: 78.3)
RTOP Design: 17.7
RTOP Content: 28.0
RTOP Culture: 32.7

DCI: Structures and Processes. HS-LS1-1.

Design an organ that meet the following functional criteria:

a. Ability to provide secretion
b. Strength
c. Contractibility
d. Smooth inner lining of a hollow organ
e. Controlled by impulse from the brain/megacomputer.

SEP: 2; 8. Students have to construct an explanation to justify their choice of cell materials used is relation to providing organ with the given criteria. Students need to use text and visual information from their previous labs and textbook to illustrate and annotate their design on a poster.

CCC: Structure and function. Teacher provides students with various functions cell must perform, students must match and justify the appropriate cell structure with each function.

Post-RET (RTOP score: 85.3)
RTOP Design: 17.7
RTOP Content: 32.7
RTOP Culture: 35.0

DCI: Structures and Processes. HS-LS1-1. Explore how skin temperature can change under various conditions.

SEP: 1; 3. After reading on skin temperature and exploring temperature on 10 different parts of their body, students write a question they can answer through an investigation related to skin temperature and then wrote an accompanying hypothesis, procedure, and materials list. Students will eventually implement their procedure.

CCC: Structure and function. Teacher explicitly directs students to explore how the skin, as an organ with connections to muscles and blood vessels within a larger system, acts to regulate body temperature. The explicit focus is on how the skin functions to regulate temperature. The focus in this lesson is more on the how the skin is structured to be interconnected with other systems (circulatory and muscular) rather than the specifics of skin independent of these other systems.

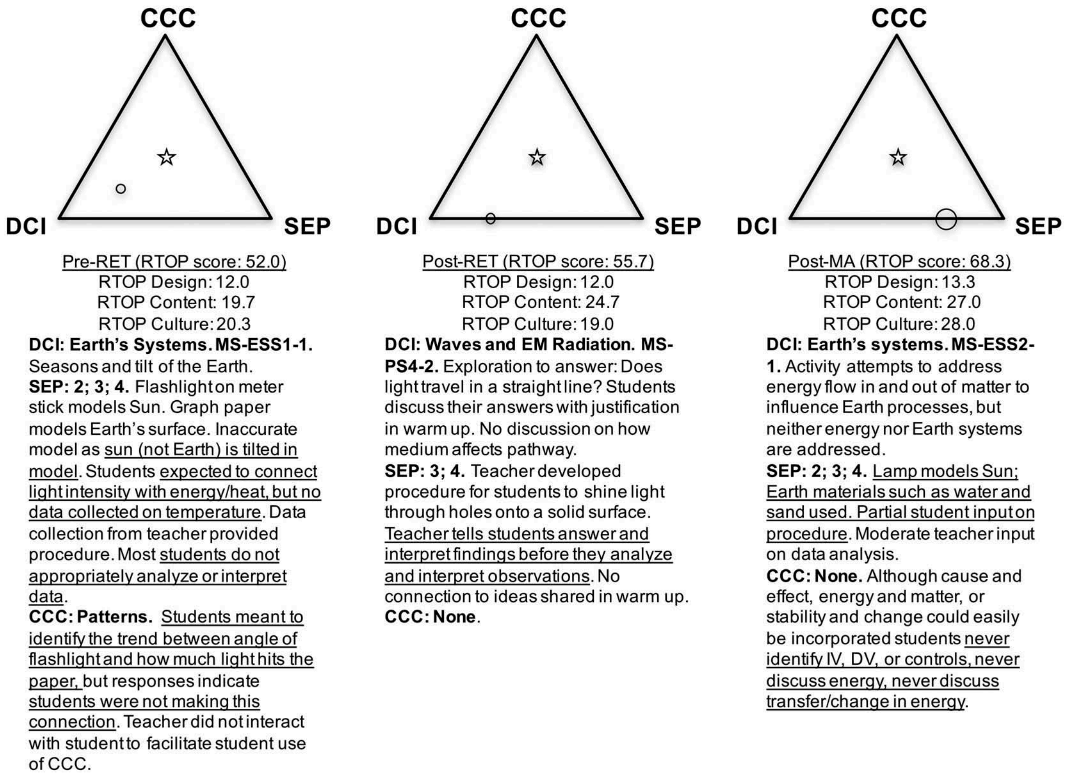
Post-MA (RTOP score: 66.3)
RTOP Design: 13.3
RTOP Content: 25.7
RTOP Culture: 27.3

DCI: Structures and Processes. HS-LS1-1. Match and apply the functions of different structures of the brain to map and describe sensory input and motor output pathways for someone experiencing a fire and running and yelling for help.

SEP:2. Students model pathway to illustrate and explain one sensory input and one motor output pathway for a person who smells, smoke, sees a fire, and feels the heat of the flames and then runs out of a room yelling for help

CCC: Structure and function. Teacher explicitly directs students to explore the specific functions of different structures of the brain then apply this knowledge to map how different structures work together to help a person respond to a fire.

Figure 7. Teacher 5's constant, ideal instruction (explicit and coherent student engagement with 3D occurs only post-RET and maintained post-MA). Some key features of lesson used to place lesson on SQ-EQuIP are underlined.



Pre-RET (RTOP score: 52.0)
RTOP Design: 12.0
RTOP Content: 19.7
RTOP Culture: 20.3

DCI: Earth's Systems. MS-ESS1-1. Seasons and tilt of the Earth.
SEP: 2; 3; 4. Flashlight on meter stick models Sun. Graph paper models Earth's surface. Inaccurate model as sun (not Earth) is tilted in model. Students expected to connect light intensity with energy/heat, but no data collected on temperature. Data collection from teacher provided procedure. Most students do not appropriately analyze or interpret data.
CCC: Patterns. Students meant to identify the trend between angle of flashlight and how much light hits the paper, but responses indicate students were not making this connection. Teacher did not interact with student to facilitate student use of CCC.

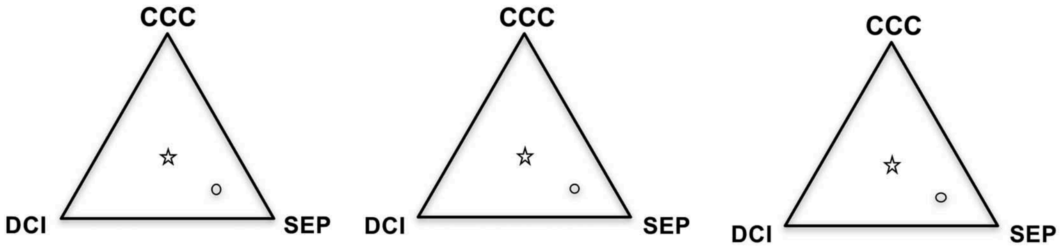
Post-RET (RTOP score: 55.7)
RTOP Design: 12.0
RTOP Content: 24.7
RTOP Culture: 19.0

DCI: Waves and EM Radiation. MS-PS4-2. Exploration to answer: Does light travel in a straight line? Students discuss their answers with justification in warm up. No discussion on how medium affects pathway.
SEP: 3; 4. Teacher developed procedure for students to shine light through holes onto a solid surface. Teacher tells students answer and interpret findings before they analyze and interpret observations. No connection to ideas shared in warm up.
CCC: None.

Post-MA (RTOP score: 68.3)
RTOP Design: 13.3
RTOP Content: 27.0
RTOP Culture: 28.0

DCI: Earth's systems. MS-ESS2-1. Activity attempts to address energy flow in and out of matter to influence Earth processes, but neither energy nor Earth systems are addressed.
SEP: 2; 3; 4. Lamp models Sun; Earth materials such as water and sand used. Partial student input on procedure. Moderate teacher input on data analysis.
CCC: None. Although cause and effect, energy and matter, or stability and change could easily be incorporated students never identify IV, DV, or controls, never discuss energy, never discuss transfer/change in energy.

Figure 8. Teacher 18's lack of clear progression to explicitly and coherently engage students in 3D learning post-RET to post-MA.



Pre-RET (RTOP score: 54.0)
RTOP Design: 10.3
RTOP Content: 24.3
RTOP Culture: 19.3

DCI: Space systems. MS-ESS1-1. Log observations of phases of moon for 6 weeks, study of vocabulary related to moon topography followed by lab exploring crater impact on moon with flour, cocoa, and ball model. Moon predominantly explored in isolation of Sun and Earth.

SEP: 2. Teacher demonstrated and directed procedure and data collection with model, gives away what will happen. Teacher explicitly ignores the steps of students writing a hypothesis with their supporting ideas present in activity handout. Thus, no ideas about what will happen developed being tested as intended by SEP 2.

CCC: Patterns. The materials are designed to have students identify cause and effect between size and speed of impactor on the size of the crater and the nature of ejecta; teacher ignores the cause and effect features of lesson. Students will instead look for patterns to answer questions on handout. Handout questions and data table address energy transformation, but energy is never addressed.

Post-RET (RTOP score: 49.7)
RTOP Design: 10.7
RTOP Content: 20.7
RTOP Culture: 18.3

DCI: Earth's systems. HS-ESS2-4. Intended for students to answer: *How does the tilt of the Earth influence weather in northern hemisphere*? However, teacher *tells* students how tilt influences seasons. Questions that walking students through simulation address relationship b/w *seasons* and length of day. Energy is never addressed and teacher communicates to students Earth's orbit is always round, and Earth's orbit does not influence Earth's weather or climate.

SEP: 2. Simulation used for students to answer individual questions about length of day in different seasons. No evidence students use or will use data to describe phenomenon or explicitly answer question. Teacher gives away some of what will happen.

CCC: Pattern. The teacher discusses pattern b/w tilt of Earth and seasonal temperatures and simulation explores tilt of Earth and length of day. This created an implicit pattern of greater tilt and the Earth receiving more sunlight.

Post-MA (RTOP score: 48.7)
RTOP Design: 9.0
RTOP Content: 22.0
RTOP Culture: 17.7

DCI: Weather and climate. MS-ESS2-5. Create a poster with a scaled replica of a model to illustrate cloud type, altitude, and weather conditions with which they are associated. In warm up, teacher tells students how interacting air masses cause wind; this information is never connected to main focus of lesson where students work on making poster.

SEP: 2; 8. Teacher developed procedure to replicate diagram from textbook. Students will eventually research and include on poster conditions associated with each cloud type. Teacher provides sources for student research. Students replicate rather than develop, use, or revise model. Although students research and must present weather conditions associated with each cloud type, the presentation is not used to evaluate merit or validity of concepts or methods.

CCC: Scale, proportion, and quantity. Students must scale poster and place clouds accurately at altitude they are known to form. Teacher makes sure students are following directions, but never facilitates why an accurately scaled representation is needed or important; students mindlessly follow instructions.

Figure 9. Teacher 16's implementation of pre-RET materials was out of alignment with lesson materials; lesson transformed from an inquiry lesson to a verification lab. Some key features of lesson used to place lesson on SQ-EQuIP are underlined.

observed in two separate lessons. In these two lessons, although the teachers had an ideal placement on the SQ-EQuIP, their RTOP scores were relatively low for these ideal lessons (e.g., post-MA lesson in Figure 7). RTOP scores in these lessons were likely lower because although students were highly engaged in a complex synthesis of scientific content and ways of thinking, the lessons did not involve students designing or engaging in investigations.

Finally, the percentage frequency with which teachers in each group implemented lessons that included grade appropriate dimensions — as indicated by the NGSS — across the three time points was compared. We considered a dimension as grade appropriate if the identified DCI, SEP, or CCC was at or above the NGSS grade band of the students observed in the lesson. At both pre-RET and post-RET comparison, teachers were more likely than PD teachers to include at least one grade-appropriate DCI in the observed lessons (Figure 10).

However, at post-MA 90% of PD teachers included at least one grade-appropriate DCI, compared with 77.7% of comparison teachers. Across all three time points, PD teachers were more likely to include at least one grade-appropriate SEP. When percentage frequencies were compared between pre-RET lessons and post-MA lessons, neither group showed any change in the frequency of their inclusion of grade-appropriate SEPs. Pre-RET, the inclusion of at least one grade-appropriate CCC was observed in 30% and 33.9% for PD and comparison teachers, respectively. Post-RET, the percentage of PD teachers who implemented a lesson with at least one grade-appropriate CCC more than doubled, to 70%, while there was no change in the percentage of comparison teachers whose lessons included at least one grade-appropriate CCC. At post-MA, 80% of PD teachers included a grade- appropriate CCC, versus 66.7% of comparison teachers.

Discussion

The focus of this study was to determine how the SQ-EQuIP enabled a longitudinal comparison of teachers' use of 3D instruction, to compare how SQ-EQuIP evaluations performed relative to the RTOP, and to ascertain whether a PD program designed to increase teachers' ability to teach with inquiry-based practices also increased their ability to implement grade appropriate 3D instruction. This discussion section is structured within the context of the research questions, the results of the data analysis, and existing literature.

How does PD teachers' instruction change in relation to grade appropriate 3D instruction and in relation to comparison teachers as characterized by the SQ-EQuIP?

After two years PD program participation, 80.0% of PD teachers had progressed or maintained an ideal of near ideal placement of the SQ-EQuIP, compared with 22.2% of comparison teachers. When grade appropriateness of all three dimensions was taken into account along with ideal or near-ideal placement on the SQ-EQuIP, 60% of PD teachers achieved both, while no comparison teacher had this achievement, post-MA. The inclusion of below grade-level dimensions occurred in a minority of observed lessons, but is notable because they occurred in two post-MA lessons of PD teachers and the only two comparison teachers who were at or near ideal placement on the SQ-EQuIP. It is notable because it signals that some teachers, although comfortable with explicitly and coherently

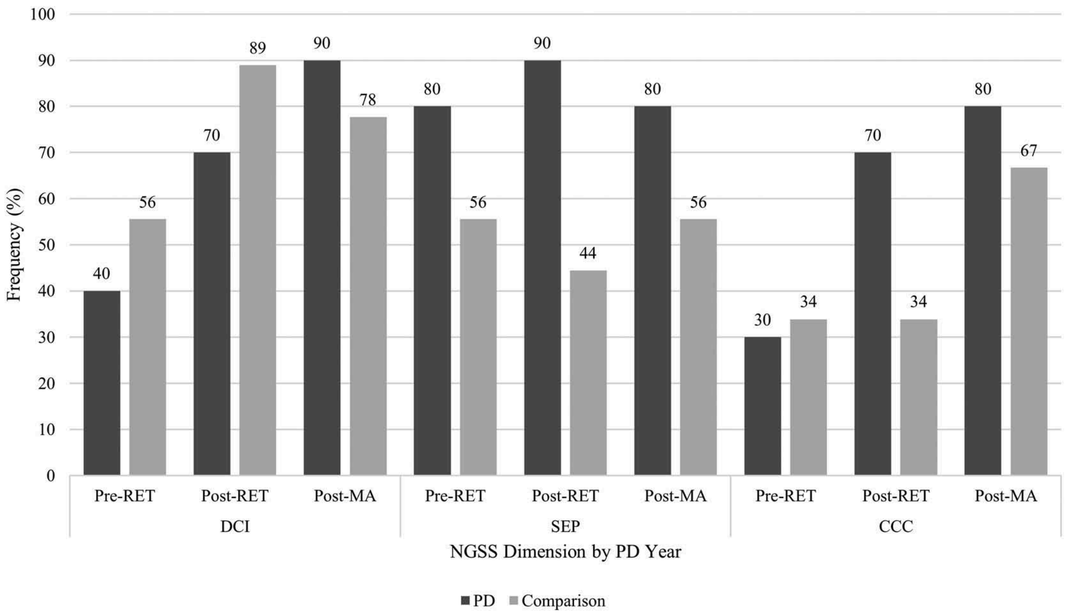


Figure 10. Percentage frequency at which each teacher group included grade-level appropriate dimensions across the three time points of the PD program. *Note*. SEPs were only included in the frequency count if authentically used.

engaging students with the three dimensions, are not selecting and engaging students with more in-depth and sophisticated ways of thinking of scientific and engineering concepts as students advance through grade levels as intended by *A Framework for K-12 Science Education* (NRC, 2012) and the NGSS (NGSS Lead States, 2013). Further, CCCs were the least likely of the three dimensions to be observed in both PD and comparison teachers pre-RET. Pre-RET, only about one-third of observed lessons of both PD and comparison teachers included a CCC. This was not an unexpected finding, as these unifying concepts were not overtly included in previous science education standards (NRC, 2012). Therefore, teachers likely had less awareness of a need to or knowledge of how to incorporate these concepts in their instruction. However, after one year, nearly three-quarters of observed PD teachers' lessons included a grade-appropriate CCC while only one-third of comparison teachers included a grade-appropriate CCC. This comparatively rapid increase in PD teacher inclusion of a grade-appropriate CCC one year after PD indicates that the RET, a content-focused experience strengthened by the modeling of effective pedagogical skills, also supported teachers' ability to incorporate these unifying concepts as intended by the NGSS (NGSS Lead States, 2013).

In the lessons where one or more dimensions were implicitly present on the SQ-EQuIP rubric (Achieve, 2014) and the SEPs authentically used, one or more of the following reasons accounted for the placement: (a) teacher did not in any way engage students with inquiry-based features that were present in lesson materials; (b) teacher gave away the observations or conclusions before student investigation; (c) *students* failed to explicitly and coherently engage with one or more dimensions despite teacher attempts to address the dimension(s) in a lesson. As a result, the quality of the classroom materials a teacher designs or chooses to use may not be indicative of their ability to effectively implement the materials in relation to 3D instruction. Therefore, we reinforce the assertion by CADRE (2016) and Roseman et al. (2015) that one weakness of the EQuIP rubric (2014) is that it does not account for 3D instructional coherency from the student's point of view. This weakness has not been addressed in the final, quantitative version of the EQuIP rubric (2016). Therefore, although the EQuIP rubric is currently the only tool that is explicitly aligned to the NGSS and available to PD program developers, PD developers should be encouraged to adapt the EQuIP rubric in such a way that it enables them to evaluate how the teachers and students interact with lesson materials within the real-world contexts of science classrooms. The SQ-EQuIP offers one model of how the EQuIP (2014, 2016) can be adapted for observational use.

How does the change in teachers' instruction as characterized by the SQ-EQuIP compare with changes tracked by the RTOP?

Overall, the RTOP scores tracked teacher performance on the SQ-EQuIP well. High RTOP scores (typically 75 or higher) tended to be only be associated with lessons that were at or near the ideal on the SQ-EQuIP. However, the RTOP failed when a DCI was absent, but students were actively engaged with SEP 3 — *Planning and Carrying Out Investigations* (see pre-RET lesson in Figure 6). We believe this occurred because students planning and carrying out investigations seemed to easily allow teachers to demonstrate valuing and facilitating student ideas as well as facilitating student communication of their ideas with peers. This finding confirmed our expectation that appropriate use of the SQ-EQuIP would reveal this failure, as Marshall et al. (2011) previously identified this global constructivist emphasis of the RTOP. However, we were surprised that the RTOP also failed when students were actively engaged with DCIs, SEPs, and CCCs in sophisticated ways, but the lesson did not specifically include SEP 3. Upon examination, four of the 25 items (items # 4 and 11–13; Sawada et al., 2002) on the RTOP or 16 of the 100 possible points could be directly linked to planning and carrying out investigations. This latter failure of the RTOP in this study highlights that teachers can use complex 3D activities that do not involve students planning and carrying out investigations. An example of such a complex 3D activity is the post-MA lesson in Figure 7, which was a review for a classroom exam addressing the structure, function, and interactions of the brain with other biological systems such as the respiratory system in humans.

Statistical analyses indicated that the PD teacher group entered the PD program with higher mean scores on all RTOP scales, and these scores remained higher and consistently increased throughout the PD program relative to the comparison teacher group. This consistent increase occurred despite the PD teachers' lower potential to make gains given their higher RTOP mean scores upon PD program entry. The comparison teacher group showed either a decrease or no change in their mean RTOP scores pre-RET to post-RET, but an overall gain pre-RET to post-MA. This overall gain made by the comparison teachers is most plausibly explained by comparison teachers attending district required PD events as the state began its transition to the NGSS during the delivery of the PD program.

Further, despite PD teachers' lower potential for growth as indicated by their higher mean RTOP scores upon PD program entry, PD teachers' gain on the RTOP content scale was 8.5 times higher than the gain made by comparison teachers and their gain on the RTOP classroom culture scale was twice the gain made by comparison teachers. It is notable that most of the gain PD teachers had on the content scale score occurred post-RET while there was no observed change in comparison teachers mean score for this scale. This post-RET gain by PD teachers on the content scale occurred alongside their more than doubling the percentage frequency of the number of lessons that included a grade appropriate CCC from pre-RET to post-RET lesson while no change was observed in this frequency for comparison teachers. These trends support the findings of previous qualitative studies (Herrington et al., 2016; Herrington, Bancroft, Edwards, & Tanis, 2017) that the RET was likely the core PD experience with the greatest impact on PD teachers' instructional practices.

Teachers in the gradual progression pathway entered the PD program with the lowest RTOP scores (average total RTOP score of 58) and were also the teachers who took two years to implement a lesson that was at or near ideal on the SQ-EQuIP. In comparison, teachers who entered with higher scores (average total RTOP score 73) were typically placed at or near ideal at pre-RET or post-RET. The overall alignment between RTOP scores and the placement of an observed lesson on the SQ-EQuIP implies that RTOP scores at pre-RET in conjunction with placement on the SQ-EQuIP can serve as one baseline indicator that PD developers can use to identify teachers who will likely need more time and support to reform their instruction. Therefore, despite its weaknesses and lack of alignment to 3D instruction, the RTOP still seems to be capable of offering PD developers useful insight into teacher practices. Further, the SQ-EQuIP's visual representation of what dimensions were included in an observed lesson, the extent to which each dimension was present in the lesson, and how well the dimensions were integrated allowed for easy tracking of each teacher's progress toward 3D instruction across all three time points included in this longitudinal study. We believe this visual representation of the Alignment to NGSS category on the EQuIP over time presents nuanced snapshots about changes in teacher instructional practices that neither the qualitative descriptions on version 2 of the EQuIP (Achieve, 2014) nor the quantitative scores or Likert scale (inadequate = 0 to extensive = 3) on version 3 of the EQuIP (Achieve, 2016) can provide.

Conclusions

The ideal or near ideal placement of 80.0% of PD teachers' lessons and 22.2% of comparison teachers' lessons on the SQ-EQuIP after two years (post-MA) indicates that that the features of this PD program, aligned to the NSES (NRC, 1996) and best practices for science PD programs (Garet et al., 2001; Loucks-Horsley, Stiles, Mundry, Love, & Hewson, 2010; Supovitz & Turner, 2000), increased teacher ability to explicitly and coherently engage middle- and high-school students with 3D learning. The two post-MA lessons from PD teachers that were not at ideal or near ideal included materials that explicitly and coherently integrated all three dimensions. However, these teachers did not support the students' explicit and coherent 3D engagement. This struggle was observed in most pre-RET lessons in both teacher groups, half of the PD teachers' post-RET lessons, and most comparison teachers' post-RET and post-MA lessons. This finding is a caution to teachers and PD developers that lesson materials that are rated high on the EQuIP rubrics (Achieve, 2014, 2016) will likely only translate into high-quality science instruction if teachers possess the content and pedagogical content knowledge to support student connections among the three dimensions within those materials.

Placements on the SQ-EQuIP triangulated well with total RTOP scores. The SQ-EQuIP facilitated the evaluation and long-term tracking of middle- and high-school science teachers' ability to implement 3D instruction with an emphasis on the student's perspective of this implementation. Ultimately, we found that a long-term PD program designed around the NSES (NRC, 1996) increased middle- and high-school science teacher ability to implement three-dimensional instruction compared with a comparison group of teachers.

Study limitations

Although the sampling of 19 teachers from suburban and rural western Michigan schools restricts both claims of significance and the generalizability of the results of this study, results may be transferable to other longer-term PD opportunities closely aligned to the NSES that used best practices for science PD programs, and for which teachers volunteered. Additionally, this study did not focus on teacher ability to provide instructional support and to monitor student progress. Therefore, we recognize as a limitation of the study our inability to offer insight on the extent to which PD teachers connected classroom science to students' lives; connections addressed in the instructional supports and monitoring student progress categories of the EQuIP (2014, 2016). As classrooms become increasingly diverse in terms of students' heritage and language, such connections are crucial to promote student engagement with and interest in science (NGSS Lead States, 2013; Rodriguez, 2015).

Therefore, as PD program developers move toward the NGSS, incorporation of PD experiences that support equitable 3D science experiences for all students must be a priority.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the NSF EHR (Directorate for Education & Human Resources) Grant no: 1118658.

ORCID

Senetta F. Bancroft http://orcid.org/0000-0002-7498-1169
Deborah G. Herrington http://orcid.org/0000-0001-6682-8466

References

Achieve. (2014). *EQuIP rubric for lessons and units: Science*. Retrieved from http://www.nextgenscience.org/resources/equip-rubric-lessons-units-science

Achieve. (2016). *EQuIP rubric for lessons and units: Science*. Retrieved from https://www.nextgenscience.org/sites/default/files/EQuIPRubricforSciencev3.pdf

Akkus, R., Gunel, M., & Hand, B. (2007). Comparing an inquiry-based approach known as the science writing heuristic to traditional science teaching practices: Are there differences? *International Journal of Science Education*, *29*(14), 1745–1765. doi:10.1080/09500690601075629

American Association for the Advancement of Science. (1989). *Project 2061—Science for all Americans*. Washington, DC: Author.

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

Blanchard, M. R., Southerland, S. A., Osborne, J. W., Sampson, V. D., Annetta, L. A., & Granger, E. M. (2010). Is inquiry possible in light of accountability? A quantitative comparison of the relative effectiveness of guided inquiry and verification laboratory instruction. *Science Education*, *94*(4), 577–616. doi:10.1002/sce.v94:4

Bodzin, A. M., & Beerer, K. M. (2003). Promoting inquiry-based science instruction: The validation of the science teacher inquiry rubric (STIR). *Journal of Elementary Science Education*, *15*(2), 39–49. doi:10.1007/BF03173842

Borko, H., & Putnam, R. (1995). Expanding a teacher's knowledge base: A cognitive psychological perspective on professional development. In T. Guskey & M. Huberman (Eds.), *Professional development in education: New paradigms and practice* (pp. 35–61). New York, NY: Teachers College Press.

Community for Advancing Discovery Research in Education. (2016). *Teaching and learning under the Next Generation Science Standards*. Retrieved from http://cadrek12.org/resources/stem-smart-brief-teaching-and-learning-under-next-generation-science-standards

Crawford, B. A. (2007). Learning to teach science as inquiry in the rough and tumble of practice. *Journal of Research in Science Teaching*, *44*(4), 613–642. doi:10.1002/(ISSN)1098-2736

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199.

Enderle, P., Dentzau, M., Roseler, K., Southerland, S., Granger, E., Hughes, R., & Saka, Y. (2014). Examining the influence of RETs on science teacher beliefs and practice. *Science Education*, *98*(6), 1077–1108. doi:10.1002/sce.21127

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*(4), 915–945. doi:10.3102/00028312038004915

Gess-Newsome, J. (1999). Pedagogical content knowledge: An introduction and orientation. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining pedagogical content knowledge* (pp. 3–17). New York, NY: Kluwer Academic Publishers.

Hayes, K. N., Lee, C. S., DiStefano, R., O'Connor, D., & Seitz, J. C. (2016). Measuring science instructional practice: A survey tool for the age of NGSS. *Journal of Science Teacher Education*, *27* (2), 137–164. doi:10.1007/s10972-016-9448-5

Heath, B., Lakshmanan, A., Perlmutter, A., & Davis, L. (2010). Measuring the impact of professional development on science teaching: A review of survey, observation and interview protocols. *International Journal of Research & Method in Education*, *33*(1), 3–20. doi:10.1080/17437270902947304

Herrington, D. G., Bancroft, S. F., Edwards, M. M., & Schairer, C. J. (2016). "I want to be the inquiry guy!" How research experiences for teachers change beliefs, attitudes, and values about teaching science as inquiry. *Journal of Science Teacher Education*, 27(2), 183–204.

Herrington, D. G., Bancroft, S. F., Edwards, M. M., &, Tanis, S. (2017, April). *Changing teacher values about science instruction: Cumulative influences of a research experience and materials development*. National Association for Research in Science Teaching, San Antonio, TX.

Kimberlin, C. L., & Winetrstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276–2284. doi:10.2146/ajhp070364

Krajcik, J. (2015). Project-based science: Engaging students in three-dimensional learning. *The Science Teacher*, *82*(1), 25–27. doi:10.2505/4/tst15_082_01_25

Krajcik, J., Codere, S., Dahsah, C., Bayer, R., & Mun, K. (2014). Planning instruction to meet the intent of the next generation science standards. *Journal of Science Teacher Education*, *25*(2), 157–175.

Loucks-Horsley, S., & Stiegelbauer, S. (1991). Using knowledge to guide staff development. In A. Lieberman & L. Miller (Eds.), *Staff development for education in the 90's: New demands, new realities, new perspectives* (pp. 15–36). New York, NY: Teachers College Press.

Loucks-Horsley, S., Stiles, K. E, Mundry, S., Love, N., & Hewson, P.W. (2010). *Designing professional development for teachers of science and mathematics*. Corwin Press.

Marshall, J. C., Smart, J., & Horton, R. M. (2009). The design and validation of EQUIP: An instrument to assess inquiry-based instruction. *International Journal of Science and Mathematics Education*, *8*(2), 299–321. doi:10.1007/s10763-009-9174-y

Marshall, J. C., Smart, J., Lotter, C., & Sirbu, C. (2011). Comparative analysis of two inquiry observational protocols: Striving to better understand the quality of teacher-facilitated inquiry-based instruction. *School Science and Mathematics*, *111*(6), 306–315. doi:10.1111/j.1949-8594.2011.00091.x

Miner, D., & DeLisi, J. (2012). *Inquiring into science instruction observation protocol (ISIOP): Data collection instrument*. Retrieved from http://ltd.edc.org/resource-library/inquiring-science-instruction-observation-protocol-isiop

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

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: The National Academies Press.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, *62*(3), 307–332. doi:10.3102/00346543062003307

Phelps, A., & Lee, C. (2003). The power of practice: What students learn from how we teach. *Journal of Chemical Education*, *80*, 829–832. doi:10.1021/ed080p829

Piburn, M., Sawada, D., Turley, J., Falconer, K., Benford, R., Bloom, I., & Judson, E. (2000). *Reformed teaching observation protocol (RTOP): Reference manual* (ACEPT Technical Report No. IN00-3). Tempe, AZ: Arizona Collaborative for Excellence in the Preparation of Teachers. Retrieved from ERIC database. (ED 447 205).

Rodriguez, A. J. (2015). What about a dimension of engagement, equity, and diversity practices? A critique of the next generation science standards. *Journal of Research in Science Teaching*, *52* (7), 1031–1051. doi:10.1002/tea.21232

Roseman, J., Fortus, D., Krajcik, J., & Reiser, B. J. (2015). *Curriculum materials for Next Generation Science Standards: What the science education research community can do*. Paper presented at NARST Annual International Conference, Chicago, IL.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics*, *102*(6), 245–253. doi:10.1111/ssm.2002.102.issue-6

Schultz, S. E., & Pecheone, R. L. (2014). Assessing quality teaching in science. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 444–492). San Francisco, CA: Jossey-Bass.

Shumba, O., & Glass, L. (1994). Perceptions of coordinators of college freshman chemistry regarding selected goals and outcomes of high school chemistry. *Journal of Research in Science Teaching*, *31*(4), 381–429. doi:10.1002/tea.3660310407

Supovitz, J. A., & Turner, H. M. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching*, *37*(9), 963–980. doi:10.1002/(ISSN)1098-2736

Wilson, C. D., Taylor, J. A., Kowalski, S. M., & Carlson, J. (2010). The relative effects and equity of inquiry-based and commonplace science teaching on students' knowledge, reasoning, and argumentation. *Journal of Research in Science Teaching*, *47*(3), 276–301.

Wong, S., & Luft, J. (2015). Secondary science teachers' beliefs and persistence: A longitudinal mixed-methods study. *Journal of Science Teacher Education*, *26*(7), 619–645. doi:10.1007/s10972-015-9441-4

Yezierski, E. J., & Herrington, D. G. (2011). Improving practice with target inquiry: High school chemistry teacher professional development that works. *Chemistry Education Research and Practice*, *12*(3), 344–354.