Online Video-Based Lesson Analysis Professional Development: A Course for High School Science Teachers

Susan M. Kowalski, Betty Stennett, Mark Bloom, Karen Askinas

BSCS

Austin Lukondi

Colorado College

Pamela Van Scotter

BSCS

This research was supported by a grant from the National Science Foundation (1118643).

We gratefully acknowledge the efforts of Catherine Stimac, Heather Young, and MacGregor Campbell at Oregon Public Broadcasting.

Correspondence concerning this article should be addressed to Susan M. Kowalski, BSCS, 5415 Mark Dabling Blvd., Colorado Springs, CO 80918. E-mail: skowalski@bscs.org

Online Video-Based Lesson Analysis Professional Development:

A Course for High School Science Teachers

Analyzing classroom video to enhance teachers' knowledge and practice is becoming more common in professional development (PD). Mathematics PD has long used video as a focal point for teachers' collaborative examination of the central activities of mathematics teaching (Borko, 2011). Using video to analyze practice "allows one to enter the world of the classroom without having to be in the position of teaching in-the-moment" (Sherin, 2004, p.13). Slowing down the teaching so that teachers can dig deeply into student thinking, content storyline, and their own ideas about specific content is emerging as an effective PD paradigm.

Building on the early video analysis work of mathematics education researchers (e.g., Seago & Mumme, 2000; Seago, 2003; Seago, 2004; Sherin and van Es, 2002; Sherin & Han, 2004; van Es & Sherin, 2002), Ball and Cohen's (1999) call for practice-based teacher education, and her own work as the principal investigator of the 1999 Third International Math and Science Video Study (Roth et al., 2006), Kathleen Roth developed *Science Teachers Learning through Lesson Analysis* (STeLLA; Roth et al., 2011; National Academies of Sciences, Engineering, and Medicine, 2015). Roth and her colleagues have shown that STeLLA can be effective with elementary science teachers (Roth et al., 2011; Taylor et al., 2016). But taking the effective STeLLA model to scale would require buy-in and significant funding from school districts. The high cost of face-to-face PD makes it unlikely that teachers in small rural districts, remote areas, and even larger districts with limited PD budgets would have access to STeLLA; yet, all teachers should have access to effective PD no matter where they live or how well funded their district. An online lesson analysis PD based on the STeLLA model would decrease costs, decrease personnel requirements, and allow more teachers to participate, including these teachers from small, remote districts or economically disadvantaged districts.

BSCS in collaboration with Oregon Public Broadcasting (OPB), the National Teachers Enhancement Network (NTEN) at Montana State University, the National Renewable Energy Laboratory (NREL), and the Great Lakes Bioenergy Research Center (GLBRC) developed and studied the use of an innovative online, multimedia, professional development course focused on energy-related concepts within a context of the production and use of alternative energy. The course, *Energy: A Multidisciplinary Approach for Teachers* (EMAT), takes advantage of the affordances of a multimedia environment, incorporating animations, classroom videos, and interactive learning experiences as part of the overall instruction related to energy concepts.

Concepts related to energy are fundamental to understanding science and they weave through each of the science disciplines. A basic understanding of key energy concepts is an important component of science literacy. Science educators have long recognized the importance of energy as a core organizing concept, as reflected in the high number of energy-related concepts in the National Science Education Standards (NRC, 1996), American Association for the Advancement of Science (AAAS) benchmarks (AAAS, 1993), and Next Generation Science Standards (NGSS; NGSS Lead States, 2013). Teachers who struggle with this content are not able to effectively address student misconceptions and help students see how energy concepts are connected and relate to their lives. Energy-related concepts have always been essential to making informed decisions about resource use and management. Current concerns over climate change have underscored the importance of using a systems approach to understanding energy and matter concepts.

Despite the centrality of energy concepts across science disciplines and with standards, students and teachers have difficulty in developing sound conceptual understandings of energy (Black & Solomon, 1983; Doménech et al., 2007; Sağlam-Arslan & Kurnaz, 2009). In fact, student difficulties in understanding energy promoted the development of learning models such

as the conceptual-change learning model (Posner, Strike, Hewson, & Gertzog, 1982; Driver & Oldham, 1985). The challenge of teaching energy literacy remains a priority for science educators.

Theoretical Framework for the Study

The facilitated online PD course uses a two-faceted theoretical approach as shown in Figure 1: situated cognition (using video-based lesson analysis) and constructivist learning theories (using the BSCS 5E Instructional Model). Evidence indicates that both facets of the course can enhance teachers' knowledge and practice (Bybee et al., 2006; Roth et al., 2011; Taylor et al., 2016). While these two facets of PD have been used and studied independently, EMAT has merged them in an online course such that they may synergistically promote teacher learning.

According to situated cognition theory, learning is naturally tied to authentic activity, context, and culture (Brown, Collins, & Duguid, 1989). The situated cognition approach to PD provides teachers contexts in which they can integrate the many complex aspects associated with teaching. The contexts "enable teachers to see content and teaching issues embedded in real classroom contexts; … treat content as central and intertwined with pedagogy; [and] focus on the specific content … teachers are teaching" (Roth et al., 2011). Grounding teachers' learning in the context of teaching practice raises teachers' motivation as they grapple with ideas within

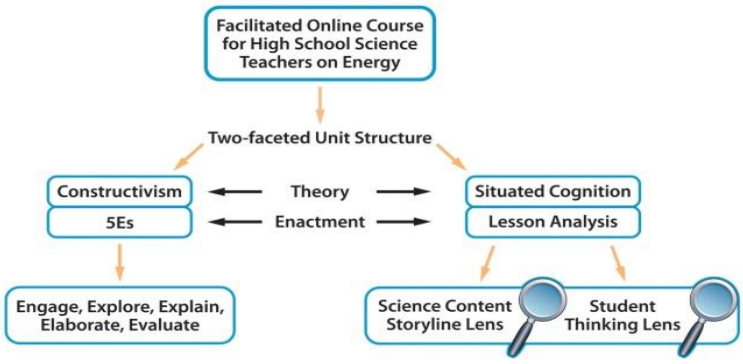

Figure 1. **Course structure by unit.**

situations where they will be using the knowledge (Garet et al., 2001). The situated cognition approach contrasts sharply with traditional PD programs that are typically short term and isolated from teachers' classrooms and approach content deepening and pedagogy as distinct objectives (Ball & Cohen, 1999).

The constructivist approach to PD allows teachers to grapple with their own prior conceptions about energy, develop explanations from evidence within a coherent conceptual framework, and reflect on their thinking as their conceptual understanding develops. Constructivist approaches have been shown to be effective for science content learning for both students (Wilson, Taylor, Kowalski, & Carlson, 2010) and teachers (Raya-Carlton, Weaver, & Krebs, 2010). The course, targeted at high school science teachers, aims to enhance teacher knowledge and practice to ultimately improve student learning.

Methods

The goal of our research was to examine the promise of efficacy of the various components of the course. In spite of the rapid expansion of online PD, little is known about best practices for its design and implementation (Dede, 2006; Dede et al., 2006). In particular, there remains a dearth of evidence linking PD of any kind to student learning(National Academies of Sciences, Engineering, and Medicine, 2015). Most PD research is evaluative in nature and fails to attend to longitudinal effects of the PD. Our work counters this trend by addressing the following research questions:

1. Do teachers demonstrate improved content knowledge about energy concepts after participating in the EMAT course? If so, is the difference statistically and practically significant?

2. Do teachers demonstrate improved ability to analyze lessons for evidence of student thinking and coherence of science content after participating in the EMAT course? If so, is the difference statistically and practically significant?
3. Do teachers demonstrate improved teaching practice through the appropriate use of key strategies as described in the STeLLA lesson analysis framework (Figure 2) in the year following participation in the EMAT course? If so, is the difference statistically and practically significant?
4. After participating in EMAT, do teachers help students attain higher posttest scores (pretest adjusted) than they did for their prior year's students, taught before teachers' participation in EMAT? If so, is the difference statistically and practically significant?
5. Which components of the EMAT course are most effective in enhancing teacher knowledge and practice, and why do those components seem to be effective?

The questions about teacher knowledge are associative—our design for the teacher outcomes is a pretest-posttest design. We will not make causal claims about our teacher outcomes. However, our research examines the extent to which we can attribute changes in student learning to teachers' participation in the course. As such, we used a design that would allow us to make a causal link between the intervention and student outcomes—a design that requires a comparison group as well as baseline measures to enhance the confidence we can have in making causal claims. For student outcomes our study uses a cohort-control quasi-experimental design. Teachers in the research project participated over two school years. Teachers' students the first year constituted the comparison group, and their students in the second year constituted the treatment group. Teachers took the course in the summer between the two school years. We conducted the entire quasi experiment *twice* over two national field tests. We recruited 35 teachers in the first field test and 39 teachers for the second field test. Teachers

from across the US participated in the project, with a preference given to those teachers who taught students from racial/ethnic groups that are typically underrepresented in the sciences (including African American, Latino, American Indian, and Native Hawaiian/Pacific Islander students).

We organize our paper as follows. First we provide a full description of the EMAT course and how it compares to its parent PD, STeLLA. Second, we examine teacher knowledge and practice outcomes. Third, we examine student outcomes. Fourth, we share our findings on which elements of the course seem to be the most effective. Finally, we reflect on our findings for teachers and students, consider the strengths and limitations of the EMAT course, and consider implications for taking the STeLLA PD model to high school teachers online.

Design of the EMAT Course

Course Structure: Overview

EMAT is a 10-week summer graduate course for high school science teachers. Teachers received three hours of graduate credit for their successful completion of the course. EMAT situates energy concepts within alternative energy contexts and embeds lesson analysis PD throughout the course. Using select strategies from the STeLLA conceptual framework, teachers learn to reveal, support, and challenge student thinking and create coherent science content storylines with their lessons. The Student Thinking (ST) Lens and the Science Content Storyline (SCS) Lens make up the two-lens approach to science teaching and lesson analysis that is a hallmark of the STeLLA model of PD.

The EMAT course was developed according to the BSCS 5E Instructional Model (Bybee et al., 2006). The 5E Instructional Model allows learners (whether those learners are teachers or students) to *engage* in a concept and express their current understanding. They *explore* it in a range of ways before beginning to construct an *explanation*. The model then provides an

opportunity for learners to *elaborate* their understanding either by delving deeper into a concept or applying the concept in a new situation and to *evaluate* their growing understanding of the concept before encountering a new one. These five phases in the model (Engage, Explore, Explain, Elaborate, and Evaluate) provide teachers opportunities to build their understanding through carefully structured experiences. Both the content and pedagogy portions of the course rely on the 5E Instructional Model.

Course Structure: Energy Concepts

Teachers develop understanding of key energy concepts and the processes of electrical energy generation across six units. The first unit, Coal, provides a foundation for teachers to learn about our country's most common energy source and to use it as a reference for comparison with other electrical energy technologies. Participants learn about electromagnetic induction and steam turbine generators and use a systems approach to consider both monetary and environmental costs for a coal-fired power plant. The remaining five units present alternative sources for generating electricity, heating and cooling homes, or producing fuel for vehicles. The units include Nuclear, Wind, Geothermal, Biofuels, and Solar. The course emphasizes three key ideas in each unit:

1. Energy can be neither created nor destroyed.
2. Tracking energy and matter inputs and outputs within systems helps promote understandings about the system's potentials and limitations.
3. Energy transfer is never 100% efficient; some energy always leaves the system as heat.

These three key ideas are associated with two of the crosscutting concepts identified by the NGSS (NGSS Lead States, 2013), namely, *Energy and Matter: Flows, Cycles, and Conservation* and *Systems and System Models*.

Course Structure: Pedagogy

Each of the six units also includes a set of STeLLA-based lessons aimed at enhancing teachers' pedagogical knowledge and practice. Within the lesson analysis portion of each unit, teachers have opportunities to

- examine video examples of students learning energy concepts,
- analyze videos for student thinking and coherence of science content storylines,
- collaborate with colleagues in synchronous discussions while analyzing video,
- reflect on the STeLLA strategies, and
- consider ways to apply the STeLLA strategies to their own teaching.

Due to course time limitations, we were unable to include all 17 of the STeLLA strategies in EMAT. Based on consultation with Roth and based on her earlier findings (Roth et al., 2011), we selected a subset of 12 high-leverage strategies for the course as indicated in Figure 2. The highlighted strategies are those included in the EMAT course.

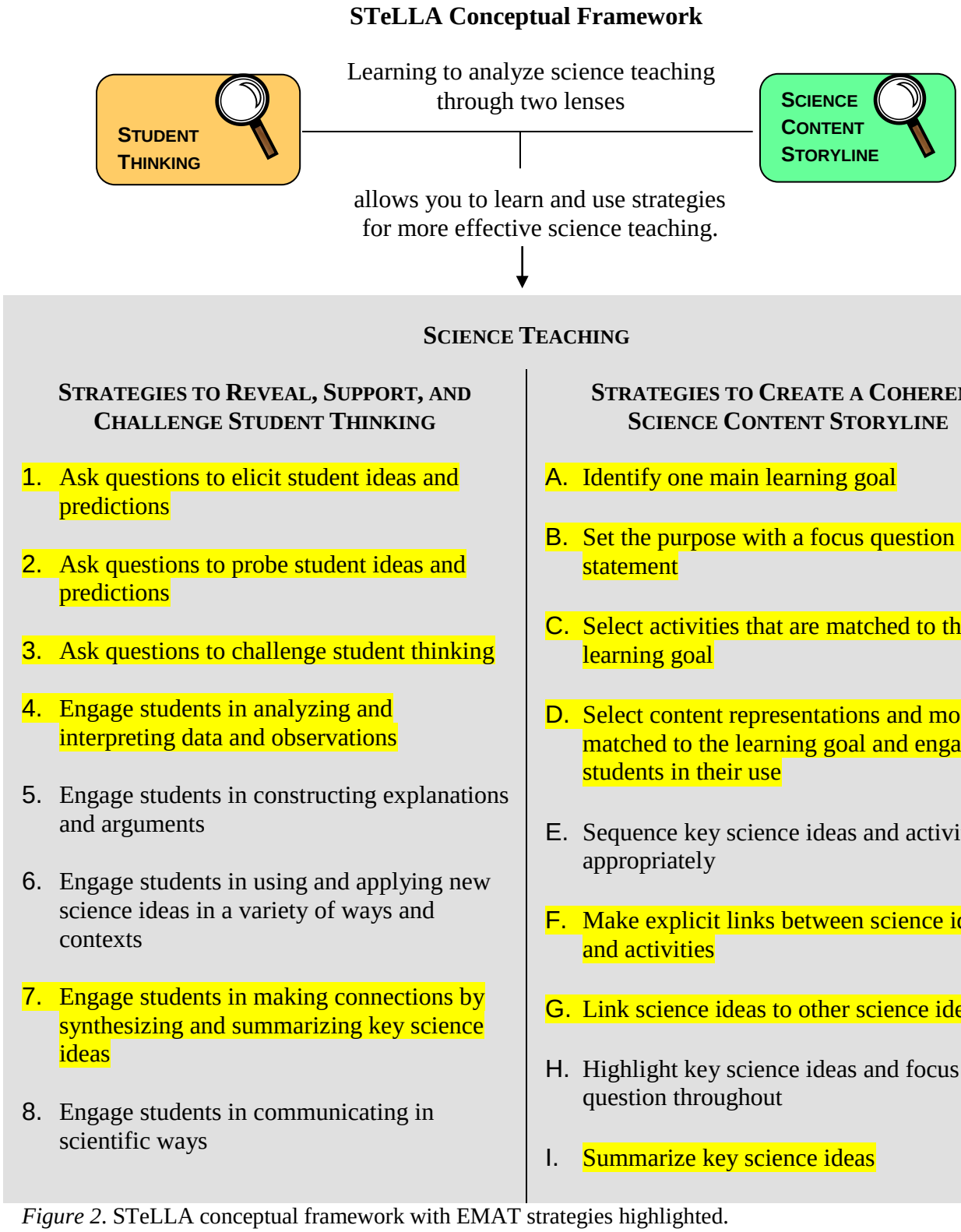

Figure 2. STeLLA conceptual framework with EMAT strategies highlighted.

An integral part of lesson analysis is small-group discussion around videos. Teachers participated in one two-hour synchronous online discussion near the end of each unit (a total of six discussions for each group of teachers). During these synchronous discussions, teachers used protocols to analyze videos for strategies to reveal, support, or challenge student thinking. The teachers also analyzed videos for coherent science content storylines by engaging in tasks such as examining lessons for one main learning goal, considering links between science ideas and activities, and considering links between science ideas and related science ideas.

Comparison of EMAT to STeLLA PD with elementary teachers

We made significant modifications of the STeLLA PD model in designing EMAT. Some of these modifications were to accommodate high school teachers of various science disciplines. Others were related to the online delivery mechanism and the condensed amount of time with teachers. Table 1 compares the elementary teacher face-to-face PD program with EMAT. We discuss elements of the course that were successful or posed challenges in the section that follows.

Table 1
Comparison of STeLLA with EMAT

	STeLLA	**EMAT**
Audience	Elementary teachers	High school science teachers
Science content	Water cycle, food webs, Earth's changing surface, and Sun's effect on climate	Energy (in the context of alternative energy sources)
Delivery	Face-to-face	Asynchronous and synchronous online
Hours	88.5 hours (31.5 hours of science content deepening + 57 hours of lesson analysis)	120 hours (average) (72 hours of science content deepening + 48 hours of lesson analysis)
Curriculum	Student lessons initially provided with strategies embedded. Teachers ultimately developed their own student lessons as part of the PD.	No student curriculum provided. Teachers had assignments to modify existing lessons and embed strategies from the course.
Strategies	Full suite of strategies from both lenses (Figure 2)	Selected strategies from both lenses (highlighted in Figure 2)

Outcomes

Teacher Content Knowledge, Research Question 1

In this section we address the first research question of the EMAT project: *Do teachers demonstrate improved content knowledge about energy concepts after participating in the EMAT course? If so, is the difference statistically and practically significant?*

To assess changes in teacher content knowledge, the teacher participants completed a pretest before and a posttest following each unit. Each test consisted of approximately 20–25 questions. Most questions were multiple choice, but each test also included several open-ended response items. Regardless of item type, each item was worth 1 point and was scored as either correct (1 point) or incorrect (0 points). At the start, 35 teachers took the course. We collected data from the 28 teachers who finished at least one unit pretest and one unit posttest. Each unit test included items directly related to each of the three key energy concepts as well as unit-specific items. For example, all units included items assessing the idea that energy always leaves a system in a nonuseful form as heat, but only the Solar unit included items related to the photovoltaic effect.

A summary of the changes in teacher content knowledge for all six of the units is seen in Table 2. Teachers displayed significant gains ($p < .001$) in content knowledge for the Coal, Nuclear, Biofuels, and Solar units. The effect sizes are in the range of 1.09 for Nuclear to 1.82 for Coal. Gains also were seen for the Wind and Geothermal units but they were not statistically significant.

Table 2

Teacher Content Pretest and Posttest Scores

Assessment	N	Pre mean (SD)	Post mean (SD)	t-statistic	p-value	Pre-post effect size (d)	Confidence interval around effect size	
							Lower	Upper
Unit 1: Coal	28	14.3 (3.2)	20.4 (3.2)	10.81	< .001	1.82	1.28	2.36
Unit 2: Nuclear	26	16.6 (4.8)	22.2 (4.5)	5.52	< .001	1.09	0.60	1.58
Unit 3: Wind	24	13.7 (3.3)	14.9 (3.2)	1.27	.216	0.23	-0.13	0.60
Unit 4: Geothermal	25	14.9 (4.1)	16.0 (2.0)	1.65	.111	0.30	-0.06	0.67
Unit 5: Biofuels	25	19.0 (2.9)	23.3 (2.2)	9.54	< .001	1.56	1.08	2.03
Unit 6: Solar	24	17.1 (4.1)	21.9 (3.6)	8.43	< .001	1.23	0.85	1.61
Total across units (Rasch person measures)	28	0.20 (0.5)	1.1 (0.6)	16.48	< .001	1.71	1.39	2.03

We carried out a separate analysis to assess changes in teacher content knowledge regarding the three energy-related themes. For each unit test, selected questions that aligned with an energy-related theme were included in the analysis. The results in Table 3 show significant gains ($p < .001$) in knowledge about each of the three energy-related themes. The effect sizes are 0.94 for conservation of energy, 0.81 for energy efficiency, and 0.81 for systems thinking.

Table 3

Energy Themes Pretest and Posttest Scores

Assessment	N	Pre mean (SD)	Post mean (SD)	t-statistic	p-value	Pre-post effect size (d)	Confidence interval around effect size	
							Lower	Upper
Conservation of energy	20	6.8 (2.5)	9.0 (2.0)	5.99	< .001	0.94	0.63	1.25
Efficiency	17	15.1 (1.4)	18.4 (3.4)	4.71	< .001	0.81	0.51	1.11
Systems	22	21.4 (0.5)	25.0 (4.4)	4.67	< .001	0.81	0.46	1.16

For the total score across units we used Rasch person measures (scale scores). This approach allowed us to use data from all teachers to create a single person measure, provided the teacher had completed at least one pretest and one posttest. It also allowed us to use the teachers'

content scale scores in multilevel modeling to predict student achievement (also measured in scale scores). The unit of measurement for the Rasch scale scores is in logits. The mean person measure of 0.2 logits on the pretest corresponds with a total score of approximately 75 points on the combined measure. The mean person measure of 1.13 logits on the posttest corresponds with a total score of approximately 110 on the combined measure. Thus, the change from pretest to posttest that is just under 1 logit corresponds to a change from pretest to posttest of about 35 points (out of 154 possible).

Feedback from teacher participants gave some indications as to why significant knowledge gains were not seen with the Wind and Geothermal units. The content of the Wind unit used more mathematics relative to other units. Several teachers remarked that the math made the Wind content challenging to master. Many of the teachers expected the content of the Geothermal unit to focus on power generation from thermal vents. Although this idea was included, teachers were surprised that most of the content focused on geothermal heat exchange in the context of heating and cooling individual buildings. The assessments for both the Wind and Geothermal units were relatively difficult and not as well aligned with the three energy-related themes as compared to the other units.

We wanted to determine whether teacher characteristics such as highest degree or years of science teaching experience were influential in predicting teacher posttest scores (adjusting for pretest score). We found that teacher pretest score was strongly predictive of teacher posttest score but that teacher years of science experience and highest degree did not significantly predict posttest score. In general, all teachers gained just under one logit from pretest to posttest, and that gain was independent of pretest score, years of teaching experience, or highest degree. Table 4 provides the results from the regression analysis.

Table 4

Predicting teacher content posttest score as a function of pretest score, years of science teaching (YrsSci), and highest degree (HighDeg).

	B	SE	β	t-statistic	p-value	Confidence Interval around B	
						Lower	Upper
Intercept	0.998	0.128		7.813	< .001	0.734	1.262
Pretest	0.859	0.119	0.837	7.186	< .001	0.612	1.105
YrsSci	-0.004	0.011	-0.046	-0.410	.685	-0.027	0.018
HighDeg	0.012	0.126	0.011	0.091	.928	-0.249	0.273

Taken together, the study data demonstrate that the EMAT course was associated with enhancement of teacher content knowledge about key energy and matter concepts. Teachers displayed significant gains in their knowledge of the three energy-related themes that are essential organizing core concepts. This study provides preliminary evidence that an online course that integrates constructivism (using the BSCS 5Es) and lesson analysis can serve as a useful resource for teachers needing to enhance their knowledge of challenging energy- and matter-related content. Due to the pre-post design on teacher outcomes (with no comparison group), we cannot make causal inferences. Rather, these data suggest that the PD model has strong potential for supporting the enhancement of teachers' content knowledge.

Teacher Ability to Analyze Video, Research Question 2

Teachers' ability to analyze classroom videos and reflect on the use of key teaching strategies is emerging as an important skill—one that has shown promise in leading to the transformation of teacher practice and enhancing student learning (Roth et al., 2011; Kersting, Givvin, Thompson, Santagata, & Stigler, 2012). In the STeLLA PD program, Roth and colleagues (2011) and Taylor and colleagues (2016) found that the PD enhanced teachers' ability to reflect deeply on teaching and learning through video using two lenses: the Science Content

Storyline Lens and the Student Thinking Lens. We have built upon the success of the STeLLA PD model by implementing STeLLA strategies and video analysis as part of the EMAT course.

In this section we examine the extent to which an online PD course can support teachers in learning to analyze classroom video to recognize the use of strategies (1) to reveal, support, and challenge student thinking and (2) to construct a coherent science content storyline for students. As teachers analyze video for the use of effective practice, they access knowledge that helps them determine how to further learning in the classroom (Kersting et al., 2012; Roth et al., 2011; Taylor et al., 2016). Also in this section we address the following research question from the larger project: *Do teachers demonstrate improved ability to analyze lessons for evidence of student thinking and coherence of science content after participating in the EMAT course? If so, is the difference statistically and practically significant?*

Design and Procedure

Throughout the course, teachers learned about the strategies as they watched video examples of the strategies in use, analyzed videos on their own, and participated in facilitated, online synchronous discussions about the videos.

We measured teachers' ability to analyze videos through their written reflections as they watched video clips. Teachers completed a pretest prior to taking the course and a posttest at the end of the course. The identical pretest and posttest asked teachers to analyze four video clips. Each clip was between five and nine minutes long and involved upper elementary science instruction in authentic classrooms. EMAT teachers also had access to the transcripts from each video clip. We made the strategic decision to select clips used by the face-to-face elementary STeLLA PD project (Taylor et al., 2016) in order to facilitate cross-project comparisons. The clips included use of the strategies that participants would learn during EMAT and included

energy concepts (e.g., energy concepts in the context of the water cycle). However, the video clips were not specifically about the three key energy concepts of focus in EMAT; nor were they in a high school setting. We determined that the teaching practices shown would be apparent despite the age and content differences. We checked our assumptions with pilot teachers prior to the EMAT field tests, and pilot teachers confirmed that the use of upper elementary classrooms did not inhibit one's ability to comment on the teaching strategies in the lesson.

Each of the four clips included a brief description of the classroom context to read before starting the clip. In video clip 1, students shared posters (created in an earlier lesson) illustrating the water cycle. Student groups obtained feedback and fielded questions from the teacher and other students about their posters. In clip 2, student groups received a scenario that focused on the water cycle and each group worked to explain how water molecules were moving in relation to temperature changes, condensation, and evaporation. The teacher questioned groups of students about their thinking while they worked. Clip 3 exhibited a discussion of students evaluating latitude and temperature data at different times of the year in different locations. The clip included a teacher presentation, student group work, and students sharing ideas after their group discussions. Clip 4 showcased a lesson about the relative importance of the angle of the Sun hitting Earth and the distance from the equator on the temperature in different locations of Earth. It included a teacher presentation about the prior day's activity and students' conclusions about what they learned in that activity. For each of the four clips, EMAT teachers provided open-ended comments responding to the following prompt:

> For each video clip, spend about 5–10 minutes describing and analyzing anything you notice about the teaching, the science content, the students, and/or the classroom environment. Explain/analyze the issues and/or questions that the video

raised for you. Feel free to comment about things that are missing from the lesson, as well as things you observe. Your explanations and analyses should be in the form of complete sentences or questions. Do not use phrases or bulleted lists. Watch and analyze all four video clips.

Teachers had the opportunity to comment on the full range of strategies included in the EMAT course (Figure 1). Three out of four of the video clips included opportunities to comment on all 12 of the strategies participants were learning in EMAT. The remaining clip exhibited all but one strategy that participants were learning.

Roth, Askinas, and Gardner (2013) developed a rubric to score teachers' written video analysis responses as part of the recently completed STeLLA PD efficacy trial (National Science Foundation [NSF] award# 0918277; Taylor et al., 2016). The rubric includes definitions of each strategy and guidelines for scoring teachers' comments on each strategy. Using the STeLLA rubric, we scored teachers' comments by strategy, assigning 0, 1, or 2 points depending on the teacher's apparent depth of understanding about the strategy. Generally, comments that showed correct understanding of strategies and more analytical use of the strategies generated higher scores while a lack of comments about a strategy or incorrect use of a strategy generated lower scores. For example, some comments written on the pretest included statements about classroom management or the size of groups seen in activities. Pretest comments such as these often attended to neither the coherence of the science content storyline nor to reflections on student thinking apparent in the video. These types of comments generated a score of 0. On the posttest, teachers' comments tended to extend beyond classroom management issues. Comments related to lost opportunities or that suggested alternative methods of instruction within the contexts of

lesson coherence and student thinking were often scored a 2 as they showed more in-depth analysis and understanding of specific strategy uses.

Two coders initially jointly coded and discussed their scores on 20 responses. Coders then divided and scored the remaining responses, including an additional 20 overlapping responses to measure interrater agreement. The final interrater reliability statistics reveal that the coders remained well calibrated throughout coding. However, there were two items on which coders could not achieve agreement in spite of extensive negotiation and discussion. Dropping the two items made the most sense given these limitations. We used two measures of interrater agreement: the intraclass correlation coefficient (ICC; two-way mixed effects, absolute agreement) was 0.898, and Cohen's kappa was 0.738. Both measures show highly satisfactory levels of interrater agreement.

Analyses and Findings. We scored both pre and post responses upon completion of the course to blind the coders to time point. We examined changes in teachers' ability to analyze classroom video and also used Ordinary Least Squares (OLS) regression to examine teacher characteristics that predicted post video analysis scores. We used Rasch person measures in our analyses. Rasch person measures are true scale scores (whereas raw point totals on an assessment are not) and allow us to place person ability and item difficulty on the same logit scale. A negative person score indicates that an individual's ability to analyze video was below the mean item difficulty. All average Rasch person measures for the EMAT teachers (both pre and post) were negative, indicating that the average EMAT teacher's ability to analyze video was below the average item difficulty. The assessment was extremely difficult for the teachers, even at posttest, with an average score per item of just 0.65 out of 2 points. Nevertheless, we found significant improvement from pretest to posttest for the overall measure ($p < .001$; $d = 1.38$) as

well as for the student thinking ($p < .001$; $d = 1.13$) and science content storyline ($p < .001$, $d = 1.23$) subscales. Effect sizes make sense only in context (Hill, Bloom, Black, & Lipsey, 2008). For context, we can compare the pre-post effect sizes on the video analysis task for the EMAT teachers to the pre-post effect size for the STeLLA PD teachers (Taylor et al., 2016). We selected a subset of treatment STeLLA teachers who completed identical pre-post video analysis tasks as the EMAT teachers. The STeLLA pre-post pedagogical content knowledge (PCK) effect size was $d = 2.607$ ($p < .001$) with an effect size confidence interval of [1.940, 3.274]. In other words, the STeLLA pre-post video analysis effect size was more than twice as large as the EMAT video analysis effect size. Table 5 highlights the video analysis findings.

Table 5

Video analysis scores, Rasch person measures. N (EMAT) = 23; N (STeLLA) = 44

Video analysis measure	**Pre mean (SD)**	**Post mean (SD)**	**t diff score**	**SD diff score**	**p-value**	**Pre-post effect size (d)**	**Confidence interval around effect size**	
							Lower	**Upper**
EMAT overall	-1.82 (0.64)	-0.89 (0.70)	6.90	0.59	< .001	1.38	0.85	1.92
Student thinking score	-2.03 (0.92)	-0.92 (1.03)	5.31	0.91	< .001	1.13	0.61	1.66
Science content storyline score	-1.75 (0.62)	-0.95 (0.65)	5.71	0.58	< .001	1.23	0.68	1.78
STeLLA overall	-1.72 (0.30)	-0.75 (0.42)	16.06	0.41	< .001	2.61	1.94	3.27

There are several important similarities and differences between the STeLLA and EMAT video analysis data. First, the EMAT and STeLLA teachers started with similar video analysis ability, but the STeLLA teachers finished with higher mean scores. Second, both EMAT and STeLLA teachers had negative mean Rasch posttest scores (indicating that the assessment was

difficult for both groups). We have considered the open-ended nature of the prompt as a possible source of the difficulty of the assessment. Addressing all strategies in a response (without any explicit prompt to address the use of strategies emphasized in the course) likely placed a fairly high cognitive demand on teachers.

Third, the difference in effect size between the EMAT teachers and STeLLA teachers is only partly accounted for by larger gains by the STeLLA teachers. Another important factor is that the standard deviation of the difference score for the EMAT teachers was larger than that for the STeLLA teachers by almost 50% (0.586 vs 0.405). That is, the changes for the EMAT teachers were more variable than the changes for the STeLLA teachers. Although participation in EMAT was associated with enhanced ability to analyze videos for the student thinking and science content storyline strategies that are part of the STeLLA framework, the changes associated with the EMAT online PD model for high school teachers were lower and more variable than the changes associated with the STeLLA face-to-face model for elementary teachers.

Factors influencing teachers' post video analysis scores

We considered each teacher participant's highest degree (HighDeg), years of science teaching experience (YrsSci), pretest video analysis person measure (PreVA), and post content person measure (PostCont) as predictors of post video analysis scores (Y_i). We used the following ordinary least squares regression model to examine the relationships:

$$Y_i = \beta_0 + \beta_1 \text{PreVA} + \beta_2 \text{YrsSci} + \beta_3 \text{HighDeg} + \beta_4 \text{PostCont} + \varepsilon_i$$

We grand mean centered all predictors. Thus, we interpret the intercept to be the average video analysis posttest score for the entire sample.

In our analysis, the pre video analysis scores (p = .010) and the post content scores (p = .021) were significant predictors of post video analysis scores. A teacher's highest degree and years of science teaching experience were not predictive of their post video analysis scores (Table 6).

Table 6

Predicting post video analysis person measures (N = 23).

Predictor	B	β	SE	t-statistic	p-value	Confidence interval for B	
						Lower	Upper
Intercept	-0.996		0.111	-8.946	< .001	-1.230	-0.762
YrsSci	0.003	0.029	0.019	0.170	.867	-0.037	0.043
HighDeg	0.262	0.193	0.227	1.155	.263	-0.215	0.739
PreVA	0.730	0.474	0.255	2.858	.010	0.193	1.267
PostCont	0.591	0.444	0.233	2.532	.021	0.100	1.081

In addition to predicting the overall post video analysis person measure, we examined each subscale (student thinking and science content storyline) separately. The student thinking component included one case that met several criteria to be categorized as an outlier. The teacher's Rasch person measure score on the post video analysis student thinking subscale was 3.85 standard deviations below the mean of his colleagues on this subscale. The unstandardized residual value (-2.64), the studentized residual value (-3.56), and the unstandardized change in the HighDeg and PreVA_ST coefficients (0.36 and 0.44) all support the case that this teacher is unduly biasing the regression coefficients. The results of the regression excluding this case are shown in Table 7.

Table 7

Post Video Analysis Student Thinking (ST) outcome (N = 22; omitting one outlier).

Predictor	B	β	SE	t-statistic	p-value	Confidence interval for B	
						Lower	Upper
Intercept	-0.756		0.108	-6.991	< .001	-0.984	-0.528
YrsSci	0.027	0.295	0.019	1.449	.165	-0.012	0.067
HighDeg	0.076	0.068	0.226	0.339	.739	-0.399	0.552
PreVA_ST	0.204	0.229	0.178	1.140	.270	-0.173	0.580
PostCont	0.561	0.505	0.228	2.464	.025	0.081	1.042

Within the student thinking component of the video analysis score, content learning was significant (p = .025) in predicting post video analysis scores, but the pre video analysis score was not (p = .270).

Table 8

SCS Lens Component; PostVA_SCS outcome (N = 23).

Predictor	B	β	SE	t-statistic	p-value	Confidence interval for B	
						Lower	Upper
Intercept	-1.067		0.111	-9.577	< .001	-1.301	-0.833
YrsSci	0.004	0.041	0.019	0.232	.820	-0.035	0.043
HighDeg	0.205	0.161	0.161	0.929	.365	-0.259	0.669
PreVA_SCS	0.670	0.423	0.423	2.475	.023	0.101	1.239
PostCont	0.612	0.491	0.491	2.730	.014	0.141	1.083

When analyzing teachers' comments about the coherence of the instruction, the post content test score (p = .014) and the pre video analysis score (p = .023) were both predictive of the science content storyline component of the post video analysis. Thus, their years of science teaching and highest degree were not predictive of the science content storyline component of the video analysis task.

Teacher Practice, Research Question 3

Transformation of science teaching and learning involves transforming teaching practice. Although teacher content knowledge and pedagogical content knowledge (Kersting et al., 2012) have both been shown to predict student achievement, teacher practice is almost certainly an important mediator. In their cluster randomized trial of the STeLLA PD program, Roth and her colleagues found that teacher practice does, in fact, mediate the relationship between the professional development intervention and student achievement (manuscript in preparation). In anticipation of scoring teachers' classroom practice, we asked teachers to record their teaching in the year prior to participation in EMAT and record their teaching once again following their participation in EMAT. We transcribed the videos prior to coding.

As part of the STeLLA efficacy study, Roth and principal investigator for the EMAT project (Kowalski) developed a video analysis coding protocol to score individual classroom sessions (approximately one hour in length) for the teacher's use of the STeLLA strategies (Roth & Kowalski, 2015). Language for the protocol and scoring rubric emerged from the STeLLA conceptual framework. The coding protocol was extensive, requiring six to eight hours to code one hour of recorded classroom instruction. Roth and Kowalski initially used and refined the rubric to jointly score six master videos that showcased a wide array of teaching practices, and using discussion to come to consensus on all scores across the six videos. Kowalski later coded a seventh master video. We used two master videos for training purposes and the remaining five for calibration. We developed a team of six coders for the STeLLA efficacy project, and three of those coders went on to code EMAT videos. Coders identified information from watching the videos and reading transcripts and pulled segments of transcript into predefined nodes using NVivo software (v. 10.0). The nodes corresponded to evidence of strategy use. Using evidence

from transcript elements collected into nodes, coders scored the videos. Each strategy was associated with three items: The first was a dichotomous item indicating presence or absence of the strategy; the remaining two items were scored from 0 to 3 and reflected the quality of strategy use.

We coded 30 pre videos and 20 post videos of the EMAT teachers (reflecting attrition we experienced over the two-year participation expectation). We created overall Rasch person measures as well as Rasch person measures for each of the two STeLLA lenses (ST and SCS). We anchored all pretest scores to posttest.

Our initial analyses examine the changes in teacher practice from pretest to posttest that were associated with teacher completion of the EMAT course (Table 9).

Table 9

Teacher Classroom Practice Measure. N (EMAT) = 20; N (STeLLA) = 51.

Classroom practice measure	**Pre mean (SD)**	**Post mean (SD)**	**SD of diff score**	**t-statistic for diff score**	**p-value**	**Pre-post effect size (d)**	**Confidence interval around effect size**	
							Lower	**Upper**
Total score	-1.06 (0.69)	-0.64 (0.81)	0.96	1.98	.063	0.57	-0.04	1.17
Student Thinking Lens	-1.61 (1.06)	-0.95(1.00)	1.57	1.88	.076	0.64	-0.09	1.38
Science Content Storyline Lens	-0.91 (0.92)	-0.46 (1.12)	1.18	1.72	.102	0.44	-0.09	0.97
STeLLA (total)	-0.71 (0.80)	1.05 (1.13)	1.76	10.00	< .001	2.09	1.36	2.82

Once again, the Rasch person measures for EMAT are negative, even for the post practice measure, indicating that the measure was difficult for EMAT teachers. This is not the case for the STeLLA teachers. The STeLLA teachers' post score was 1 logit above the mean item difficulty. In addition, although EMAT mean post practice scores are all higher than mean pre practice

scores, the changes from pre to post only approach significance (p = .063 for the total score). The effect size for change in practice associated with the EMAT course is about one-quarter that of the STeLLA PD program (d = 0.57 for EMAT; d = 2.09 for STeLLA). In this case, the difference in effect sizes rests almost entirely with the difference in means. EMAT has a smaller standard deviation of the difference score than STeLLA, but STeLLA has the larger mean difference and the larger effect. Finally, it is interesting to note that the elementary STeLLA teachers had higher mean pre practice scores than EMAT teachers.

To better understand which strategies teachers tended to adopt and which were most challenging for EMAT teachers, we examined the effect sizes for changes in teacher practice at the strategy level. As with the teacher content knowledge and video analysis outcomes, we wanted to examine how teacher characteristics predicted teacher post practice score. We initially used the following model but found that there were multicollinearity issues, particularly for the highest degree variable.

$$Y_i = \beta_0 + \beta_1 \text{PrePractice} + \beta_2 \text{YrsSci} + \beta_3 \text{HighDeg} + \beta_4 \text{PostCont} + \beta_5 \text{PostVA} + \varepsilon_i$$

We revised the model to exclude highest degree. We were comfortable with this decision because we felt that the post content score and the post video analysis score were capturing information that was redundant with highest degree (with post content picking up the overlap with science degrees, and post video analysis picking up the overlap with education degrees). Post content and post video analysis were also highly correlated with each other (bivariate correlation r = 0.681). Following Cohen, Cohen, West, and Aiken (2003), we computed z-scores for both the post content measure and the post video analysis measure, then averaged and grand mean centered the result (PostCont/VA_z). Thus, we are predicting teachers' post practice scores using an amalgam measure that is indicative of both their content knowledge and their ability to

analyze classroom practice video for key strategies. We decided that using the amalgam measure was potentially more appropriate than arbitrarily dropping either the content measure or the video analysis measure.

$$Y_i = \beta_0 + \beta_1 \text{PrePractice} + \beta_2 \text{YrsSci} + \beta_3 \text{PostCont/VA\_z} + \varepsilon_i$$

Table 10

Predicting teachers' post practice scores (N = 17)

Predictor	B	β	SE	t-statistic	p-value	Confidence interval for B	
						Lower	Upper
Intercept	-0.06		0.40	-0.15	.883	-0.92	0.80
YrsSci	-0.04	-0.32	0.03	-1.27	.226	-0.11	0.03
PrePractice	0.03	0.03	0.26	0.12	.906	-0.53	0.60
PostCont/VA_z	0.41	0.49	0.19	2.23	.044	0.01	0.81

Here we find that, once again, years of science teaching experience is not a significant predictor of teacher practice. In addition, it is surprising to note that the pre practice measure is not at all predictive of the post practice measure. This replicates the work that Roth and her colleagues found in the STeLLA efficacy trial (manuscript under development). The STeLLA strategies that form the STeLLA conceptual framework are new to teachers. Although in many ways they reflect what is known about "good science teaching," teachers historically have not had the necessary scaffolds to think about using a complex set of strategies. The STeLLA strategies create a structure for teachers to really learn to *do* good science teaching. As a result, pre videos have almost no relationship to post videos—teachers' initial practice is uniformly lacking in use of the STeLLA strategies for both elementary and high school teachers and for teachers with both high and low post practice scores.

Our amalgam measure (the average of content knowledge with ability to analyze classroom video) is a significant predictor of teachers' post practice video scores. The amalgam measure may be capturing the construct that others have called pedagogical content knowledge (PCK) (Shulman, 1986; Kersting et al., 2012). In that work researchers have found that knowledge of the content and how to teach that specific content is a key attribute of effective teachers. Our exploratory work shown here supports that prior work. This amalgam measure has elements of a PCK measure in that we are assessing not only what content teachers know but the extent to which they can apply that content to classroom situations.

It should be noted that in all of these analyses we have very few degrees of freedom. As a result, the parameter estimates may be unstable. These findings are exploratory and are of interest chiefly as they relate to what others have found (Kersting et al., 2012; Roth et al., 2011; Taylor et al., 2016).

Student Achievement

All teacher results are based on a pre-post design, and we provided some context for interpreting those findings by comparing the EMAT results with the STeLLA results. We now turn to the quasi-experimental study of the impact of EMAT on student achievement. We initially planned to use only teachers and students from the second field test in our analyses. We had the content scores, video analysis scores, and practices scores for teachers in the second field test but lacked practice measures in the first field test due to limited resources for coding. However, the significant attrition of EMAT teachers across the two years of the program left us with far too few degrees of freedom for our hierarchical regression. Students in the first field test had completed a pretest and posttest with items that overlapped to a great extent with the student assessment for the second field test. We selected items in common across both test

administrations and pooled the students in the analysis. Our inability to use the teacher outcome measures in the model and the added power of pooling students and their teachers across two field tests convinced us of the merits of the pooled analysis.

The student assessment consisted of 35 multiple choice questions related to the same three key energy concepts that teachers were learning. The items were situated within the same unit contexts that the teachers were learning, but we were careful to provide enough information that the students did not need to know anything about the energy generation system (e.g., generating electricity from coal) in order to answer the energy concept questions.

By comparing unconditional models to full models we were able to estimate the variance explained by class and by teacher in our analytic model. We found the percent of variance on the intercept at the teacher level to be nearly 56%,

$$\text{on intercept: } \frac{\tau_{\beta 00}}{\tau_{\beta 00} + \tau_{\pi 00}} = (0.0428)/(0.0428 + 0.03427) = 0.555$$

and the percent of variance on the slope between teachers was 42%.

$$\text{on slope: } \frac{\tau_{\beta 11}}{\tau_{\beta 11} + \tau_{\pi 11}} = (0.0138)/(0.0138 + 0.0188) = 0.42$$

That is, a very large proportion of our variance is accounted for by knowing which teacher students had. Multisite cluster trials and analyses are appropriate when there is reason to believe that the treatment effect may vary in important ways across the sites of the experiment (in this case, each teacher is a site of a mini-experiment with one treatment and one control class). The high variance on the slope and intercept for the teacher level validate our use of a multisite cluster analysis with students at level 1, class at level 2, and teachers (the site of each mini-experiment) at level 3. Our complete analytic model is shown below.

Complete Analytic Model

Level 1

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}\text{Gen} + \pi_{2jk}\text{FRL} + \pi_{3jk}\text{Grd10} + \pi_{4jk}\text{Grd11} + \pi_{5jk}\text{Grd12} + \pi_{6jk}\text{ELL} + \pi_{7jk}\text{Race01} + \pi_{8jk}\text{Pre} + e_{ijk}$$

Level 2

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}\text{Trt} + r_{0jk}$$

Level 3

$$\beta_{00k} = \gamma_{000} + \gamma_{000}\text{MnPre} + u_{00k}$$

$$\beta_{01k} = \gamma_{010} + u_{01k}$$

Table 11

Test of Main Effect of Treatment on Student Achievement. Data combined across first and second field tests.

Variable	**Coefficient**	**Standard error**	**t-ratio**	**d.f.**	**p-value**
Level 3 (teacher)					
Intercept	-0.330	0.028	-11.717	60	< .001
MnPre	0.512	0.071	7.201	60	< .001
Level 2 (class)					
γ_{010} (avg. Trt effect)	0.080	0.054	1.479	61	.144
Level 1 (student)					
Gender	-0.136	0.026	-5.338	2,451	< .001
Grade10	-0.052	0.049	-1.067	2,451	.287
Grade11	-0.106	0.052	-2.049	2,451	.040
Grade12	-0.119	0.054	-2.204	2,451	.027
ELL01	-0.098	0.034	-2.896	2,451	.004
Race01	-0.117	0.030	-3.955	2,451	< .001
FRL	-0.009	0.033	-0.287	2,451	.774
Pre	0.628	0.023	27.754	2,451	< .001

Table 11 showcases our findings from the quasi-experiment. These data show that although the treatment group of students outperformed the comparison group of students. We interpret the treatment coefficient (0.08) as follows: on average, the mean class student posttest score for the treatment group was 0.08 logits higher than for the comparison group, controlling for pretest and other demographic factors. The difference was not significant at the alpha = 0.05 level (p = .144). Across the two treatment groups, we see that girls, English language learners, and students

from racial/ethnic groups traditionally underrepresented in the sciences had lower achievement scores, but this is a measure across both groups and not particular to the EMAT group. The effect size for the intervention was d = 0.13, variance of effect size = 0.20, SE_d = 0.452, and the lower and upper confidence interval values for the effect size were [-0.757, 1.016]. The fact that the effect size had such a large variance and we see a confidence interval for the effect size with such a wide range of values is an indication that *effects of participating in EMAT varied drastically from teacher to teacher*. This finding is consistent with elements of the analysis we have seen earlier (e.g., the high standard deviation on teachers' video analysis scores and the high amount of variation in intercept and slope at the teacher level in the hierarchical linear model). Examination of individual teachers' practice scores also supports the finding: Three teachers had lower post practice scores compared with their pre practice scores, while the remaining teachers had higher post practice scores. This finding means that for some students, their teacher's participation in EMAT coincided with increased student achievement; for others, their teacher's participation seems to have coincided with reduced student achievement. The overall positive average effect of 0.13 masks these important distinctions.

Elements of EMAT that Support Achievement, Research Question 5

At this point, the million-dollar question is why is participation in EMAT associated with such varied effects? Is the variation truly a result of EMAT, or did we have a sample that included teachers who simply had a bad second year? Is there something about a teacher's personality or beliefs that would allow us to predict which teachers might do well with EMAT and which might not?

We included computer-mediated discourse analysis statements made by select teachers during the course to try to answer this million-dollar question. We identified six teachers as case

study teachers for the computer-mediated discourse analysis. To select these teachers, we first used a graph of teachers' posttest person measure content scores plotted against their pretest person measure scores for the total EMAT content test. Next, we identified a teacher with a low pretest and a low posttest (low-low), a teacher with a low pretest and a relatively high posttest (low-high), a teacher with a high pretest and a relatively low posttest (high-low), and a teacher with a high pretest and a high posttest (high-high). Finally, we identified two additional teachers based on student outcomes. The first teacher's treatment students greatly outscored the comparison students after controlling for pretest (large positive effect); the second teacher's comparison students outscored the treatment students after controlling for pretest (negative effect). It is interesting to note that the case study teacher with the large positive treatment effect is also the teacher with uncharacteristically uniform student responses on the treatment posttest. Under consultation from our external evaluator, we decided to drop this teacher from our analyses as this teacher's student data are not meaningful. Unfortunately, we discovered the anomaly only after we had undertaken the effort to code the teacher's comments.

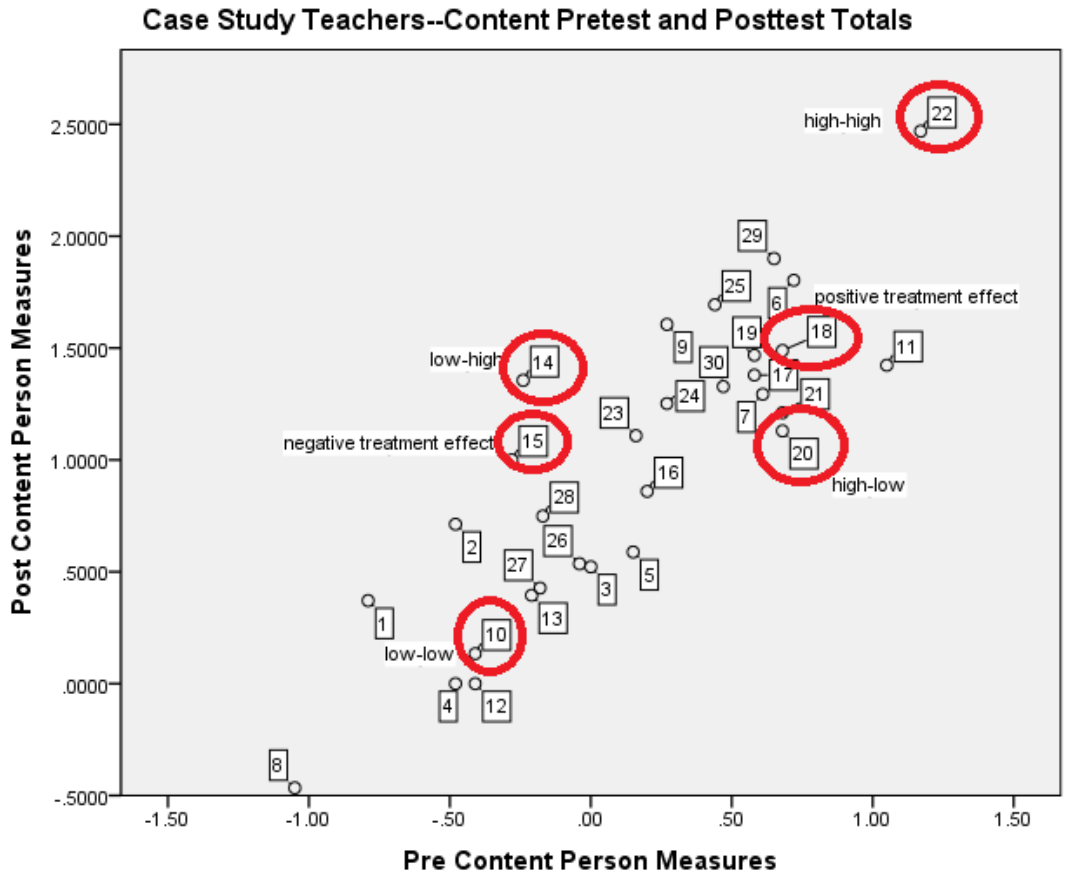

Figure 3. Case study teachers total content pretest and posttest scores.

After identifying our six case study teachers, we examined all of their comments throughout the entire EMAT course, including their comments in their online notebooks, survey comments, course assignments including end-of-unit reflections, and any discussion board comments. We coded the statements by course element and by teacher as transformative, positive, neutral, or negative. We ended with frequency counts for each teacher, each course element, and each type of statement. To understand the relationships between teachers better, we calculated a value to reflect the overall positive or negative nature of the comments as a percentage of the total number of comments:

$$\text{net \% comments} = \frac{(\#\,\text{transformative} + \#\,\text{positive statements}) - \#\,\text{negative statements}}{\text{total}\,\#\,\text{statements}}$$

Based on this equation, a positive percent indicates that the teacher made more positive statements than negative. A negative percent indicates that the teacher made more negative statements than positive. We divided by the total number of comments because some teachers made many more comments and statements than others. A net percent of 0 indicates that the teacher made the same number of positive comments as negative comments.

Table 12

Net percent comments on each type of course element (net negative comments in grey).

Case	Animation	Interactive	Reading	Content	Classroom video	Lesson analysis	Synch. disc.	Net % content	Net % pedagogy
negative treatment effect	40%	0%	0%	0%	17%	6%	0%	7%	5%
high-high	60%	24%	0%	29%	33%	68%	none	29%	46%
high-low	none	none	none	67%	none	89%	100%	67%	81%
low-high	69%	38%	50%	11%	36%	50%	62%	31%	35%
low-low	52%	24%	57%	30%	-29%	-33%	33%	34%	-5%

It is difficult to identify any patterns in these data. The teacher with the negative treatment effect and the teacher categorized as low-low had the least positive opinions of the pedagogy portions of the course (including watching classroom videos, participating in lesson analysis, and participating in synchronous discussions). By comparison, the teacher with the most positive treatment effect had moderately positive comments about the pedagogy portion, and the high-high, high-low, and low-high teachers all had very positive comments about the pedagogy portion. It is unfortunate that our case study teachers did not all have complete student data as there are a variety of reasons that some of our teachers did not complete data collection.

The short answer to our million-dollar question is, *we don't know*. Further analyses of teacher comments, selection of a different group of teachers, or conducting additional data collection and research in the future may shed more light on the issue. For now, all we can say is that teachers had generally favorable opinions of the course with a small number of exceptions. The elements of the course that teachers tended to either love or distinctly NOT enjoy were the lesson analysis elements. Further study is needed.

Conclusions

In summary, the EMAT online video-based lesson analysis course supported teachers' enhancement of knowledge and practice, with stronger evidence for enhanced knowledge than for enhanced practice. Although the quasi-experimental study showed that a teacher's participation in EMAT may have had a small positive effect on student learning, these findings are dwarfed by the finding that there is very large variability in effect sizes by teacher.

Successes and Challenges

The successes and challenges of this new approach to using video analysis of practice in an online environment with high school teachers go beyond the statistical measures and relate to the elements of comparison in the previous section. While we did see statistically significant gains in teacher content knowledge and teachers' ability to analyze video, we did not see significant gains in teacher practice or student achievement. Successes and challenges related to EMAT were varied and are discussed in the following paragraphs.

Audience. The typical high school science teacher has a degree in a field of science and is likely to teach that science for his or her entire career. The typical elementary teacher, as in the STeLLA PD program, has a degree in teaching and has expertise in facets of teaching other than those related to teaching science. This fundamental difference in the science preparation of the

teachers affects the buy-in for learning new content and the willingness to engage in content-related activities. An elementary teacher has to be prepared to teach at a variety of grade levels, each with its own set of science standards to address. The high school science teacher has a specific subject area (e.g., biology, chemistry, or physics) that has one set of standards that is consistent for several years. In EMAT, these high school teachers were reluctant and out of their comfort zone for several of the units. The units in the course spanned biology content as well as physics, geoscience, and chemistry content. Teachers in the course often commented that they "would never teach this content," a comment that was rarely heard with elementary teachers.

Delivery. Online delivery may often be convenient for the participants, but it is not without its challenges. There are the technical difficulties of watching video during the study group sessions with participants having varying connectivity speeds. This sometimes caused issues during the study groups when we watched and analyzed video. To help with this issue, we chose to use audio-only for discussion and turn the video chat feature off during the study groups (maintaining the ability to see video for analysis). However, there were some advantages of using online delivery. Through online delivery we were able to recruit teachers from across the country and had a variety of types of schools and student demographics. This type of delivery attracted teachers from rural areas in Alaska, Hawaii, and other parts of the continental US. These teachers, due to their lack of proximity to other US locales, often can't access any PD other than that from their own district. However, it was often difficult to schedule the synchronous discussions for small groups of teachers that live in a variety of time zones. Teachers have very full summer schedules with other PD, family vacations, and life responsibilities. Even though online courses offer some anonymity and safety when

participating in a synchronous discussion, the richness of interaction diminished in the online delivery system.

Curriculum. A major difference between the elementary program and EMAT is that the elementary teachers were provided an exemplar set of lessons incorporating the teaching strategies that were the focus of the PD. Additionally, the teachers co-developed a second set of lessons in another but related subject area. In doing so, they developed lessons with embedded strategies and created lessons specifically for highlighting these high-leverage teaching strategies. In contrast, the EMAT teachers were not given curriculum. They did not receive lessons with the strategies embedded and were not able to take these lessons and use them with their students. Because the high school teachers took the course in the summer, they were not even able to practice the strategies with students until the course was over. The contact time for the elementary teachers (88.5 hours over a year) compared to the contact time for the high school teachers (120 hours over 10 weeks) seems to reflect that the high school teachers had some advantage with a greater number of hours. However, the elementary teachers were better able to build their knowledge of the strategies over time than the EMAT teachers. Also, the elementary teachers had many of their hours during the school year and were able to practice the strategies with students and use the provided curriculum. The EMAT teachers did not have this opportunity.

Aside from logistical issues of compression of time for the course and lack of student materials, consideration of theory may shed light on why EMAT demonstrated a smaller impact than STeLLA. Brown, Collins, and Duguid (1989) and Lave (1988) identified authentic context, activity, and culture as essential components of learning. By situating learning opportunities for teachers within authentic classroom contexts, teachers are more likely to obtain usable

knowledge—knowledge they can access meaningfully and purposely on a day-to-day basis with their students. By this argument, professional learning opportunities should be as tightly tied to classroom contexts as possible. The STeLLA professional learning program (Roth et al., 2011; Taylor et al., 2016) did just that: Elementary teachers were engaged in learning key science concepts related to the water cycle, food webs, Earth's changing surface, and the Sun's effect on climate; they considered student alternative conceptions within these same four science contexts; and they developed capacity (within a cognitive apprenticeship structure; see Collins, 2006 and Collins, Brown, & Newmann, 1989) to write a coherent series of lessons related to these same four science contexts. Every element of the professional learning experience mapped directly to what the teachers would do in the classroom.

In designing the EMAT course, we made every attempt to situate teachers' learning within authentic classroom contexts, but we were thwarted in part by the nature of our audience: high school science teachers teaching energy concepts across science disciplines. From the beginning, we were not able to situate every learning experience in an authentic context for every teacher in our study. At best, teachers found themselves learning within truly authentic contexts only part of the time. To mitigate the effects of working with teachers across science disciplines (across contexts), we identified three key crosscutting energy concepts and designed learning activities for teachers around them. We highlighted each key concept (using icons and reflection questions) throughout the course in hopes that teachers would begin to understand the connections woven within and throughout the course, starting with a unit on electromagnetic induction and ending with a unit on photosynthesis and the production of biofuels.

The STeLLA PD model builds on both situated cognition and cognitive apprenticeship theories. We were thoughtful in our attention to integrating situated cognition into EMAT

(challenges notwithstanding); however, we were less overt in our use of cognitive apprenticeship. There was no opportunity to scaffold lesson development. Because teachers hailed from a variety of disciplines, we did not provide them with model lessons to use in the classroom—there was no "one size" that would fit all. Furthermore, because the EMAT course took place during the summer, there was no opportunity for teachers to try what they were learning and then share back with the group on their successes and challenges. Thus, although the teacher materials showcased what high quality curriculum materials could be and the professional learning leaders modeled the kind of teaching we hoped teachers would use in their classrooms, the participating teachers did not acquire the usable knowledge they needed to develop their own coherent lessons and enact the practices they were learning, at least not to the same extent as the STeLLA teachers did.

There are a number of possible implications from our findings. First, is it possible to extend the STeLLA PD model to a multidisciplinary high school audience? EMAT had modest success, but the high variability of the treatment effect on students is worrisome. If PD providers (online or face-to-face) cannot situate teacher learning in relevant contexts and provide opportunities for mentorship and growth, it may be difficult to translate the PD for such a diverse audience as high school teachers.

Another implication arises from the question, *How big of an effect is big enough?* This question raises cost/benefit issues. The EMAT course is associated with about half the effect on teacher learning and one-fourth the effect on teacher practice as the STeLLA PD model. Is that enough? What should a PD provider do in the face of effect size variability? Is it worth it to take a promising model and risk helping some but harming others? Is the high variability in effect size related to an implementation dip (practice may get worse before it improves)? This may be a

reasonable explanation: if teachers are working to better uncover student thinking during their instruction, and as they do, students bring up ideas that may be tangential to the original main learning goal, teachers may actually have less coherence in their instruction in their early efforts to use the STeLLA strategies. This may particularly be true if they do not have model curriculum materials with strategies embedded, and have not had the opportunity to learn to plan their own lessons using STeLLA strategies.

Dede (2006) called for more research linking PD to student impacts and more research on the elements of successful PD. Our study attempted to follow that guidance but perhaps raised more questions than we answered. Further study is needed.

References

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

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education In G. Sykes & L. Darling-Hammond (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3-32). San Francisco: Jossey Bass.

Black, P., & Solomon, J. (1983). Life world and science world: Pupils' ideas about energy. In G. Marx (Ed.), *Entropy in the school: Proceedings of the 6th Danube seminar on physics education* (pp. 43-55). Budapest: Roland Eotvos Physical Society.

Borko, H., Koellner, K., Jacobs, J., & Seago, N. (2011). Using video representations of teaching in practice-based professional development programs. *ZDM, 43*(1), 175-187.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-41.

Bybee, R. W., Taylor, J. A., Gardner, A. L., Van Scotter, P., Carlson Powell, J., Westbrook, A., & Landes, N. M. (2006). The BSCS 5E instructional model: Origins and effectiveness. Washington, DC: National Institutes of Health Office of Science Education.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regressions/correlation analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Collins, A. (2006). Cognitive apprenticeship. In R. K. Sawyer (Ed.) *The Cambridge Handbook of the Learning Sciences*. Cambridge University Press.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics (M. Lipman, Trans.). In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 233-243). Hillsdale, NJ: Lawrence Erlbaum Associates. (Reprinted from: Thinking children and education, Dubuque, IA: Kendall Hunt Publishers, pp. 453-494, 1993).

Dede, C. (2006). *Online professional development for teachers: Emerging models and methods*. Cambridge, MA: Harvard Education Press.

Dede, C., Ketelhut, D., Whitehouse, P., Breit, L., & McCloskey, E. (2006). *A research agenda for online teacher professional development. Harvard Graduate School of Education.*Unpublished manuscript, Cambridge, MA.

Doménech, J., Gil-Pérez, D., Gras-Martí, A., Guisasola, J., Martínez-Torregrosa, J., Salinas, J., et al. (2007). Teaching of Energy Issues: A Debate Proposal for a Global Reorientation. *Science & Education*, *16*(1), 43-64.

Driver, R., & Oldham, V. (1985). A constructivist approach to curriculum development. *Studies in Science Education, 13,* 105-122.

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.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172-177.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal, 49*(3), 568-589.

Lave, J. (1988). *Cognition in practice: Mind, mathematics, and culture in everyday life*. Cambridge, UK: Cambridge University Press.

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

National Academies of Sciences, Engineering, and Medicine. (2015). *Science Teachers Learning: Enhancing Opportunities, Creating Supportive Contexts*. Committee on Strengthening Science Education through a Teacher Learning Continuum. Board on Science Education and Teacher Advisory Council, Division of Behavioral and Social Science and Education. Washington, DC: The National Academies Press.

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

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*(2), 211-227.

Raya-Carlton, P., Weaver, D., & Krebs, L. (2010). Across the sciences: Course efficacy evaluation final report Retrieved from http://edmedia.opb.org/research_eval/AcrosstheSciencesEvaluationFinalReport.pdf

Roth, K. J., Druker, S. D., Garnier, H. E., Lemmens, M., Chen, C., Kawanaka, T., Rasmussen, D., Trubacova, S., Warvi, D., Okamoto, Y., Gonzales, P., Stigler, J., & Gallimore, R. (2006). Teaching science in five countries: Results from the TIMSS 1999 video study (NCES 2006-011). Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/timss.

Roth, K. J., Garnier, H., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. Z. (2011). Videobased lesson analysis: Effective science PD for teacher and student learning. *Journal of Research in Science Teaching, 48*(2), 117-148. doi: 10.1002/tea.20408

Roth, K. J., & Kowalski, S.M. (2015). *STeLLA II Lesson Video Coding Manual*. Biological Sciences Curriculum Study (BSCS), Colorado Springs, CO.

Sağlam-Arslan, A., & Kurnaz, M. A. (2009). Prospective Physics Teachers' Level of Understanding Energy, Power and Force Concepts. *Asia-Pacific Forum on Science Learning and Teaching, 10*(1).

Seago, N. (2003). Using video as an object of inquiry for mathematics teaching and learning. In J. Brophy (Ed.), *Advances in research on teaching* (Vol. 10, pp. 259-286). New York, NY: JAI Press.

Seago, N. (2004). Using video as an object of inquiry for mathematics teaching and learning. In J. Brophy (Ed.), *Advances in research on teaching: Using video in teacher education* (Vol 10, pp. 259-286). New York, NY: Elsevier JAI.

Sherin, M. G. (2004). New perspectives on the role of video in teacher education. In J. Brophy (Ed.), *Advances in research on teaching: Using video in teacher education* (Vol 10, pp. 1-27). New York, NY: Elsevier JAI.

Sherin, M., & Han, S. (2004). Teacher learning in the context of a video club. *Teaching and Teacher Education, 20*, 163-183. doi: 10.1016/j.tate.2003.08.001

Sherin, M. G. & van Es, E. A. (2002). Using video to support teachers' ability to interpret classroom interactions. In *Society for Information Technology and Teacher Education: Information Technology and Teacher Education Annual (4)*, 2532 – 2536. Norfolk VA: Association for the Advancement of Computing Education.

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

Taylor, J., Roth, K., Wilson, C., Stuhlsatz, M., & Tipton (2016). The Effect of an Analysis-of-Practice, Videocase-Based Teacher Professional Development Program on Elementary Students' Science Achievement. Published online in the *Journal of Research on Educational Effectiveness*, 2 February 2016.

van Es, E. A., & Sherin, M. G. (2002). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education*, *10*(4), 571-596.

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. *Journal of Research in Science Teaching, 47*(3), 276-301. doi: 10.1002/tea.20329