3-29-2016

Tool Trouble: Challenges With Using Self-Report Data to Evaluate Long-Term Chemistry Teacher Professional Development

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

Ellen J. Yezierski
Miami University - Oxford

Senetta F. Bancroft
Southern Illinois University Carbondale

JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 53, NO. 7, PP. 1055–1081 (2016)

Research Article

Tool Trouble: Challenges With Using Self-Report Data to Evaluate Long-Term Chemistry Teacher Professional Development

Deborah G. Herrington,[1] Ellen J. Yezierski,[2] and Senetta F. Bancroft[3]

[1]*Department of Chemistry, Grand Valley State University, 1 Campus Drive, Allendale, Michigan 49401*
[2]*Department of Chemistry and Biochemistry, Miami University, Oxford, Ohio*
[3]*Department of Curriculum and Instruction and Department of Chemistry and Biochemistry, Southern Illinois University, Carbondale, Illinois*

Received 27 April 2015; Accepted 29 February 2016

Abstract: The purpose of this study was to compare the ability of different instruments, independently developed and traditionally used for measuring science teachers' beliefs in short-term interventions, to longitudinally measure teachers' changing beliefs. We compared the ability of three self-report instruments (Science Teaching Efficacy Belief Instrument Form A [STEBI], Teaching of Science as Inquiry instrument [TSI], Inquiry Teaching Beliefs instrument [ITB]) and one observational instrument (Reformed Teaching Observation Protocol [RTOP]) to appropriately measure high school chemistry teachers' beliefs as they engaged in a two and a half year professional development program. Collectively our findings from these four instruments, across three separate cohort of teachers ($N = 16$), indicated conflicting changes in teacher beliefs. For example, the STEBI indicated teachers' self-efficacy remained unchanged or increased while the TSI indicated a concurrent decrease in self-efficacy throughout the PD program. Additionally, the ITB seemed to indicate a decrease in teachers' knowledge of inquiry while their interview data and RTOP scores indicated a concurrent increase in their knowledge of and ability to enact inquiry-based practices. We reconcile these conflicting results and discuss the implications these findings have for validly and reliably measuring science teacher belief changes within longer duration PD. © 2016 Wiley Periodicals, Inc. J Res Sci Teach 53: 1055–1081, 2016

Keywords: long-term professional development; science teacher beliefs; self-report instruments; observational instruments

There has been a longstanding national call for inquiry to be embedded in K-12 science classrooms (American Association for the Advancement of Science, 1993; National Research Council, 1996, 2000, 2012). Inquiry instruction emphasizes teacher facilitation of students pulling from their schemas and engaging in evidence-based argument and explanation about investigations (National Research Council, 1996, 2012). Further, when inquiry is central in science classrooms, students of all abilities and backgrounds are more capable and more likely to engage with science as argument and explanation (Basu & Barton, 2007; Seiler, 2001). Despite the

Contract grant sponsor: National Science Foundation; Contract grant numbers: 0553215, 1118658, 1118759.
Correspondence to: D. G. Herrington; E-mail: herringd@gvsu.edu
DOI 10.1002/tea.21323
Published online 29 March 2016 in Wiley Online Library (wileyonlinelibrary.com).

benefits of and persistent calls for reform, most evidence indicates that science instruction in US classrooms is not nor has ever been significantly inquiry centered (Crawford, 2007; Crippen, 2012). Constraints that influence this disparity between what is needed in science instruction and what teachers implement in K-12 science classrooms can include traditional habits ingrained from teachers' prior experiences as students (de Vries, Jansen, Helms-Lorenz, & van de Grift, 2014); limited or compartmentalized subject knowledge (Flores, Lopez, Gallegos, & Barojas, 2000; Roehrig & Luft, 2004); limited knowledge of inquiry (Wallace & Kang, 2004); heavy dependence on curriculum materials such as textbooks (Devetak & Vogrinc, 2013); and a positivist view of science, (Abd-El-Khalick, Bell, & Lederman, 1998; Crawford, 2007). However, these constraints are not insurmountable for teachers since providing them with professional development (PD) opportunities to deepen their conceptual and practical understanding of science, inquiry, and pedagogical content knowledge can transform their beliefs about science instruction (Wallace & Kang, 2004; Wilson, Floden, & Ferrini-Mundy, 2002).

PD opportunities that deepen teachers' conceptual and practical understanding of science, inquiry, and related pedagogy can transform their beliefs about science instruction reform (Herrington, Yezierski, Luxford, & Luxford, 2011; Lumpe, Vaughn, Henrikson, & Bishop, 2014; Wallace & Kang, 2004). Further, transformation of teachers' beliefs about teaching and learning science can help them to overcome the constraints commonly associated with implementation of reformed instructional practices (van Driel, Meirink, van Veen, & Zwart, 2012; Wallace & Kang, 2004). PD capable of transforming teachers' beliefs, not just their knowledge, about science, inquiry, and related pedagogy is crucial to science instruction reform because beliefs are "far more influential than knowledge in determining how individuals organize and define tasks and problems and are stronger predictors of behavior" (Pajares, 1992, p. 311). Nonetheless, teachers' beliefs about teaching and learning are idiosyncratic and can be very resistant to change (Brownlee, Boulton-Lewis, & Purdie, 2002; van Driel et al., 2012). The unique set of teacher beliefs that must be transformed to enable teachers' enactment of inquiry-based practices makes understanding how PD programs promote these transformations vital to national reform (Enderle et al., 2014).

Fundamental to understanding how PD programs work to transform teachers' instructional practices are instruments that generate valid and reliable data that empirically capture belief transformations within the contexts associated with these transformations (Bleicher, 2004). However, measuring changes in science teachers' beliefs is problematic (Mansour, 2009) because "there are no clear logical rules for determining the relevance of beliefs to real-world events and situations" (Nespor, 1987, p. 321). This problem is compounded by a rising need for PD interventions with increased duration (both total number of contact hours and time span over which the PD takes place) to effectively reform teacher practice (Garet, Porter, Desimone, Birman, & Yoon, 2001; Klassen & Chiu, 2010; Loucks-Horsley, Stiles, Mundry, Love, & Hewson, 2010; Yezierski & Herrington, 2011) while there remains a lag in the specific development of tools to longitudinally measure teachers' beliefs about science, inquiry, and reformed instructional practices. Further complicating this problem is the plethora of instruments that have been designed to measure related but different outcomes of science teachers' beliefs and have been developed independently from each other; therefore, when used together measurements can lack consistency and coherency (Heath, Lakshmanan, Perlmutter, & Davis, 2010). As a part of an ongoing exploration of the effects of the two and half year long Target Inquiry (TI) PD program on in-service chemistry teacher beliefs, a primary purpose of this study was to detail the challenges encountered in longitudinally measuring teachers' changing beliefs using instruments developed and traditionally used for measuring science teachers' beliefs in short-term interventions.

Theoretical Framework and Background Literature

Conceptualizing Beliefs

There are various definitions of beliefs currently in use in science education literature (Blömeke, 2014; Brown & Cooney, 1982; Haney, Lumpe, & Cerniak, 2003; Nespor, 1987; Pajares, 1992; Rokeach, 1968). We found Pajares's (1992) conceptualization, shaped in part by Rokeach's (1968) work, to be most useful in methodologically framing our understanding and measurement of teacher beliefs. Pajares' (1992) synthesis of seminal works on beliefs resulted in his distilled definition of belief as "an individual's judgment of the truth or falsity of a proposition, a judgment that can only be inferred from a collective understanding of what human beings say, intend, and do" (p. 316). Several studies indicate that the judgmental or evaluative nature of beliefs and how they connect to teachers' practice support the need to overlay teachers' beliefs with teachers' content knowledge, views on the nature of science, and pedagogical knowledge to fully understand what teachers ultimately do in their classrooms (Crawford, 2007; Richardson, 1996; Savasci & Berlin, 2012; Vermunt, 2014). Teacher beliefs can be broadly conceptualized as a bridge between their knowledge and teaching (Blömeke, 2014). Within this conceptualization of a bridge, teacher beliefs, practices, and knowledge are cast as connected yet separate from each other. In contrast, Rokeach's (1968) conceptualizations of beliefs ascribes a cognitive, affective, and behavioral dimension to a belief (Figure 1). We accordingly further conceptualize each of these three dimensions as a component of a teacher's beliefs rather than separate from them. Even further, the action an individual may enact is dictated by the knowledge and feelings the individual holds about a particular object or context (Rokeach, 1968). That is, an individual cannot enact behavior for which they possess no related knowledge. Consequently, the relationship among what an individual knows, feels, and ultimately does (or does not do) within a particular context are so intertwined it is difficult to meaningfully affect one component without affecting another component (Rokeach, 1968). Thus, a capture of teachers' knowledge, feelings, and enactment of inquiry is needed for a holistic understanding of the beliefs that influence their classroom practice as opposed to a capture of their knowledge or other belief components in isolation (Crawford, 2007).

Beliefs are also organized in terms of life events or episodes directly derived from personal experiences or indirectly from sociocultural contexts and institutional rules (Mansour, 2009). This episodic dimension means beliefs often derive their subjective power, authority, and legitimacy from vivid experiences or "critical episodes" in individuals' lives (Nespor, 1987). Long-term PD

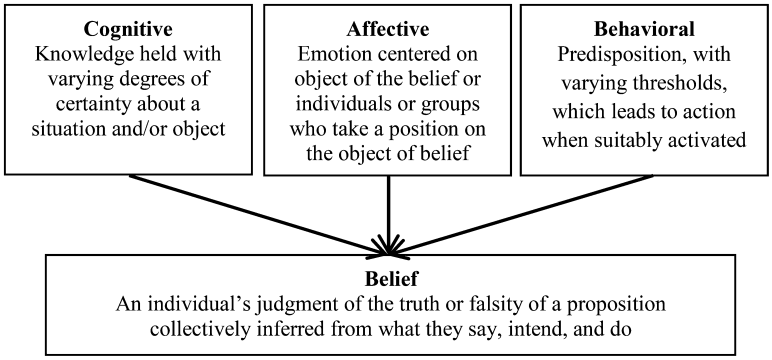

Figure 1. Model of a belief based on from Pajares's (1992) to Rokeach's (1968) conceptualizations.

programs can move science teachers' beliefs about teaching and learning that typically underlie traditional instruction towards reformed beliefs by incorporating experiences that create opportunities for inquiry-related critical episodes. Experiences that have been shown to have elements of critical episodes (i.e., they provide teachers with vivid experiences that are linked to enduring changes in instructional practice), include subject-specific research, pedagogical research, community, and reflection (Garet et al., 2001; Loucks-Horsley et al., 2010).

Therefore, for in-service science teachers, the degree to which their beliefs about inquiry are able to change during PD has a complex interdependency on their predisposition towards reformed practices based on previous critical classroom episodes, their experiences within the PD program, and the sociocultural environment of their school (Lotter, Harwood, & Bonner, 2007; van Driel et al., 2012). However, longitudinally tracking and documenting how PD experiences shape beliefs, if at all, as teachers engage with these experiences is challenging. This challenge is a product of both the multidimensional, episodic, and sociocultural nature of beliefs and also science education researchers' limitation to measuring beliefs solely through making inferences from teachers' statements, intentions, and actions related to inquiry in science teaching and learning.

Measuring Science Teachers' Beliefs

Beliefs must be measured inferentially, as an individual's underlying state is "fraught with difficulty because individuals are often unable or unwilling, for many reasons, to accurately represent their beliefs" (Pajares, 1992, p. 314). This poses significant challenges as science education researchers seek to develop not only tools to measure teachers' beliefs about inquiry, but also tools that can longitudinally and holistically capture how those beliefs change through teacher education or PD programs (Crawford, 2007; Luft & Roehrig, 2007). Resultantly, science education literature contains numerous methods to assess specific aspects of teachers' beliefs (Haney, Lumpe, Czerniak, & Egan, 2002). These methods and related instruments draw from a multitude of epistemologies and conceptualizations of teacher beliefs. No one instrument exists that can holistically capture teachers' beliefs; therefore, the following four instruments were selected: Science Teaching Efficacy Belief Instrument Form A (STEBI) (Riggs & Enochs, 1990); Teaching of Science as Inquiry instrument (TSI) (Smolleck, Zembal-Saul, & Yoder, 2006; Smolleck & Yoder, 2008); Inquiry Belief Teaching instrument (ITB) (Harwood, Hansen, & Lotter, 2006); and Reformed Teaching Observation Protocol (RTOP) (Sawada et al., 2002). Collectively, the tools selected for this study addressed the cognitive (STEBI, TSI, ITB, RTOP), affective (STEBI, TSI, ITB), and behavioral (RTOP) dimensions of participating teachers' beliefs within the context of science instruction. We focus our review on the constructs or rationale behind the four instruments used in this study. Details related to the administration of each instrument as well as the validity and reliability of the data obtained will be discussed in the Methods section.

In our attempt to capture the cognitive and affective components of science teachers' beliefs, we chose to use the STEBI and TSI; two Likert-type instruments built on Bandura's (1977) constructs of self-efficacy and outcome expectancy. The TSI was modeled after items on the STEBI and STEBI Form B (Enochs & Riggs, 1990; Smolleck & Yoder, 2008). Self-efficacy is "concerned with judgments of how well one can execute courses of action required to deal with prospective situations" (Bandura, 1982, p. 122) and is situated within social learning theory (Bandura, 1977). Bandura (1977) asserted that self-efficacy is a powerful tool to understand and predict behavior when applied appropriately. Gibson's and Dembo's (1984) findings from the development of the 30-item Teacher Efficacy Scale, which was among the first instruments found to provide valid and reliable data to quantitatively measured teachers' affect and cognition via their self-efficacy on a Likert scale, supported Bandura's assertion. Subsequent studies have found

that self-efficacy beliefs have a powerful influence on "thought patterns, emotional reactions, and the orchestration of performance through adroit use of subskills, ingenuity, resourcefulness, and so forth" (Gist & Mitchell, 1992, p.186). However, Bandura (1977, 1986, 2006) has continually emphasized that self-efficacy measurements have explanatory and predictive power only when they are specific to a task and the context in which the task must be performed. This specificity requirement has been shown to be needed in the measurement of science teachers' self-efficacy (Raudenbush, Rowan, & Cheong, 1992; Ross, Cousins, Gadalla, & Hannay, 1999). However, because microscopically defined measures gain predictive power but lose generalizability, instruments must strike a balance between specificity of measures and applicability of those measures to other settings (Pajares, 1996). Although balance is difficult to achieve (Tschannen-Moran, Woolfolk Hoy, & Hoy, 1998), the interest in science teachers' self-efficacy persists because there seems to be a strong connection between their self-efficacy beliefs and how much effort they choose to put forth, their persistence when faced with constraints, and how they cope with failure with respect to instructional tasks (Mansour, 2009).

Behavior is also influenced by outcome expectancy, which is a second construct situated within social learning theory (Bandura, 1977). Outcome expectancy regulates behavior based on an individual's judgment of whether their actions will produce a desirable outcome (Bandura, 1995). Although self-efficacy and outcome expectancy both emerge from the level of competence a person expects they can bring to a specific task within a specific context, they are distinct constructs (Bandura, 1986, 1997, 2006). Self-efficacy is a judgment of capability to execute specific tasks while outcome expectancy is a judgment about the likely outcomes from performing these tasks (Bandura, 2006). Further, self-efficacy beliefs precede and shape outcome expectancies (Bandura, 2006). Bandura's (1977) theory predicts that "teachers who believe student learning can be influenced by effective teaching [outcome expectancies] and who also have confidence in their own teaching abilities [self-efficacy beliefs] should persist longer, provide a greater academic focus in the classroom, and exhibit different types of feedback than teachers who have lower expectations concerning their ability to influence student learning" (Gibson & Dembo, 1984, p. 570 as cited in Riggs & Enochs, 1990). Despite this connection, outcome expectancies seem to add little to the predictive power of self-efficacy measures (Bandura, 1986; Tschannen-Moran et al., 1998). Yet, self-efficacy mediates both task performance and outcome expectancy (Bandura, 1986). Therefore, the inclusion of items addressing teachers' outcome expectancies can enhance an instrument's ability to explain (but not necessarily predict) behavior (Tschannen-Moran et al., 1998). However, given the inferential nature of beliefs discussed earlier, these quantitative measures on their own offer incomplete insights into teachers' beliefs without a further eliciting of teachers' internal models of science instruction.

To further elicit the internal conceptions of beliefs about inquiry teaching in the science classroom we selected the ITB (Harwood et al., 2006). The ITB uses a structured set of prompts to elicit these beliefs. Methods using instruments or protocols that help make teachers' cognitive and affective conceptualizations of teaching and learning with inquiry evident through some form of scaffolding, usually through a consistent set of prompts, is an essential facet of measuring teachers' beliefs. Consistent scaffolding is essential because direct questions, that for example ask teachers to describe their philosophy of teaching, are ineffective and sometimes counterproductive in eliciting their beliefs (Kagan, 1992). Some of these methods include asking teachers to think aloud as they analyze classroom vignettes (Akerson & Hanuscin, 2007) or videotaped performances (Lee, Hart, Cuevas, & Enders, 2004); semi-structured interviews (Luft & Roehrig, 2007); asking teachers to draw concept maps to depict their understanding of particular terms; discourse analysis of teacher questioning used in their classrooms (Oliveira, 2010); and close analysis of the language teachers use in their classrooms and descriptions of their thoughts and

actions (Kagan, 1992). The ITB instrument is grounded in a phenomenographic perspective that people have internal models of the world and base their behavior on those models (Harwood et al., 2006). The ITB generates both quantitative and qualitative data representative of teachers' beliefs about teaching and learning science with inquiry at the time of their engagement with the instrument through a card sorting activity and follow-up interview (Harwood et al., 2006). Although the STEBI, TSI, and ITB capture various dimensions of teachers' cognitive and affective components of their beliefs about science teaching and learning, they do not explicitly capture the crucial behavioral component of teachers' beliefs in the "rough and tumble" of classroom practice (Crawford, 2007).

We attempted to capture the cognitive and behavioral component of teachers' enactment of inquiry-based practices with the RTOP. The RTOP is an observational protocol designed to measure the extent to which teachers' instructional practices align with research proven practices (Sawada et al., 2002). Given that a definitive goal of science education reform is changing teachers' classroom practices, teacher educators, researchers, and PD programs need a protocol to determine whether interventions have ultimately supported teachers' achievement of this goal. In the absence of evidence of this achievement, traditional instructional practices are likely to persist necessitating continued calls for reform (Corcoran, Mosher, & Rogat, 2009). Although effective science instruction is often difficult to define (Haney et al., 2002), there are observational tools aligned with the national frameworks for science instruction that can aid researchers' identification of reformed classrooms. One example of this type of tool is the *Local Systemic Change Revised Classroom Observation Protocol* (Horizon Research, Inc., 1998). This protocol is a criterion-referenced instrument that uses trained observer judgments of teacher lessons and is accompanied by pre and post classroom observation interviews to elicit teachers' beliefs about science instruction. The RTOP, based in part on the *Local Systemic Change Revised Classroom Observation Protocol* (Horizon Research, Inc., 1998), is also a criterion-referenced instrument that uses trained observer judgments. However, it is designed to more specifically align with reformed science instruction and lacks an accompanying interview protocol.

The RTOP measures instructional quality through researcher observation while the STEBI, TSI, and ITB measure teachers' instruction via teachers' self-report. The self-report tools used in this study were developed and tested for short-term interventions (typically one or two semesters). The purpose of our study was to examine how these self-report tools designed for relatively short-term contexts can be appropriately used to measure teachers' beliefs and subsequent classroom practice before, during, and after the long-term (2.5 years) TI PD program. The following research questions framed the study:

1. How do teachers' self-report scores on teaching beliefs instruments about science teaching and learning with inquiry change during TI PD program?
2. How do teachers' changes in practice over time as measured by the RTOP correspond with their self-report scores on teaching beliefs instruments?
3. What are the implications for measuring teacher beliefs within PD that meets the call for programs which are long in duration?

Methods

Context of the Study

A quantitative, quasi-experimental, repeated measures, longitudinal design (Shadish, 2002) was used to compare teachers' scores on each of the four instruments over the duration of the

2.5 year PD program. Participating teachers were from a wide variety of schools in the Western Michigan area. Sampling was not randomized and was restricted to teachers who were personally motivated to increase their use of inquiry in their science teaching. The sample of teachers included in this study limits the generalizability of our results. However, the transferability of results may be suitable to other longer-term PD opportunities for which teachers volunteer. Three cohorts of teachers participated in 2.5 years of PD, which had three core experiences designed to offer teachers inquiry-related critical episodes. The model of TI PD is shown in Figure 2.

Core experiences of PD chronologically include research experiences for teachers (RET), materials adaptation (MA), and action research (AR). Each of these experiences was preceded by preparatory experiences during the regular academic year and was delivered primarily during consecutive summers. The RET component models the scientific process for teachers where for 6 weeks during the summer they work closely with chemistry mentors reviewing literature, mastering laboratory techniques, collecting and analyzing data, and presenting their findings at a regional or national science conference. At the end of the RET, teachers make small modifications to two of their existing classroom activities so the activities better reflect the scientific process modeled during their RET. Teachers spend the succeeding months strengthening their understanding of reformed science teaching and learning though group discussions and engaging with the chemistry education research literature. They then use this strengthened understanding in the process of MA and designing their AR. Through the MA process teachers develop and/or adapt classroom instructional materials to include inquiry, pilot the adapted activities with peers within their cohort, and revise activities based on peer feedback. Through AR teachers collect data to evaluate their adapted materials, present their findings, and write a scholarly text for a journal submission or a master's program thesis requirement.

PD occurred at Grand Valley State University (GVSU) and the delivery schedule for each cohort is shown in Table 1 below. For more detailed descriptions of PD delivery, timeline, preparatory, core, and application experiences, see Yezierski and Herrington (2011).

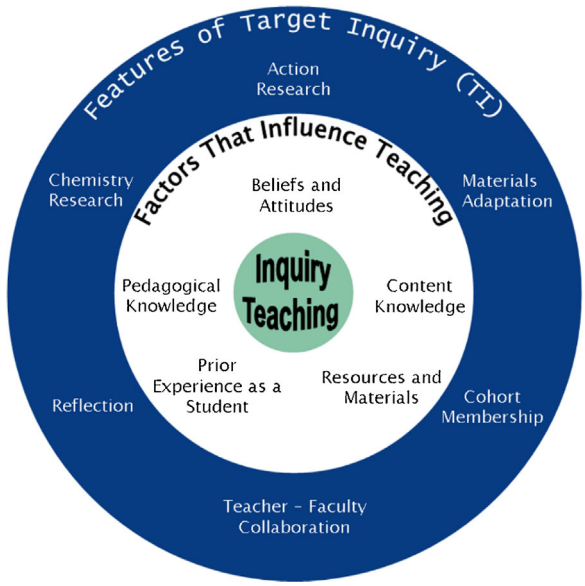

Figure 2. TI PD model. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

Table 1
Timeline for delivery of TI PD for cohorts 1–3

	Cohort		
Calendar Year	1	2	3
Fall 2005	Baseline data collection		
2006	Pre-RET		
	RET		
2007	Pre-MA and Pre-AR		
	MA	Baseline data collection	
Mid-2008	AR	Pre-RET	
		RET	
2009		Pre-MA and Pre-AR	
		MA	Baseline data collection
Mid-2010		AR	Pre-RET
			RET
2011			Pre-MA and Pre-AR
			MA
Mid-2012			AR

Participants

To be accepted into the program, teachers had to have a major or minor in chemistry, be currently teaching, and meet GVSU's graduate admission requirements of a 3.0 GPA. The program could accommodate up to 10 teachers in each cohort. Some teachers inquired about the program but then opted to not apply after learning that the PD was 2.5 years in duration. Each applicant was interviewed to ascertain their reason for wanting to participate in the program and willingness to reform their classroom instruction, and to gauge peer and administrator support for such reforms at their institution. Across the three cohorts, 24 chemistry in-service teachers from area high schools and two area colleges participated in PD. Eight of these teachers, however, had incomplete datasets and were subsequently excluded from this study. Demographic data for all teachers with complete data sets are summarized by cohort in Table 2.

Instruments

The STEBI was designed to measure elementary school teachers' beliefs about science teaching and learning through their self-efficacy and outcome expectancy (Riggs & Enochs, 1990). It has since been used with middle and high school science teachers (Enderle et al., 2014; Khourey-Bowers & Simonis, 2004). Validity analyses of STEBI data resulted in 13 items correlated with a self-efficacy subscale and 12 items correlated with an outcome expectancy subscale. Internal consistency tests of the two subscales yielded a Cronbach's alpha reliability coefficient of 0.92 for self-efficacy and 0.77 outcome expectancy (Riggs & Enochs, 1990). Teachers score each item using a five-choice Likert scale (1 = strongly disagree to 5 = strongly agree). Individual scores on the STEBI can range from 25 to 125 where high scores indicate high efficacy. Table 3 shows sample items from the STEBI.

The TSI measures pre-service teachers' self-efficacy and outcome expectancy towards teaching and learning science as inquiry (Smolleck & Yoder, 2008). The 69-items on the TSI are aligned with the five features of inquiry-based instruction as defined in National Science Education Standards (National Research Council, 2000). Although based on STEBI and STEBI B (Smolleck & Yoder, 2008), the TSI's alignment with National Research Council (2000) standards likely gives it increased specificity for measuring beliefs related to reformed teaching compared to

Table 2
TI PD participant and school data

Teacher Data				Teachers' School Demographics				Teachers' School Achievement[f]	
Teacher#	Teaching Experience[a]	Gender	School#	Economically Disadvantaged (%)	Non-White (%)	Student Population	AYP[e] (Y/N)	Math (%)	Science (%)
			MI	34.5	27.5			59.3	63.9
Cohort 1[b]									
1	17	M	1	22.3	5	544	Y	75.4	80.5
2	10	M	2	7.7	3.7	1,088	Y	63.5	66.6
3	20	M	3	32.5	41.8	1,640	Y	52.1	58
4	10	F	4	7.3	5.5	1,599	Y	78.2	73.8
5	3	F	5*						
6	6	M	6	19.2	27.1	1,755	N	68.3	74.4
Cohort 2[c]									
7	2	M	7	13.0	10.0	1,869	Y	69	75
8	14	M	8*			977			
9	3	M	9	7.0	9.0	1,495	Y	70	77
10	3	F	10	13.0	13.0	1,348	Y	64	76
11	2	F	11	75.0	99.0	576	N	4	10
Cohort 3[d]									
12	10	M	12	14.6	7.0	1,138	Y	48.9	36.0
13	2	M	13	28.7	13.0	1,234	Y	31.6	27.1
14	1	M	14	64.9	22.2	387	Y	11.8	11.6
15	13	F	15	14.3	3.2	833	Y	54.8	48.7
16	7	M	16	38.6	9.0	458	Y	36.1	22.5

[a]Years of teaching experience upon entry to the program.
[b]Cohort 1 participant and school data from 2005. School data were obtained, but no longer available, from http://www.ses.standardsandpoors.com/
[c]Cohort 2 participant and school data from 2007. School data were obtained, but no longer available, from http://www.ses.standardsandpoors.com/
[d]Cohort 3 participant and school data from 2009. School data were obtained from http://www.michigan.gov/mde
[e]AYP, adequate yearly progress.
[f]MI administered a new set of standardized exams in 2007. Student scores shown for Cohort 3 teachers were lower on these new exams compared to scores in previous years.
*Denotes private school, some data not published.

Table 3
Sample items from STEBI

4. When the science grades of students improve, it is often due to their teacher having found a more effective teaching approach.
5. I know the steps necessary to teach science concepts effectively.
13. Increased effort in science teaching produces little change in some students' science achievement.
24. I do not know what to do to turn students on to science.

Note: Items written in the first person are within the self-efficacy subscale and items in the third person are within the outcome expectancy subscale.

Table 4
Sample items from TSI

When I teach science . . .
1. I will be able to offer multiple suggestions for creating explanations from data.
5. I have the necessary skills to determine the best manner through which children can obtain scientific evidence.
21. I will be able to play the primary role in guiding the identification of scientific questions.
29. My students will derive scientific evidence from instructional materials such as a textbook.
41. My students will refine their explanations using possible connections to scientific knowledge that have been provided.
52. My students will analyze teacher provided data in a particular manner.
61. I will expect students to use internet based resources or other materials to further develop their investigations.

Note: Items written as "I" statements are within the self-efficacy subscale and items written as "My students" statements are within the outcome expectancy subscale (Dira-Smolleck, 2004).

the STEBI. Validity analyses resulted in 34 of these items correlated with a self-efficacy subscale and 35 items correlated with an outcome expectancy subscale (Smolleck & Yoder, 2008). Internal consistency tests of the two subscales associated with each of the five essential features yielded Cronbach's alpha reliability coefficients of 0.50 or higher (Smolleck & Yoder, 2008). Teachers score each item based on a five-point Likert scale (1 = strongly disagree to 5 = strongly agree) and individual scores can range from 69 to 345 where high scores indicate high efficacy. Table 4 shows sample items from the TSI.

The ITB (Harwood et al., 2006) is a blended qualitative and quantitative instrument consisting of 18 cards describing classroom activities as inquiry oriented (eight cards), non-inquiry (six cards), and neutral (four cards). Developers Harwood et al. (2006) settled on the 18 descriptions after testing three versions of the instrument in various settings. The sorted cards are a visual representation of teachers' internal model of inquiry. A follow-up interview allowing teachers to explain their internal model of science instruction and inquiry serves a validity check for researchers' interpretations of the visual ITB models. Table 5 shows sample activities described on the ITB.

The RTOP (Piburn et al., 2000; Sawada et al., 2002) is a 25-item classroom observation protocol used to measure changes in teachers' inquiry related classroom practices. Highly aligned with national science standards to reflect reformed instructional practices (American Association for the Advancement of Science, 1993; National Research Council, 1996), the RTOP is subdivided into three subscales. The lesson design and implementation subscale contains five items to measure teachers' pedagogical ability to create a classroom setting as a community that engages in exploration before explication. The content subscale contains ten items with five items

Table 5
Sample items from ITB

Inquiry Activities
H. Students collaborating with one another.
P. Students using evidence to defend their conclusions.
Neutral Activities
M. Students asking questions.
D. Students reading assignments in textbooks.
Non-Inquiry Activities
C. Students listening to instructor lecture.
O. Students completing worksheets.

Table 6
Sample items from RTOP

I. Lesson Design and Implementation
2. The lesson was designed to engage students as members of a learning community.
3. In this lesson, student exploration preceded formal presentation.
II. Content
Propositional knowledge
6. The lesson involved fundamental concepts of the subject.
7. The lesson promoted strongly coherent conceptual understanding.
Procedural knowledge
13. Students were actively engaged in thought-provoking activity that often involved the critical assessment of procedures.
14. Students were reflective about their learning.
III. Classroom Culture
Communicative interactions
16. Students were involved in the communication of their ideas to others using a variety of means and media.
17. The teacher's questions triggered divergent modes of thinking.
Student/teacher relationships
21. Active participation of students was encouraged and valued.
22. Students were encouraged to generate conjectures, alternative solution strategies, and ways of interpreting evidence.

correlating with propositional and procedural pedagogical content knowledge, respectively. The classroom culture subscale contains ten items with five items correlating with communicative interactions and student/teacher interactions respectively. Tests for internal consistency from data collected from 141 public school, college, and university classrooms yielded a standardized α of 0.97 for the entire instrument and a Cronbach's alpha reliability coefficient of at least 0.80 for each subscale (Sawada et al., 2002). Each item is evaluated on a five point scale (0 = never occurred to 4 = very descriptive). Individual teacher scores can range from 0 to 100 where high scores indicate high alignment with reformed instructional practices. Table 6 shows sample items on the RTOP.

Data Collection and Analysis

The STEBI was administered four times (pre-RET, post-RET, post-MA, post-AR/TI) for Cohort 1 and twice (pre-RET, post-RET) for Cohort 2. Baseline data collection for Cohort 1 occurred prior to TSI publication. Resultantly, Cohort 1 teachers' beliefs were first measured by the TSI post-RET. Additionally, due to funding, data collection ended in 2012 after Cohort 3's completion of the RET. Therefore, the TSI was administered thrice for Cohorts 1 and 2

(Cohort 1: post-RET, post-MA, post-AR/TI; Cohort 2: baseline, post-RET, post-MA) and twice (baseline, post-RET) for Cohort 3. The ITB was administered four times (pre-RET, post-RET, post-MA, post-AR/TI) for Cohort 1, thrice (pre-RET, post-RET, post-MA) for Cohort 2, and twice (pre-RET, post-RET) for Cohort 3. For the ITB, teachers were asked to place cards they perceive as describing inquiry activities nearest to the "classroom" card, located in the center of a 17" by 17" square, and descriptions they perceived less descriptive of inquiry activities farther from the "classroom" card. A brief interview followed the card sorting activity which allowed teachers to explain their model and served as an internal validity check. The STEBI, TSI, and ITB were all administered to teachers within a 1 week block at the beginning of each summer.

Teachers invited researchers to video record one to two lessons in their classrooms at times they would be doing what they believed to be inquiry-based lessons. Following the developers' intended use of the RTOP (Piburn et al., 2000), our 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. Any lesson-related materials such as lesson plans and student handouts were also collected. One lesson for every teacher was recorded prior to PD (or pre-RET) serving as a baseline measurement. Videos and lesson materials were used to independently assign RTOP scores by three trained raters. The three independently assigned scores were compared, and if total scores differed by more than five points the raters negotiated every individual item that differed by more than one point to decide consensus scores. 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 one point. After individual item negotiations were complete sub-scores and total scores were recalculated by each rater. Final consensus was considered to be found when all three total scores were within five points of each other after negotiation. An average of these three consensus scores was assigned to each observed lesson. The biannually staggered recruitment of PD participants provided five rated lessons (pre-RET, post-RET, post-MA, post-AR/PD, 2 years post-PD completion) for Cohort 1, four (pre-RET, post-RET, post-MA, post-AR/PD) for Cohort 2, and two (pre-RET and post-RET) for Cohort 3. As the PD program progressed, previously scored lessons were revisited by raters to verify consistent interpretations of teachers' practices and rater scoring; scoring was considered to be consistent both within and between cohorts. These annual measurements allowed us to track teachers' beliefs, as manifested by their enactment of their knowledge of reformed instructional practices, before, during, and after exposure to TI PD. Therefore, classroom observations did not attempt to randomly capture a lesson in each teacher's classroom, but rather a lesson *the teacher believed* to be a best representation of their use of inquiry based practices. Our use of RTOP scores as a result aligns with how we conceptualized a belief earlier in the literature review. Figure 3 summarizes a timeline for implementation of core PD experiences and data collection.

To track how scores on the self-report tools (STEBI, TSI, and ITB) changed over the course of the long-term PD experience (Research Question 1) and how they performed with respect to the observational tool (RTOP) (Research Question 2) we report measurements made over time by each instrument by cohort. Further, since we compared teacher scores over time by instrument by cohort, we regard the teacher as the unit of measurement and the cohort as the unit of analysis. Accordingly, the results section is divided into three parts. For each of the three cohorts, the findings from the self-report tools are followed by the findings from the RTOP. Data trends for quantitative tools are displayed graphically and any statistically significant differences between years are reported. Using the statistical software package R (version 4.1), we ran non-parametric Friedman tests to compare teacher scores by instrument within each cohort. Friedman (1937) tests were employed because Likert type scales, used in the STEBI, TSI and RTOP, often violate the

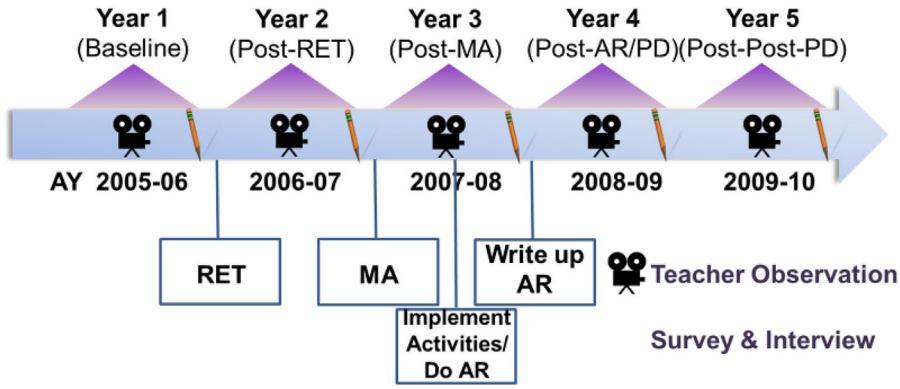

Figure 3. Implementation of core experiences and data collection timeline for Cohort 1. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

normality assumption of ANOVA. If Friedman tests indicated significance and there were data for more than 2 years, pairwise comparisons of scores between years were conducted using the Nemenyi *post hoc* test (Pohlert, 2014). The qualitative findings from the ITB are also presented to determine how teachers' ITB models and related statements of beliefs converge with quantitative findings for each cohort.

The quantitative portion of the ITB was abandoned due to persistent inconsistencies and we instead subjected teachers' models to qualitative analysis (see Herrington et al., 2011). To perform qualitative analysis of teachers' ITB models, digital photographs of models were used to create scaled figures with inquiry, non-inquiry, and neutral activity cards color coded for visual analysis. Any cards that teachers interpreted differently from the developers' intended meaning were marked with a star and the color changed to reflect teachers' interpretations. Final qualitative analysis was performed by chronologically compiling each model of each teacher's ITB model for rating. Raters independently viewed each teacher's compiled ITB models and rated their emergent models of inquiry as better, worse, or same between consecutive years and over the course of PD. The following criteria were used for rating: placement of inquiry and non-inquiry cards in reference to "classroom" card; number of tiers (fewer tiers = better rating); and mixing of inquiry and non-inquiry cards in the same tier (separate tiers for these two card categories = better). Inter-rater agreement for teachers' compiled ITB models averaged 85%. Additionally, in the interviews following their latest ITB construction, teachers were presented with their previous ITB models and were asked to explain any perceived differences among their models. Interviews were transcribed *verbatim*.

Results

Cohort 1

There were no significant differences found in Cohort 1 teachers' STEBI self-efficacy (χ^2 (3, $N = 6$) = 4.8, $p > 0.05$) and outcome expectancy (χ^2 (3, $N = 6$) = 2.0, $p > 0.05$) scores. Therefore, Cohort 1 teachers' median STEBI scores on each of the sub-scales remained relatively unchanged throughout TI PD (see Figure 4).

Thus, as a group Cohort 1 entered PD with a confidence in their ability to teach science that remained relatively unchanged during and after PD. However, a similar stability in self-efficacy and outcome expectancy beliefs was not tracked as measured by the TSI.

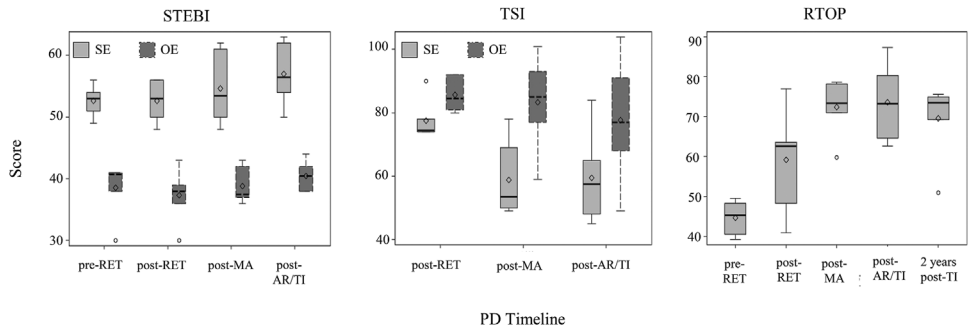

Figure 4. Comparison of measurement of Cohort 1 teachers' beliefs ($N = 6$). SE, self efficacy; OE, outcome expetancy; ◇, mean score.

The Friedman test indicated a significant difference in Cohort 1 teachers' TSI self-efficacy scores (χ^2 (2, $N = 6$) = 9.0, $p < 0.05$). Follow-up pairwise comparisons with a Nemenyi test indicated a statistically significant decrease in self-efficacy post-RET (Mdn = 74.50) to post-MA (Mdn = 53.50), $p = 0.03$. Additionally, as shown in Figure 4, their decreased TSI self-efficacy scores remained relatively unchanged post-AR/PD (Mdn = 57.50). Therefore, while Cohort 1 teachers' median TSI score remained declined post-AR, the absence of a significant difference in this decline post-MA to post-AR indicates that their self-efficacy towards teaching science with inquiry may have stabilized during their third and final year of PD. Further as shown in Figure 4, the TSI tracked relatively stable outcome expectancy for Cohort 1, $\chi^2\,(2, N = 6) = 1.9, p > 0.05$.

As we previously reported (Herrington et al., 2011), most Cohort 1 teachers displayed increased ability to organize their ITB models (fewer tiers and increased separation of inquiry and non-inquiry cards) as they progressed through PD along with a deepened conceptual and practical understanding of science as inquiry. Further, a majority of teachers made statements in their follow-up ITB interview that indicated they had moved away from viewing inquiry as theoretical to a practice they had personally appropriated for use in their classroom (see Herrington et al., 2011 for more details). However, some teachers who expressed this deepened understanding and appropriation of teaching science with inquiry created ITB models that were less separated (Herrington et al., 2011). According to the ITB developers (Harwood et al., 2006), less separated models indicate teachers are less able to discriminate between inquiry and non-inquiry activities. However, this interpretation conflicted with the statements by teachers who had less separated ITB models. Thus, the validity check via interviews revealed that all Cohort 1 teachers had an increased understanding of and ability to apply inquiry-related practices in their classrooms.

A significant difference in Cohort 1's RTOP scores was found, $\chi^2\,(4, N = 6) = 16.6, p < 0.01$. Follow-up pairwise comparisons with a Nemenyi test indicated a statistically significant increase in their median RTOP scores pre-RET (Mdn = 45.30) to post-MA (Mdn = 73.42), $p = 0.007$ (see Figure 4). The increase in RTOP scores post-RET (Mdn = 62.60) and then post-MA (Mdn = 73.42) indicate that as teachers sequentially engaged with PD activities their instruction also sequentially grew to incorporate more characteristics of reformed practices (Herrington et al., 2011). As shown in Figure 4 these gains then leveled off as their post-AR/PD (Mdn = 73.00) and 2 years post-PD (Mdn = 73.50) remained relatively unchanged to their post-MA scores. Therefore, the RTOP measured a sustained, longitudinal increase in Cohort 1 teachers' ability to enact reformed instructional practices which is supported by findings from the ITB, but conflicts with findings from the STEBI and TSI as shown in Figure 4.

Cohort 2

Cohort 2's STEBI self-efficacy scores significantly increased pre-RET (Mdn = 41.00) to post-RET (Mdn = 56.00), χ^2 (1, $N = 5$) = 5.0, $p < 0.05$. There was no statistically significant difference in their STEBI outcome expectancy scores pre-RET (Mdn = 33.00) to post-RET (Mdn = 42.00), χ^2 (1, $N = 5$) = 0.2, $p > 0.05$. However, the Friedman test indicated a significant difference in TSI self-efficacy scores (χ^2 (2, $N = 5$) = 8.4, $p < 0.05$). Follow-up pairwise comparisons with a Nemenyi test indicated a statistically significant decrease in TSI self-efficacy scores pre-RET (Mdn = 81.00) to post-MA (Mdn = 55.00), $p = 0.01$. The Friedman test also indicated a significant difference in TSI outcome expectancy scores, χ^2 (2, $N = 5$) = 8.4, $p < 0.05$. Follow-up pairwise comparisons with a Nemenyi test indicated a statistically significant decrease in TSI outcome expectancy scores pre-RET (Mdn = 108.00) to post-MA (Mdn = 84.00), $p = 0.01$. Cohort 2 teachers' STEBI and TSI self-efficacy and outcome expectancy scores are shown in Figure 5. Given these conflicting trends in measurement of similar constructs on the STEBI and TSI within Cohort 2 and compared to Cohort 1, the STEBI was no longer administered to teachers. Ideally, we would have continued to administer the STEBI to Cohort 2 and three teachers. However, the intensive nature of TI PD coupled with the significant demands on participants' time as in-service teachers led to the decision that it was inappropriate to ask teachers to spend additional time completing an instrument that was yielding inconsistent data.

Teacher 7's ITB models, shown in Figure 6, illustrate a typical trend seen in Cohort 2 teachers. Teacher 7's baseline ITB model of inquiry, though not well organized (four tiers), displayed his high ability to separate inquiry-based and non-inquiry activities, as these cards are clearly separated with inquiry activity cards grouped together and closest to the classroom card and non-inquiry cards farthest from the classroom card.

Teacher 7's post-RET models became better organized (two tiers) with no obvious change in his ability to discriminate among inquiry and non-inquiry activities baseline to post-RET as shown in Figure 6. However, Teacher 7 reintroduced four tiers in his post-MA model, the same number of tiers in the baseline model, indicating that the teacher's post-MA model of inquiry seemed to regress to a less organized model of inquiry and with a worsened ability to separate inquiry-based activities from non-inquiry activities. Yet, Teacher 7's explanation of his ITB models' progression reveals a clear and appropriated understanding of inquiry. He explains:

> [The baseline] one here seems very structured, I think I had a fairly good idea of what I wanted my classroom to be and in terms of the ideal situation, and I think this is more the last one is more of this is what it has become, so this is maybe what I had hoped, and this is maybe where it's actually at . . . [The post-MA model] I think this is more of now I know a lot about what inquiry instruction looks like . . . I think at some point my thinking was how do I take everything that I do in the classroom and jam it into this thing that we call inquiry, and realizing that that doesn't work . . .. So, this [post-RET model] was maybe the idealized, after I knew more. (post-MA interview)

Teacher 7's excerpt reveals his clarified understanding of inquiry resulted in more informed critiques of his use of inquiry-related instructional practices in his classroom. Further, Teacher 7's increasing RTOP scores, shown below each corresponding year's ITB model in Figure 6, reveal Teacher 7 was able to increasingly incorporate behaviors associated with reformed practices. Thus, the ITB and RTOP instruments together indicate Teacher 7 moved away from thinking about inquiry as an idealized, abstract notion to a well-understood set of practices he strove to incorporate into his instruction. As with some teachers in Cohort 1, this deeper understanding of inquiry, which led to more informed critiques of teachers' inquiry-related instructional practices

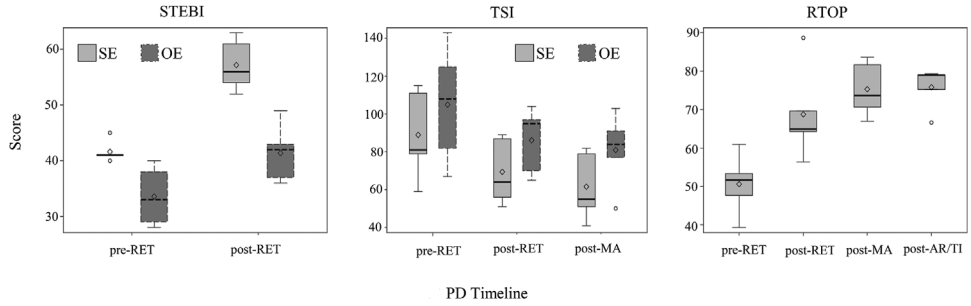

Figure 5. Comparison of measurement of Cohort 2 teachers' beliefs ($N = 5$). SE, self efficacy; OE, outcome expetancy; ◇, mean score.

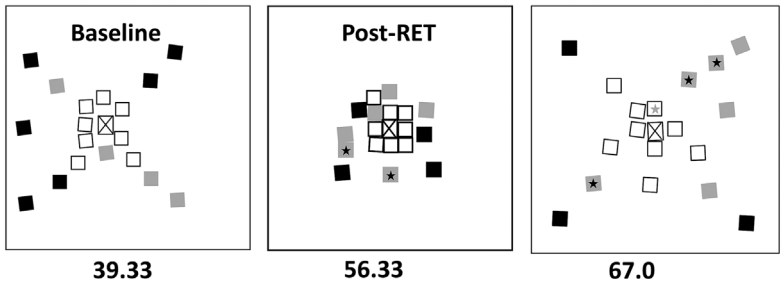

Figure 6. Changes in Teacher 7's (Cohort 2) ITB models and RTOP scores. White, inquiry; gray, neutral; black, non-inquuiry; X, classroom card; ★, color code for card changed to align with teacher's interpretation of card.

as described in his excerpt, also seemed to result in teachers' reduced self-efficacy towards teaching science with inquiry as indicated by Cohort 2's overall decrease in TSI self-efficacy scores. The increasing trend in Cohort 2 teachers' RTOP scores was similar to the trends seen in Teacher 7's RTOP scores.

There was a significant difference in Cohort 2's median RTOP scores, χ^2 (3, $N = 5$) = 10.0, $p < 0.05$. Follow-up pairwise comparisons with a Nemenyi test indicated a statistically significant increase pre-RET (Mdn = 51.67) to post-MA (Mdn = 73.66), $p = 0.02$. Further, as shown in Figure 5, overall teachers' gains in their RTOP scores were maintained 1 year beyond PD (Mdn = 79.00). Thus, regardless of the confusion of whether Cohort 2 teachers' self-efficacy beliefs increased (as measured by the STEBI) or decreased (as measured by the TSI) or that their outcome expectancies increased (as measured by the STEBI) or decreased (as measured by the TSI), overall ITB and RTOP measurements indicated teachers displayed increased ability to enact behaviors aligned with reformed teaching practices. However, it is notable that despite indications of deepened conceptual and practical understanding and greater enactment of reformed practices similar to Cohort 1, by their final year of PD decreased TSI outcome expectancy scores indicated that overall Cohort 2 teachers' were less likely to believe that effective inquiry-based science instruction influences student achievement. We did not see this same pattern for Cohort 1 teachers.

Cohort 3

As shown in Figure 7, Cohort 3 teachers' TSI self-efficacy did not statistically significantly differ pre-RET (Mdn = 80.00) to post-RET (Mdn = 71.00), χ^2 (1, $N = 5$) = 1.8, $p > 0.05$. Their

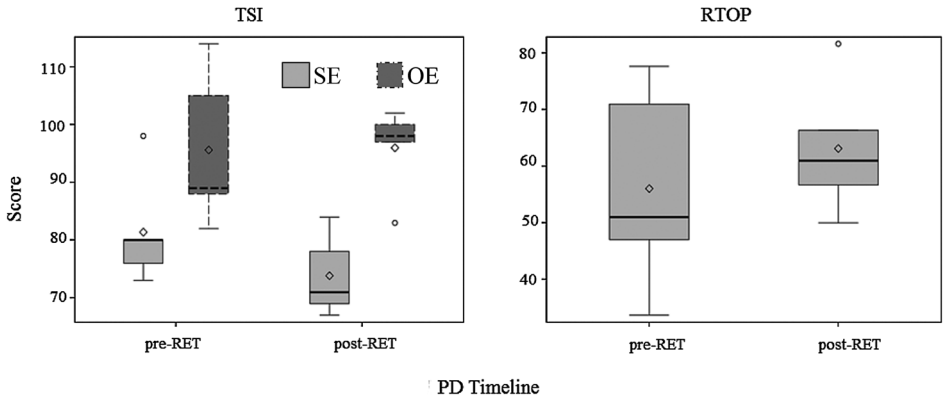

Figure 7. Comparison of measurement of Cohort 3 teachers' beliefs ($N = 5$). SE, self efficacy; OE, outcome expetancy; ◇, mean score.

related median outcome expectancy scores increased from 89.00 to 98.00, χ^2 (1, $N = 5$) = 0.2, $p > 0.05$. Despite the lack of a statistically significant difference, the trend of a decrease in TSI self-efficacy scores is consistent with the significant decreasing trend seen both in Cohorts 1 and 2. However, their increased median TSI outcome expectancy score is opposite to the decreasing trends seen with Cohorts 1 and 2 teachers during PD. As with Cohorts 1 and 2 teachers, Cohort 3 teachers tended to have well separated, but disorganized baseline ITB models. However, unlike Cohorts 1 and 2 teachers, Cohort 3 teachers' ITB models did not show a tendency towards greater organization and/or separation post-RET.

A sample of a Cohort 3 teacher's ITB model shows his progression from baseline to post-RET in Figure 8. This shows Teacher 14's ITB model was less organized as indicated by an increased number of tiers (from four to five) and decreased separation between inquiry and non-inquiry cards. Teacher 14 explains part of his process for arranging some of the ITB cards in Figure 8:

> . . . my students do a lot of the activities that aren't what I would consider inquiry that they do a lot of collaboration, but it's just sitting down, and you answer 1–5, I'll answer 5–10. It's technically collaboration, but it's not really. It [collaboration] doesn't have to be inquiry, and I think I just got a little more picky with those things. I can't picture somebody doing any of the stuff around the center [activities described on cards clustered in columns four and five of post-RET ITB model] this year . . . if my classroom was exactly the way I wanted it to be, and it was inquiry-based there would be collaborating with one another. But I could make a class where there's tons of collaboration and no inquiry, so I think that's why it got moved down to there. (post-RET interview)

Teacher 14's stated ability to be "a little more picky" about how collaboration can be shaped by the teacher to either be inquiry-based or devoid of inquiry reflects a deepened conceptual understanding of inquiry post-RET versus pre-RET. Further, his statement addressed hypothetical classroom contexts, which was also typically found in Cohorts 1 and 2 teachers post-RET ITB interviews. Thus, his statement lacks the stronger indications of the personal appropriation of inquiry present in Cohorts 1 and 2 teachers' post-MA interviews. The lack of appropriation can likely be linked to Cohort 3 not yet experiencing the MA component of PD.

As shown in Figure 7, Cohort 3 had an increase in their median RTOP scores from pre-RET (Mdn = 51.00) to post-RET (Mdn = 61.00), χ^2 (1, $N = 5$) = 1.8, $p > 0.05$. Thus, Cohort 3 teachers'

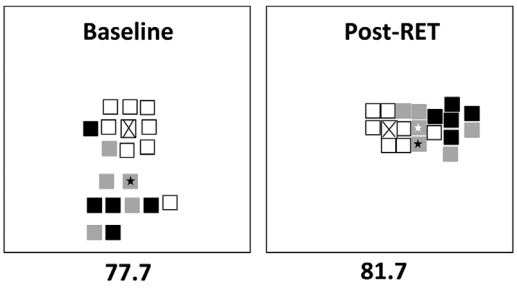

Figure 8. Changes in Teacher 14's (Cohort 3) ITB models and RTOP scores. White, inquiry; gray, neutral; black, non-inquuiry; X, classroom card; ★, color code for card changed to align with teacher's interpretation of card.

increase in median RTOP scores and decrease in median TSI self-efficacy scores, though not statistically significant, reflect trends analogous to Cohorts 1 and 2. These consistent trends indicate that overall all three PD teacher cohorts experienced a decreased self-efficacy towards teaching science with inquiry after a year of engaging with PD, but a concurrent increased enactment of inquiry-related practice. Despite these consistent trends in self-efficacy, there was no trend seen across cohorts in outcome expectancy in relation to their self-efficacy, their understanding of inquiry, nor their ability to enact inquiry-related practices.

Discussion and Conclusions

In response to Research Question 1, we found disaggregating the self-efficacy and outcome expectancy subscales on the STEBI and TSI important in identifying any clear trends in scores. STEBI self-efficacy scores for Cohort 1 teachers remained essentially stable across the three administrations while the STEBI self-efficacy scores for Cohort 2 teachers showed a statistically significant gain after their first year in PD. A plausible explanation for this may be related to differences in the cohort's years of teaching experience. Four of the five teachers in Cohort 2 entered PD with 3 years of experience or less experience (see Table 3) while four of the six teachers in Cohort 1 entered PD with 10 or more years of experience (see Table 2). Bandura (1997) proposed that self-efficacy beliefs remain relatively stable once established. Previous studies support Bandura's (1997) assertion and our findings that teacher efficacy appears to be more stable for more experienced teachers (Ross, 1994). Therefore, it is probable that Cohort 1's stable self-efficacy towards teaching and learning science in general may be a result of their greater years of teaching experience compared to Cohort 2. However, with respect to the use of the STEBI in long-term PD, Cohort 1 teachers' stability in self-efficacy beliefs was not necessarily indicative of their enactment of reformed instructional practices.

The TSI, the second instrument used to measure teachers self-efficacy beliefs, though yielding a clear pattern across cohorts, also seemed to not be indicative of teachers' classroom practices. The theoretical underpinnings of self-efficacy beliefs predict that increased self-efficacy related to a specific task and domain likely increases the transformation of beliefs related to that task and domain as well as the individual selecting to engage in those tasks (Bandura, 1986; Craighead & Nemeroff, 2001; Pajares, 1992). However, all cohorts showed a decrease in their median TSI self-efficacy scores at some point in PD, for Cohorts 1 and 2 overall decreases were statistically significant. Further, Cohort 2 showed successive decreases in their median TSI self-efficacy scores post-RET and post-MA despite concurrent increases in each cohort's RTOP scores. This unexpected finding is likely linked to teachers' more relativistic and sophisticated construction of inquiry and their subsequent more critical judgments during PD versus pre-RET.

During the development and validation of the TSI instrument, researchers found that participating pre-service teachers had inflated self-efficacy perceptions with regard to teaching of science as inquiry (Smolleck et al., 2006). Our consistent findings of decreased TSI scores across all three cohorts indicate that in-service high school chemistry teachers may also have inflated perceptions of their self-efficacy with respect to inquiry instruction. McDonald (1991) linked in-service teachers inflated self-efficacy with a false sense of certainty in their skills which increased as they gained experience. Wheatley (2000) found new in-service teachers were "overly optimistic" about their efficacy. However, teachers' gaining new and more effective practices is often offset by their revised definitions of good teaching (Tschannen-Moran et al., 1998). Some studies have found in-service teachers implementing new practices initially indicated lowered self-efficacy beliefs, but these eventually rebounded, usually once they saw positive changes in student learning (Ross, 1994; Stein & Wang, 1988). However, the trends in TSI scores indicated TI PD teachers did not experience a similar rebound of self-efficacy. Instead, we saw Cohort 1 teachers' declining TSI self-efficacy scores stabilize in their final year in PD. TSI scores 2 years after PD were indicative of continued stabilization rather than a rebound. Wheatley (2000, 2002) argued that teacher reframing of their self-efficacy are beneficial to their professional growth towards reformed instruction. Benefits are derived from this reframing because context specific doubts foster disequilibrium in teachers' thinking and "transformative change, genuine learning, happens only through disequilibrium, through the discovery that what I thought I knew isn't enough to deal with this new situation" (Jones & Nimmo, 1999, as cited in Wheatley, 2002). Wheatley's (2002) argument conflicts with much of the literature on teachers' self-efficacy and their subsequent classroom practices. However, our findings of decreased TSI scores yet increased enactment of inquiry-based practices and increasingly complex understandings of those practices indicate that TI PD may provoke beneficial reframing of self-efficacy for teachers of varying years of experience. These findings related to the concept of this beneficial reframing signify that predominant understandings and measurements of teachers' self-efficacy may be unable to appropriately capture changing teacher ability to teach science as inquiry within a long-term intervention context.

The discrepancy between the STEBI's measurement of stable self-efficacy for more experienced teachers and increasing self-efficacy for less experienced teachers versus the TSI's measurement of a consistent decrease across the three cohorts may be a result of the TSI's close alignment with reformed science instruction. With its close alignment to inquiry-based instruction (Smolleck & Yoder, 2008) based on National Research Council (2000) standards, the items on the TSI have greater context specificity to reformed practices than the STEBI. Therefore, a reasonable explanation for the self-efficacy measurement discrepancies between the two instruments is the difference in the tasks and/or context they were designed to measure (Bandura, 1977, 1986, 2006). However, despite the TSI's greater context specificity compared to the STEBI, it likely lacks enough task related specificity as it consistently measured decreasing self-efficacy scores as teachers' RTOP scores revealed their increased ability to enact the tasks associated with reformed practices. The relationship between self-efficacy measurements and task specificity will be discussed in our response to Research Question 2. No relationship between self-efficacy and outcome expectancy beliefs on either the STEBI or TSI across the three cohorts was found. The absence of any relationship was not surprising given that outcome expectancy been found to be uncorrelated to teachers' self-efficacy as discussed earlier (Tschannen-Moran et al., 1998).

The ITB follow-up interview elicited teacher statements that were consistent with teacher's deepened conceptual and practical understanding of reformed practices. Despite the quantitative aspect of the ITB not yielding valid data within the context of this study due to misinterpretations of the meaning of the activities described on cards (Herrington et al., 2011), we find the ITB

instrument to be a valuable qualitative tool to assess teachers' changing understanding of teaching and learning science with inquiry over multiple years. The ITB's indication of a teacher's deepened conceptual and practical understanding of reformed practices seemingly conflicted with the decreased self-efficacy scores on the TSI. Nevertheless, when this deepened conceptual and practical understanding of reformed practices is framed within the perspective of teachers' critical redefinition and subsequent re-evaluation of good science teaching (Tschannen-Moran et al., 1998; Wheatley, 2000, 2002), this apparent conflict was reconciled. However, the reconciliation between cognitive measurements of teachers' beliefs would have been unlikely without the ITB's provision of a structured set of prompts to elicit teachers' internal conceptualizations of their beliefs about reformed practices. Thus, the ITB's value seems rooted in the descriptions on the cards, the disaggregation of inquiry activities teachers perform during the card sorting process based on those descriptions, and the follow-up interview protocol. Each of these aspects of the ITB instrument generated a varied and consistent set of prompts for teachers to explicate their changing perception of the role and the use of inquiry in their classrooms. Most notably, the ITB instrument's card sorting process and accompanying interview revealed teachers' shift over time from viewing inquiry as a theoretical or abstract concept to an instructional strategy they appropriated and subsequently practiced in their classrooms.

In addressing Research Question 2, the relationship between change in practice as measured by the RTOP and self-report measures, we found that the trends in self-report measures somewhat conflicting with the RTOP's observational measures. Across all cohorts, the RTOP tool measured a consistent and sequential change in the conceptual, procedural, and pedagogical knowledge of inquiry of teachers throughout their experiences in TI PD. Further, RTOP scores revealed that these changes endured one and two years post-PD for Cohorts 2 and 1 teachers, respectively. Martin and Hand (2009) also found the RTOP to offer clear, consistent, multi-year measurements of teachers' implementation of reformed practices. Capturing teachers' enactment of inquiry within the classroom is key to measuring teachers' belief changes as they engage in PD because this enactment is a "linchpin" goal of reform policies (National Research Council, 2012). Therefore, we found the RTOP to be an essential tool to measure the behavioral dimension of teacher beliefs.

Our findings from the RTOP data revealed that collectively Cohorts 1–3 teachers' behavior moved away from teaching and learning strategies typically associated with traditional beliefs towards reformed beliefs of science instruction. Further, qualitative data from interviews following teachers' construction of their ITB models in this study indicate that some teachers' beliefs about inquiry in relation to their own instructional practice became more complex. This increased complexity suggests their exposure to PD may have resulted in a reconstruction of a simplistic epistemology of inquiry into more relativistic, sophisticated beliefs about inquiry (Brownlee et al., 2002). Therefore, when comparing the performance of the self-report tools to the observational tool, the ITB and RTOP indicated that teachers' cognitive and behavioral components of their beliefs about science instruction became more aligned with reformed practices. No meaningful trends emerged from the outcome expectancy scores on the STEBI and TSI in relation to the corresponding self-efficacy scores on these instruments or in relation to the ITB or RTOP. The trends in STEBI and TSI self-efficacy scores appeared to be in opposition to trends in teachers' RTOP scores over time; however, the ITB data allowed these opposing trends to be reconciled. This reconciliation linked teachers' decreasing self-efficacy (related to their affective component of their beliefs) with more critical perspectives (related to their cognitive component of their beliefs) of reformed science instruction.

Haney et al. (2002) also found inconsistent findings between the ratings from their observational tool, the *Local Systemic Change Classroom Observation Protocol* (Horizon

Research, Inc., 1998), and their measurements of teachers' self-efficacy using the self-report instruments which included a portion of the STEBI. Our results related to teachers' STEBI and TSI scores in relation to their RTOP scores and interview data reinforce Haney et al. (2002) findings that self-report, self-efficacy tools may not be appropriately capture teachers' knowledge of or enactment of reformed classroom practices. Studies (Mone, Baker, & Jeffries, 1995; Stumpf, Brief, & Hartman, 1987) outside of the science education community suggest that the ability of self-efficacy tools to capture an individual's performance on complex tasks is weaker when compared to measuring self-efficacy in relation to simple tasks.

Each requirement of PD bears the hallmark of a complex task as described by Campbell (1988). PD required teachers to incrementally learn, synthesize, and implement new knowledge about science, inquiry, and pedagogical content knowledge. Further, our teachers were required to implement this new knowledge as they navigated their unique personal and cultural constraints to classroom reform over 2.5 years while also catering to various students' needs within and between classes. These requirements when combined have dynamic task complexity, the most multivariate type of task complexity (Wood, 1986). Dynamic task complexity have several subordinate functions including: the number of distinct acts that need to be executed in the performance of the task, the number of distinct information cues that must be processed in the performance of those tasks, the judgments about timing, frequency, intensity, and location requirements for task performance, and the adaptation of task performance to changes in the environment over time (Wood, 1986). As task complexity increases, assessment of task requirements and individual and situational resources or constraints for these tasks must also be measured to increase the accuracy of self-efficacy measurement and therefore enhance validity (Gist & Mitchell, 1992). Further, changing self-efficacy may require individuals changing the way they process information and their subsequent selection of behaviors in which they choose to engage. Our findings from the ITB data clearly indicated that teachers across all three cohorts consistently changed their processing of information about teaching and learning with inquiry from abstract and simple to practical and complex. Additionally, data from the RTOP instrument indicated they subsequently modified their classroom behavior to become more aligned with reformed instructional practices. These findings indicate that previous studies and/or current instruments designed to measure science teacher efficacy may not sufficiently delineate the subordinate functions of the complex task of reforming science instruction.

Implications

In response to Research Question 3, our findings clearly have implications for measuring teacher change within long-term PD programs. Though classroom observations have been shown to be effective measures of teacher change, these are expensive in terms of time and resources. Therefore, in evaluating the effectiveness of PD programs many researchers rely on teacher self-report measures as they are able to capture data from a lager sample size in a less invasive manner while also requiring less time and fewer resources. However, available teacher self-report instruments largely measure some components of teachers' beliefs about teaching and learning with little regard to complex multidimensionality of those beliefs. Resultantly, these measures may work well for short-term PD which lack the necessary time or critical activities to cause a shift in teacher beliefs, but long-term PD programs aimed at affecting lasting instructional reform by necessity must change aspects of all three components of teachers' beliefs about teaching and learning. This holistic change in beliefs, in turn, changes the lens through which teachers are viewing the instrument, thereby reducing the validity and reliability of the data obtained. The apparent absence of delineation of the specific functions or tasks science teachers must perform to effectively reform their instruction on Likert-type, self-report instruments designed to measure

their changing self-efficacy beliefs supports a need to investigate the applicability of these instruments across different populations of teachers within the specific context of long-term PD. While the balance between specificity and generalizability must be considered (Pajares, 1996), this apparent absence creates a need for the development of new instruments explicitly designed to incorporate teachers' capability to perform specific tasks expected of teachers in reformed classrooms.

Gist and Mitchell (1992) argue that psychometric measurement of complex tasks, such as those expected in science instruction reform, should probe individuals about specific behaviors needed to perform those tasks. Within the context of reformed science instruction these instruments can include items that ask teachers to self-report on performing tasks such as the types of question they pose to students (van Zee, Iwasyk, Kurose, Simpson, & Wild, 2001), the types of student group interactions they use (Schneider, Krajcik, & Blumenfeld, 2005), their accuracy in presenting real-world science and the process of science to students (Pop, Dixon, & Grove, 2010), and their adaption of lessons to reflect real-world science and the process of science (Schneider et al., 2005). The need for the exploration of refined or new psychometric instruments to measure teacher beliefs about reformed science instruction longitudinally highlights the important role of concurrent use of qualitative and observational tools (Klassen, Tze, Betts, & Gordon, 2011).

Additionally, researchers must provide teachers with a self-report tool, such as the ITB in this study, which elicits teachers' internal representations of inquiry through a consistent set of prompts. The inferential nature of belief measurement demands a tool that makes teachers' internal representations of inquiry, and the extended time of long-term PD demands this tool contain a consistent frame of reference for teachers to respond. While researchers' interpretations of teachers' representations remain indirect, this consistent frame of reference allows researchers to make direct comparisons of these representations in pre and post measurements. Further, this type of self-report tool allows teachers to co-construct knowledge about science teaching and learning. This co-construction of knowledge has been cited as an important facet in understanding teacher beliefs (Keys & Bryan, 2001). Further, we have found, as have Crawford (2007) and Richardson, Anders, Tidwell, and Lloyd (1991), that it is insufficient for researchers to use representations of teachers' beliefs in isolation from how teachers enact these representations in real-world classroom practice. Therefore, an observational protocol that allows researchers to examine and evaluate how teachers enact their internal representations of inquiry in their classrooms is essential for use in conjunction with a qualitative self-report tool.

Finally, PD providers who seek to transform teacher beliefs about science teaching and learning should be willing to provoke doubt in novice and experienced teachers' initial assessments of self-efficacy related to their conceptual understanding and practice of reform science instruction. Doubt is essential to teachers' own pursuit of inquiry (Gabella, 1995) and when supported can aid teachers in becoming more critical of their teaching skills as their conceptual and practical understanding of teaching with reformed practices increase (Wheatley, 2000, 2002). However, to constructively engage this doubt and support transformation, PD providers must also provide teachers with opportunities for reflection (Schön, 1983) and a community of support (Friedman, 1997). Constructively engaging teachers with their doubts and providing them with opportunities to develop and implement new strategies require PD developers use long-term interactions (van Driel et al., 2001). Therefore, PD designed to reform science instruction should build into its model continual opportunities for teachers to reflect and collaborate (Loucks-Horsley et al., 2010; Wheatley, 2002). Reflection and collaboration are particularly effective when teachers have structured discussions, facilitated by an external expert, about implementing similar tasks, but have different experiences performing these tasks in their

classrooms (Ryan, 1999). These types of discussions increase their own practical knowledge as well as promote teachers' willingness to experiment with ideas shared by their colleagues (van Driel et al., 2001).

The authors thank Dr. Sango Otieno and Grand Valley State University's Statistical Consulting Center for their assistance in conducting the statistical analyses of the quantitative data as well as all of the TI teachers for their participation in the program. This material is based upon work supported by the National Science Foundation under Grant No. (0553215, 1118658, 1118759).

References

Abd-El-Khalick, F., Bell, R. L., & Lederman, N. G. (1998). The nature of science and instructional practice: Making the unnatural natural. Science Education, 82(4), 417–436.

Akerson, V. L., & Hanuscin, D. L. (2007). Teaching nature of science through inquiry: Results of a 3-year professional development program. Journal of Research in Science Teaching, 44(5), 653–680.

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

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review, 84, 191–215.

Bandura, A. (1982). Self-efficacy mechanism in human agency. American Psychologist, 37(2), 122–147.

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. Journal of Social and Clinical Psychology, 4(3), 359–373.

Bandura, A. (1995). Self-efficacy in changing societies. Cambridge, UK: Cambridge University Press.

Bandura, A. (1997). Self-efficacy: The exercise of control. New York, NY: Freeman.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), Self-efficacy beliefs of adolescents (pp. 307–337). Charlotte, NC: Information Age Publishing.

Basu, S. J., & Barton, A. C. (2007). Developing a sustained interest in science among urban minority youth. Journal of Research in Science Teaching, 44(3), 466–489.

Bleicher, R. E. (2004). Revisiting the STEBI-B: Measuring self-efficacy in preservice elementary teachers. School Science and Mathematics, 104(8), 383–391.

Blömeke, S. (2014). Framing the enterprise: Benefits and challenges of international studies on teacher knowledge and teacher beliefs—Modeling missing links. In S. Blömeke, F. Hsieh, G. Kaiser & W. H. Schmidt (Eds.), International perspectives on teacher knowledge, beliefs and opportunities to learn (pp. 3–17). New York, NY: Springer.

Brown, C. A., & Cooney, T. J. (1982). Research on teacher education: A philosophical orientation. Journal of Research and Development in Education, 15(4), 13–18.

Brownlee, J. M., Boulton-Lewis, G. M., & Purdie, N. M. (2002). Core beliefs about knowing and peripheral beliefs about learning: Developing a holistic conceptualisation of epistemological beliefs. Australian Journal of Educational and Developmental Psychology, 2, 1–16.

Campbell, D. J. (1988). Task complexity: A review and analysis. Academy of Management Review, 13(1), 40–52.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). Learning progressions in science: An evidence-based approach to reform. CPRE Research Report# RR-63. Consortium for Policy Research in Education.

Craighead, W. E., & Nemeroff, C. B. (2001). The concise Corsini encyclopedia of psychology and behavioral science. New York, NY: John Wiley & Sons.

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.

Crippen, K. J. (2012). Argument as professional development: Impacting teacher knowledge and beliefs about science. Journal of Science Teacher Education, 23(8), 847–866.

de Vries, S., Jansen, E. P., Helms-Lorenz, M., & van de Grift, W. J. (2014). Student teachers' beliefs about learning and teaching and their participation in career-long learning activities. Journal of Education for Teaching, 40(4), 344–358.

Devetak, I., & Vogrinc, J. (2013). The criteria for evaluating the quality of the science textbooks. In M. S. Khine (Ed.), Critical analysis of science textbooks: Evaluating instructional effectiveness (pp. 3–15). New York, NY: Springer.

Dira-Smolleck, L. (2004). The development and validation of an instrument to measure preservice teachers' self-efficacy in regard to the teaching of science as inquiry. (Doctoral dissertation). Retrieved from https://etda.libraries.psu.edu/search/1/50/31/author/term=smolleck

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.

Enochs, L., & Riggs, I. (1990). Further development of an elementary science teaching efficacy belief instrument: A preservice elementary scale. School Science and Mathematics, 90, 694–706.

Flores, F., López, A., Gallegos, L., & Barojas, J. (2000). Transforming science and learning concepts of physics teachers. International Journal of Science Education, 22(2), 197–208.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32(200), 675–701.

Friedman, V. J. (1997). Making schools safe for uncertainty: Teams, teaching, and school reform. Teachers College Record, 99(2), 335–370.

Gabella, M. S. (1995). Unlearning certainty: Toward a culture of student inquiry. Theory Into Practice, 34(4), 236–242.

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.

Gibson, S., & Dembo, M. (1984). Teacher efficacy: A construct validation. Journal of Educational Psychology, 76, 569–582.

Gist, M. E., & Mitchell, T. R. (1992). Self-efficacy: A theoretical analysis of its determinants and malleability. Academy of Management Review, 17(2), 183–211.

Haney, J. J., Lumpe, A. T., & Czerniak, C. M. (2003). Constructivist beliefs about the science classroom learning environment: Perspectives from teachers, administrators, parents, community members, and students. School Science and Mathematics, 103(8), 366–377.

Haney, J. J., Lumpe, A. T., Czerniak, C. M., & Egan, V. (2002). From beliefs to actions: The beliefs and actions of teachers implementing change. Journal of Science Teacher Education, 13(3), 171–187.

Harwood, W. S., Hansen, J., & Lotter, C. (2006). Measuring teacher beliefs about inquiry: The development of a blended qualitative/quantitative instrument. Journal of Science Education and Technology, 15(1), 69–79.

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.

Herrington, D. G., Yezierski, E. J., Luxford, K. M., & Luxford, C. J. (2011). Target inquiry: Changing chemistry high school teachers' classroom practices and knowledge and beliefs about inquiry instruction. Chemistry Education Research and Practice, 12(1), 74–84.

Horizon Research, Inc. (1998). Core evaluation manual: Classroom observation protocol. Chapel Hill, NC: Author.

Jones, E., & Nimmo, J. (1999). Collaboration, conflict, and change: Thoughts on education as provocation. Young Children, 54(1), 5–10.

Kagan, D. M. (1992). Implication of research on teacher belief. Educational Psychologist, 27(1), 65–90.

Keys, C. W., & Bryan, L. A. (2001). Co-constructing inquiry-based science with teachers: Essential research for lasting reform. Journal of Research in Science Teaching, 38(6), 631–645.

Khourey-Bowers, C., & Simonis, D. G. (2004). Longitudinal study of middle grades chemistry professional development: Enhancement of personal science teaching self-efficacy and outcome expectancy. Journal of Science Teacher Education, 15(3), 175–195.

Klassen, R. M., & Chiu, M. M. (2010). Effects on teachers' self-efficacy and job satisfaction: Teacher gender, years of experience, and job stress. Journal of Educational Psychology, 102(3), 741.

Klassen, R. M., Tze, V. M., Betts, S. M., & Gordon, K. A. (2011). Teacher efficacy research 1998–2009: Signs of progress or unfulfilled promise? Educational Psychology Review, 23(1), 21–43.

Lee, O., Hart, J. E., Cuevas, P., & Enders, C. (2004). Professional development in inquiry-based science for elementary teachers of diverse student groups. Journal of Research in Science Teaching, 41(10), 1021–1043.

Lotter, C., Harwood, W. S., & Bonner, J. J. (2007). The influence of core teaching conceptions on teachers' use of inquiry teaching practices. Journal of Research in Science Teaching, 44, 1318–1347.

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

Luft, J. A., & Roehrig, G. H. (2007). Capturing science teachers' epistemological beliefs: The development of the teacher beliefs interview. Electronic Journal of Science Education, 11(2), 38–63.

Lumpe, A., Vaughn, A., Henrikson, R., & Bishop, D. (2014). Teacher professional development and self-efficacy beliefs. In R. Evans, J. Luft, C. Czerniak & C. Pea (Eds.), The role of science teachers' beliefs in international classrooms (pp. 49–63). Boston, MA: Sense Publishers.

Mansour, N. (2009). Science teachers' beliefs and practices: Issues, implications and research agenda. International Journal of Environmental & Science Education, 4(1), 25–48.

Martin, A. M., & Hand, B. (2009). Factors affecting the implementation of argument in the elementary science classroom. A longitudinal case study. Research in Science Education, 39(1), 17–38.

McDonald, J. (1991). A messy business: Experience allows teachers to cope with uncertainty, not eliminate it. Teacher Magazine, 3(3), 54–55.

Mone, M. A., Baker, D. D., & Jeffries, F. (1995). Predictive validity and time dependency of self-efficacy, self-esteem, personal goals, and academic performance. Educational and Psychological Measurement, 55(5), 716–727.

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: The National Academies Press.

Nespor, J. (1987). The role of beliefs in the practice of teaching. Journal of Curriculum Studies, 19, 317–328.

Oliveira, A. W. (2010). Improving teacher questioning in science inquiry discussions through professional development. Journal of Research in Science Teaching, 47(4), 422–453.

Pajares, F. (1992). Teachers' beliefs and education research: Cleaning up a messy construct. Review of Educational Research, 62, 307–332.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. Review of Educational Research, 66, 543–578.

Piburn, M., Sawada, D., Turley, J., Falconer, K., Benford, R., & Bloom, I. (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.

Pohlert, T. (2014). The pairwise multiple comparison of mean ranks package (PMCMR). R Package. Retrieved from https://cran.r-project.org/web/packages/PMCMR/index.html

Pop, M. M., Dixon, P., & Grove, C. M. (2010). Research experiences for teachers (RET): Motivation, expectations, and changes to teaching practices due to professional program involvement. Journal of Science Teacher Education, 21(2), 127–147.

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1992). Contextual effects on the self-perceived efficacy of high school teachers. Sociology of Education, 65(2), 150–167.

Richardson, V. (1996). The role of attitudes and beliefs in learning to teach. In J. Sikula, T. J. Buttery & E. Guyton (Eds.), Handbook of research on teacher education (pp. 102–119). New York: Macmillan.

Richardson, V., Anders, P., Tidwell, D., & Lloyd, C. (1991). The relationship between teachers' beliefs and practices in reading comprehension instruction. American Educational Research Journal, 28(3), 559–586.

Riggs, I., & Enochs, L. (1990). Towards the development of an elementary teacher's science teaching efficacy belief instrument. Science Education, 74, 625–637.

Roehrig, G. H., & Luft, J. A. (2004). Constraints experienced by beginning secondary science teachers implementing scientific inquiry lessons. International Journal of Science Education, 26(1), 3–24.

Rokeach, M. (1968). Beliefs, attitudes, and values: A theory of organization and change. San Francisco, CA: Jossey-Bass.

Ross, J. A. (1994). The impact of an inservice to promote cooperative learning on the stability of teacher efficacy. Teaching and Teacher Education, 10(4), 381–394.

Ross, J. A., Cousins, J. B., Gadalla, T., & Hannay, L. (1999). Administrative assignment of teachers in restructuring secondary schools: The effect of out-of-field course responsibility on teacher efficacy. Educational Administration Quarterly, 35(5), 782–805.

Ryan, S. (1999). Constructing knowledge together: Teacher teams as learning communities. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Savasci, F., & Berlin, D. F. (2012). Science teacher beliefs and classroom practice related to constructivism in different school settings. Journal of Science Teacher Education, 23(1), 65–86.

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.

Schneider, R. M., Krajcik, J., & Blumenfeld, P. (2005). Enacting reform-based science materials: The range of teacher enactments in reform classrooms. Journal of Research in Science Teaching, 42(3), 283–312.

Schön, D. A. (1983). The reflective practitioner: How professionals think in action. New York, NY: Basic books.

Seiler, G. (2001). Reversing the "standard" direction: Science emerging from the lives of African American students. Journal of Research in Science Teaching, 38(9), 1000–1014.

Shadish, W. R. (2002). Revisiting field experimentation: Field notes for the future. Psychological Methods, 7(1), 3.

Smolleck, L. A., & Yoder, E. P. (2008). Further development and validation of the teaching science as inquiry (TSI) instrument. School Science and Mathematics, 108(7), 291–297.

Smolleck, L. D., Zembal-Saul, C., & Yoder, E. P. (2006). The development and validation of an instrument to measure preservice teachers' self-efficacy in regard to the teaching of science as inquiry. Journal of Science Teacher Education, 17, 137–163.

Stein, M. K., & Wang, M. C. (1988). Teacher development and school improvement: The process of teacher change. Teaching and Teacher Education, 4, 171–187.

Stumpf, S. A., Brief, A. P., & Hartman, K. (1987). Self-efficacy expectations and coping with career-related events. Journal of Vocational Behavior, 31(1), 91–108.

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. Review of Educational Research, 68(2), 202–248.

van Driel, J. H., Beijard, D., & Verloop, N. (2001). Professional development and reform in science education: The role of teachers' practical knowledge. Journal of Research in Science Teaching, 38(2), 137–158.

van Driel, J. H., Meirink, J. A., van Veen, K., & Zwart, R. C. (2012). Current trends and missing links in studies on teacher professional development in science education: A review of design features and quality of research. Studies in Science Education, 48(2), 129–160.

van Zee, E. H., Iwasyk, M., Kurose, A., Simpson, D., & Wild, J. (2001). Student and teacher questioning during conversations about science. Journal of Research in Science Teaching, 38(2), 159–190.

Vermunt, J. D. (2014). Teacher learning and professional development. In S. Krolak-Scherdt, S. Glock & M. Böhmer (Eds.), Teachers' professional development (pp. 79–95). Boston, MA: Sense Publishers.

Wallace, C. S., & Kang, N. H. (2004). An investigation of experienced secondary science teachers' beliefs about inquiry: An examination of competing beliefs sets. Journal of Research in Science Teaching, 41(9), 936–960.

Wheatley, K. F. (2000). Positive teacher efficacy as an obstacle to educational reform. Journal of Research and Development in Education, 34(1), 14–27.

Wheatley, K. F. (2002). The potential benefits of teacher efficacy doubts for educational reform. Teaching and Teacher Education, 18(1), 5–22.

Wilson, S., Floden, R., & Ferrini-Mundy, J. (2002). Teacher preparation research: An insider's view from the outside. Journal of Teacher Education, 53(3), 190–204.

Wood, R. E. (1986). Task complexity: Definition of the construct. Organizational Behavior and Human Decision Processes, 37(1), 60–82.

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.