International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: http://www.tandfonline.com/loi/tsed20

Understanding science teaching effectiveness: examining how science-specific and generic instructional practices relate to student achievement in secondary science classrooms

Jamie N. Mikeska, Tamara Shattuck, Steven Holtzman, Daniel F. McCaffrey, Nancy Duchesneau, Yi Qi & Leslie Stickler

To cite this article: Jamie N. Mikeska, Tamara Shattuck, Steven Holtzman, Daniel F. McCaffrey, Nancy Duchesneau, Yi Qi & Leslie Stickler (2017) Understanding science teaching effectiveness: examining how science-specific and generic instructional practices relate to student achievement in secondary science classrooms, International Journal of Science Education, 39:18, 2594-2623, DOI: 10.1080/09500693.2017.1390796

To link to this article: https://doi.org/10.1080/09500693.2017.1390796

Published online: 06 Dec 2017.

Submit your article to this journal

Article views: 1

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=tsed20

Download by: [RMIT University Library] **Date:** 08 December 2017, At: 02:04

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION, 2017
VOL. 39, NO. 18, 2594–2623
https://doi.org/10.1080/09500693.2017.1390796

Routledge
Taylor & Francis Group

Check for updates

Understanding science teaching effectiveness: examining how science-specific and generic instructional practices relate to student achievement in secondary science classrooms

Jamie N. Mikeska [a], Tamara Shattuck[b], Steven Holtzman[a], Daniel F. McCaffrey[a], Nancy Duchesneau[c], Yi Qi[a] and Leslie Stickler[a]

[a]Research & Development, Educational Testing Service, Princeton, NJ, USA; [b]College of Arts and Sciences, Western New England University, Springfield, MA, USA; [c]College of Education, Michigan State University, East Lansing, MI, USA

ABSTRACT

In order to create conditions for students' meaningful and rigorous intellectual engagement in science classrooms, it is critically important to help science teachers learn which strategies and approaches can be used best to develop students' scientific literacy. Better understanding how science teachers' instructional practices relate to student achievement can provide teachers with beneficial information about how to best engage their students in meaningful science learning. To address this need, this study examined the instructional practices that 99 secondary biology teachers used in their classrooms and employed regression to determine which instructional practices are predictive of students' science achievement. Results revealed that the secondary science teachers who had well-managed classroom environments and who provided opportunities for their students to engage in student-directed investigation-related experiences were more likely to have increased student outcomes, as determined by teachers' value-added measures. These findings suggest that attending to both generic and subject-specific aspects of science teachers' instructional practice is important for understanding the underlying mechanisms that result in more effective science instruction in secondary classrooms. Implications about the use of these observational measures within teacher evaluation systems are discussed.

ARTICLE HISTORY
Received 20 January 2017
Accepted 8 October 2017

KEYWORDS
Science teaching; secondary teachers; value-added measures; observational measures

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Introduction

Creating conditions for students' meaningful and rigorous intellectual engagement in science classrooms is both critically important and challenging for teachers (National Research Council, 2007, 2011; NGSS Lead States, 2013; Schneider & Plasman, 2011). Science teachers make numerous pedagogical decisions and engage in a wide variety of instructional practices on a daily basis – all of which have the potential to positively or negatively influence students' opportunities to learn. Providing high-quality learning opportunities in science classrooms requires attention to a myriad of factors from the selection and design of instructional activities or tasks to the facilitation of them with

CONTACT Jamie N. Mikeska jmikeska@ets.org
© 2017 Informa UK Limited, trading as Taylor & Francis Group

students to the consideration of students' backgrounds and ideas (Davis, Petish, & Smithey, 2006; Kloser, 2014; Windschitl, Thompson, Braaten, & Stroupe, 2012). This research underscores the point that science teachers have an important role in fostering students' achievement.

For the most part, the body of literature examining effective instructional practices in science classrooms focuses almost exclusively on supporting teachers in learning how to enact reform-oriented science instruction (cf. Minner, Levy, & Century, 2010). These science-specific instructional practices are ones that directly target the integration of scientific practices and habits of mind, and are mostly endemic to science classrooms, such as opportunities for students to design and conduct scientific-investigations, analyze and critique scientific data, and construct scientific explanations and arguments. At the same time, literature across subject areas has focused on the importance of generic aspects of teachers' instructional practice. These generic features of teachers' instruction are not situated within particular content areas, but are ones that teachers in many different subjects leverage to positively impact students' learning, such as providing students with opportunities to take responsibility for their own learning; providing a well-structured and managed classroom environment; or engaging in productive dialogic interactions with students. Although these generic aspects of instructions interact with the subject-specific content under study, they are generic in the sense that they have been studied and used across content areas to promote students' learning. Yet, little is known about how these science-specific and generic aspects of science teachers' instructional practice can be used to create conditions for meaningful learning opportunities for students.

In this study we examined the extent to which a sample of 99 secondary biology teachers used science-specific and generic instructional practices in their classrooms, and how the use of these instructional practices related to their students' achievement, as determined by teachers' value-added measures (VAM). Understanding which instructional practices science teachers employ and how various instructional practices relate to student achievement can be useful for making decisions about which practices science teachers should prioritise in the classroom and which practices should be the focus of science teachers' professional development. The main research questions guiding this study are:

(1) How frequently do secondary biology teachers use subject-specific and generic instructional practices that are currently promoted as indicators of quality teaching in their classrooms?
(2) Which of these science-specific and generic instructional practices predict student achievement?

In the next section, we begin with a description of the study's theoretical framework. Then, we review the literature in two areas: (1) research examining the subject-specific and generic aspects of science teachers' instructional practices that are linked to positive student learning outcomes and (2) research examining the use of value-added models (VAMs) in teacher evaluation. The main part of the paper focuses on describing the study's methodology and findings in relation to the study's research questions. We end with a discussion on the study's implications regarding the use of these measures within teacher evaluation systems.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Theoretical framework

This study is grounded in the theoretical orientation of the opportunities to learn framework. Characterising and understanding the key areas that are related to students' opportunities to learn has a long history in educational policy and in the research on achievement gaps. Studies have shown that students' opportunities to learn are related to school and district factors and processes, student and teacher characteristics, and student achievement (Blömeke, 2014; Chism & Pang, 2014; Lafontaine, Baye, Vieluf, & Monseur, 2015). For the purposes of this study we bear down on one particular educational indicator that research has suggested is critical for understanding students' opportunities to learn: teachers' instructional practice. In the last few decades, there have been numerous professional development efforts to provide high-quality learning opportunities for teachers in order to directly impact changes to their knowledge and beliefs and, ultimately, lead to improved classroom practice and better student outcomes (Borko, 2004; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). A key assumption in the theory of action for teacher learning is that what teachers know is related to what they can do and the learning opportunities that they provide their students (Desimone, 2009).

Knowing how to improve student learning requires that one is able to discern the underlying mechanisms responsible for particular outcomes in an instructional system (Raudenbush, 2008). This work requires that one delves into the 'black box' of instructional practice and determine how teachers develop their students' conceptual understanding (Cohen, Raudenbush, & Ball, 2003). As Raudenbush (2009) notes, currently we have limited knowledge on how 'core elements can be combined to produce a coherent instructional system that we can train teachers to enact reliably to optimize the impact of schooling' (p. 174). This study is designed to directly address this gap and examined how the different types of instructional practices that secondary science teachers use are related to students' learning.

Examining science teachers' instructional practices

Research has indicated that there are both subject-specific and generic aspects of science teachers' instructional practices that are more likely to result in increased student outcomes. In this section, we highlight the main findings across this research and note the key features of K-12 science teachers' instruction that have been noted as hallmarks of high quality science teaching. Although there are other factors that play a role in teachers' abilities to carry out these practices – for example, much research has documented the critical importance of science teachers' content knowledge and pedagogical content knowledge in instructional decision-making (Berry, Friedrichsen, & Loughran, 2015) – here our intent is only to identify which pedagogical practices science teachers use during instruction that research has suggested are linked to students' learning.

As shown in the literature reviewed below, the majority of the studies of science teachers' instructional practices tend to be small scale in nature, focused on examining only a handful of classrooms or teachers. In addition, the few large scale studies in this area tend to use teachers' self-report data as a measure of their instructional practice and focus on only one – or a small set – of instructional practices. Finally, the studies to date typically examine instructional practices that are either specific to science

Downloaded by [RMIT University Library] at 02:04 08 December 2017

instruction (e.g. how teachers engage their students in scientific-investigations) or generic across disciplines (e.g. how teachers manage their classrooms); studies rarely investigate how both science-specific and generic instructional practices operate together within and across science lessons to support student learning. Currently the field has relatively few examples of observational studies at scale examining the connection between a wide variety of science teachers' instructional practices and their students' learning. This study will directly contribute to the emerging research in science to better understand the connection between effective teaching practices – both subject-specific and generic practices – and student achievement.

Subject-specific features of science teachers' instruction

Inquiry-based instruction has been highly advocated in K-12 science education due to the increasingly prominent goal to have students learn how scientific knowledge is constructed, and engaging in scientific-investigations gives students an opportunity to do just that (Cobern et al., 2010; Fogleman, McNeill, & Krajcik, 2011; Minner et al., 2010; Songer, Lee, & Kam, 2002). This type of instruction, which can vary on a continuum from more teacher-directed to student-directed (National Research Council, 2000), provides opportunities for students to engage directly in scientific-investigations, for example, by generating investigation questions, collecting and making sense out of data, and communicating results to multiple audiences. Student-directed activities as those in which the students have more autonomy and decision-making options within the classroom activity, while teacher-directed activities involve more guidance from the instructor (National Research Council, 2000).

Studies have suggested that engaging students in student-directed science investigation learning opportunities can lead to improved student outcomes. Student-directed investigations provide opportunities for increased student cognitive involvement where students are expected to make decisions in the investigation (Fogleman et al., 2011). Similarly, Songer et al. (2002) found that student-directed inquiry, in which students ask and generate their own questions, made science instruction more relevant and personally meaningful for students. However, other research (e.g. Hmelo-Silver, Duncan, & Chinn, 2007; Kirschner, Sweller, & Clark, 2006) suggests that scaffolding is necessary in inquiry-based instruction to prevent students' cognitive overload. For example, Stone (2014) notes that scaffolding student learning in inquiry 'by using instruction and assessment tools aimed at promoting development of specific inquiry skills' (p. 97) rather than implicitly teaching inquiry skills can significantly increase student understanding of inquiry practices. Likewise, a meta-analysis by Furtak, Seidel, Iverson, and Briggs (2012) found higher effect sizes for studies that involved teacher-led activities. In response to the debate over whether student-directed or teacher-directed activities are more beneficial, Martin-Hansen (2002) suggests that different types of inquiry (more student-directed, more teacher-directed, or a combination) may meet specific needs in science classrooms.

Although the specific nature of the inquiry-based instruction may vary, studies have linked these subject-specific features of science teachers' instructional practice to student outcomes. A few meta-analyses have shown that inquiry-based instruction can be beneficial to student learning. One such analysis by Minner and DeLisi (2010) found that 51% of 138 studies showed positive impacts of inquiry instruction on student

Downloaded by [RMIT University Library] at 02:04 08 December 2017

content learning and retention, especially when the inquiry instruction emphasised active thinking and drawing conclusions from data. Schroeder, Scott, Tolson, Huang, and Lee (2007) found in another meta-analysis of experimental and quasi-experimental studies that inquiry strategies have a significant positive effect on student achievement. Other studies have also found inquiry-based instruction to be beneficial for narrowing the achievement gap between white and minority students (Marshall & Alston, 2014) and promoting achievement among students with behavioural/emotional challenges or learning disabilities (Palincsar, Magnusson, Cutter, & Vincent, 2002). In particular, a three-year study (Geier et al., 2008) suggested that a standards-based inquiry science curriculum can lead to standardised achievement test gains when the curriculum is highly specified and aligned with professional development and administrative support. Overall this research highlights the importance of creating opportunities for students in science classrooms to engage in inquiry-based science instruction as an important characteristic of effective science teaching.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Generic features of science teachers' instruction

The literature has also pointed to the critical importance of attending to generic features of science teachers' instruction that have proved to be particularly important for promoting students' science learning. Much of these generic features of high-quality instructional practice have their roots in the process-product research of the 1980s and 1990s, which dominated studies of teaching effectiveness during that time period (Brophy, 2000; Good & Brophy, 2000; Kyriakides, Christoforou, & Charalambous, 2013; Seidel & Shavelson, 2007). As noted by Seidel and Shavelson (2007), this previous research focused on 'conditions of teaching that enhance student outcomes' (p. 456); the process variables usually consisted of various teaching approaches or interventions, such as direct instruction, time on task, cooperative learning, feedback/evaluation, or differentiation, that researchers could identify and link to student outcomes. However, this research did not focus on the underlying mechanisms by which teachers engage students within each of these particular teaching approaches or interventions. In more recent years, the focus has shifted towards identifying and analyzing various teaching practices, or components, that signal patterns in how teachers engage students in the learning process (Seidel & Shavelson, 2007). Since these generic teaching practices can be employed across disciplines, below we highlight key findings from research in both science and other content areas that suggests these generic instructional practices are key features of high-quality learning environments. These generic teaching practices focus on the ways in which teachers structure and support the social interactions within the classroom, how they organise the content storyline within and across lessons, how they provide opportunities for students to self-direct their own learning, and how they manage the classroom environment.

Creating a positive environment and supporting productive interactions

The research literature supports the importance of attending to various aspects of the classroom environment and promoting positive interactions to support student learning. For example, several observational studies confirmed the association between the emotional support that teachers provide for students and classroom quality (Malmberg, Hagger, Burn, Mutton, & Colls, 2010; Opdenakker & Van Damme, 2006; Park &

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Oliver, 2009). Malmberg et al. (2010) attributed 11% of the variance in observed classroom quality to teachers' emotional support for their students while Park and Oliver (2009) identified psychologically safe classrooms – defined as environments in which students feel accepted, understood, valued, and free to express their emotions – as a key aspect of instruction benefitting students. Opdenakker and Van Damme (2006) similarly found that a learner-centered teaching style, which included the teacher's orientation towards student personal development and trusting relationships with students, was positively associated with instructional supports and effective classroom practices. Building relationships with students, promoting their autonomy, and establishing a positive affective climate in their classrooms were the primary characteristics that differentiated award-winning educators in a large urban school district from their average or low-performing peers (Worley, Titsworth, Worley, & Cornett-DeVito, 2007). Likewise, in a synthesis of 49 empirical articles, Rolland (2012) identified a significant relationship between teacher emotional support and student achievement scores.

Research has also suggested the importance of ensuring that students have opportunities to share their ideas, and teachers use this information responsively to support students' learning (Gotwals, Cisterna, Lane, Kintz, & Ezzo, in press; Harris, Phillips, & Penuel, 2012; Kane, Taylor, Tyler, & Wooten, 2011; Park & Oliver, 2009; Worley et al., 2007). In addition to establishing a positive classroom climate, award-winning teachers in the Worley et al. (2007) study were distinguished by their use of formative assessment strategies. Gotwals et al. (in press) reviewed a broad swath of literature on formative assessment practices that reinforced the importance of questioning strategies such as eliciting student understanding, asking challenging questions, using adequate wait time, and privileging student ideas in discussions rather than chasing the 'right' answer. Similarly, a conceptual and literature review of high-leverage practices for science education identified four core practices, three of which related to discourse and student questioning: eliciting student ideas, helping students make sense, and pressing students for evidence-based explanations (Windschitl et al., 2012). Both literature reviews suggest that these sets of instructional practices are related to more positive student learning outcomes.

Finally, literature about verbal practices that teachers use in their classrooms indicate that a broad array of practices associated with high-level, or cognitively demanding, questions and discourse seem to activate student thinking and promote positive learning outcomes (Alozie & Mitchell, 2014; Bleicher, Tobin, & McRobbie, 2003; Chin, 2006, 2007; Christodoulou & Osborne, 2014; ibe, 2009; McNeill & Pimentel, 2010; Oliveira, 2010; Smart & Marshall, 2013; Stronge, Ward, Tucker, & Hindman, 2007; Walshaw & Anthony, 2008). High cognitive-level questioning, such as responsive questioning where teachers press students for deeper thinking (Chin, 2006, 2007), encourages more elaborate student responses that include reasoning. This link between the cognitive-level of teachers' questions and the cognitive-level of thinking reflected in student responses has been documented across studies (Christodoulou & Osborne, 2014; ibe, 2009; McNeill & Pimentel, 2010; Smart & Marshall, 2013; Stronge et al., 2007).

The importance of lesson organisation

Within the literature, researchers and practitioners alike advocate for the use of three lesson organisation practices to bolster student achievement outcomes. First, teachers must use learning goals to 'to cue and/or encourage them [students] to cognitively

operate on goal-relevant information in a particular way' (Jiang & Elen, 2011, p. 555). In addition, framing the goals to activate a mastery orientation among students, where students learn to develop 'competence through an iterative learning process' (Rolland, 2012, p. 397), is related to higher achievement levels. Second, in addition to learning goals, teachers who want to foster conceptual learning need to provide task overviews coupled with student reflection activities that connect back to the content (Davis, 2000; Roth et al., 2011; Thadani, Stevens, & Tao, 2009). Third, research has suggested that teachers' use of thematic units, content storylines, big ideas, and cross cutting concepts is imperative to supporting students' learning. Teachers can use these teaching structures to differentiate learning for students, as well as to make the content more relevant and concrete, which is associated with higher achievement score gains for students (Conderman & Bresnahan, 2008; Kloser, 2014; Park & Oliver, 2009; Roth et al., 2009; Roth et al., 2011).

Providing opportunities for self-directed learning

Self-directed learning is the process in which a student monitors, evaluates, and engages in his or her own learning. Paris and Paris (2001) found that successful self-directed learning allowed students to pursue their own goals in engaging and supportive environments. Engaging learning environments support learners as they actively construct meaning, set goals, and choose strategies in relation to particular achievement contexts (Pintrich, 2000). Teachers can directly influence the development of systematic patterns of thoughts, actions, and feelings in their students that enable them to attain personal learning goals (Ainley & Patrick, 2006). Students who engage in problem-based learning with the intent to develop life-long self-directed learning skills have shown superior learning outcomes (Loyens, Magda, & Rikers, 2008; McCaslin et al., 2006; Schmidt, Vermeulen, & Van Der Molen, 2006).

The growing use of value-added models in teacher evaluation

The abundance of student achievement test score data that resulted from state and federal education policy, such as the No Child Left Behind Act of 2001, and the resulting data warehouses that linked those test scores back to students' teachers and schools allowed for the use of student test scores to evaluate the effectiveness of teachers through VAMs (McCaffrey, Lockwood, Koretz, & Hamilton, 2003; McGuinn, 2012; MET, 2012). VAMs are statistical models designed to provide a measure of teacher effectiveness by attempting to identify and isolate individual teachers' contributions to student learning by using students' prior achievement scores and other background variables to adjust for difference among the students taught by different teachers (Amrein-Beardsley, Collins, Polasky, & Sloat, 2012; Darling-Hammond, 2015; Hull, 2013; McCaffrey et al., 2003).

Proponents claim that using VAMs is an efficient and cost-effective method for identifying and sorting effective and ineffective teachers because these models statistically account for key factors, such as school and student characteristics, that also impact student learning (Chetty, Friedman, & Rockoff, 2014; Hanushek & Rivkin, 2010). Most importantly, VAMs can be used to identify teachers that could benefit from focused support or feedback to improve their knowledge and skills (Koedel, Mihaly, & Rockoff, 2015; Strunk, Weinstein, & Makkonnen, 2014). Despite these affordances of using

Downloaded by [RMIT University Library] at 02:04 08 December 2017

VAMs, researchers have suggested that there are methodological and inferential concerns when evaluating teacher effectiveness with this method. These concerns relate to the instability of value-added estimates, which can fluctuate from 'year to year, class to class, and test to test, as well as across statistical models' (Darling-Hammond, 2015, p. 133; Morgan, Hodge, Trepinksi, & Anderson, 2014). In addition, research has found that school characteristics and classroom composition can influence value-added scores significantly (Rothstein, 2009). However, other researchers report that the scope of bias for VAMs is relatively small (Kane & Staigher, 2008) or can be eliminated through the use of statistical modelling and quasi-experimental research designs (Bacher-Hicks, Kane, & Staiger, 2014; Chetty et al., 2014; Koedel et al., 2015).

VAMs have also had considerable use in research on teaching, teachers, and educational policy (c.f., Braun, 2005; Goldhaber & Hansen, 2008; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Polikoff & Porter, 2014). In particular, several authors have used value-added to study the relationship between teaching practices and student learning (c.f., Grossman, Loeb, Cohen, & Wyckoff, 2013; Lockwood, Savitsky, & McCaffrey, 2015; Ruzek, Domina, Conley, Duncan, & Karabenick, 2015). In this study, we expand upon this current research by examining the relationship between teachers' value-added scores and their use of subject-specific and generic instructional practices in secondary science classrooms.

Methods

Sample selection and description

The sample of video-recorded lessons included in this study represents a subset of lessons collected from over 2,500 teachers who participated in the Measures of Effective Teaching (MET) project (MET, 2013b). All teachers in the larger MET project volunteered to participate and were located in one of six predominantly large urban school districts across the United States. Research has suggested that the reliability of classroom observations can be increased by using three or more observations, using multiple raters, and including multiple years of data (MET, 2013a). Thus, we focused on the video-recorded lessons of ninth-grade biology teachers ($N = 237$) who submitted at least three video-recorded lessons over both years of the MET study ($N = 163$). We further restricted the sample of teachers to those who had at least 15 students with scores on the standardised science assessment (ACT biology test) each year, which was the student outcome measure used to calculate teachers' value-added scores.[1] This restriction reduced the final sample for this study to 99 ninth-grade biology teachers. We then randomly selected three to five lessons per teacher, yielding 475 videos of classroom practice across the 99 secondary biology teachers.[2] For the most part, different raters coded each lesson per teacher (although if the teacher had five lessons included, then two lessons were coded by the same rater). For each participating teacher, we developed a model[3] to derive one value-added score per teacher that used data from different class sections and multiple years, and drew upon the same outcome measure (ACT biology test) and controls (e.g. student demographics and previous test scores) available in the MET database.

Biology teachers who participated in the MET study worked in five different school districts in the United States. For the most part, the background characteristics of the

Downloaded by [RMIT University Library] at 02:04 08 December 2017

99 biology teachers participating in this study are similar to the MET biology teachers who were not included in this study due to missing data or limited participation across the two study years. For example, teachers in our study sample had similar patterns of post-graduate educational attainment and participation in professional development as the non-study teachers from the MET biology pool. Comparisons revealed only one key difference in the teacher population who participated in this study and the non-study teachers. MET biology teachers included in our study were significantly more likely to be White than non-study teachers, $z = 2.629$, $p = 0.004$, while Black or African-American teachers were significantly less likely to be included in our study sample, $z = -2.554$, $p = 0.005$. Table 1 provides information on the background characteristics of teachers who participated in this study, non-study MET biology teachers, and all MET biology teachers.

Just as sample teachers were more likely to be White, their students were also more likely to be White (29% as compared with 20% of the students in non-study teachers' classes), although this difference was not statistically significant, $z = -1.525$, $p = 0.064$. Non-study biology teachers were significantly more likely to have a larger proportion of students identified as Asian in their classes as compared with teachers in our study

Table 1. Teacher characteristics as a percentage of the sample.

	MET biology teachers		
Characteristic	In study ($n = 99$)	Not in study ($n = 138$)	All ($n = 237$)
School district			
A	29	[a]	30
B	12	[a]	7
C	11	[a]	13
D	22	[a]	12
E	25	[a]	39
Gender			
Male	33	27	30
Female	67	73	70
Race/Ethnicity			
White	72	46	63
Black/African-American	14	29	22
Hispanic	7	5	6
Other	6	11	9
Total teaching experience			
Less than 1 year	2	0	1
1–3 years	29	20	24
4–9 years	39	38	38
10–14 years	7	20	14
15 years or more	22	22	22
Within-district experience			
Less than 1 year	5	4	4
1–3 years	39	28	34
4–9 years	35	39	36
10–14 years	5	16	12
15 years or more	16	13	14
Advanced degrees			
Master's or higher	42	36	39
Professional development			
10+ hours in past 2 years	60	59	59
Median class size	28	26	26
Mean age of students	14.9	15.1	15.0

[a]Data were not calculated.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Table 2. Characteristics of teacher's students as a percentage of the sample.

	MET biology teachers		
Characteristic	In study ($n = 99$)	Not in study ($n = 138$)	All ($n = 237$)
Gender			
Male	47	51	49
Female	53	49	51
Race/Ethnicity			
White	29	20	24
Black/African-American	39	38	38
Hispanic	24	29	27
Asian	5	11	9
Other	3	2	2
Gifted	9	7	8
Special education	5	6	6
English language learner	7	6	6
Free/reduced-price lunch	53	55	54

sample, $z = 1.758$, $p = 0.039$. Similar proportions of students in participating study teachers' classes (39%) and non-study teachers' classes (38%) identified as Black. Table 2 illustrates the racial/ethnic identification and language learning status of students in the classes of sample teachers, non-study teachers, and all MET biology classes.

Data collection

Primary data collection involved eight raters scoring 475 videos of the secondary biology teachers' lessons using the Inquiring into Science Instruction Observation Protocol (ISIOP). The ISIOP is based on a constructivist notion of teaching that assumes students need to be engaged intellectually in the learning process (Minner & DeLisi, 2010). This observational tool was designed to provide a compilation of quantitative scores representing various features of a science teacher's instructional practice including: (1) the nature of teachers' verbal practices, (2) the nature of classroom activities, (3) the teacher's classroom instructional leadership, (4) the science content addressed, and (5) the kinds of scientific-investigation experiences the students engaged in. The study reported here focused on analyzing the subject-specific and generic features of science teachers' instruction that occur at a broader, more holistic (macro) level across the course of a whole lesson[4]. These macro level features included two subject-specific aspects of teachers' instruction – the kinds of scientific investigation-related experiences (IRE) the students engaged in and the science content (SC) addressed – and one generic aspect – the teacher's classroom instructional leadership practices (ILP). Tables 3–5 provide a detailed summary of the items used by raters to assess the science teachers' instructional quality for each of these aspects.

In this study, raters used all three sets of codes to examine each lesson for how the teacher used these science-specific and generic instructional practices. In particular, four raters used a four-point scale to rate instructional time[5] spent on various SC and IRE during each lesson: 0 (none of the instructional time), 1 (1–10% of the instructional time), 2 (11–50% of the instructional time), or 3 (51–100% of the instructional time). In addition to providing scores on individual items, these raters provided an overall score for each of the nine IRE sections (e.g. student-directed

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Table 3. Overview of IRE codes (Minner & DeLisi, 2012, pp. 16–19).

	IRE category	IRE ID	Brief description of IRE
Student-directed activities	Questioning/Exploration	IRE 8	Students research what is already known from existing resources to generate ideas to investigate.
		IRE 9	Students generate investigation questions.
		IRE 10	Teacher helps students figure out what will make a good investigation question (i.e. testable, empirical).
		IRE 11	Students make their own predictions or formulate hypotheses as part of an investigation.
	Design	IRE 13	Students design ways to investigate research questions, including choosing appropriate variables, techniques, and tools to gather, record, and analyze data.
		IRE 14	Teacher discusses with students the role of variables and controls in investigation designs.
		IRE 15	Students identify treatment and control variables.
	Data Collection and Organisation	IRE 17	Students make descriptive observations.
		IRE 18	Students make accurate measurements using scientific tools and instruments.
		IRE 19	Students access and record secondary data (existing datasets or databases) using computers.
		IRE 20	Students devise and use their own organisational scheme for recording data.
	Analysis	IRE 22	Students use mathematics to transform, organise, or interpret data.
		IRE 23	Students use physical models or simulations to assist with the analysis and interpretation of data/evidence.
		IRE 24	Students assess the reliability and/or validity of the knowledge generated in an investigation by critiquing methodological flaws and how well procedures were followed.
	Conclusions/ Communication/ Evaluation	IRE 26	Students build logical arguments about the cause-and-effect relationships between variables.
		IRE 27	Students share investigation results and their own thinking/ conclusions/interpretations about the meaning of those results.
		IRE 28	Students plan and/or deliver a presentation of results to the class.
		IRE 29	Students evaluate and revise their explanations/predictions in light of alternative explanations posed by the teacher, other students' investigations or other sources of existing scientific knowledge.
Teacher-directed activities	Questioning/Exploration	IRE 31	Teacher tells the students the questions they will investigate.
	Design	IRE 33	Teacher provides the variables to investigate.
		IRE 34	Teacher provides the procedures to follow in the investigation.
	Data Collection and Organisation	IRE 36	Students record data on worksheets or in science notebooks with a format prescribed by the teacher.
		IRE 37	Teacher provides data for the students.
	Analysis and Conclusion	IRE 39	Teacher tells students the analysis procedures.
		IRE 40	Teacher provides data analysis for students.
		IRE 41	Teacher tells the students what to conclude from an investigation.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

questioning/exploration; teacher-directed data collection) and the four SC sections (e.g. life science domain; earth science domain) signifying the instructional time spent in these overall categories for each lesson. The other four raters scored each lesson on the ILP items using the following criteria described in the ISIOP: 0 = does not describe the lesson/not at all; 1 = slightly characteristic of the lesson; 2 = somewhat characteristic of the lesson; 3 = very characteristic of the lesson. In general, higher scores correspond with higher quality instruction, albeit for a few ILP items that were reverse coded for data analysis purposes.

Table 4. Overview of SC codes (Minner & DeLisi, 2012, p. 11).

SC category	SC ID	Brief description of SC
Life science	SC 88	The cell
	SC 89	Molecular basis of heredity
	SC 90	Biological evolution
	SC 91	The interdependence of organisms in ecosystems and energy follow in ecosystems
	SC 92	Matter, energy, and organisation in living systems
	SC 93	Biological governance of organism behaviour
Earth science	SC 95	Energy in the earth system
	SC 96	Geochemical cycles
	SC 97	The origin and evolution of the earth system
	SC 98	The origin and evolution of the universe
Physical science	SC 100	Structure of atoms, nuclear forces, and radioactive isotopes
	SC 101	Structure and properties of matter
	SC 102	The nature of chemical reactions
	SC 103	Motions and forces
	SC 104	Conservation of energy and the increase in disorder
	SC 105	Interactions of energy and matter
Science and technology	SC 107	Technological design cycle
	SC 108	Understanding about science and technology

Table 5. Overview of ILP codes (Minner & DeLisi, 2012, pp. 8–9).

ILP domains	ILP items
Teaching style (10 items)	44. Teacher had welcoming teaching style
	45. Teacher pushed student thinking forward
	53. Teacher used adequate wait time
	60. Teacher used formative assessment
	62. Teacher exhibited enthusiasm
	63. Teacher utilised student thoughts
	64. Teacher solicited prior ideas from students
	68. Teacher asked students to expand on ideas
	69. Students asked substantive questions
	70. Teacher exhibited openness to ideas
Support for self-directed learning (6 items)	52. Teacher encouraged students to work together
	54. Teacher encouraged students to respond to classmates
	57. Teacher monitored students' progress
	58. Teacher encouraged students to take responsibility
	61. Teacher facilitated self-pacing
	66. Students worked cooperatively
Lesson organisation (7 items)	43. Teacher facilitated learning-conducive environment
	46. Teacher stated learning goals
	47. Teacher provided activities overview
	48. Teacher stated lesson expectations
	49. Teacher situated lesson
	50. Teacher connected key science ideas
	56. Short transitions
Dealing with distractions (5 items)	51. 2+ students exhibited distracting behaviour*
	55. Students remained on task
	59. Interruptions derailed goals*
	65. Students asked irrelevant questions*
	67. Students were attentive

Note: Items marked with asterisks were reverse coded to create the overall score for that ILP domain.

These eight raters – four raters for IRE and SC and four raters for ILP – focused on scoring the 475 videos for these features of the biology teachers' instructional practice. A master coder led each group of four raters through an intensive two-week online training programme using the online resources created by the ISIOP developers (see isiop.edc.org). The goal was for each rater to learn how to use the ISIOP's coding system and how

Downloaded by [RMIT University Library] at 02:04 08 December 2017

to apply the rubrics and scoring methods reliably, which is a critical step for consistent and accurate scoring (Joe, Tocci, Holtzman, & Williams, 2013). All raters were required to obtain acceptable inter-rater reliabilities with the master-coded certification lessons before moving forward; these scores were based on reliability analyses conducted as part of the ISIOP field test (Minner & DeLisi, 2012). Lessons were assigned to individual raters using a balanced cross-design that accounted for potential sources of error (Shavelson & Webb, 1991). Throughout the data collection process, raters double-scored a random subset of these videos (about 15% of the videos) and the master coder engaged each rater group in weekly calibration exercises to ensure that they were maintaining a shared understanding of how to code for these measures that was aligned with the instruments' codebook. In total, each rater scored approximately 136 lessons during the 12 week scoring period, with most raters coding approximately 12 videos per week. Prior to data analysis, each lesson had a compilation of distinct scores representing teachers' use of both science-specific (IRE and SC) and generic (ILP) instructional practices within the lesson.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Data analysis

This study involves quantitative analysis to examine how frequently these secondary biology teachers used these instructional practices and to determine which instructional practices are predictive of students' science achievement. The first step in the analysis involved calculating quantitative scores for each of the instructional practice features using protocols and rating guidelines developed by Karelitz, Hirsch, DeLisi, and Minner (2010). We calculated each score first at the lesson/video level and then aggregated across all lessons to the teacher level. For this analysis, we calculated item-level scores for each of the IRE and SC items separately, as well as for the nine overall IRE sections and the overall life SC area. In addition, for the IRE coding, we calculated lesson/video and teacher level scores for the student-directed IRE, the teacher-directed IRE, and the overall IRE by averaging across the IRE section scores. For the ILP analysis, we calculated item-level scores for each of the ILP items separately, as well as calculated four overall ILP domain scores (averaged across the item-level scores within that category) at the lesson/video and teacher levels. In the next step, we used descriptive statistical analyses to examine the distribution of the scores with a focus on: (1) the kinds of IRE the students engaged in (3 scores – teacher-directed activities; student-directed activities; overall), (2) the SC addressed (1 score – overall alignment to life science standards), and (3) the teacher's classroom ILP (4 scores – teaching style; support for self-directed learning; lesson organisation; dealing with distractions). This part of the analysis targeted the first research question by providing empirical evidence on the extent to which these secondary science teachers exhibited these subject-specific and generic features of high-quality science teaching in their practice.

In the final step, we examined the relationship between the subject-specific measures and the generic measures by calculating the correlations among the different instructional features within each group (e.g. correlations between the overall IRE, student-directed IRE, and teacher-directed IRE scales) and across groups (e.g. correlations between the IRE and ILP scales). We also used regression to determine which variables significantly predicted teachers' VAM, which directly addressed the second research question regarding

which instructional practices are predictive of students' achievement. For this analysis, we used a separate regression model – one for each different instructional feature (e.g. student-directed IRE; SC; dealing with distractions) as the predictor – to answer the second research question. Each regression model followed a similar structure:

$$\text{VAM} = B_0 + B_1{*}\text{Predictor} + B_2{*}\%\text{Free Lunch} + B_3{*}\text{Last Year ELA} + B_4{*}\text{Last Year Math}$$

In each model, we used students' average prior year achievement scores[6] and their average free/reduced lunch status as covariates, and teachers' value-added score as the outcome measure.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Results

Science teachers' use of subject-specific instructional practices

Investigation-related experiences

Overall, findings show that the distribution of scores was quite limited and condensed near one end of the scale for a majority of the items focused on students' opportunities to engage in IRE. In general, items in the IRE scale for both student-directed activities and teacher-directed activities were positively skewed, indicating that scores clustered at the lower end of the scale. Comparatively, the skewness is more evident for student-directed activities, indicating these experiences were less likely to occur within these teachers' classrooms. Tables 6 and 7 provide the descriptive statistics for the items across the student- and teacher-directed scales.

These findings suggest that, on average, these secondary biology teachers provided their students with limited opportunities to engage in IRE. When they did so, it was typically the teacher who provided the investigation question, the procedures, and the format for recording the collected data, although students were usually the ones making descriptive observations to address the teacher-provided investigation question. It was rare for students to generate their own investigation questions or engage in extended analysis or scientific sense-making to generate conclusions based on investigation results. Figure 1 shows the distribution of investigation-related overall scores across the 475 videos in the study. Note that Table 3 provides a brief description of the specific IRE comprising each of these larger categories.

Science content

Most biology lessons in this study showed a strong emphasis on SC (refer to Figure 2 below). Approximately 75% of the videos received a rating of three – the highest emphasis score – for the extent to which the lesson activities focused on key scientific concepts and engaging students in thinking about the science in the lesson. Across this set of lessons teachers were most likely to focus their lesson activities on helping to develop students' understanding about the cell (e.g. structure, functions, cellular growth and differentiation) and the molecular basis of heredity (e.g. role of chromosomes in reproduction, mutations). Overall findings show that these biology teachers frequently taught science lessons aligned with the content standards.

Table 6. Descriptives for video level scores for IRE student-directed activities ($n = 475$).

Student-directed activities	IRE item	Min	Max	Mean	Std. Dev.
Questioning/Exploration	IRE 8	0.00	1.00	0.01	0.09
	IRE 9	0.00	1.00	0.01	0.11
	IRE 10	0.00	1.00	0.01	0.11
	IRE 11	0.00	1.00	0.08	0.27
	IRE 12*	0.00	2.00	0.10	0.31
Design	IRE 13	0.00	3.00	0.07	0.30
	IRE 14	0.00	1.00	0.09	0.27
	IRE 15	0.00	2.00	0.03	0.17
	IRE 16*	0.00	3.00	0.14	0.40
Data collection and organisation	IRE 17	0.00	3.00	0.64	0.86
	IRE 18	0.00	2.00	0.07	0.28
	IRE 19	0.00	2.50	0.01	0.15
	IRE 20	0.00	2.00	0.02	0.16
	IRE 21*	0.00	3.00	0.64	0.87
Data analysis	IRE 22	0.00	3.00	0.11	0.38
	IRE 23	0.00	3.00	0.05	0.29
	IRE 24	0.00	1.00	0.03	0.17
	IRE 25*	0.00	3.00	0.18	0.48
Conclusions/Evaluation	IRE 26	0.00	3.00	0.18	0.41
	IRE 27	0.00	2.00	0.07	0.26
	IRE 28	0.00	3.00	0.02	0.19
	IRE 29	0.00	2.00	0.01	0.12
	IRE 30*	0.00	3.00	0.23	0.48

Note: Overall ratings are marked with an asterisk.

Science teachers' use of generic instructional practices

When examining the teachers' scores across the four ILP domains (refer to Figure 3 below), there is a much wider distribution across the first three ILP domains in comparison to the last ILP domain. On average, teachers in this study showed strong classroom management skills, but had more variability in their teaching style, support for self-directed learning, and lesson organisation. In general, findings show that these biology teachers were more likely to: project a welcoming teaching style; actively monitor students' progress and encourage self-pacing; facilitate a learning-conducive physical environment; have short transitions between activities; and have students who remained on task and attentive. However, these biology teachers were less likely to: use wait time when eliciting students' ideas; solicit their students' background knowledge; encourage students to respond to one another's ideas; allow students to make decisions about what to do during the lesson; and situate the lesson within the context of previous lessons. The

Table 7. Descriptives for video level scores for IRE teacher-directed activities ($n = 475$).

Teacher-directed activities	IRE item	Min	Max	Mean	Std. Dev.
Questioning/Exploration	IRE 31	0.00	1.00	0.43	0.48
	IRE 32*	0.00	2.00	0.43	0.49
Design	IRE 33	0.00	3.00	0.33	0.57
	IRE 34	0.00	3.00	0.54	0.78
	IRE 35*	0.00	3.00	0.60	0.80
Data collection and organisation	IRE 36	0.00	3.00	0.53	0.80
	IRE 37	0.00	3.00	0.07	0.32
	IRE 38*	0.00	3.00	0.56	0.82
Analysis and conclusions	IRE 39	0.00	3.00	0.09	0.31
	IRE 40	0.00	1.00	0.01	0.10
	IRE 41	0.00	2.00	0.05	0.25
	IRE 42*	0.00	3.00	0.13	0.38

Note: Overall ratings are marked with an asterisk.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

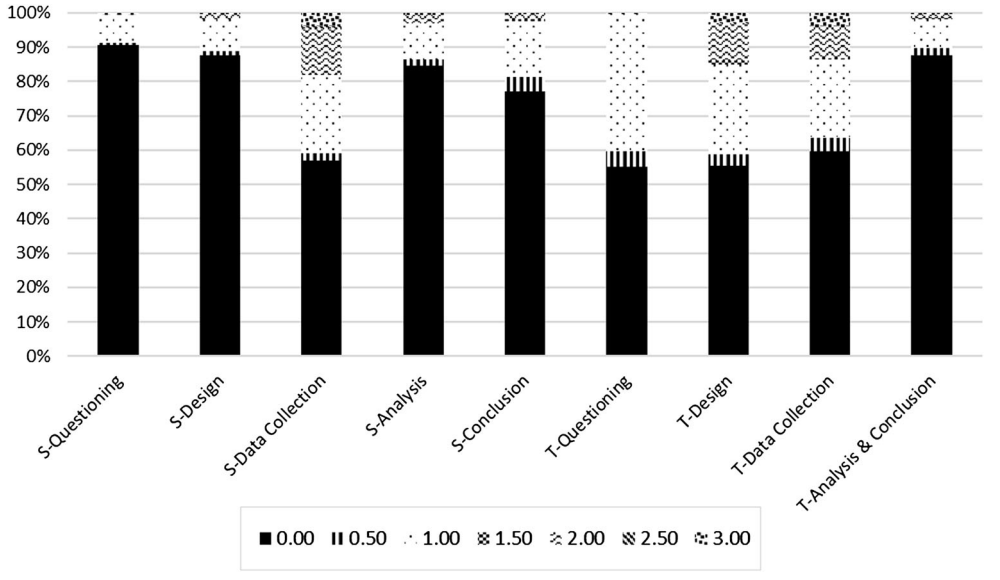

Figure 1. Distribution of video level scores for IRE overall ratings.

teaching practices that focused on more challenging aspects of instruction, such as engaging students in interactive discourse, were observed less frequently in this data set.

Relationships between measures of science teaching effectiveness

Investigation-related experiences

First, we examined the relationship between these subject-specific measures by calculating the correlations among the overall IRE, student-directed IRE, and teacher-directed IRE

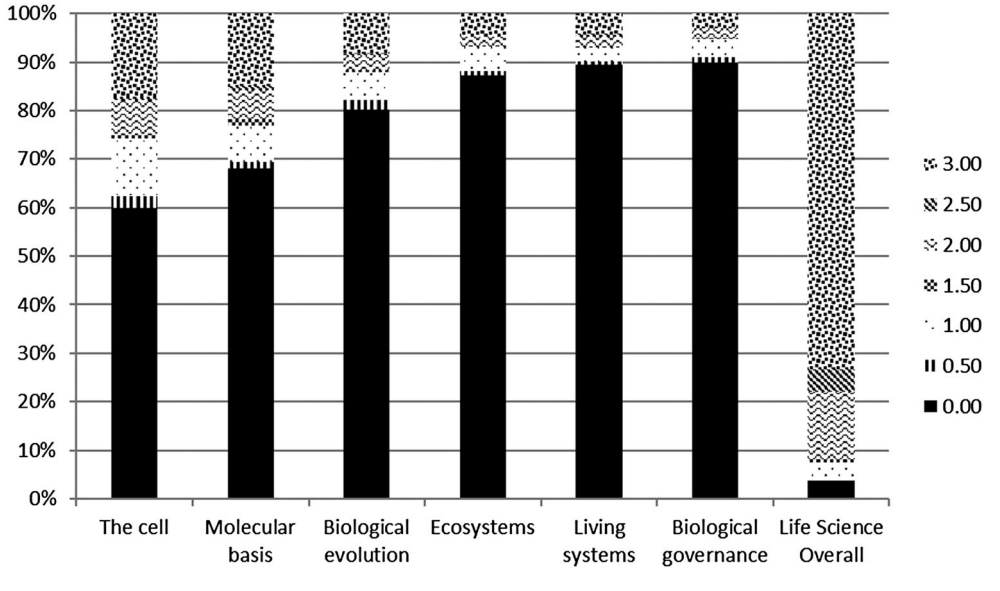

Figure 2. Distribution of video level scores for SC ratings in life science.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

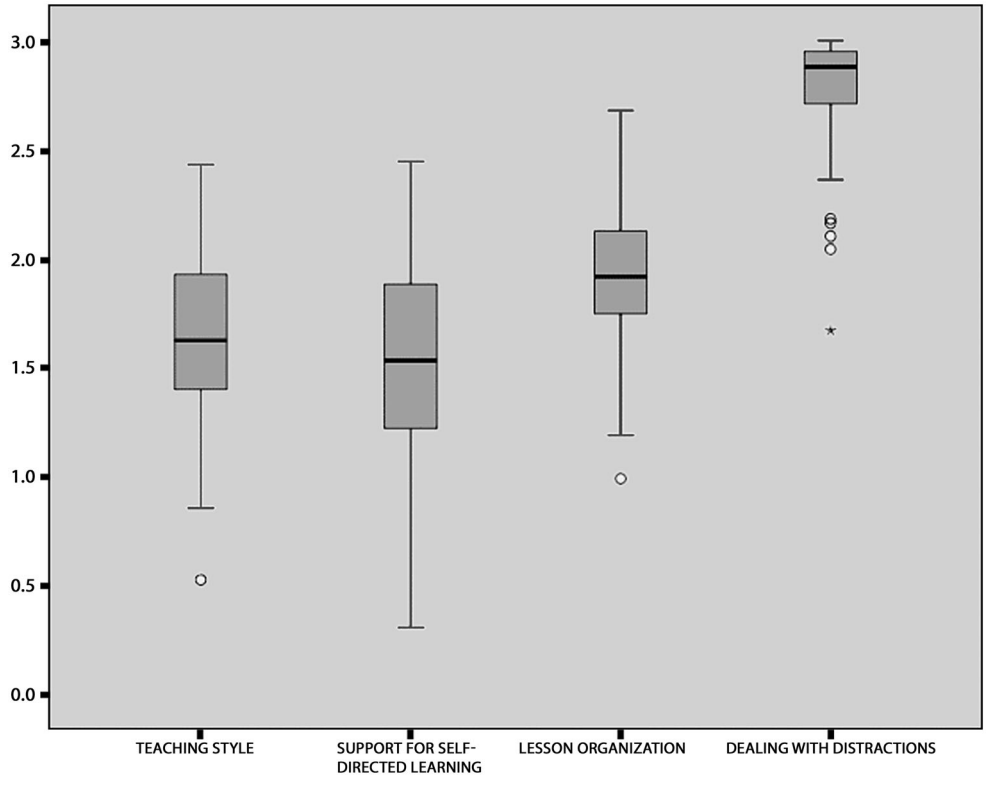

Figure 3. Distribution of teachers' scores in the four ILP domains.

scales. Table 8 shows the correlations between these three scales. Findings showed that the student-directed IRE and teacher-directed IRE scales are moderately correlated, suggesting that teachers who engage their students in scientific-investigation are likely to do so by using a combination of both student and teacher-directed IRE. Across many lessons, raters tended to code for both types of IRE in the same lesson – typically for teachers providing the investigation question, procedures, and variables (three teacher-directed IRE) in combination with students making descriptive observations (one student-directed IRE).

Second, we used regression to test if there are any significant relationships between teachers' IRE scores and value-added scores. In the model, we used the average prior year achievement scores and the average free/reduced lunch status for the teachers' students as covariates. In terms of using these instructional variables (student-directed IRE, teacher-directed IRE, and SC) to predict teachers' VAM, the most noteworthy finding

Table 8. IRE overall scale correlations.

	Student-directed IRE	Teacher-directed IRE	Overall IRE
Student-directed IRE	1		
Teacher-directed IRE	.687**	1	
Overall IRE	.905**	.931**	1

**Correlation is significant at the 0.01 level (2-tailed).

Downloaded by [RMIT University Library] at 02:04 08 December 2017

is that only one subject-specific instructional feature – student-directed IRE – significantly and positively predicts VAM [($R^2 = .155$, $F(4, 71) = 3.250$, $p < .05$]. Teachers who are more likely to provide opportunities for their students to take the lead in implementing scientific-investigations are more likely to have higher VAM.

Instructional leadership practices

To examine the relationship between the ILP measures, we began by calculating correlations among the ILP domain scales. Table 9 highlights the correlations across the ILP domain scales. Findings showed that the four ILP scales are positively, but weakly, related, with correlations ranging from a minimum of .212 (between ILP domains two and four) to a maximum of .445 (between ILP domains one and three). These findings suggest that these four ILP domains are likely measuring distinct aspects of these science teachers' instructional practices.

We then used regression to test if there were any significant relationships between teachers' ILP domain scores and their value-added scores. Similar to the above regression model, we used students' average prior year achievement scores and their average free/reduced lunch status as covariates. In terms of using the ILP instructional variables (teaching style; self-directed learning; lesson organisation; and dealing with distractions) to predict teachers' VAM, only one of the ILP domains – 'Dealing with Distractions' – significantly predicted VAM [($R^2 = .159$, $F(4, 71) = 3.344$, $p < .05$].

Across observational measures

We also examined the correlations between the science-specific IRE measures and the generic ILP observational measures. As shown in Table 10, the main pattern is that IRE student- and teacher-directed activities show small correlations with these generic instructional practices. The most noteworthy relationship is between the ILP support for self-directed learning domain and teachers' use of student-directed IRE; teachers who are

Table 9. ILP domain score correlations.

ILP domains	1	2	3	4
1. Teaching style	1			
2. Self-directed learning	.444**	1		
3. Lesson organisation	.445**	.369**	1	
4. Dealing with distractions	.215*	.212*	.280**	1

*Correlation is significant at the 0.05 level (2-tailed).
**Correlation is significant at the 0.01 level (2-tailed).

Table 10. Correlations between science-specific (IRE) and generic (ILP) observational measures.

	IRE student-directed activities	IRE teacher-directed activities
Teaching style	.301**	.136
Self-directed learning	.496**	.347**
Lesson organisation	.204*	.226*
Dealing with distractions	.165	.135
IRE student-directed activities	1	.687**
IRE teacher-directed activities	.687**	1

*Correlation is significant at the 0.05 level (2-tailed).
**Correlation is significant at the 0.01 level (2-tailed).

Downloaded by [RMIT University Library] at 02:04 08 December 2017

more likely to provide opportunities for students to direct their own scientific inquiry are also more likely to provide support for self- directed learning. This finding is not surprising given that student-directed activities provide opportunities for students to make decisions about their own learning. Although some of the other correlations are significant, all of them are small in nature, suggesting that these generic and subject-specific teaching practices were not always present in combination in these teachers' classrooms.

Discussion

This study contributes to a growing body of research that is moving beyond the use of surveys to measure the relationship between teaching and learning components. In particular, this study examined the ways in which science teachers engage their students in the learning process in order to understand the mechanisms that are most relevant for promoting student learning. To date most studies of teaching effectiveness have examined a small set of teaching practices. Moreover, studies seldom examine both generic and subject-specific practices simultaneously within lessons. This study examined how the use of multiple teaching practices – both generic and science-specific teaching practices – related to student learning, as measured by VAM. Following one of the recommendations in a recent meta-analysis on studies of teaching effectiveness (Seidel & Shavelson, 2007), this study investigated a 'more comprehensive set of teaching components relevant to the complete cycle of learning' (p. 484), a gap that was missing in the current research literature, especially in the content areas.

Major contributions of this study stem from the investigation of a comprehensive set of instructional practices that focus on the interactional nature of teaching and the ways in which science teachers engage their students in the learning process. Results from this study have implications for both teacher development and teacher evaluations. In particular, findings highlight concerns related to the limited evidence of high-quality science teaching instruction, the limited evidence of alignment between measures of science teaching quality, and the prominent attention to mainly generic observational measures in teacher evaluation systems.

Limited evidence of high quality science teaching instruction

Despite the literature suggesting that it is important for students to be cognitively engaged in inquiry-based science instruction (Chen & She, 2015; Cobern et al., 2010; Fogleman et al., 2011; Hmelo-Silver et al., 2007; Marshall & Horton, 2011; McNeill & Krajcik, 2008; Minner et al., 2010), this study found that IRE tended to be scored at the lower end of the scale, suggesting that these secondary biology teachers were conducting minimal inquiry-based instruction. We also found that the skew is more evident for student-directed activities than for teacher-directed activities, indicating that student-directed experiences were more limited. This is in line with the literature suggesting that teachers are hesitant to conduct inquiry in the classroom (Brand & Moore, 2011; Marshall & Smart, 2013), have difficulty managing inquiry-based science (Capps & Crawford, 2013; Harris & Rooks, 2010), and may be more comfortable conducting teacher-directed inquiry than student-directed inquiry (Ozel & Luft, 2013). However, our results showing that scores for student-directed IRE were positively related with teachers' value-added

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Downloaded by [RMIT University Library] at 02:04 08 December 2017

scores is consistent with the literature suggesting that student-directed inquiry instruction is critical to effective science education (Fogleman et al., 2011).

These findings add to the empirical research confirming what science teacher educators have long suggested – science teachers struggle to engage in more ambitious teaching practices whereby they provide opportunities for students to engage in substantial sense-making, critique, and explanation of scientific data and concepts. While research supports the effectiveness of a student-centered teaching style that promotes social and cognitive engagement, a variety of challenges contribute to the relative scarcity of high-quality implementation in practice (Furtak & Kunter, 2012; Herman, Osmundson, & Silver, 2010; Kock, Taconis, Bolhuis, & Gravemeijer, 2013; Kolodner et al., 2003; Smart & Marshall, 2013; Wu & Huang, 2007). Teacher questioning overall tends to focus on recall of facts and procedures – not the high-level questions that promote cognitive engagement (Smart & Marshall, 2013). In addition, Herman et al. (2010) found that, while teachers initiated formative assessment by using tools to gauge student understanding, they struggled to analyze and interpret student responses and rarely formulated next steps to address gaps in student understanding. Similarly, teachers and students alike felt overwhelmed and 'overtaxed' by shifting from more traditional, teacher-centered instruction to a minimally-structured inquiry-based model (Furtak & Kunter, 2012, p. 309; Kolodner et al., 2003). In terms of student-teacher interactions, there appears to be a 'sweet spot' of balancing cognitive autonomy support with structure to help students learn. In several studies, instruction intended to be cognitively engaging and autonomy-supportive lacked sufficient structure and routines to benefit students (Furtak & Kunter, 2012; Kock et al., 2013). Taken together, the literature suggests that enacting a student-centered teaching style is a very important – but also very difficult – thing to do. Although this study cannot speak to the reasons why these more ambitious teaching practices are limited across this data set, it does provide empirical support to the argument that professional development opportunities to increase science teachers' abilities to engage students intellectually and socially in the learning process are imperative. In addition, this research study suggests that a fruitful next step would be to recruit a purposeful sample of teachers who primarily engage in student-directed scientific inquiry and examine the instructional practices they use and how those practices relate to their students' achievement.

Limited alignment between measures of science teaching quality

In the last decade, VAM have been gaining traction as an efficient method for identifying effective and ineffective teachers (Hull, 2013; McCaffrey et al., 2003). However, the empirical evidence presented in this study suggests that VAM and subject-specific and generic observational measures of science teaching instruction show limited relationships. Only one science-specific teaching practice – the use of student-directed IRE – and one generic feature – well-managed classroom environments – were significant and positive predictors for teachers' VAM. As a part of multiple-measure teacher evaluation systems, VAM have been studied in their relationship with other measures of teacher effectiveness for two major reasons. One reason is that researchers are interested in examining the validity of one measure given its relationship with other measures (e.g. Bell et al., 2012; Hill, Kapitula, & Umland, 2010). The other reason is that practitioners need to know

if the different measures send consistent signals to teachers and principals in schools about teacher effectiveness (Grossman et al., 2013; Strunk et al., 2014).

In general, researchers reported low to moderate positive correlations between VAM and various classroom observational measures, both in research contexts and practical settings (Kane & Staiger, 2012). One common explanation for such limited and weak relationships is that each measure provides unique information about the teacher's effectiveness, meaning that they provide different information to administrators and teachers about teachers' instructional practice, strengths and areas for growth. Polikoff and Porter (2014) also found weak associations between teachers' instructional alignment and VAM and speculated that this may suggest that state tests are not up to the task to differentiate effective teaching from ineffective teaching. This may indicate that certain components of an observation protocol better capture teachers' abilities to improve student achievement, while others may not be as helpful or sensitive in raising test scores. Taken together, these findings suggest that it is wise to be critical of using VAM to make high-stakes personnel decisions. Instead, VAM could be one of the measures used formatively to provide information about teachers' effectiveness and identify teachers who could benefit from additional support or feedback to improve their knowledge and skills in particular areas (Koedel et al., 2015; Strunk et al., 2014).

Connection to teacher evaluations

Across the nation, states and school districts have been tasked with creating valid and reliable teacher evaluation systems that can be used for two key purposes: (1) to provide critical feedback for improving teachers' instructional practice and (2) to make personnel decisions about teacher retention, promotion, and placement (Darling-Hammond, 2012; McGuinn, 2012; National Council on Teacher Quality, 2012). Observational measures have been widely-used within these systems and the majority of districts have opted to use generic observational tools for assessing teachers' practice (Cohen & Goldhaber, 2016). However, generic observational protocols are designed to measure domains, or components, of teachers' instruction that are not aligned with the subject area. For example, the Classroom Assessment Scoring System (CLASS), a well-known and widely-used generic observation protocol, assesses teachers' effectiveness in terms of their support for students' emotional and academic development. These generic measures potentially limit what one can learn about how teachers engage in the content-intensive work of teaching students specific content and discipline-based practices.

More recently, researchers and teacher educators have argued persuasively for the importance of attending to the discipline-specific aspects of teachers' instructional practice. Greater efforts to develop subject-specific observational tools, such as the Protocol for Language Arts Teaching Observation (PLATO), the Mathematical Quality of Instruction (MQI), and the Quality Science Teaching (QSI) protocols from the MET project (MET, 2010), over the last decade signal the increasing prominence of assessing key characteristics of subject matter teaching. Results from this study reveal that the secondary science teachers who had well-managed classroom environments and who provided opportunities for their students to engage in student-directed IRE were more likely to have increased student outcomes, as determined by teachers' value-added measures. These findings suggest that attending to both subject-specific and generic aspects of

Downloaded by [RMIT University Library] at 02:04 08 December 2017

science teachers' instructional practice is important for understanding the underlying mechanisms that result in more effective science instruction in secondary classrooms.

Limitations

This study has a few important limitations that should be taken into consideration. First, the sample of teachers participating in the overall MET study were volunteers from various large urban school districts. This sample of convenience means that the sample of ninth-grade biology teachers participating in the MET study, and thus the sample of teachers in this study, are not a random sample of all ninth-grade biology teachers across the United States. Additionally, we found that many of the participating biology teachers either did not participate across both years of the MET study and/or failed to submit at least three video-recorded lesson samples. The implication was that our study had a reduced sample size than originally anticipated (99 of the 237 biology teachers who participated in the overall MET study). It is unknown to what extent there may have been some systematic difference in the biology teachers who participated in this study and secondary biology teachers nationwide.

Second, as the data indicated, there was limited variability in some features of these secondary science teachers' classroom instruction, which may have limited the relationships noted across measures. Since this secondary analysis was based on data collected from the MET study, we cannot account for this limitation in this specific study. However, one implication for future research is to be strategic in recruiting participants that may be more likely to engage in ambitious teaching practices to examine the relationship between these different teaching quality measures. Any attempts to reduce this potential selectivity bias in the sample would be important to consider in order to expand the field's understanding of how science teachers' instructional practices, especially the subject-specific ones, relate to student outcomes.

Another limitation is related to the outcome measure used in this study. Although VAM have been touted as statistical models that account for key factors, such as school and student characteristics, that also impact student learning (Chetty et al., 2014; Hanushek & Rivkin, 2010), critics suggest that there are significant methodological and inferential concerns when evaluating teacher effectiveness with this method. For example, decisions regarding the choices of parameters such as tests, years of data included, and modelling specifications has been shown to influence value-added scores (Koedel et al., 2015).

In this study, we were limited to the data collected on the MET research project, which included the use of only one student outcome in biology (the ACT biology test), up to two years of data, and a particular VAM used on the project. It is unclear to what extent the decisions regarding these parameters have influenced the findings. For example, some research has suggested that there are differences in the sensitivity of various measures and it is unknown to what extent the ACT biology test was sensitive to changes in students' achievement. In addition, there are various ways in which teaching can effect student outcomes; in this study we focused exclusively on the effect on students' cognitive growth. However, other research has suggested that teaching can also effect students' motivation and learning processes (Seidel & Shavelson, 2007). The limited significant relationships found in this study may be related to the exclusive focus on students' cognitive growth as an outcome.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

A final limitation of this study is related to the use of an observational protocol that required scorers to rate the frequency by which teachers employed a wide variety of instructional practices during one lesson period. As noted in the methods section, the ISIOP was designed to collect data on different aspects of science teachers' instructional practices. As many of these practices were noted as hallmarks of high-quality instruction, the ISIOP developers designed the protocol to measure how frequently each practice was used within the lesson. However, it may not be feasible or practical for a teacher to score on the higher end of the rating scale for each instructional practice within one lesson, In other words, it is likely unrealistic to expect teachers to address every aspect included on an observation protocol during one observed lesson. While we tried to address this concern by including multiple observations per teacher, it may be that certain lessons lend themselves to the use of particular instructional practices more readily (Grossman, Cohen, & Brown, 2014; Mikeska, Holtzman, McCaffrey, & Shattuck, 2017) and there may be potential biases associated with collecting data based on the time frequency of these instructional practices at the lesson level.

Conclusion

By examining features of teachers' instruction and their relationship to VAM, we can better understand which practices have more and less potential to improve or hinder student learning. In this study, findings suggest that secondary science teachers who had well-managed classroom environments and who provided opportunities for their students to engage in student-directed IRE were more likely to have increased student outcomes, as determined by teachers' value-added measures. These findings are an important contribution to the field as they suggest that attending to both subject-specific and generic features of science teachers' instructional practice is critical. The current environment of teacher evaluation tends to privilege the more generic features of teachers' instruction, with limited attention to the content-specific aspects of teaching in the disciplines. This study suggests that continuing to do so would be missing key elements of instruction that are directly linked to student outcomes. Moreover, despite the fact that many of these subject-specific and generic instructional features are noted as hallmarks of high-quality science teaching, the majority of these observational measures did not exhibit a strong relationship to VAM. Also, limited variation in practice was captured supporting other findings that ambitious instruction is challenging for science teachers to enact (Furtak & Kunter, 2012; Kolodner et al., 2003; Reinsvold & Cochran, 2012). Further research is warranted to better understand why there is limited empirical support for the assumption of alignment between measures in teacher evaluation systems and to better understand the influence of other factors, such as the affordances and limitations associated with the outcome measures used, when examining the relationship between these different teaching quality measures in science.

Notes

1. Project researchers decided on a minimum threshold of 15 student test scores per year because, as found by McCaffrey, Sass, Lockwood, and Mihaly (2009), the standard errors of estimated value added decline very little with increasing sample size beyond 15 students.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

2. Due to the random selection of videos within teachers, our sample includes three biology teachers for whom no videos were observed in Year 1 (i.e. only videos submitted in Year 2 were selected).
3. Section-level value-added scores were computed as part of the original MET study and were retrieved from archived MET datasets. The procedures used to compute these scores can be found in 'Have we identified effective teachers? Validating measures of effective teaching using random assignment' (MET, 2013b). Teacher-level value-added scores for each teacher in our sample were computed from the section-level value-added scores using the following formula:

$$V_{\text{total}} = \frac{((V_{11}/SE_{11}^2) + (V_{12}/SE_{12}^2) + (V_{21}/SE_{21}^2))}{((1/SE_{11}^2) + (1/SE_{12}^2) + (1/SE_{21}^2))},$$

where V_{tj} = VA estimate from section j for a teacher in year t and SE_{tj} = corresponding standard error for the VA estimate. The ISIOP developers designed this instrument to associate a specific time frequency for rating the use of these instructional practices (e.g., student-directed investigation related experiences). The use of these time frequencies supported raters in consistently rating the extent to which these particular instructional practices are present within a specific lesson.
4. These three components included the most reliable measures in our analyses. Of the remaining two components of the ISIOP, the first component (nature of classroom activities) is largely descriptive. The second component (nature of teachers' verbal practices) had significant difficulty with its measurement properties.
5. The use of the four-point rating scale was designed to allow raters to provide a measure of how often teachers employed these particular practices within one lesson.
6. In this study, we used students' prior achievement in the regression model as controls because we wanted to be able to sort out the potential of different practices used with students of different backgrounds. Although we would have preferred to include students' prior achievement in science as a control in the regression model, we had to use their prior achievement in math and ELA only due to extensive missing data in students' prior science scores in the MET database.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Acknowledgements

The authors would like to express their appreciation for the raters on this project who devoted numerous hours to carefully observing and analyzing video-recorded lessons from these secondary science teachers' classrooms.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Jamie N. Mikeska http://orcid.org/0000-0002-8831-2572

References

Ainley, M., & Patrick, L. (2006). Measuring self-regulated learning processes through tracking patterns of student interaction with achievement activities. *Educational Psychology Review*, *18*(3), 267–286.

Alozie, N., & Mitchell, C. (2014). Getting students talking: Supporting classroom discussion practices in inquiry-based science in real-time teaching. *The American Biology Teacher*, *76*(8), 501–506.

Amrein-Beardsley, A., Collins, C., Polasky, S. A., & Sloat, E. F. (2012). Value-added model (VAM) research for educational policy: Framing the issue. *Education Policy Analysis Archives*, *21*(4). doi:10.14507/epaa.v21n4.2013

Bacher-Hicks, A., Kane, T.J., & Staiger, D.O. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles* (NBER Working Paper No. 20657).

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62–87.

Berry, A., Friedrichsen, P., & Loughran, J. (2015). *Re-examining pedagogical content knowledge in science education*. New York, NY: Routledge.

Bleicher, R. E., Tobin, K. G., & McRobbie, C. J. (2003). Opportunities to talk science in a high school chemistry classroom. *Research in Science Education*, *33*(3), 319–339.

Blömeke, S. (2014). *International perspectives on teacher knowledge, beliefs and opportunities to learn: TEDS-M results*. Dordrecht: Springer.

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

Brand, B. R., & Moore, S. J. (2011). Enhancing teachers' application of inquiry-based strategies using a constructivist sociocultural professional development model. *International Journal of Science Education*, *33*(1), 889–913.

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Policy Information Center: ETS.

Brophy, J. (2000). *Teaching. Educational practice series 1*. Brussel: International Academy of Education. Retrieved from http://www.ibe.uneso.org

Capps, D. K., & Crawford, B. A. (2013). Inquiry-based instruction and teaching about nature of science: Are they happening? *Journal of Science Teacher Education*, *24*(3), 497–526.

Chen, C. T., & She, H. C. (2015). The effectiveness of scientific inquiry with/without integration of scientific reasoning. *International Journal of Science and Mathematics Education*, *13*(1), 1–20.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632.

Chin, C. (2006). Classroom interaction in science: Teacher questioning and feedback to students' responses. *International Journal of Science Education*, *28*(11), 1315–1346.

Chin, C. (2007). Teacher questioning in science classrooms: Approaches that stimulate productive thinking. *Journal of Research in Science Teaching*, *44*(6), 815–843.

Chism, B., & Pang, V. O. (2014). Transforming education and supporting equity through opportunity to learn standards. *National Forum of Applied Educational Research Journal*, *27*(1/2), 19–30.

Christodoulou, A., & Osborne, J. (2014). The science classroom as a site of epistemic talk: A case study of a teacher's attempts to teach science based on argument. *Journal of Research in Science Teaching*, *51*(10), 1275–1300.

Cobern, W. W., Schuster, D., Adams, B., Applegate, B., Skjold, B., Undreiu, A., & Gobert, J. D. (2010). Experimental comparison of inquiry and direct instruction in science. *Research in Science & Technological Education*, *28*, 81–96.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*(6), 378–387.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*(2), 119–142.

Conderman, G., & Bresnahan, V. (2008). Teaching big ideas in diverse middle school classrooms. *Kappa Delta Pi Record*, *44*(4), 176–180.

Darling-Hammond, L. (2012). *Creating a comprehensive system for evaluating and supporting effective teaching*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L. (2015). Can value added add value to teacher evaluation? *Educational Researcher*, *44*(2), 132–137.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Davis, E. A. (2000). Scaffolding students' knowledge integration: Prompts for reflection in KIE. *International Journal of Science Education*, *22*(8), 819–837.
Davis, E. A., Petish, D., & Smithey, J. (2006). Challenges new science teachers face. *Review of Educational Research*, *76*(4), 607–651.
Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*, 181–199.
Fogleman, J., McNeill, K. L., & Krajcik, J. (2011). Examining the effect of teachers' adaptations of a middle school science inquiry-oriented curriculum unit on student learning. *Journal of Research in Science Teaching*, *48*(2), 149–169.
Furtak, E. M., & Kunter, M. (2012). Effects of autonomy-supportive teaching on student learning and motivation. *The Journal of Experimental Education*, *80*(3), 284–316.
Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, *82* (3), 300–329.
Geier, R., Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. *Journal of Research in Science Teaching*, *45*(8), 922–939.
Goldhaber, D., & Hansen, M. (2008). *Is this just a bad class? Assessing the stability of measured teacher performance* (Working paper #2008-5). Seattle: Center on Reinventing Public Education, University of Washington.
Good, T., & Brophy, J. (2000). *Looking in classrooms* (8th ed.). New York: Teachers College.
Gotwals, A. W., Cisterna, D., Lane, J., Kintz, T., & Ezzo, D. (in press). Distinguishing observable formative assessment practices: A synthesis of the literature. *Educational Assessment Journal* (forthcoming special issue).
Grossman, P., Cohen, J., & Brown, L. (2014). *Understanding instructional quality in English language arts: Variations in the relationship between PLATO and value-added by content and context. Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. San Francisco, CA: John Wiley & Sons.
Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education*, *119*(3), 445–470.
Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*, 267–271.
Harris, C. J., Phillips, R. S., & Penuel, W. R. (2012). Examining teachers' instructional moves aimed at developing students' ideas and questions in learner-centered science classrooms. *Journal of Science Teacher Education*, *23*(7), 769–788.
Harris, C. J., & Rooks, D. L. (2010). Managing inquiry-based science: Challenges in enacting complex science instruction in elementary and middle school classrooms. *Journal of Science Teacher Education*, *21*(2), 227–240.
Herman, J. L., Osmundson, E., & Silver, D. (2010). *Capturing quality in formative assessment practice: Measurement challenges* (CRESST Report 770). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Hill, H. C., Kapitula, L., & Umland, K. (2010). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 1–38.
Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, *42*(2), 99–107.
Hull, J. (2013). *Trends in teacher evaluation: How states are measuring teacher performance*. Alexandria, VA: Center for Public Education.
ibe, H. N. (2009). Metacognitive strategies on classroom participation and student achievement in senior secondary school science classrooms. *Science Education International*, *20*(1/2), 25–31.
Jiang, L., & Elen, J. (2011). Why do learning goals (not) work: A reexamination of the hypothesized effectiveness of learning goals based on students' behaviour and cognitive processes. *Educational Technology Research and Development*, *59*(4), 553–573.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Joe, J. N., Tocci, C. M., Holtzman, S. L., & Williams, J. C. (2013). *Foundations of observation: Considerations for developing a classroom observation system that helps districts achieve consistent and accurate scores*. Princeton, NJ: Educational Testing Service.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (research paper). Measures of Effective Teaching (MET) project, Bill and Melinda Gates Foundation.

Kane, T. J., & Staigher, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper No. 14607).

Kane, T., Taylor, E., Tyler, J., & Wooten, A. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*(3), 587–613.

Karelitz, T. M., Hirsch, L., DeLisi, J., & Minner, D. (2010). *ISIOP technical report: Initial psychometric testing results*. Waltham, MA: Education Development Center.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86.

Kloser, M. (2014). Identifying a core set of science teaching practices: A Delphi expert panel approach. *Journal of Research in Science Teaching*, *51*(9), 1185–1217.

Kock, Z. J., Taconis, R., Bolhuis, S., & Gravemeijer, K. (2013). Some key issues in creating inquiry-based instructional practices that aim at the understanding of simple electric circuits. *Research in Science Education*, *43*(2), 579–597.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195.

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., & Ryan, M. (2003). Problem-based learning meets case-based reasoning in the middle-school science classroom: Putting learning by design(tm) into practice. *Journal of the Learning Sciences*, *12*(4), 495–547.

Kyriakides, L., Christoforou, C., & Charalambous, Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, *36*, 143–152.

Lafontaine, D., Baye, A., Vieluf, S., & Monseur, C. (2015). Equity in opportunity-to-learn and achievement in reading: A secondary analysis of PISA 2009 data. *Studies in Educational Evaluation*, *471*(11), doi:10.1016/j.stueduc.2015.05.001

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of Bayesian exploratory factor analysis without restrictions. *The Annals of Applied Statistics*, *9*(3), 1484–1509.

Loyens, S. M., Magda, J., & Rikers, R. M. (2008). Self-directed learning in problem-based learning and its relationships with self-regulated learning. *Educational Psychology Review*, *20*(4), 411–427.

Malmberg, M., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology*, *102*(4), 916–932.

Marshall, J. C., & Alston, D. M. (2014). Effective, sustained inquiry-based instruction promotes higher science proficiency among all groups: A 5-year analysis. *Journal of Science Teacher Education*, *25*(7), 807–821.

Marshall, J. C., & Horton, R. (2011). The relationship of teacher-facilitated, inquiry-based instruction to student higher-order thinking. *School Science and Mathematics*, *111*(3), 93–101.

Marshall, J. C., & Smart, J. B. (2013). Teachers' transformation to inquiry-based instructional practice. *Creative Education*, *4*(2), 132–142.

Martin-Hansen, L. (2002). Defining inquiry. *The Science Teacher*, *69*, 34–37.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*(4), 572–606.

McCaslin, M., Bozack, A. R., Napoleon, L., Thomas, A., Vasquez, V., Wayman, V., & Zhang, J. (2006). Self-regulated learning and classroom management: Theory, research and considerations

Downloaded by [RMIT University Library] at 02:04 08 December 2017

for classroom practice. *Handbook of Classroom Management: Research, Practice, and Contemporary Issues*.
McGuinn, P. (2012). *The state of teacher evaluation reform: State education agency capacity and the implementation of new teacher-evaluation systems*. Washington, DC: Center for American Progress.
McNeill, K. L., & Krajcik, J. S. (2008). Scientific explanations: Characterizing and evaluating the effects of teachers' instructional practices on student learning. *Journal of Research in Science Teaching*, *45*(1), 53–78.
McNeill, K. L., & Pimentel, D. S. (2010). Scientific discourse in three urban classrooms: The role of the teacher in engaging high school students in argumentation. *Science Education*, *94*(2), 203–229.
MET. (2010). *The PLATO protocol for classroom observations*. Retrieved from http://k12education.gatesfoundation.org/wp- content/uploads/2015/12/PLATO_10_29_101.pdf
MET. (2013a). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Retrieved from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf
MET. (2013b). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/12/MET_Validating_Using_Random_Assignment_Research_Paper.pdf
MET Project. (2012). *Ensuring fair and reliable measures of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation.
Mikeska, J. N., Holtzman, S., McCaffrey, D., Liu, S., & Shattuck, T. (2017). *Using classroom observations to evaluate science teaching effectiveness: Implications of lesson sampling across lab and non-lab lessons*. Manuscript submitted for publication.
Minner, D., & DeLisi, J. (2010). *ISIOP technical report: Conceptual framework*. Waltham, MA: Education Development Center.
Minner, D., & DeLisi, J. (2012). *ISIOP user manual*. Waltham, MA: Education Development Center.
Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction-what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching*, *47*(4), 474–496.
Morgan, G. B., Hodge, K. J., Trepinksi, T. M., & Anderson, L. W. (2014). The stability of teacher performance and effectiveness: Implications for policies concerning teacher evaluation. *Education Policy Analysis Archives*, *22*(95), 00–00.
National Council on Teacher Quality. (2012). *State of the states 2012: Teacher effectiveness policies*. Washington, DC. Retrieved from https://www.nctq.org/dmsView/State_of_the_States_2012_Teacher_Effectiveness_Policies_NCTQ_Report
National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.
National Research Council. (2000). *Inquiry and the national science standards*. Washington, DC: National Academy Press.
National Research Council. (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.
Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives*, *18*(23), 1–27.
NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.
Oliveira, A. W. (2010). Improving teacher questioning in science inquiry discussions through professional development. *Journal of Research in Science Teaching*, *47*(4), 422–453.
Opdenakker, M., & Van Damme, J. (2006). Teacher characteristics and teaching styles as effectiveness enhancing factors of classroom practice. *Teaching and Teacher Education*, *22*(1), 1–21.
Ozel, M., & Luft, J. (2013). Beginning secondary science teachers' conceptualization and enactment of inquiry-based instruction. *School Science and Mathematics*, *113*(6), 308–316.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Palincsar, A., Magnusson, S., Cutter, J., & Vincent, M. (2002, Jan/Feb). Supporting guided-inquiry instruction. *Teaching Exceptional Children*, *34*(3), 88–91.

Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist*, *36*(2), 89–101.

Park, S., & Oliver, S. (2009). The translation of teachers' understanding of gifted students into instructional strategies for teaching science. *Journal of Science Teacher Education*, *20*(4), 333–351.

Pintrich, P. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). San Diego, CA: Academic Press.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis*, *36*(4), 399–416.

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal*, *45*(1), 206–230.

Raudenbush, S. W. (2009). The brown legacy and the O'Connor challenge: Transforming schools in the images of children's potential. *Educational Researcher*, *38*(3), 169–180.

Reinsvold, L. A., & Cochran, K. F. (2012). Power dynamics and questioning in elementary science classrooms. *Journal of Science Teacher Education*, *23*(7), 745–768.

Rolland, R. G. (2012). Synthesizing the evidence on classroom goal structures in middle and secondary schools: A meta-analysis and narrative review. *Review of Educational Research*, *82*(4), 396–435.

Roth, K. J., Chen, C., Lemmens, M., Garnier, H. E., Wickler, N. I., Atkins, L. J., … Zembal-Saul, C. (2009). *Coherence and science content storylines in science teaching: Evidence of neglect? Evidence of effect?* Proceedings from the NARST 2009 Symposium: Science Content Storylines. Retrieved April 17, 2015, from http://phys.csuchico.edu/~ljatkins/Publications/Atkins_Coherence.pdf

Roth, K. J., Garnier, H. E., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. (2011). Video-based lesson analysis: Effective science PD for teacher and student learning. *Journal of Research in Science Teaching*, *48*(2), 117–148.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Ruzek, E. A., Domina, T., Conley, A. M., Duncan, G. J., & Karabenick, S. A. (2015). Using value-added models to measure teacher effects on students' motivation and achievement. *The Journal of Early Adolescence*, *35*(5–6), 852–882.

Schmidt, H. G., Vermeulen, L., & Van Der Molen, H. T. (2006). Longterm effects of problem-based learning: A comparison of competencies acquired by graduates of a problem-based and a conventional medical school. *Medical Education*, *40*(6), 562–567.

Schneider, R. M., & Plasman, K. (2011). Science teacher learning progressions: A review of science teachers' pedagogical content knowledge development. *Review of Educational Research*, *81*(4), 530–565.

Schroeder, C., Scott, T., Tolson, H., Huang, T., & Lee, Y. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching*, *44*(10), 1436–1460.

Seidel, T., & Shavelson, R. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*(4), 454–499.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizabilty theory: A primer*. Newbury Park, CA: Sage.

Smart, J. B., & Marshall, J. C. (2013). Interactions between classroom discourse, teacher questioning, and student cognitive engagement in middle school science. *Journal of Science Teacher Education*, *24*(2), 249–267.

Songer, N. B., Lee, H., & Kam, R. (2002). Technology-rich inquiry science in urban classrooms: What are the barriers to inquiry pedagogy? *Journal of Research in Science Teaching*, *39*(2), 128–150.

Stone, E. M. (2014). Guiding students to develop an understanding of scientific inquiry: A science skills approach to instruction and assessment. *Cell Biology Education*, *13*(1), 90–101.

Downloaded by [RMIT University Library] at 02:04 08 December 2017

Stronge, J. H., Ward, T. J., Tucker, P. D., & Hindman, J. L. (2007). What is the relationship between teacher quality and student achievement? An exploratory study. *Journal of Personnel Evaluation in Education*, *20*(3-4), 165–184.

Strunk, K. O., Weinstein, T. L., & Makkonnen, R. (2014). Sorting out the signal: Do multiple measures of teachers' effectiveness provide consistent information to teachers and principals? *Education Policy Analysis Archives*, *22*(100), 1–41.

Thadani, V., Stevens, R. H., & Tao, A. (2009). Measuring complex features of science instruction: Developing tools to investigate the link between teaching and learning. *Journal of the Learning Sciences*, *18*(2), 285–322.

Walshaw, G., & Anthony, G. (2008). The teacher's role in classroom discourse: A review of recent research into mathematics classrooms. *Review of Educational Research*, *78*(3), 516–551.

Windschitl, M., Thompson, J., Braaten, M., & Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Science Education*, *96*(5), 878–903.

Worley, D., Titsworth, S., Worley, D. W., & Cornett-DeVito, M. (2007). Instructional communication competence: Lessons learned from award-winning teachers. *Communication Studies*, *58* (2), 207–222.

Wu, H. K., & Huang, Y. L. (2007). Ninth-grade student engagement in teacher-centered and student-centered technology-enhanced learning environments. *Science Education*, *91*(5), 727–749.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL 2007–No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from http://ies.ed.gov/ncee/edlabs

Downloaded by [RMIT University Library] at 02:04 08 December 2017