

Testing Predictors of Instructional Practice in Elementary Science Education: The Significant Role of Accountability

KATHRYN N. HAYES,¹ CARY J. TREXLER²

¹*Department of Educational Leadership, California State University, East Bay, Hayward, CA 94542-3004;* ²*School of Education and College of Agricultural and Environmental Sciences, University of California, Davis, CA 95816, USA*

Received 22 November 2014; revised 28 August 2015; accepted 1 September 2015

DOI 10.1002/sc.21206

Published online in Wiley Online Library (wileyonlinelibrary.com).

ABSTRACT: Many resources have been committed to research on science teaching pedagogies, resulting in a robust understanding of best instructional practices. Yet, exposure to excellent science instruction in elementary school is haphazard at best and often inequitable. Although the research community has attended to the role of teacher traits, such as attitude and preparation, very little research examines the role of policy and context in constraining or supporting such best practices at the elementary level. This study attempts to fill this need through a multilevel model of how teacher traits, socioeconomic context (SE context), and accountability pressures predict students' opportunity to engage in hands-on and laboratory science education. Results indicate accountability pressure eclipsed all other predictors, including SE context, in accounting for variance in the model. Final analysis indicates that teachers at high accountability pressure schools spent about 60% less time using hands-on and laboratory teaching methods than teachers at lower accountability pressure schools, an equity issue given the distribution of underrepresented students. Teacher traits including professional development, experience, attitude, and degree were insignificant in the model. This has implications for efforts to improve elementary science education in the United States, as research and national priorities have been focused primarily on teacher development. © 2015 Wiley Periodicals, Inc. *Sci Ed* 1–24, 2015

Correspondence to: Kathryn N. Hayes; e-mail: kathryn.hayes@csueastbay.edu

INTRODUCTION

Recent policy developments in the United States invoke the economic importance of student preparation for Science, Technology, Engineering, and Math (STEM) careers as well as the need to address the underrepresentation of minorities and women in both the STEM workforce and education pipeline (NRC, 2007; PCAST, 2010). In addition, regardless of background, all students should have the opportunity to become scientifically literate and thus be able to participate more fully and democratically in society (NRC 2012a; PCAST, 2010). These dual goals (developing future scientists and science literacy for all) rely on exposure to excellent science education early in the K–12 system.

Evidence has suggested that exposure to the process of science in the elementary years helps students develop scientific habits of mind, giving rise to the interest and excitement that underlie decisions to participate in the sciences (Maltese & Tai, 2010; PCAST, 2010). Yet, exposure to science education in elementary school is often haphazard at best, and it is frequently inequitable (Dorph, Shields, Tiffany-Morales, Hartry, & McCaffrey, 2011; Fulp, 2002), resulting in deficits that follow students into middle and high school (NRC, 2012b).

Although considerable research has documented best practices in science education, comparatively little has examined the role of policy and context in constraining or supporting best practices at the elementary level. In a recent literature review, Fensham (2009) noted the dearth of policy research in science education, making a case that recognizing the role of policy greatly enhances scholarly understanding of the mechanisms that drive teaching and learning. Such an understanding is particularly salient now due to the extensive reforms required by the Next Generation Science Standards (NGSS) (NRC, 2013). This study attempts to meet this critical need through a multilevel model demonstrating the comparative roles of teacher traits, socioeconomic context (SE context), and policy factors in predicting instructional practices in elementary science education.

LITERATURE REVIEW

Science Education Instructional Practices

While some aspects remain under debate, foundational science instructional practices are well established and supported by major organizations (NRC, 2012a), resting on an inextricable link between science as a body of knowledge and science as a practice of extending and refining knowledge (Duschl, Schweingruber, & Schouse, 2007). Because of the simultaneous focus on learning content as well as participating in the process of science (Capps & Crawford, 2013), the range of instructional practices may include inquiry and experimentation as well as direct instruction, demonstration, and engagement with expository text (Duschl et al., 2007). The present study considers such a range of pedagogical practices; however, hands-on education and the opportunity for inquiry is of special concern because of its relationship to student learning (Judson, 2010; Minner, Levy, & Century, 2009).

Following several studies (e.g., Dorph et al., 2011; Fulp, 2002; Ginns & Watters, 1999), the present study uses the teacher-generated (emic) phrase, “hands-on or laboratory activities” in the survey. As noted by Biggers (2013), “science-as-inquiry requires students to explore science topics directly, through hands-on investigations” (p. 31), yet teachers’ reporting of hands-on pedagogies does not guarantee the quality of science instruction (Capps & Crawford, 2013) (i.e. “hands-on” can omit sense-making activities associated with inquiry) (Biggers, 2013). Thus, this study is focused on the necessary but not sufficient opportunity for students to engage in science-oriented activities that can include inquiry,

reported as hands-on or laboratory activities. With this caveat, we situate the study in the existing literature on both inquiry and hands-on approaches to science education.

Inquiry, project-based learning, and various forms of experiential learning have deep roots in educational practice and literature, starting with Dewey, Kilpatrick, and turn of the 20th-century progressives (Dewey, 1916; Montgomery, 1994). Inquiry specifically has been a hallmark of excellent science education (Abd-El-Khalick et al., 2004; Anderson, 2012; Marshall, Horton, Igo, & Switzer, 2009). Inquiry was defined by the NRC (1996) as involving students in investigation and experimentation activities to “develop knowledge and understanding of scientific ideas, as well as an understanding of how scientists study the natural world” (p. 23).

Definitions of inquiry have evolved to include students conducting data collection and analysis, engaging in reasoning, explanation and argumentation, and communicating results (Abd-El-Khalick et al., 2004; Duschl & Osborne, 2002), all of which served as a foundation for the NRC (2012a) framework for K–12 Science Education of the NGSS science and engineering practices. In specifying the practices of science, the framework (NRC, 2012a) laid out three spheres of activity: investigating, evaluating, and developing explanations and solutions. In both literature and the present study, teachers’ descriptions of hands-on and laboratory science correspond most closely to the first, ranging from “cookbook labs” to investigation activities that engage students in critical thinking and meaning construction (Dorph et al., 2011; Ginns & Watters, 1999; NRC, 2012a).

Arguably, children should have the opportunity to participate in the full range of science education activities (Duschl et al., 2007), including direct instruction, demonstration, and inquiry or laboratory activities. Yet evidence suggests that inquiry and the opportunity for inquiry provided by hands-on, lab-based activities are neglected in many elementary classrooms, particularly in high-poverty contexts (Capps & Crawford, 2013; Dorph et al., 2011; Fulp, 2002). Inequities in children’s exposure to hands-on learning may limit science career preparation and their ability to participate as full citizens in an increasingly technoscientific society. Differential distribution of science pedagogical practices at the elementary level, however, is not well documented.

Differentiating the Role of Teachers, Social Context, and Policy Milieu

Scholars who attend to multiple factors that guide instructional practices have described a combination of internal elements (a teacher’s content preparation, confidence, attitude, beliefs about students, classroom management, and other elements of individual discretion) and external elements (resources, materials, student population, leadership support, and policy directives) (Biggers, 2013; Lee & Houseal, 2003; Valli & Buese, 2007). The present study draws from these elements in demonstrating the role of teacher traits (internal; Level 1), including experience, attitude, hours of science professional development (PD), and degree; it also draws from policy/contextual factors (external, Level 2), including accountability pressures and SE context. This review briefly describes each of these in turn before turning to the model.

Teacher Traits: Well-Researched but Still Uncertain. Certainly there have been valid concerns regarding the lack of science content and pedagogical preparation among elementary teachers. Preservice elementary teachers tend to take few science courses in college or during teacher preparation (Fulp, 2002; Lee & Luykx, 2005; Ramey-Gassert, Shroyer, & Staver, 1996), and they may lack preparation in teaching using inquiry pedagogies (Schneider & Plasman, 2011). Yet the relationship of this lack of preparation to inquiry or

4 HAYES AND TREXLER

hands-on practice is inconsistent. Lack of scientific content knowledge may affect teacher confidence, resulting in more textbook-driven teaching and an overreliance on didactic pedagogy (Supovitz & Turner, 2000); yet, content knowledge does not always predict greater investment in inquiry (Joern, 2009). Likewise, although there may be a relationship between knowledge of inquiry pedagogies and inquiry practice (Capps & Crawford, 2013), some elements of inquiry are easier for teachers to internalize and practice (e.g., collecting data) than others (e.g., posing questions) (Schneider & Plasman, 2011).

Low confidence, low self-efficacy, and negative attitude toward science are also cited as substantive barriers to excellent science teaching (Dorph et al., 2011; Murphy, Neil, & Beggs, 2007); however, evidence is again somewhat mixed. Some studies (Harlen, Holroyd, & Byrne, 1995; Lee & Houseal, 2003; Supovitz & Turner, 2000) contend teachers tend to implement more hands-on or inquiry instruction when they feel efficacious or have a positive attitude toward science; while others find attitudes and self-efficacy to be uncorrelated with inquiry or hands-on learning (Ginns & Watters, 1999; Munck, 2007).

In both federal reports (e.g., NRC, 2012a; PCAST, 2010) and research literature, PD is often suggested as a key reform to address the above deficiencies, and researchers have generated a considerable body of evidence about the impact of PD on preservice and in-service elementary teachers (e.g., Banilower, Heck, & Weiss, 2007; Duschl et al., 2007; Penuel, Fishman, Gallagher, Korbak, & Lopez-Prado, 2008; Supovitz & Turner, 2000). Findings on science PD outcomes are consistent with PD studies across other content areas; these point to the importance of training that takes place over a long period of time, is aligned with curriculum and teacher goals, and involves teachers in active learning (Desimone, Porter, Garet, Yoon, & Birman, 2002; Supovitz & Turner, 2000). More specifically, inquiry practices may be influenced by PD that allows teachers many opportunities to practice and demonstrate competence (Jeanpierre, Oberhauser, & Freeman, 2005), although the effects may be mediated by other variables such as attitudes and context (Banilower et al., 2007). Many school districts, however, neglect to support teachers in these ways, offering primarily didactic and short-term science PD activities that do not support shifts in teacher practice (Desimone et al., 2002; Penuel et al., 2008).

Despite the inconsistent nature of many of these results, most quantitative studies do not look beyond teacher traits to test policy or contextual predictors of the instructional approaches teachers use in their science classrooms (Supovitz and Turner (2000) is an exception). However, an emergent body of literature is beginning to address this pressing need by examining the role of accountability in shaping science education practice.

Accountability Pressures: Best Practice Suffers. Accountability is a strongly incentivized set of policies that attach rewards or sanctions to schools, teachers, or students according to the outcomes of standardized tests (Cuban, 2004). Many states began implementing test-based accountability measures in the mid-to-late 1990s, and in 2001 the U.S. government moved to institutionalize accountability throughout the states by signing into law the No Child Left Behind (NCLB) act, which sought to close the achievement gap and improve academic achievement (U.S. Department of Education, 2007). The act accomplishes these goals through requiring schools to demonstrate students (and student subgroups) are making adequate yearly progress (AYP), determined primarily through the percent of students deemed proficient on standardized tests, called the academic performance index (API). In many states, including California, where this study took place, sanctions accumulate with each consecutive year a school does not achieve both API growth and AYP in a given subject area (CDE, 2012). In California, AYP was determined primarily through math and English language arts (ELA) scores. Science was tested three times over

a student's K–12 career, but at the elementary level science test scores were counted as only 6% of elementary school API based on 5th-grade tests.

At both state and national levels, there have been indications of coming changes in accountability systems, including a shift to more critical thinking and performance tasks on exams aligned with the Common Core State Standards (CCSS) (Smarter Balanced Assessment Consortium [SBAC], 2013) and possibly greater inclusion of science test scores in the calculation of API (Yamamura, 2012). Most current policies, however, still focus on math and ELA as the standards for student achievement and college readiness (SBAC, 2013). In addition, there appears to be little change in the underlying structure of accountability, which relies on external rewards and sanctions as incentives for shifting teaching practice.

Because of the focus on math and language arts as well as the broad and shallow scope of tested material (Smith & Southerland, 2007), science education may have been uniquely affected by accountability systems, leading to declines in time for science education as well as hands-on pedagogies (CEP, 2007; Dorph et al., 2011). Since NCLB, teachers report their science teaching practice has become increasingly teacher centered and fact based (Anderson, 2012; Lee & Luykx, 2005; Settlage & Meadows, 2002; Smith & Southerland, 2007). Accountability may also limit teachers' abilities to follow student interests and to express enthusiasm, thus impacting student curiosity (Wideen, O'Shea, Pye, & Ivany, 1997). In addition, reductions in available time resulting from accountability-driven initiatives can severely impact teachers' ability to conduct inquiry or hands-on activities (Dorph et al., 2011; Lee & Houseal, 2003). Inquiry science requires commitment of both time and resources to carry out experiments and to develop students' scientific thought processes; therefore, constraints on preparation and instructional time greatly impact inquiry in the elementary classroom (Biggers, 2013; Lee & Houseal, 2003).

In sum, research demonstrates that accountability pressure primarily had a negative impact on elementary science education in general and inquiry, or hands-on, pedagogies specifically. However, much of this research relies on qualitative data or on teacher and administrator perceptions; very few studies quantitatively examined the role of accountability pressure in relation to other mechanisms (the exceptions include the primarily descriptive studies listed above: CEP, 2007; Dorph et al., 2011).

SE Context: Understudied. Although there is a clear and well-documented relationship between SE context and student test scores (Crane, Huang, & Barrat, 2011; Powers, 2003; Sirin, 2005), research on the role of school SE context in reference to science instructional practice is sparse. Studies on the influence of SE context tend to fall into two areas: the first regards the relationship between instructional practice and the availability of resources, and the second pertains to teacher beliefs and behaviors toward students in poverty. Regarding the first, teachers need resources and materials to implement hands-on science education, and studies conducted in some districts report funds and materials to be inadequate (Penuel et al., 2008), with teachers citing “broke” districts and reporting spending their own money on supplies (Dorph et al., 2011; Lee & Houseal, 2003).

In terms of the second, there are mixed results. High poverty context can impact use of inquiry pedagogies and critical thinking activities when teachers have lower expectations of students (Duschl et al., 2007; Warburton & Torff, 2005), and teachers may engage in less hands-on education because they worry students cannot “handle it,” or because class sizes are unwieldy (Lee & Houseal, 2003). However, in other studies teachers cited the benefits of critical thinking, hands-on and inquiry science for underrepresented students, implicating accountability as constraining hands-on education (Carlone, Haun-Frank, & Kimmel, 2010;

Shaver, Cuevas, Lee, & Avalos, 2007; Warburton & Torff, 2005). The spotty and mixed results in this area indicate a clear need for modeling the influence of student SE context on teacher instructional strategies in elementary science.

Bringing It All Together: What Is Known of the Predictors of Hands-On or Inquiry Practice

Most studies on factors influencing the use of inquiry or hands-on science treat the variables in relative isolation or focus on teacher traits (e.g., Capps & Crawford, 2013; Joern, 2009; Munck, 2007). Of the studies that quantify the percent of time teachers spend on various pedagogies in elementary science instruction, Fulp (2002) noted teachers reported spending 29% of instruction time on hands-on science (from the 2000 National Survey of Science and Mathematics Education), and Dorph et al. (2011) found elementary teachers reported 42% of the time always/often doing hands-on science. However, neither modeled the predictors for this distribution, other than grade level (significant, Fulp, 2002) and program improvement status (insignificant, Dorph et al., 2011). This may be due to sampling design (either too few teachers per school or too few schools per district); an issue we mitigate in the present study.

A few studies test multiple predictors. For example, Supovitz and Turner (2000) conducted a rigorous study of predictors of classroom inquiry practice, focusing on the effect of the NSF program, the Local Systemic Change Initiative. The study determined several teacher factors to be significant, including hours of PD, attitude toward science education, and content preparation. Additionally, a significant relationship was found between SE context and inquiry practice: teachers used inquiry practices 20% less frequently for each standard deviation increase in percent Free and Reduced Lunch (FRL) assistance (a measure of student poverty). However, the study was conducted pre-NCLB and thus included no measure of accountability pressure.

RESEARCH PURPOSE AND HYPOTHESES

A robust body of research attends to the roles of various teacher traits in determining science teacher instructional practices (e.g., Lee & Luykx, 2005; Supovitz & Turner, 2000), but neglects the potential role of accountability pressure (demonstrated across many primarily qualitative or descriptive studies, e.g., Anderson, 2012; CEP, 2007), and SE context (demonstrated in the literature on educational policy, e.g., Supovitz & Turner, 2000). To test the role of accountability and SE context in contrast with teacher traits, we used a multi-level model to simultaneously examine teacher and school predictors of elementary science instructional practices across schools in one district, addressing the following hypotheses:

- Hypothesis 1.* Owing to the institutional and sociological leverage embedded in accountability policies and community SE context, accountability pressure and SE context will have a greater relationship with elementary science instructional practices than teacher traits (experience, attitude, PD, and degree), as teachers make their instructional choices within the constraints imposed by policy and context.
- Hypothesis 2.* Based on nascent research that NCLB-based accountability mechanisms are affecting teacher practices through strongly incentivized assessment policies, accountability pressure will have a greater relationship to elementary science education instructional practices than community SE context.

If hypotheses one and two are substantiated in the primary model, teacher traits would have a less substantial relationship to instructional practices than accountability pressure. If that is the case, teacher preference for certain practices should not differ significantly across accountability pressure, but their ability to carry out those pedagogies would differ. It follows that:

Hypothesis 3. Accountability pressure will predict the difference between reported and preferred instructional practices to a greater extent than teacher traits or SE context.

METHODS

Sampling Procedure

District Role. Districts play a major role interpreting and setting policy by allocating time, supporting PD, setting priorities, and choosing curriculum (Hamilton et al., 2007). Thus, a typical sampling procedure employed when sampling many districts—often only one to five schools per district and only a few teachers at each school—although more generalizable, is less able to delineate teacher and school effects within a given district policy context. Because the present study focuses on school level effects of SE context and policy, we sampled half the schools in one carefully selected district to control for district curriculum and policy interpretations. The sampling design thereby allowed for a clearer analysis of the relevant factors than a broad, but shallow, sampling design.

Valley district was selected because it was representative of California districts in the following ways: (a) the district means are quite close to California state means in API, percent English language learners, and percent FRL (Figure 1), and (b) it spans both urban and suburban areas of a mid-sized city, thus schools vary widely in FRL, ethnicity, and accountability measures (Appendix, Table A1). Due to the focus on one district, generalizability is a limitation. Nonetheless, the results of this study lay the foundation for additional studies, as well as providing verification for qualitative findings regarding accountability pressure.

To obtain a minimal sample size that would produce accuracy within $\pm 5\%$ of the district teacher population at a 95% confidence interval (Rea & Parker, 2005), we sampled 231 of 580 valley district elementary teachers (Grades K–5). A random stratified sampling procedure was used to select 20 schools from the set of 42 elementary schools (Rea & Parker, 2005). Six schools were randomly selected from the lowest and highest API quartiles and four each from the middle quartiles. Selection was more heavily weighted at the ends of the spectrum to have adequate sampling representation for the Level 2 predictors, accountability pressure, and socioeconomic status. This represents a limitation in calculating instructional practice averages, but it has negligible effect on HLM statistics.

Sampling response bias examined on FRL and API were within a reasonable range (Table 1). Two schools out of the original sample that opted not to participate were replaced with the school with the most similar API. We asked all K–5th-grade teachers in each sampled school to complete the survey. Average teacher participation rate across schools was 71%, a total of 182 teachers. Of these teachers, 84% completed the entire survey and are represented in the full model. Pearson's chi-squared and z -tests were used to determine teacher response bias through differences in the numbers of teachers at each grade level per API quartile. Differences were statistically insignificant.

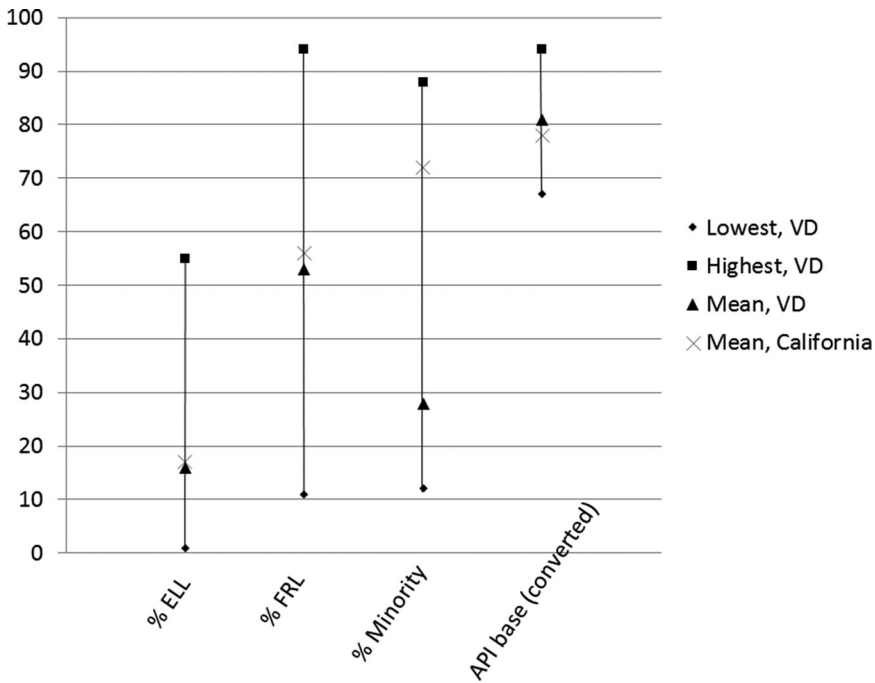


Figure 1. Mean and range of the percent of English language learners, free and reduced lunch, minority, and API base score (2011) for schools in valley district (VD), as compared to California state means (see Table A1 in the Appendix for numerical data).

TABLE 1
School Response by API Quartile, With Analysis of Sampling Bias in the Two Schools That Chose Not to Participate

API Quartile	Total Schools in Quartile	Selected to Participate	Participated	Sampling Response Bias (Deviation of the Nonrespondent From the Mean of Schools in Quartile)	
				FRL	API
First	11	6	5	0.83 SDs	~1 SD
Second	11	4	4	N/A	N/A
Third	10	4	3	~1 SD	1.2 SDs
Fourth	10	6	6	N/A	N/A
Sum	42	20	18		

Note: Total school response (90%); 18 of 20.

Instrumentation

Survey. The Science Instructional Time and Pedagogy (SITP) survey consisted of seven sections, four of which were used for the models presented in this paper (see Table 2). Either the researchers or principal presented the survey to the teachers with an electronic link. In

TABLE 2
Description of Survey Sections

Section	Area	Question Types
1	Demographics	Three closed-response questions
2	Instructional time	Two numerical answer questions regarding science education time, divided into options
3	Pedagogy distribution	Two numerical answer questions asking for time attribution as a percentage
4	Thoughts on science education	Three Likert-type questions with a total of 10 statements Two demographic questions One question asking for details on hours of professional development One open-ended question asking for explanation of science teaching methods

all cases the survey was optional, with the incentive of a school site staff gift certificate if 80% of the teachers responded.

Validation. SITP was adapted from a survey used by Dorph et al. (2011) to document elementary science implementation across California, modified to specifically test the constructs of interest in this study. Validity of the SITP survey was established through a three-step process. First, the questions were tested and modified during six focus group interviews of 34 teachers and administrators (Rea & Parker, 2005) that focused on elementary science education instructional practices, barriers, and supports (Krueger, 1994). Second, to establish face validity, the resulting SITP draft was subjected to cognitive interviews with two teachers (Desimone & Le Floch, 2004), resulting in additional minor modifications. Two academic content experts also reviewed the instrument. Finally, criterion validity for key constructs was established through comparisons with open-ended SITP survey questions and sampled teacher interviews. Reliability was tested through internal consistency (described by variable).

Variables

For a summary of variables, see Table 3; details are discussed in the following section.

Dependent Variables. The survey asked teachers to report (a) the *minutes per week* they spent teaching science (mean of 68 minutes, Table 5); (b) the percent of this time students engaged in each of the *instructional approaches*, summarized as hands-on, textbook, lecture, audiovisual, and demonstration (out of 100%) (see Table 4); and (c) the percent of time teachers would *prefer* to spend on each instructional approach. The main outcome variable in this study focused on instructional approaches. Of the five approaches, only the percent of time teachers reported “students reading chapters or other material and answering textbook or worksheet questions” (*percent textbook*) and percent of time teachers reported students “doing hands-on or laboratory activities” (*percent hands-on*) demonstrated significant relationships to predictors; since *percent textbook* acted as an inverse of *percent hands-on*, we primarily focus on *percent hands-on* in the following analyses. On average, teachers reported using hands-on or lab activities more than any other individual approach (42%). However, in sum, the other approaches focus on science instructional practices that

TABLE 3
Description of Main Variables

Variable Label	Role	Type	Description
Science instruction time	Descriptive	Continuous	Minutes per week teaching science education
Percent hands-on	H1 and H2 dependent (primary focus)	Continuous	Percent of time teachers report students doing hands-on or laboratory activities
Percent textbook	H1 and H2 dependent (secondary focus)	Continuous	Percent of time teachers report students reading chapters or other material and answering textbook or worksheet questions
Preference hands-on	H3 dependent	Continuous	Preferred percent of time doing hands-on or laboratory activities
Preference textbook	H3 dependent	Continuous	Preferred percent of time reading chapters or other material and answering textbook or worksheet questions
SE context	Second-level predictor	Composite, continuous	Socioeconomic context: A composite of school FRL and percent of underrepresented students
AYP pressure	Second-level predictor	Continuous (also converted to binary)	Measure of testing pressure, years out of the last six the school did not meet adequate yearly progress
GRADE	Covariate	Series of dummy variables	Grade level taught
PD hours	First-level predictor	Ordinal	Science professional development hours over the last 3 years, reported by teacher (0, less than one hour, 1–3 hours, 4–10 hours, 10–20, more than 20), aggregate over four sources
Years taught	First-level predictor	Ordinal	Reported years taught (1–3, 4–6, 7–9, 10–15, 15 or more)
Degree	First-level predictor	Binary	Whether they have a science degree
Attitude	First-level predictor	Continuous	Composite variable, average of Likert ratings (scale 1–5): how comfortable they are with science, how important science is at their grade level, how much they liked science, and whether they would like to teach more science (Cronbach's alpha = .817)

are likely less active and more teacher centered (Dorph et al., 2011). Reliability was tested in two ways: Reliability was measured between *percent hands-on* and percent teachers *preferred* for each instructional approach (Cronbach's alpha; Table 4); and *percent hands-on* was compared with interview data from eight of the surveyed teachers (sampled from four schools), Pearson's correlation $r = .842$.

The SITP survey used the somewhat general phrase “hands-on or laboratory activities” rather than specifying elements of inquiry for the following reasons: in focus groups, the term “hands-on” and “labs” were used by teachers almost exclusively to describe the

TABLE 4
Reliability Measure (Cronbach's Alpha) Between Percent Teachers Reported and Preferred for Each Pedagogy as well as Average Percent Reported

Variable (Reported and Preferred Percent)	Item: Percent of Time Teachers Report . . .	Cronbach's Alpha	Average Percent Reported (%)
Hands-on	Students doing hands-on or laboratory activities	.80	42.19
Textbook	Students reading chapters or other material and answering textbook or worksheet questions	.70	21.64
Lecture	Students listening and taking notes during a presentation by the teacher	.87	7.96
Audiovisual	Students watching audiovisual presentations	.47	13.43
Demonstration	Students watching a science demonstration	.62	14.08

opportunity for investigation. Thus, the terms have greater face validity than terms less familiar to teachers. Because this study focused on activities that provided an opportunity inquiry rather than identifying with specific elements of inquiry, it was deemed appropriate to use the emic (teacher language) terms for the sake of greater validity. The terms were further validated through an open-ended question on each survey asking teachers to describe excellent elementary science education. Of respondents, 77% used the word hands-on, associating “hands-on” with experiments or labs (32%) or investigations or discovery (30%). Teachers thus generally considered hands-on to indicate students involved in labs, experiments, or some form of investigation or discovery.

Predictor Variables. Variables at Level 1 (teacher traits) included grade level, PD hours, experience (years taught), and whether or not they had a bachelors degree in science or science education, as well as a measure of attitude (see Tables 3 and 5). *Grade level* (0–5) was used as a control and treated as a series of dummy variables; Grades K–1 and 2–3 are placed together, with Grades 4 and 5 separated because—due to testing in 5th grade—teachers at the fourth- and fifth-grade level are more likely to demonstrate substantive differences in their science instructional practices than teachers in the primary grades. In the models, K–1 serves as the reference variable (it becomes the intercept as it is left out of the model), because there are no accountability-based tests in K–1. Grade level was relatively evenly distributed amongst teachers in schools (Table 5). Teachers reported *PD hours* in matrix format (0, less than 1 hour, 1–3 hours, 4–10 hours, 11–20 hours, more than 20 hours), noting the number of hours they had received over 3 years by subject: math, language arts, and science, as well as by source: county or district, college or university, informal institution, or site based. The variable was a composite of reported science education PD. Teachers obtained PD from a variety of sources (documented in interview data); there was little district-sponsored PD in the last 3 years. Because the purpose of this study was not to establish the results of particular PDs, but rather the relationship of typical PD to science education in a realistic context, we did not ask for the content or nature of the PD. The distribution exhibited some issues of normality; it was

TABLE 5
Descriptive Statistics for Predictor and Dependent Variables

	<i>N</i>	Minimum	Maximum	Mean/Proportion	SD
School-level variables					
AYP pressure	18	0.00	6.00	2.61	1.94
SE context	18	11.30	89.95	42.21	24.07
Teacher-level variables					
Science instruction time	182	0.00	200	68.37	42.93
Percent hands-on	161	0.00	100	42.19	24.70
Percent textbook	161	0.00	100	21.64	23.44
Preference hands-on	159	0.00	100	54.58	20.58
Preference textbook	159	0.00	70.00	13.03	11.85
Grade 2–3	161	0.00	1.00	0.33	0.47
Grade 4	161	0.00	1.00	0.20	0.40
Grade 5	161	0.00	1.00	0.16	0.37
PD hours (bins)	162	0.00	17.00	4.86	2.86
Years taught (bins)	162	2.00	5.00	4.56	0.80
Degree (binary)	153	1.00	2.00	1.12	0.33
Attitude (Likert)	155	1.25	4.75	4.04	0.65

Note: For teacher-level variables, *N* reflects the number of teachers. For school-level variables, *N* reflects the number of schools.

TABLE 6
Intercorrelations Among Teacher-Level Variables (Level 1)

	Grade (Ordinal)	Degree	PD Hours	Attitude	Years Taught
Grade (as ordinal, 1–5)	1	.163*	.052	–.005	–.012
Degree		1	–.042	.147	.038
PD hours			1	.200*	.101
Attitude				1	.076
Years taught					1

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

positively skewed due to a few teachers reporting many hours of PD and many reporting zero hours. However, transformations of the variable did not result in more accurate modeling or shifts in significance.

Years taught allowed for five choices (1–3, 4–6, 7–9, 10–15, and 15 or more); the mean (4.6 out of 5 choices, Table 5) indicated a relatively high level of experience across teachers. *Degree* was a binary variable; 13% of teachers reported a bachelor's degree in science or science education. *Attitude* was calculated by averaging teachers' ratings of items related to their attitude toward teaching science; these items were similar to those in the Horizon National Survey of Science and Mathematics Education Science Questionnaire (van Aalderen-Smeets & Walma van der Molen, 2013) (see Table 3). Reliability for this construct was high (Cronbach's alpha = .817). The distribution was negatively skewed, with most teachers at the high end of the Likert-type scales (mean of 4.04 out of 5). However, neither binning nor transformation altered the variable's significance within the model, thus it was left as a composite scale. Correlations between Level 1 variables are depicted in Table 6.

Level 2 Variables

SE Context. The *SE context* variable was a composite of school FRL percent and percent of students underrepresented in science (not White or Asian) (Cronbach's $\alpha = .923$) (Ed-Data, 2013). We used percent underrepresented as part of SE context rather than percent minority because Asian students are highly represented in both science majors and careers, and thus percent minority would be misleading in terms of equity (PCAST, 2010).

AYP Pressure. Under NCLB, whether or not a school makes AYP each year for each subject (math and ELA) and subgroup of students is used to determine sanctions; in California these pressures accumulate; at Year 2 schools must notify parents of being out of compliance—by Year 5 schools are subject to restructuring and alternative governance. They do not reset unless the school makes AYP 2 years in a row. Because “not making AYP” has been a key element of pressure on schools and teachers (Dorph et al., 2011; Hamilton et al., 2007; Judson, 2013; Penuel et al., 2008), and because sanctions accumulate up to Year 6 and are continual and iterative thereafter (California Department of Education, 2012), the measure for this construct was calculated by adding the cumulative years each school did not make AYP in either math or LA out of the last 6 years (Ed-Data, 2013). Following state policy, our measure of accumulated pressure was not reset unless the school made AYP in the given subject 2 years in a row. Because non-Title 1 schools still receive sanctions (although less specified) for not making AYP or API growth in California, both Title 1 and non-Title 1 schools were included in this measure. Cumulative *AYP pressure* was validated as a measure through teacher interviews regarding the source of their feelings of accountability pressure. *SE context* was significantly correlated with *AYP pressure* ($r = .781, p < .001$); issues with colinearity are discussed in the results section.

Variables were tested for assumptions violations prior to HLM modeling. Normality was assessed through an examination of frequency distributions and standardized skew and kurtosis. In cases of standardized skew or kurtosis values above 2.0, the log transformed or squared variable was tested in the model; none were retained due to lack of significance (Allison, 1999). Linearity and homoscedasticity were assessed through a visual inspection of the scatterplots of each independent variable on studentized residuals. By taking account of the clustering effect at school level, the multilevel model ensured that the residuals of the dependent variables were independent and normally distributed at Level 2. In addition, the model was tested for independent distribution of residuals at Level 1. Data were examined for sensitivity to outliers through a consideration of Cook's D. A z -score was calculated for outlying data points, resulting in the removal of one data point due to a z -score exceeding 3.3 (Tabachnick & Fidell, 2007).

Model

A multilevel model was used to examine the comparative role of teacher-level and school-level factors in predicting science instructional practices using HLM7 (Raudenbush & Bryke, 2002). The model employs a teacher-level (1) and school-level (2) equation, each with an error term uncorrelated across levels (Raudenbush & Bryk, 2002). The two-level approach alleviates aggregation bias (wherein standard errors may be inflated) and accounts for similarities in teacher response within schools, as well as reducing error independence violation across levels, which may be present in ordinary least-squares regression. That is, the two-level model accounts for how teachers are nested within schools and allows the delineation of school-level effects from within school (teacher-level) effects.

Other than grade dummies, teacher-level traits were centered around the group (school) mean to clearly delineate within and between school effects. Because teacher traits can

be interrelated with school-level factors, prior to centering, all variables were tested for interactions and whether they accounted for Level 2 variance. All interactions were insignificant. Only one variable, *attitude*, accounted significantly for Level 2 variance. When group centered, each measure of teacher traits represented the distance of that teacher trait from the school mean, with the school mean set at zero. The intercept (β_{0j}) then became the mean for each school j at Level 2. Level 2 variables were left uncentered. Slopes of Level 1 variables were fixed at Level 2 to maintain a focused model (Maerten-Rivera, Myers, Lee, & Penfield, 2010), and because tests of homogeneity of Level-1 variance (the variance of Level 1 slopes across Level 2) were insignificant. Thus, this model portrayed how group means (Level 1 intercepts) varied across schools rather than variance in slope coefficients across schools (Raudenbush & Bryk, 2002). In other words, the model tests variance in the mean percentage of hands-on practices across schools rather than variance in the relationship (slope) between hands-on practices and teacher trait (Level 1) variables across schools. All models used restricted maximum likelihood; α was set to .05.

As advocated by Raudenbush and Bryk (2002), after estimating the null model, this study compared Level 1 and Level 2 models separately, then added Level 2 predictors to the final Level 1 model. Because there was little theoretical foundation for the order of adding variables to the model, Level 1 variables were each added to the model individually first, then sequentially. The use of the deviance statistic to evaluate model fit is inappropriate in this case due to sample size (Raudenbush & Bryk, 2002). Although Level 1 variables were insignificant, they were retained in the final model for theoretical purposes.

Level 1 model (each variable except grade centered on school mean):

$$\begin{aligned} \%Hands\text{-}on_{ij} = & \beta_{0j} + \beta_{1j} \times (\text{Yrs_teach}_{ij}) + \beta_{2-4j} \times (\text{Gradedum}_{ij}) \\ & + \beta_{5j} \times (\text{Degree}_{ij}) + \beta_{6j} \times (\text{PD}_{ij}) + \beta_{7j} \times (\text{Attitude}_{ij}) + r_{ij} \end{aligned}$$

Akin to a basic linear regression, the Level 1 model specifies the predicted percent of hands-on (Y_{ij}) for individual teacher i in school j . β_{0j} is the intercept, or grand mean of all schools for kindergarten and first-grade teachers (the omitted grade variable) when all others are centered around the school mean. β_{1j} through β_{7j} are fixed coefficients identifying the vector of hands-on practices for each teacher at each school, j , based on *years taught*, *grade*, *degree*, *PD hours*, and *attitude*. In sum, the teacher's percent of time teaching hands-on is predicted as a function of their experience, grade level, BA degree, PD hours, and attitude, along with error unexplained by these variables. Interschool variation is represented by Level 2 models.

At Level 2, the constant from the level 1 model (β_{0j}) is a function of the grand mean across schools (γ_{00}) plus a coefficient representing the effect of accountability pressure on the portion of variance in *percent hands-on* (γ_{01}) and a school-level error term (u_{0j}). In other words, the intercept from Level 1 (the mean of the school) is predicted as a function of AYP pressure and error. The fixed coefficients vectors (γ_{10} – γ_{70}) represent the constant coefficient for each Level 1 variable.

Level 2 model:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} \times (\text{AYPPressure}_j) + u_{0j} \\ \beta_{1j} &= \gamma_{10}; \beta_{2j} = \gamma_{20}; \text{etc.} \end{aligned}$$

The full model accounts for Level 1 and Level 2 variables; an error term at each level allows for independent variation across levels (Raudenbush & Bryk, 2002).

TABLE 7
Variance Components of the Baseline Model (ICC)

	Variance	Percent of Variance	χ^2	p
Variance between schools (Level 2)	153.77	24.9	67.91	<.001
Variance within schools (Level 1)	464.21	75.1		
Total variance	617.98	100		

Full model:

$$\begin{aligned} \%Hands\text{-}on_j = & \gamma_{00} + \gamma_{01} \times (AYPPressure_j) + \gamma_{10} \times (Yrs_teach_{ij}) + \gamma_{20-40j} \\ & \times (Gradedum_{ij}) + \gamma_{50} \times (Degree_{ij}) + \gamma_{60} \times (PD_{ij}) + \gamma_{70} \\ & \times (Attitude_{ij}) + u_{0j} + r_{ij} \end{aligned}$$

The modeling process first generates an intraclass correlation coefficient (ICC) estimating the proportion of variance explained at each level with no additional predictors (the decomposition of variance) (Table 7). The proportion of *within* and *between* school variance explained by each variable as well as the full models (the proportional reduction in variance) was calculated by subtracting the remaining variance from the variance of the baseline model, then dividing by that variance, consistent with Raudenbush and Bryk (2002). Comparison of the base-level variance with subsequent models affords an understanding of how each model accounts for variance beyond the baseline at both Level 1 and Level 2 (a measure of effect size), and thus an idea of the comparative influence of school-level versus teacher-level traits. This calculation is noted as “additional variance accounted for” in the results text. In addition to HLM models, descriptive statistics were used to compare both instructional time and instructional practices across accountability context.

RESULTS

The following section first reports analysis of Level 1 variables, then Level 2 variables, and finally the results of the three hypotheses. Baseline decomposition of variance calculations (Table 7 and Figure 2) showed that nearly one-quarter of the variance in hands-on and laboratory approaches to science was due to between school differences; 75.1% was attributed to within school differences (teachers within schools).

Level 1

All group-centered Level 1 variables were insignificant solely and in the full models (Table 8). However, *attitude* as a single *uncentered* predictor significantly predicted *percent hands-on*, accounting for 9% additional variance at Level 2 (between school) but only 3% additional variance at Level 1 (within school) compared to the baseline decomposition of variance of within and between schools (Figure 2). Despite this noteworthy finding, *attitude* was included as group centered in the final model to delineate within school versus between school variance. The model accounted for very little of the variance within schools (4% or less; see Figure 2), indicating that the differences between teachers, although important, were either random or related to predictors not accounted for in this study.

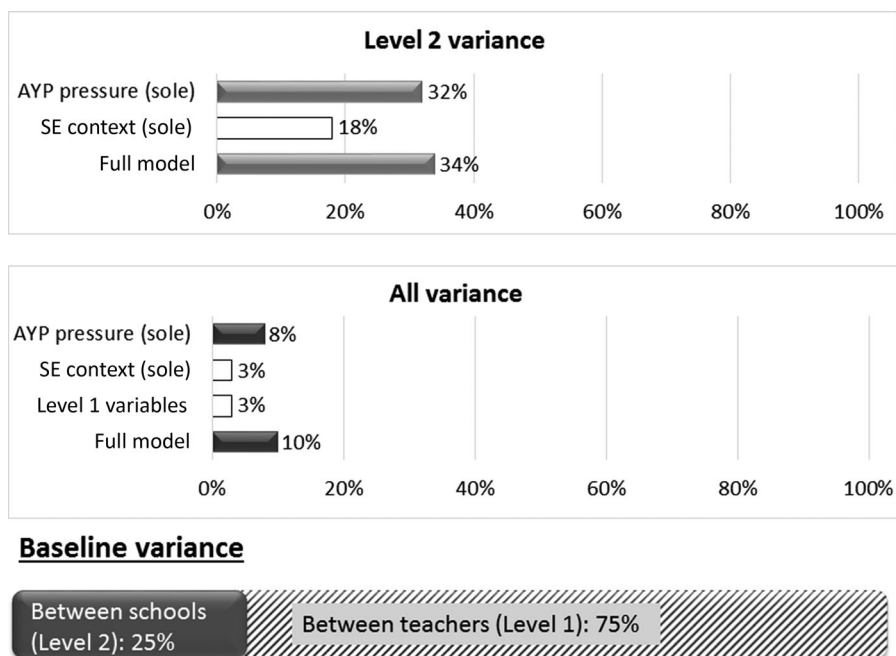


Figure 2. Variance accounted for (beyond baseline ICC) by Level 2 predictors (individually), Level 1 predictors, and the full model. Models with significant predictors are indicated with a solid bar.

TABLE 8
Two-Level Model Predicting Average Teacher Reported Percent Hands-On

Predictor	Model 1	Model 2	Model 3	Model 4
AYP pressure	-4.147* (1.610)			-4.267* (1.535)
SE context		-.280 (0.136)		
Grade 2-3			-7.099 (4.685)	-6.546 (4.532)
Grade 4			-8.168 (5.347)	-7.967 (4.681)
Grade 5			-10.121 (5.765)	-9.740 (5.209)
PD hours			-.226 (0.692)	-.225 (0.690)
Years taught			2.355 (2.632)	2.325 (1.777)
Degree			-4.918 (5.951)	-4.989 (6.719)
Attitude			5.563 (3.327)	5.538 (3.314)
Intercept	51.695*** (5.716)	52.952*** (6.484)	47.545*** (4.398)	58.275*** (6.346)
Observations	161	161	150	150

Note: * $p < .05$, ** $p < .01$, *** $p < .001$.

Slope coefficient and standard errors in parentheses (fixed effect).

Level 2

As a sole Level 2 predictor, *AYP pressure* was significant ($p < .05$) in predicting *percent hands-on*, and accounted for 32% of Level 2 variance and 8% of overall model variance compared to the baseline ICC (Table 8 and Figure 2). Similarly, with *percent textbook* as the dependent variable, *AYP pressure* (independently) accounted for 23% of the Level 2 variance and 5% of full model variance (not shown), significantly predicting 3.2 percentage points more text use for every consecutive year the school did not make AYP. Because *percent textbook* acted to some extent as a mirror to *percent hands-on*, the latter is the focus of the models and discussion.

In contrast to *AYP pressure*, *SE context* was insignificant as a sole predictor in the model. This difference is noteworthy given the high correlation between *SE context* and *AYP pressure* ($r = .781, p < .001$). To reduce overspecification, *AYP pressure* was selected as the sole Level 2 variable in the full model. In the full model (Table 8, Model 4), *AYP pressure* predicts that for every consecutive year the school did not make AYP in either math or ELA, teachers on average reduced their use of hands-on and laboratory instruction 4.3 percentage points. Thus, schools that did not make AYP 0–4 years (i.e., made AYP most years) out of the last six averaged 47% hands-on and laboratory activities; students in schools that did not make AYP 5–6 consecutive years out of the last six (high accountability pressure) averaged 26%. We will use this measure of high accountability pressure descriptively throughout the results as an interpretive tool because the strongest sanctions are applied the fifth consecutive year the school does not make AYP.

Hypotheses

Regarding the first hypothesis—community socioeconomic and accountability pressure have a greater relationship to science instructional practices than teacher traits—the results were positive. Although much of the variance was between teachers at Level 1 (75%), no tested teacher traits accounted for any of this variance. Training (*PD hours*), preparation (*degree*), *attitude*, and experience (*years taught*), often the focus of research and policy, were all insignificant in the final model (Table 8), and these variables explained little of the variation in instructional practice over the baseline decomposition of variance (less than 4%; Figure 2). Conversely, measures of *AYP pressure* were significant solely and in the full model and explained substantive variance. In addition, *attitude* when uncentered explained 9% of Level 2 variance, indicating the possibility of a relationship between school type and attitude.

The second hypothesis—accountability pressure has a greater relationship to elementary science education instructional practices than community SE context—was also demonstrated in the model. As a sole Level 2 predictor as well as in the full model, *AYP pressure* was significant, whereas *SE context* was not. In addition, *AYP pressure* accounted for nearly double the Level 2 variance as *SE context*.

The third hypothesis was constructed to further clarify whether accountability pressure was predicting teacher *ability* to carry out particular practices rather than teacher *preference* for particular practices. For this hypothesis, we tested the relationships between the predictor variables and the difference between preferred and reported practices. The two variables of interest, *percent text* and *percent hands-on*, showed a substantial gap between predicted and reported (Figure 3).

For each respondent, the percent reported was subtracted from the percent preferred (Figure 4). On average, teachers in both high and low accountability pressure schools preferred to use hands-on or lab pedagogies around half of their science instruction time (49%

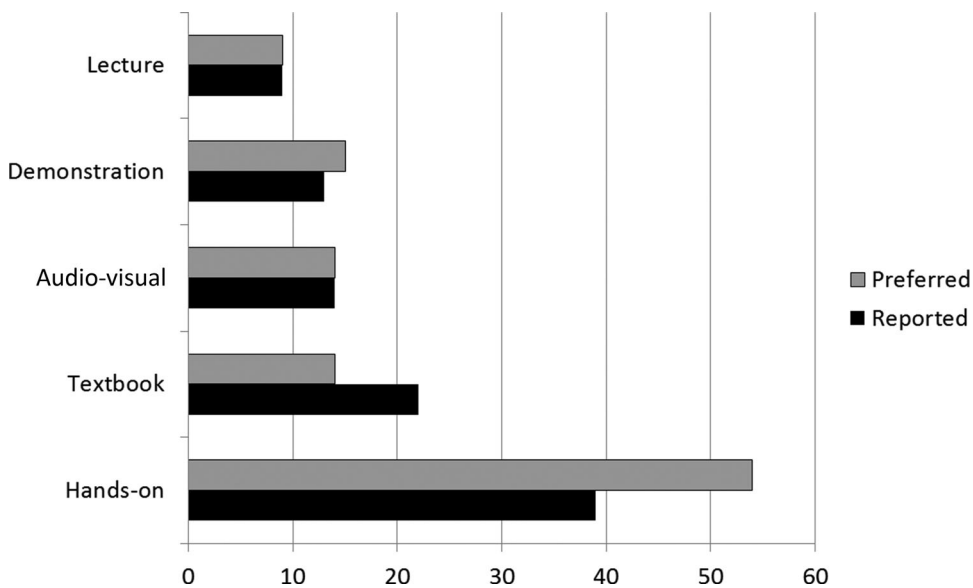


Figure 3. Comparison of teachers' average reported and preferred instructional practices (percent).

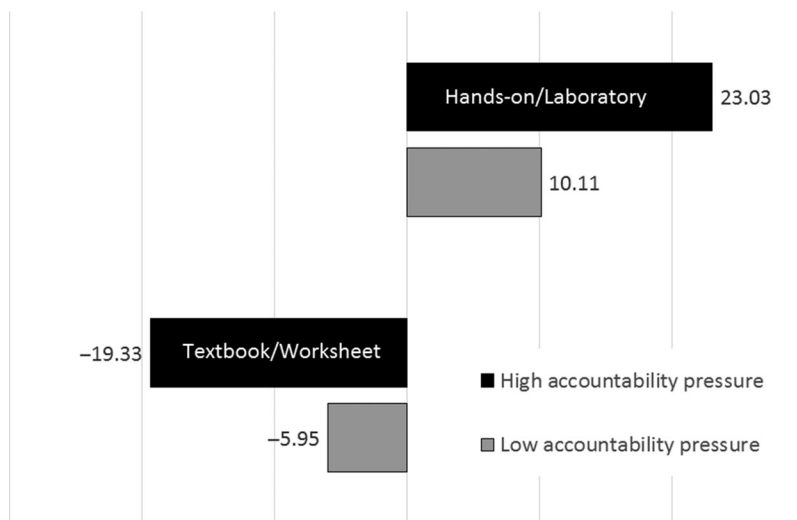


Figure 4. Difference between reported and preferred percent textbook and hands-on, high accountability pressure schools (AYP 56), and lower accountability pressure schools (AYP 0-4).

and 55%, respectively; the difference was insignificant). However, there was a 23 percentage point difference between preferred and reported for teachers at high accountability pressure schools, versus a 10 percentage point difference at low accountability pressure schools. Likewise, low accountability pressure teachers preferred textbook use on average 6 percentage points less than they reported, versus 19 percentage points less at high accountability pressure schools.

To determine whether accountability pressure predicts the difference between reported and preferred instructional practices to a greater extent than teacher traits or SE context, we conducted multilevel models with each predictor solely and with all other predictors;

TABLE 9
HLM Coefficients Reporting the Relationship of High Accountability Pressure to the Difference Between Preferred and Reported Hands-On and Textbook Pedagogies

Predictor	Hands-On	Textbook
AYP pressure	2.838* (0.994)	-2.301* (1.067)
Intercept	5.778 (3.125)	-2.865 (3.360)
Observations	159	159

Note: * $p < .05$.

Slope coefficient and standard errors in parentheses (fixed effect).

the dependent variables were the difference between preferred and reported hands-on and textbook pedagogies. For both *textbook* and *hands-on*, all Level 1 variables were insignificant and were omitted from the model. For *percent hands-on*, *AYP pressure* and *SE context* were each significant as the sole predictor, but *AYP pressure* accounted for the most additional variance (45% at Level 2, 9% of the full model). For *percent textbook*, only *AYP pressure* was significant as the sole predictor, accounting for 25% additional variance at Level 2 and 4% at the full model (Table 9). This model thus sustained the third hypothesis, further substantiating the influence of accountability pressure in predicting instructional practice.

DISCUSSION

The results of this study emerge at a pivotal time in policy and science education reform. The implementation of NGSS and the CCSS and concurrent reformulation of accountability systems provide a key opportunity to consider how to leverage equity and excellence in science instructional practice. The constellation of mechanisms related to teachers' instructional practice should be thoughtfully considered in research, assessment, and PD decisions.

Equity in Science Instructional Practice

In the focal district, teachers in high accountability pressure schools (did not make AYP 5 or 6 consecutive years) on average used hands-on pedagogies 21 percentage points less, and textbooks and worksheets 16 percentage points more, than their colleagues in lower accountability pressure schools. This is particularly important in light of the difference in science education time reported at each of these levels: At high accountability pressure schools teachers reported an average of 51 minutes of science education per week versus an average of 72 minutes at lower accountability pressure schools. This means that students at high accountability pressure schools receive on average 34 minutes of hands-on and laboratory science education per week, and students at low accountability pressure schools receive on average only 13 minutes per week—about a third of the exposure.

These findings point to substantial equity issues: Schools that did not meet AYP 5–6 out of the last 6 years averaged 80% FRL and 60% underrepresented students, as opposed to schools with lower accountability pressure that averaged 44% FRL and 25% underrepresented students (2010–2011 reporting year). Thus, in this district, poor and

underrepresented students had considerably less chance of being exposed to excellent and engaging science pedagogies than wealthier White and Asian students.

This study's use of "percent hands-on and laboratory activities" does not guarantee quality of science instruction. Hands-on was defined by teachers generally as involving labs, experiments, investigation, and discovery, but may not have included specific practices such as those defined in NGSS (i.e., constructing explanations using models). Additionally, we do not argue that science education should consist solely of hands-on or inquiry pedagogies. However, the measure is useful as a proxy for active student engagement in science learning; the other instructional practices in the SITP survey all indicated less active engagement: textbook and worksheets, lecture, demonstrations, and audiovisual.

Mechanisms Responsible for Variation

Although more of the variance in the model existed between teachers than between schools, attitude, teacher training, degree, and experience were insignificant in the final model, *and all the Level 1 (teacher trait) predictors together accounted for no more than 4% of the variance between teachers*. Conversely, measures of accountability pressure were consistently significant across models; *AYP pressure* accounted for 32% of the variance between schools and 8% of the overall variance in percent hands-on and laboratory science teaching. In addition, when not group centered *attitude* accounted for a portion of the between schools variance, implying that school culture or context may play a role. Finally, *AYP pressure* was the most substantive predictor of the gap between preferred and reported pedagogies.

These results contribute to a mixed literature on the relationship of PD and other teacher traits to hands-on or inquiry practices. Although intensive, long-term PD may contribute to teacher use of inquiry or hands-on pedagogies, our findings confirm studies that indicate typical, nonintensive PD activities make little difference in the use of hands-on and laboratory instructional practices (Marshall et al., 2009; Penuel et al., 2008; Supovitz & Turner, 2000). In terms of teaching experience, Capps and Crawford (2013) found experience made a difference in the incorporation of inquiry, whereas Supovitz and Turner (2000) and Marshall et al. (2009) found little relationship between teacher experience and inquiry practice; the latter was substantiated in these findings. Additionally, Supovitz and Turner (2000) found attitude significantly predicted inquiry, but not all research supports the salient role of attitude, as was the case in the present study (Munck, 2007). Because of the unaccounted for variance between teachers in this study, it is imperative that future research considers additional variables to parse the role of teacher traits in science education practice.

This study's results confirm and extend qualitative scholarship regarding the influence of accountability on teachers' instructional practices by demonstrating the quantitative nature of the relationship as well as the comparative effects of different mechanisms (i.e., Anderson, 2012; Lee & Luykx, 2005; Settlage & Meadows, 2002; Smith & Southerland, 2007). For teachers in this district, the role of accountability as a policy mechanism eclipsed that of SE context, which is insignificant solely and in full models. Despite the high correlation between these two variables, results indicated that AYP pressure had a specific relationship with science pedagogies separate from the influence of poverty and student population. If the substantive role of policy and insignificant role of many teacher traits are substantiated in other districts, this becomes a very important consideration for policy makers attempting to instill equity and incentivize excellent science teaching at the elementary level.

IMPLICATIONS

These results have implications regarding science education reform efforts. Research and national priorities have up to this point been focused primarily on teacher development. Teachers are often the “convenient objects of criticism,” but within an institutional structure driven by external policies, their choices may be constrained (Cuban, 2004). As demonstrated in the PD literature, shifts in practice often require intensive, long-term PD, well integrated into schools and reliant on a shared vision (Desimone et al., 2002; Elmore, Peterson, & McCarthey, 1996). Leadership, school capacity, and resources also play a role (Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010; Goetz Shuler, Backman, & Olson, 2009). However, although resource-intensive PD may make a difference (Desimone et al., 2002; Supovitz & Turner, 2000), without a consideration of the incentives embedded in the larger policy milieu, calls for teachers with greater content knowledge and professional learning opportunities may do little to increase students’ exposure to excellent science education.

Currently, a policy window exists in which to change tack. As implementation of the CCSS and NGSS proceed, greater inclusion of science in accountability measures may increase emphasis on science and possibly improve test scores (Judson, 2010, 2013). However, because of the strong, external authority of accountability mechanisms, and the complexity of policy interpretation and implementation, reformulation of assessment and accountability should be approached with great caution. NGSS reforms are dependent on teacher professional skill in both teaching (i.e., formulating lessons based on questions arising from phenomena [Reiser, 2013] and responding to changing demands in ways that are consistent with the principles of reform [Coburn, 2003]). Thus, punitive accountability policies that have been shown to erode such professionalism may undermine the very reforms sought by the policy (Darling-Hammond, 2004).

Further Research

Although the district was chosen because of the wide variability in student population and test scores, as well as similarity to California mean demographics and scores, this study should be repeated across more districts to establish greater generalizability and to test whether particular district policies may support more equitable distributions. In addition, to make full use of current national policy changes, a regression-discontinuity design measuring the impact of current policy shifts could further demonstrate policy influences. Finally, in this model much of the variance remained unaccounted for, particularly between teachers. Inclusion of specific measures of teacher self-efficacy, belief systems, and motivation may account for some of the variance and should be tested in future research.

Limitations

Based on the sampling procedure, this study is only technically generalizable to the district level. The choice of focusing on one district was deliberate, and we argue that this design allowed for insights missed by other sampling procedures. Despite concern regarding generalizability, the results of this study lay the foundation for additional studies, as well as verifying qualitative studies regarding accountability pressure. With regard to causality, if this paper concerned math or ELA instruction, it is possible that test scores and thus accountability pressure may have resulted from the distribution of instructional practices, rather than being the predictor of such practices. Without conducting quasi-experimental research causality is difficult to determine; however, in this case, science instructional

practices were unlikely to predict AYP pressure as science test scores account for less than 6% of school AYP calculation.

APPENDIX

TABLE A1
Range and Mean for Schools in Valley District, 2011 (Rounded For Confidentiality)

	Range	Mean
ELL	0–55%	16%
FRL	10–95%	50%
Percentage of underrepresented students	10–85%	30%
API (on a scale of 1000)	665–950	790

REFERENCES

- Abd-El-Khalick, F., Boujaoude, S., Duschl, R. A., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., & Tuan, H.-L. (2004). Inquiry in science education: International perspectives. *Science Education*, 88, 397–419.
- Allison, P. D. (1999). *Multiple regression: A primer*. Thousand Oaks, CA: Sage.
- Anderson, K. J. B. (2012). Science education and test-based accountability: Reviewing their relationship and exploring implications for future policy. *Science Education*, 96, 104–129.
- Banilower, E. R., Heck, D. J., & Weiss, I. R. (2007). Can professional development make the vision of the standards a reality? The impact of the National Science Foundation's local systemic change through teacher enhancement initiative. *Journal of Research in Science Teaching*, 44(3), 375–395.
- Biggers, M. S. (2013). Elementary teachers' ideas about, planning for, and implementation of learner-directed and teacher-directed inquiry: A mixed-methods study. Unpublished doctoral dissertation, Iowa City, IA: University of Iowa.
- Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: University of Chicago Press.
- California Department of Education. (2012). API requirements. Retrieved from <http://www.cde.ca.gov/ta/ac/ti/schoolpireq.asp>
- 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, 497–526.
- Carlone, H. B., Haun-Frank, J., & Kimmel, S. C. (2010). Tempered radicals: Elementary teachers' narratives of teaching science within and against prevailing meanings of schooling. *Cultural Studies of Science Education*, 5, 941–956.
- Center on Education Policy. (2007). *Choices, changes, and challenges: Curriculum and instruction in the NCLB era*. Washington, DC: Author.
- Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher*, 32(6), 3–12.
- Crane, E. W., Huang, M., & Barrat, V. X. (2011). Comparing achievement trends in reading and math across Arizona public school student subgroups (REL Technical Brief, REL 2012–019). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from <http://ies.ed.gov/ncee/edlabs>.
- Cuban, L. (2004). *The blackboard and the bottom line*. Cambridge, MA: Harvard University Press.
- Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record*, 16(6), 1047–1085.
- Desimone, L. M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis*, 26(1), 1–22.
- Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, 24(2), 81–112.
- Dewey, J. (1916). *Democracy and education*. New York, NY: Macmillan Press.

- Dorph, R., Shields, P., Tiffany-Morales, J., Hartry, A., & McCaffrey, T. (2011). High hopes, few opportunities: The status of elementary science education in California. Sacramento, CA: The Center for the Future of Teaching and Learning at WestEd.
- Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education*, 38(1), 39–72.
- Duschl, R. A., Schweingruber, H. A., & Schouse, A. W. (2007). Taking science to school: Learning and teaching science in grades K-8. Washington, DC: National Research Council.
- Ed-Data. (2013). Retrieved from <http://www.ed-data.k12.ca.us/Pages/Home.aspx>
- Elmore, R. F., Peterson, P. L., & McCarthy, S. J. (1996). Restructuring in the classroom: Teaching, learning, and school organization. San Francisco, CA: Jossey-Bass.
- Fensham, P. (2009). The link between policy and practice in science education: The role of research. *Science Education*, 93, 1076–1095.
- Fulp, S. L. (2002). Status of elementary science teaching. Chapel Hill, NC: Horizon Research.
- Ginnis, I. S., & Watters, J. J. (1999). Beginning elementary school teachers and the effective teaching of science. *Journal of Science Teacher Education*, 10(4), 287–313.
- Goetz Shuler, S., Backman, J., & Olson, S. (2009). The role of assessments and accountability. In B. B. Berns & J. O. Sandler (Eds.), *Making science curriculum matter: Wisdom for the reform road ahead* (pp. 49–59). Thousand Oaks, CA: Corwin Press.
- Hamilton, L. S., Stecher, B., Marsh, J. A., McCombs, J. S., Robyn, A., & Russell, J. L. (2007). Standards-based accountability under no child left behind: Experiences of teachers and administrators in three states. Santa Monica, CA: RAND.
- Harlen, W., Holroyd, C., & Byrne, M. (1995). Confidence and understanding in teaching science and technology in primary schools. Edinburgh, Scotland: Scottish Council for Research in Education.
- Jeanpierre, B., Oberhauser, K., & Freeman, C. (2005). Characteristics of professional development that effect change in secondary science teachers' classroom practices. *Journal of Research in Science Teaching*, 42(6), 668–690.
- Joern, W. T. (2009). Investigating the relationships between seventh and eight grade science teachers' background, self-efficacy toward teaching science as inquiry, and attitudes and beliefs on classroom control. Unpublished doctoral dissertation, Missoula, MT: University of Montana.
- Judson, E. (2010). Science education as a contributor to adequate yearly progress and accountability programs. *Science Education*, 94, 888–902.
- Judson, E. (2013). The relationship between time allocated for science in elementary schools and state accountability policies. *Science Education*, 97(4), 621–636.
- Krueger, R. A. (1994). *Focus groups: A practical guide for applied research*. Thousand Oaks, CA: Sage.
- Lee, C. A., & Houseal, A. (2003). Self-efficacy, standards, and benchmarks as factors in teaching elementary school science. *Journal of Elementary Science Education*, 15, 37–56.
- Lee, O., & Luykx, A. (2005). Dilemmas in scaling up innovations in elementary science instruction with non-mainstream students. *American Educational Research Journal*, 42(3), 411–438.
- Maerten-Rivera, J., Myers, N., Lee, O., & Penfield, R. (2010). Student and school predictors of high-stakes assessment in science. *Science Education*, 94, 937–962.
- Maltese, A. V., & Tai, R. H. (2010). Eyeballs in the fridge: Sources of early interest in science. *International Journal of Science Education*, 32(5), 669–685.
- Marshall, J. C., Horton, R., Igo, B. L., & Switzer, D. M. (2009). K-12 science and mathematics teachers' beliefs about and use of inquiry in the classroom. *International Journal of Science and Mathematics Education*, 7(3), 575–596.
- Minner, D. D., Levy, A. J., & Century, J. (2009). 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.
- Montgomery, S. L. (1994). *Minds for the making: The role of science in American education, 1750–1990*. New York, NY: The Guilford Press.
- Munck, M. (2007). Science pedagogy, teacher attitudes, and student success. *Journal of Elementary Science Education*, 19(2), 13–24.
- Murphy, C., Neil, P., & Beggs, J. (2007). Primary science teacher confidence revisited: Ten years on. *Educational Research*, 49(4), 415–430.
- National Research Council (NRC). (1996). *National science education standards*. Washington, DC: The National Academies Press.
- National Research Council (NRC). (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. Washington, DC: The National Academies Press.
- National Research Council (NRC). (2012a). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

- National Research Council (NRC). (2012b). *Monitoring progress toward successful K-12 STEM education: A nation advancing?* Washington, DC: The National Academies Press.
- National Research Council (NRC). (2013). *Next generation science standards: For states, by states.* Washington, DC: The National Academies Press.
- Penuel, W., Fishman, B., Gallagher, L., Korbak, C., & Lopez-Prado, B. (2008). Is alignment enough: Investigating the effects of state policies and professional development on science curriculum implementation. *Science Education, 93*, 656–677.
- President's Council of Advisors on Science and Technology (PCAST). (2010). *Report to the President: Prepare and inspire: K-12 education in science, technology, engineering, and mathematics (STEM) for America's future.* Retrieved from <http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-stemed-report.pdf>
- Powers, J. M. (2003). Analysis of performance-based accountability: Factors shaping school performance in two school districts. *Educational Policy, 17*(5), 558–585.
- Ramey-Gassert, L., Shroyer, M. G., & Staver, J. R. (1996). A qualitative study of factors influencing science teaching self-efficacy of elementary level teachers. *Science Teacher Education, 80*(3), 283–315.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Rea, L. M., & Parker, R. A. (2005). *Designing and conducting survey research: A comprehensive guide.* San Francisco, CA: Jossey-Bass.
- Reiser, B. (2013). What professional development strategies are needed for successful implementation of the Next Generation Science Standards? Washington, DC: Paper presented at the Invitational Research Symposium on Science Assessment. K-12 Center at ETS.
- 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.
- Settlage, J., & Meadows, L. (2002). Standards-based reform and its unintended consequences: Implications for science education within America's urban schools. *Journal of Research in Science Teaching, 39*, 114–127.
- Shaver, A., Cuevas, P., Lee, O., & Avalos, M. (2007). Teachers' perceptions of policy influences on science instruction with culturally and linguistically diverse elementary students. *Journal of Research in Science Teaching, 44*(5), 725–746.
- Sirin, S. R. (2005). Socio-economic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453.
- Smarter Balanced Assessment Consortium. (2013). *Achievement descriptors.* Retrieved from <http://www.smarterbalanced.org/achievement-level-descriptors-and-college-readiness/>
- Smith, L. K., & Southerland, S. A. (2007). Reforming practice or modifying reforms? Elementary teachers' response to the tools of reform. *Journal of Research in Science Teaching, 44*(3), 396–423.
- Supovitz, J. A., & Turner, H. M. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching, 37*(9), 963–980.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental designs using ANOVA.* Belmont, CA: Thompson.
- U.S. Department of Education. (2007). *Federal government press release.* Retrieved from <http://www2.ed.gov/news/pressreleases/2006/02/02062006.html>
- Valli, L., & Buese, D. (2007). The changing roles of teachers in an era of high-stakes accountability. *American Educational Research Journal, 44*(3), 519–558.
- van Aalderen-Smeets, S., & Walma van der Molen, J. (2013). Measuring primary teachers' attitudes toward teaching science: Development of the dimensions of attitude toward science (DAS) instrument. *International Journal of Science Education, 35*(4), 577–600.
- Warburton, E., & Torff, B. (2005). The effect of perceived learner advantages on teachers' beliefs about critical-thinking activities. *Journal of Teacher Education, 56*(1), 24–33.
- Wideen, M. F., O'Shea, T., Pye, I., & Ivany, G. (1997). High-stakes testing and the teaching of science. *Canadian Journal of Education, 22*(4), 428–444.
- Yamamura, K. (2012). Gov. Jerry Brown to call for less state testing [Web log post]. Retrieved from <http://blogs.sacbee.com/capitolalert/latest/2012/01/gov-jerry-brown-to-call-for-less-state-testing-in-schools.html>