The Effect of a Leadership Development Program for High School Science Reform on Student Achievement in Science: A Retrospective Quasi-Experiment

Joseph A. Taylor
BSCS Science Learning/University of Colorado, Colorado Springs
Molly A. M. Stuhlsatz
BSCS Science Learning
Jody Bintz
BSCS Science Learning

Abstract

This longitudinal impact study examined the effects of a leadership development program for district-based teams of educators leading science education reform efforts. Propensity score matching was used to construct a comparison group of demographically similar districts who received either none or extant leadership development programming. Treatment effects were estimated at one-year intervals for eight years after the onset of the treatment program. The longitudinal pattern of treatment effects suggests that outcomes were similar for students in treatment and comparison districts in the early phases of implementation followed by a dip in outcomes for treatment districts that eventually recovers to be higher than comparison districts by 7 years from treatment onset. Analyses of data from treatment districts suggest a strong positive association between higher student outcomes and both the existence of highly effective professional learning communities and higher retention of leadership team members.

Using US economic and democratic perspectives, a significant number of policy documents and commissioned reports in the last decade (e.g., Holdren, Lander, & Varmus, 2010; Kelly, Nord, Jenkins, Chan, & Kastberg, 2013; NRC, 2010) have strongly suggested the need to improve K-12 students' STEM outcomes. Collectively, these reports recommend a variety of initiatives including the development of programs that improve both teaching and education leadership structures.

Meeting the goal of improving science education programs requires a systemic approach that considers all program elements of influence, including the professional learning of science teachers and district leaders. Designers of district-level science programs have much more information about the effectiveness of professional development (PD) for teachers than was available 15 years ago with subsequent studies of particular PD programs for student learning (e.g., Vascio, Ross, & Adams, 2008; Yoon, et al., 2007) and/or studies determining mediating variables that influence student learning (Desimone & Hill, 2017; Supovitz & Turner, 2000). Among the studies in this literature base, student effects can be linked to teacher practice (Fishman, Marx, Best, & Tal, 2003; Loucks-Horsley, Stiles, Mundry, Love, & Hewson, 2009), classroom and school culture (Supovitz & Turner, 2000), and the use of research-based instructional materials (Cervetti, Kulikowich, & Bravo, 2015; Taylor et al., 2015).

However, while the aforementioned researchers have linked science teacher learning opportunities, teacher outcomes, and student outcomes in science, few researchers have examined the effectiveness of *leader* learning opportunities for students' outcomes, with a few notable exceptions including within this journal (Jacob, Goddard, Kim, Miller, & Goddard, 2015; Leithwood, Louis, Anderson, & Wahlstrom, 2004; Louis, Leithwood, Wahlstrom, & Anderson, 2010). While these studies made significant contributions, none combined a focus on district-level leadership teams and the effects of those teams on science outcomes. The current study will amend this prior literature by doing just that, reporting on the effects a team-based leadership development program that is intended to increase students' science outcomes through systemic science program improvement.

Study Purpose

The research described in this paper is intended for two purposes: (1) to address a clear gap in the research literature by studying the impact and implementation of a leadership development intervention for district leaders

Keywords: leadership development; science education; efficacy, curriculum implementation; effective leadership.

on students' science achievement and (2) to inform decisions about the allocation of scarce resources to support professional learning. To estimate impact we take a retrospective approach that allows for the estimation of both immediate intervention effects as well as effects estimated several years after the intervention was introduced. This approach allows us to detect effects that may not have manifested soon after the intervention was completed.

Further, because several degrees of separation lie between the intervention participants and the level where student outcomes are measured, we also conducted an implementation study within the intervention districts in which we examined the effects of each of the intervention's intermediate outcomes on the ultimate outcome of student achievement. Consequently, this study (National Science Foundation, Award DRL 1316202) includes two sets of analyses. The impact analysis was based on several years of cross-sectional, intent-to-treat effects for an intensive, three-year professional development program, the NACL, on students' science achievement in the state of Washington. The second set of analyses examined the relationship between the NACL's district-level intermediate outcomes and students' achievement outcomes, also aggregated to the district level. In sum, this study sought to answer the following research questions:

1. What was the effect of the NACL on students' science achievement compared to that of business-as-usual (BaU) practices in matched districts?
2. To what extent did the intervention effect on student achievement (NACL vs. BaU comparison) change over time?
3. In intervention districts, what was the relationship between district-level intermediate outcomes and student achievement?

Theoretical and Empirical Support for the NACL

The NACL has four district-level outcomes that it is intended to promote. These intended outcomes have a strong theoretical base and empirical evidence of their potential for improving student learning. The four outcomes are as follows:

- The use of **research-based instructional materials** supported by high quality professional development.
- The development of effective **professional learning communities** that focus specifically on the implementation of research-based instructional materials.
- The development of **leadership capacity.**
- The garnering of **organizational support** to ensure sustainability of curriculum reform efforts.

Research-based instructional materials (RBIMs) and associated professional development. During the late 1990s there was a movement in science education, primarily through NSF-funded instructional materials development, to promote materials that were developed using an iterative process that focused on research around how people learn (NRC, 1999), based on national science standards (NRC, 1996), and tested in classrooms during their development. Further, the NSF supported this work through implementation centers charged with disseminating best practices for supporting the use of these materials in classrooms. (Lawton, Berns, & Sandler, 2009). The focus of the NACL on curriculum materials was further influenced by a significant literature base that documented their importance to the teaching and learning process (Ball & Cohen, 1999; Schmidt, McKnight, & Raizen, 1997; Usiskin, 1985). Specifically, the ubiquitous placement and unique ability of curriculum materials to define both what and how teachers teach (NRC, 2002) made them a clear choice as a systemic mechanism for improving student outcomes.

While this choice of mechanism was supported by a small evidence base for the efficacy of curriculum programs (e.g., Ball & Cohen, 1996), more recent syntheses, such as that from Cheung, Slavin, Kim, and Lake (2015), suggest that the effects of curriculum programs, when unsupported by professional development, are modest. This contemporary finding supports what NACL developers theorized at the time without the support of synthesis research -- that the maximum impact of instructional materials would be realized when they are supported by on-going, high quality professional development focused on implementation. Effective professional development is an important component of a system focused on continuous improvement (Hord & Boyd, 1995). Furthermore, numerous studies have been conducted that provide evidence around how curriculum materials combined with curriculum-based PD can promote effective instruction and improve student achievement (August et al., 2014; Ball & Cohen,1996; Davis & Krajcik, 2005; Lara-Alecio et al., 2012; Lee, Deaktor, Hart, Cuevas, & Enders, 2005; Lynch, Kuipers, Pyke, & Szesze, 2005; Roth, Anderson, & Smith,1987; Schneider & Krajcik, 2002; Supovitz & Turner, 2000).

However, the mere existence of curriculum-based professional development is no guarantee that implementation and student outcomes will improve. Significant research and theoretical work has explored the characteristics of effective professional development, and this work guided the design of the NACL. Specifically, the NACL program sought to help school districts design curriculum-based and other related professional development that was consistent with the vision for science teaching and learning promoted by national standards (NRC, 1996) and the work of Loucks-Horsley and her colleagues (2003) who defined effective professional development experiences as

1. being driven by a clear, well-defined vision of effective classroom learning and teaching;
2. providing teachers with opportunities to develop knowledge and skills and broaden their teaching approaches so they can create better learning opportunities for students;
3. using instructional methods to promote learning for adults that mirror the methods to be used with students;

4. building or strengthening the learning community of science and mathematics teachers;
5. preparing and supporting teachers to serve in leadership roles if they are inclined to do so;
6. providing links to other parts of the educational system; and
7. including continuous assessment.

Professional learning communities (PLCs). The development of professional learning communities for teachers implementing common instructional materials is an important outcome of the NACL, in part because their existence, persistence, and effectiveness is indicative of a learning organization conducive to change and engaged in ongoing improvement efforts (Boyd, 1992; Boyd & Hord, 1994; Gamoran et al., 2003; Hall & Hord, 2011; Supovitz, 2002; Supovitz & Christman, 2003). Further, there is evidence of a relationship between existence of PLCs and student outcomes. For example, in a meta-analysis of PLCs, of the 11 studies that met the inclusion criteria, eight indicated that student outcomes improved as a result of PLCs (Vescio, Ross, & Adams, 2008).

The design of the NACL drew heavily on the synthetic work of educational researchers studying the influence of communities of practice on student achievement (e.g., DuFour & Eaker, 1998; DuFour, DuFour, Eaker, & Karhanek, 2004; Hord, 2007; Kruse, Louis, & Bryk, 1994; Louis, 2006; Roy & Hord, 2007). Kruse, Louis, and Bryk (1994) synthesized data from schools studied by the Center on Organization and Restructuring of Schools (CORS, 2016) and found evidence for six characteristics that can contribute to strong professional learning communities. These are described below.

First, *shared norms and values* form the foundation for all aspects of developing a professional learning community (Garmston & Wellmann, 2009; Hall & Hord, 2011; Hord & Sommers, 2008; Kennedy, Slavit, & Nelson, 2009; Kruse & Louis, 2009; McLaughlin & Talbert, 2006). Teachers and administrators should reinforce their own understandings about children and learning, teaching, and teachers' roles and attend to the nature of human needs, activities, and relationships within the organization. Second, strong professional learning communities place sustained *attention on students*, thereby emphasizing how pedagogy is linked to the process of student learning (Bolam et al., 2005; Cochran-Smith & Lytle, 1999; Louis & Marks, 1998; McLaughlin & Talbert, 2006; Supovitz, 2002; Supovitz & Christman, 2003). A PLC's focus on new or existing curricular materials provides a strong context for a collective focus on student learning. Third, *collaboration* is characterized by mutual learning that comes from joint planning of future teaching activities and ways to support improved learning (Garmston & Wellman, 2009; Hall & Hord, 2011; Hord & Boyd, 1995; Hord & Sommers, 2008; Kennedy, Slavit, & Nelson, 2009). Collaboration should involve the co-development of skills related to new practice and the generation of knowledge, ideas, and programs that advance expertise and school performance. Fourth, *deprivatized practice* involves teachers openly practicing their craft through sharing roles as mentors and advisors in order to provide help to their peers (Hall & Hord, 2011; Hord & Sommers, 2008; Kruse & Louis, 2009). Deprivatized practice within schools involves the development of peer coaching relationships or critical friends and teachers resolving issues by bringing real teaching problems to a discussion as well as engaging in classroom observations to provide evidence for discussions. Fifth, through *reflective dialogue*, or the practice of drawing out deep thinking about pedagogical practice, PLCs encourage self-awareness within the community about praxis, closing the gap between theory and practice, reducing isolation, and becoming students of their craft (Hall & Hord, 2011; Kruse & Louis, 2009). Finally, *relational trust* between PLC members, in students, and in parents is the glue that holds a PLC community together (Garmston & Wellman, 2009; Stoll et al., 2006; Bryk & Schneider, 2003). Elements of trust include respect, personal regard for others, competence, and integrity. While the establishment of a PLC is good for the adults in the school system, the ultimate outcome of interest is improvement in student learning (Carroll, Fulton, & Doerr, 2010; Fulton & Britton, 2011; Vescio, Ross, & Adams, 2008).

Leadership and organizational support. Organizational support for education innovation as well as strong teacher-, school-, and district-level leadership are paramount in sustaining district-level efforts such as those promoted by the NACL (Fullan, 2014; Hall & Hord, 2011; Loucks-Horsley, Love, Stiles, Mundry, & Hewson, 2003). In a review of the literature around sustaining educational reforms, Coburn (2003) notes that while successful adoption of new practices is relatively common, it is the sustainability and ultimately the scaling of those practices that is most difficult to achieve, especially when considering the school system as a constantly changing institution pulled in many competing directions. Coburn further notes that the arrival of resources to support the new practices is an important first step, but over time the infrastructure must also be supported if the reform effort is expected to be sustained.

One essential aspect of teacher leadership development is the sharing of the responsibilities of the reform effort. Teacher leaders provide instructional expertise and leadership and use a variety of leadership practices (e.g., take part in school decisions, share ideas with colleagues) to improve instruction beyond their own classrooms (Pellicer & Anderson, 2001). Yet, too often teacher leaders must "figure out how to lead on their own" (NRC, 2014, p. 36). Without a coherent system that supports ongoing development of science teacher leaders, there is little to draw from for insight or guidance into how science teachers develop as leaders (Lieberman & Friedrich, 2010) or their effectiveness in the classroom (Center for Comprehensive School Reform and Improvement, 2005).

Prior Evidence of NACL Effectiveness

The NACL was originally designed and tested as a PD model in response to the call from NSF to increase the awareness and implementation of NSF-funded curriculum materials (Lawton, Berns, & Sandler, 2009). The first iteration of the NACL was supported by NSF and was implemented nationally with 20 district teams in 13 states from 2000 to 2004. Findings from prior research and evaluation of the NACL showed promise of efficacy for improving leadership outcomes. However, while these prior studies of the NACL investigated development, implementation, and dissemination of the program, they did not estimate effects on student achievement.

Evidence of the program's quality and influence on district reform practices was collected and analyzed during the NSF-supported research and development phase from 2000 to 2007 and in a regional enactment of the NACL PD model from 2004 to 2010. Findings from these 10 years of work with more than 40 secondary science leadership teams reveal a number of patterns and themes. Evaluation data from the national NACL model indicated that improving the curriculum requires "*resources, leadership, time, and a well-tested and engineered process*" (St. John et al., 2006, p. 8). It requires expertise at all levels in the school system, including those who have a deep understanding of the processes of selecting and implementing curriculum materials, the capacity to work with a range of teachers and others throughout the change process, and tools to inform their work (St. John et al., 2006). Based on a pre/post self-report survey, the NACL significantly impacted the understandings, beliefs, and skills of the participants with respect to (1) science teaching and learning, (2) the role of instructional materials, (3) selection and implementation of research-based instructional materials, (4) professional development design and the development of a professional development plan, (5) leadership qualities and practices, (6) leading change, and (7) team development (BSCS, 2011). Further, the NACL helped districts develop greater capacity to provide ongoing, curriculum-based professional development (Bintz & Martin, 2007; BSCS, 2011) and provided evidence that teachers who participated in curriculum-based professional development used research-based instructional materials with higher fidelity (i.e., in ways consistent with the designers' intention).

Method

Participants

In the current study, twenty-seven school districts in two cohorts formed the intervention group. Data from both cohorts were combined for the impact analysis as the programmatic experience in the intervention group did not differ significantly across cohorts. The districts were required to apply to participate in the NACL. Representatives from the NACL development and implementation teams screened applications and made acceptance decisions based on a set of readiness factors for district-based curriculum reform. These readiness factors included the following: inclination to select and implement instructional materials, a leader in the district prepared to recruit and form a team willing to participate in the NACL, support from district leaders in the form of in-kind monetary commitment for participation, a district pledge to cover out-of-school event time, a signed contract of commitment to the program, and an a clear charge to improve in the area of secondary science reform. All school districts in the state of Washington that had not participated in the NACL were eligible to be in the comparison group.

Within school districts in the state of Washington, 10th grade state science achievement scores (the study outcome measure) were collected for nearly all students who remained enrolled at the time of testing. The only exceptions to this requirement were special education students for which the standard state test (or an alternative assessment) was not required as part of their individualized education plan.

Measures

Student assessment. The outcome measure for this study was each district's *percent met standard* on the grade 10 state standardized test in science, and the baseline achievement measure was the percent met standard on the grade 8 state test in science. Over the course of the study, the student assessment had three different names: the *Washington State Assessment of Student Learning* (WASL), the *High School Proficiency Exam* (HSPE), and the *Biology End of Course* (EoC) *Assessment.* Grade 8 percent met standard on the state reading test was also used in this study (see Table 2). *Met standard* in the Washington State testing context means that a student has achieved proficiency (level 3 of 4 achievement levels), where proficiency is defined as demonstrating adequate understanding of/ability to apply skills in the required testing domains.

These district-level achievement data are publicly available on the Office of Superintendent of Public Instruction (OSPI) website (OSPI, 2016). The state test for 10th grade science was referred to as the WASL between 2004 and 2009, the HSPE between 2010 and 2011, and the Biology EoC test since 2012. All versions of the test demonstrated favorable psychometric properties (see Table 1), with internal consistency reliabilities ranging from 0.87 to 0.92 for the years relevant to the present study.

Administrative data. The OSPI website also provides a comprehensive set of district-level administrative data, including demographics. The district-level variables used in this study include total enrollment, percent black, percent free or reduced-price lunch, and percent reading met standard.

Leadership activity. The frequency and depth of each team's work to support curriculum reform was assessed using a rating scale responded to by each member of the district leadership team (see rating scale in Supplementary File S1). The ratings from each leadership team member were averaged to form an overall district score for use in the regression models of the implementation study.

District-level (organizational) support. District-level support for curriculum reform was assessed using an Organization Support rating scale (see Supplementary File S2) developed by Guskey (2000). The ratings from each leadership team member were averaged

Table 1. Internal Consistency Reliability (alpha) for the Washington State Science Test

Year	Grade 10 test	Internal consistency (Cronbach's alpha)
2004	WASL Science	.92
2005	WASL Science	.91
2006	WASL Science	.89
2007	WASL Science	.89
2008	WASL Science	.89
2009	WASL Science	.91
2010	HSPE Science	.90
2011	HSPE Science	.87
2012	EoC Biology	.89
2013	EoC Biology	.89
2014	EoC Biology	.88

to form an overall district score for use in the regression models of the implementation study.

Engagement in professional learning communities. The frequency and depth of district-level involvement in professional learning communities were assessed using a survey completed by everyone in each district leadership team (see survey questions in Supplementary File S3). The ratings from each leadership team member at year 3 in their involvement were averaged to form an overall district score for use in the regression models of the implementation study.

Exposure to research-based instructional materials (%RBIMs). The percentage of students who had access to research-based instructional materials was assessed using a brief survey completed by district leadership. For each outcome year, we combined the percentage of 10th graders with RBIM exposure in that year with the percentage of 8th and 9th graders with exposure to RBIMs in each of the previous two years. For example, in analyses that estimate the relationship between RBIM exposure and students' year 2 science outcome score, the RBIM score for a given cohort is a composite of the 10th grade percent exposure to RBIMs (for year 2), the 9th grade percent exposure to RBIMs for the same cohort in the prior year (year 1), and the 8th grade percent exposure to RBIMs for that same cohort (two years prior or baseline year).

Team retention. The number of leadership team members who remained with the district as of 2015 was assessed using a single-item questionnaire responded to by the highest ranking remaining member of the district leadership team.

Procedure

The intervention condition. The mission of the three-year NACL program was to develop leadership teams that assist schools and districts in building the capacity to design, implement, and sustain effective high school science education programs using research-based instructional materials. The ultimate intention of the program was to improve student outcomes by first helping districts attain the four intermediate outcomes described above:

- Use of **research-based instructional materials** supported by effective professional development
- Development of **professional learning communities** focused on student learning
- Development of **leadership capacity**
- Increased **organizational support** to ensure sustainability of the other related efforts

Each leadership team consisted of a key administrator, a coach, and at least three teacher-leaders who participated in NACL institutes and on-site technical assistance. The coach and key administrator of each team took part in a 4½-day leadership institute. The purpose of this institute was for team leaders to gain insight into the overall program for that year. The whole team then took part in two institutes, several months apart, with the first spanning 5 days and the second 2½ days. In between the institutes, program staff offered on-site technical assistance specified by the needs of a team or group of teams. The NACL included over 88 contact hours of PD plus a variable amount of technical assistance at the district-site annually for three years.

Leadership teams were supported in utilizing effective communication strategies, such as norms of collaboration (Garmston & Wellman, 2009) and protocols for examining classroom artifacts. The leadership teams developed into professional learning communities through their participation in the NACL and with the intention of seeding the development or enhancement of communities back in their local context (Bintz & Martin, 2007). Leadership teams learned about, practiced, and reflected on qualities of effective leaders, leadership practices, and characteristics of effective teams as they developed, implemented, and revised their local professional development program. They learned about leading change, studying change principles, and applying lessons learned in leadership simulations and in reflecting on their advocacy and other leadership activities back in their local contexts.

Throughout the program, high school science leadership teams were expected to further develop and leverage the guiding principles to improve their high school science programs. As such, the program experienced by the participants in the study included multiple opportunities for teachers and administrators to be immersed in science lessons consistent with the National Science Education Standards (NSES: NRC, 1996; NRC, 2000), the dominant national standards at the time, and that embody research on effective science teaching and learning (NRC, 2000; NRC, 2005). Participants reflected on their experiences and studied the NSES and selected research on effective science teaching and learning. They applied this learning in two ways--through the collaborative analysis of instructional materials and classroom video of science instruction. Teams analyzed instructional materials using a specific process called

AIM (Landes, Powell, & Short, 2004; Taylor, Gardner, & Bybee, 2009). AIM is designed to help teams of stakeholders evaluate instructional materials to inform the selection and implementation of the new materials. These experiences were intended to deepen participants' knowledge and inform the shared vision of the leadership teams. This vision informed the teams' designs of their local professional development program.

Each local professional development program emerged through a comprehensive process based on a framework for designing professional development for math and science teachers recommended by Loucks-Horsley, Love, Stiles, Mundry, and Hewson, (2003). Over the course of the three-year leadership development program, teams revised their professional development plans through an iterative development and revision process that was informed by the careful analyses of student learning and other evaluation data. The feedback and technical assistance offered to teams by NACL providers focused on the consistency of the program design and plan with selected principles of high quality professional development programs.

The comparison condition. The comparison condition was business as usual at all levels of the education system. That is, while the intervention group was participating in the NACL, the 25 comparison districts proceeded with extant: professional development for district leaders, professional development for high school science teachers, and 9th and 10th grade science curriculum.

Design. In this study we made a retrospective, quasi-experimental comparison of district-level outcomes between 27 districts that completed all years of the NACL program and 25 comparison districts that were identified as comparable using a nearest neighbor matching process based on the Mahalanobis distance (Stuart, 2010). The Mahalanobis distances were estimated using the *MatchIt* subroutine within *R* (Ho, Imai, King, & Stuart, 2007). This subroutine estimates distances using a logistic regression model that regresses group assignment as a function of district-level baseline achievement and demographics. We conducted the matching separately for each of the two cohorts of NACL districts, matching each intervention district to one comparison district that looked most similar to it in the year prior to NACL implementation.

Matching results and baseline equivalence. We examined the success of the matching approach by comparing the balance (similarity) of the observed characteristics in the matched samples (Stuart, 2010). In Table 2, we provide the balance statistics for the matching process for each cohort. Each value in the table is a standardized mean difference, by each covariate, between the intervention group districts and either the entire state (original data) or the set of 1:1 nearest neighbor matches.

In accordance with the reporting recommendations of the What Works Clearinghouse (WWC), we provide in Table 3 descriptive statistics for an analytic sample of districts on the baseline achievement measure (met standard on the grade 8 state science test).

Data file construction. For both the impact and the implementation study, data for the intervention NACL group were combined across the two cohorts of districts. For example, as cohort 1 began in the 2003/2004 school year and cohort 2 in 2006/2007 school year, estimation of year 1 effects required that we combine outcomes from the 2005 and 2008 administrations of the WASL. Table 4 illustrates how the data were combined for each effect.

Statistical Methods

Intent to treat impact estimates. The effect of the NACL was estimated using a multiple regression approach (see Equation 1) that modeled district-level achievement on the 10th grade state science test (percent of students meeting standard on the test: %METSCI10) as a function of the intervention indicator (TREAT) and five district-level, grand-mean centered covariates: percent met standard on the 8th grade science achievement test (%METSCI8), percent met standard on the 8th grade reading achievement test (%METREAD8), percent of students receiving free or reduced-price lunch (%FRL), percent of students who are white (%WHITE), and the total district enrollment (ENROLL). These covariates were chosen because the empirical literature suggests that they are often highly correlated with cluster-level student outcomes in science (see Spybrook, Westine, & Taylor, 2016, Westine, Spybrook, & Taylor, 2013), thus adding precision to our estimate of the treatment effect. For example, Spybrook, Westine, and Taylor (2016) observed large cluster covariate correlations for prior year's science, prior year's reading, and demographics, with R^2 values of .90, .74, and .74 respectively.

$$\begin{aligned}\%METSCI10 = b_0 &+ b_1(\text{TREAT}) \\ &+ b_2(\%\text{METSCI8}) \\ &+ b_3(\%\text{METREAD8}) \\ &+ b_4(\%\text{FRL}) \\ &+ b_5(\%\text{WHITE}) \\ &+ b_6(\text{ENROLL}) + e \qquad (1)\end{aligned}$$

In this model, b_1, the intervention effect, is the covariate-adjusted difference across intervention groups in the percent of students meeting standards on the 10th grade state science achievement test. The effect size statistic used for all impact estimates is Hedges' g, calculated as suggested by the *What Works Clearinghouse*

Table 2. Matching Diagnostics consist of standardized mean-difference effect sizes comparing the baseline outcomes characteristics of intervention districts, by cohort, to the state at large and to the set of nearest neighbor comparison districts.

	Cohort 1		Cohort 2	
	Entire state	Matched districts	Entire state	Matched districts
Total enrollment	0.81	0.11	0.60	0.12
FRL	−0.22	0.01	0.40	0.13
Percent white	−0.34	−0.15	−0.73	−0.08
Reading percent met standard	0.01	0.03	−0.44	−0.36

Table 3. Baseline Descriptive Statistics

Group	n (districts)	Mean (percent met standard in science grade 8)	Standard deviation (percent met standard in science grade 8)	Baseline equivalence effect size (Hedges' g)
Treatment	26	37.85	15.24	**.04**
Comparison	25	37.18	15.06	

Procedures and Standards 3.0 (IES, 2016) using the intervention effect from equation 1 as the numerator of the standardized mean difference.

Implementation study. For the implementation study using only outcome and implementation data from the NACL districts, we modeled using Equation 2 the outcome %METSCI10 as a function of baseline achievement (%METSCI8) and the intermediate outcomes that we hypothesized to be in turn influential on student outcomes: use of research-based instructional materials (%RBIM), leadership practices (LEAD), use of professional learning communities (PLC), and organizational support (ORG). These intermediate outcomes were taken from the theoretical model for why the NACL was hypothesized to positively influence student outcomes.

$$\begin{aligned}\%METSCI10 &= b_0 + b_1(\%\text{METSCI8}) \\ &\quad + b_2(\%\text{RBIM}) + b_3(\text{LEAD}) \\ &\quad + b_4(\text{PLC}) + b_5(\text{ORG}) + e \qquad (2)\end{aligned}$$

Results

Impact Study (Research Questions 1 and 2)

In Table 5, we provide descriptive statistics for the intervention effects estimated for years 1–8. Adjusted means in percent met standard were derived from regression model-based estimates (intervention effect coefficient and intercept).

Using Equation 1, we estimated intervention effects at each of eight years after the onset of the NACL intervention, pooling the outcome data across two sequential cohorts of intervention and comparison districts (see Table 4). The unstandardized regression coefficients in Table 6 represent the average difference in percent proficient across intervention and comparison districts. For example, in year 8 the covariate adjusted mean difference between groups was just over three percentage points in favor of the NACL group. That is, the intervention districts were estimated to have, on average, 3% more students meeting the proficient level, controlling for the covariates in Equation 1. The effect sizes that were based on these regression coefficients ranged from –0.14 to 0.13, with the largest effects being noteworthy given the degrees of separation between the intervention and the classroom.

Implementation Study (Research Question 3)

The implementation study examined the relationship between district-level intermediate outcomes and student achievement (for NACL districts). Estimated relationships between intermediate outcomes and student achievement are discussed in the following sections.

Table 4. Combining Data Across Cohorts

Effect	Cohort 1 outcome data drawn from	Cohort 2 outcome data drawn from	Baseline achievement data combined from	Other baseline covariates combined from
Baseline	-	-	2004 8th grade WASL (Cohort 1) and 2007 8th grade WASL (Cohort 2)	Administrative data from 2004 (Cohort 1) and 2007 (Cohort 2)
Year 1	2005 WASL (10th grade)	2008 WASL (10th grade)		
Year 2	2006 WASL (10th grade)	2009 WASL (10th grade)		
Year 3	2007 WASL (10th grade)	2010 WASL (10th grade)		
Year 4	2008 WASL (10th grade)	2011 HSPE (10th grade)		
Year 5	2009 WASL (10th grade)	2012 HSPE (10th grade)		
Year 6	2010 WASL (10th grade)	2013 EoC Biology (10th grade)		
Year 7	2011 HSPE (10th grade)	2014 EoC Biology (10th grade)		
Year 8	2012 HSPE (10th grade)	2015 EoC Biology (10th grade)		

To help bridge the divide between the level of the system where the intervention was delivered and the level of the system where the outcomes were measured, we chose to test the relationship between student achievement and several intermediate outcomes of the NACL, two of which are much closer to the classroom. Descriptive statistics for these intermediate outcomes are provided in Table 7 for reference. The relationships, estimated using Equation 2, are summarized in Table 8. Found in Table 8 is a variable by variable comparison of coefficients of two regression models, one for year 3 and one for year 7, and their associated statistics, respectively. Standard regression output provides a t-statistic which is the statistic that tests whether the regression coefficient is different than zero. The value of the t-statistic is the ratio of the coefficient to the standard error and is conceptually consistent but not computationally identical to the basic t-test that compares a continuous variable across levels of a binary categorical variable.

More specifically, Table 8 includes standardized regression coefficients that estimate the relationship between district-level student outcomes and NACL intermediate outcomes, for the outcome years where sufficient data existed (years 3 and 7). As expected, baseline achievement was strongly related to outcomes, but we also note here relatively strong relationships for the Team Remaining and PLC variables. In addition to these main effects, we tested all possible two-way interactions among these intermediate outcomes and observed none to be statistically significant or otherwise noteworthy.

Discussion

Impact Study: The Magnitude of Effects

In this study, observed intervention effects ranged in magnitude from –0.14 to 0.13. Some of these effects (particularly those from years 3 and 7) are somewhat

Table 5. Descriptive Statistics for Percent Met Standard in Science (10th grade)

	Treatment group				Comparison group			
Year	Mean	Adjusted mean	Standard deviation	n (districts)	Mean	Adjusted mean	Standard deviation	n (districts)
1	36.01	36.46	14.59	27	36.63	36.36	13.10	24
2	35.24	34.98	11.92	25	34.32	34.76	12.54	25
3	36.68	37.47	13.36	27	37.59	39.49	13.55	25
4	36.47	36.85	14.25	27	36.73	38.29	15.05	25
5	46.19	46.39	17.46	26	45.70	47.71	17.58	25
6	50.83	51.11	19.63	26	50.28	51.69	17.33	25
7	55.43	54.89	17.77	26	52.36	52.29	19.21	25

larger than what synthesis studies might predict for studies of high school interventions where a state standardized test is used as an outcome measure. For example, Hill, Bloom, Black, and Lipsey (2008) conducted a synthesis of effect sizes for randomized control trials and found that for high school studies the average effect size was 0.27. However, this high school average included effect sizes computed on outcome measures that were proximal or targeted to the intervention. Based on what Hill and colleagues observed in the elementary school studies, inclusion of proximal effect sizes likely inflated the average effect size for high school interventions. Specifically, they found that the average effect size varied by the breadth of focus for the outcome measure, reporting on page 8: "within studies of elementary schools, mean effect sizes are highest for specialized tests (0.44), next-highest for narrowly focused standardized tests (0.23), and lowest for broadly focused standardized tests (0.07)." Further, a synthesis conducted by Louis, Leithwood, Wahlstrom, and Anderson (2010) found that the effects of leadership programs on student achievement tend to be small but positive. Finally, Jacob, Goddard, Kim, Miller, and Goddard (2015), in a more recent leadership impact study published in this journal, observed a range of modest effects on student achievement, with the largest just 0.04 standard deviations. Given that the effect sizes reported for the NACL were computed using scores from a broadly focused standardized test (WASL/HSPE/EoC 10 Science), we find that they are noteworthy in the context of similar studies.

Another way to interpret the magnitude of the intervention effects is to compare them to normative expectations for achievement growth (i.e., average pre-post year effect sizes for 9th and 10th grade science students). This effect size expresses students' expected gain in science achievement over the course of one year. Looking across a set of nationally normed tests, Bloom and colleagues (2008) estimated that the average pre-post year effect size for science in 9th grade is 0.22 and 0.19 for 10th grade. Thus, the two-year expected gain in achievement can be estimated as 0.41 standard deviations. For example, the effect size of 0.13 detected in this two-year intervention (for students) study is noteworthy as it corresponds to 0.13/0.41 or 32% of the two-year expected gain. Multiplying 0.32 by 18 school months for a two-year intervention, we estimate that intervention-group students emerge from the study (i.e., start 11th grade) nearly 6 school months ahead of comparison-group students in science achievement. The year 3 decrease in achievement (relative to comparison group) is also noteworthy as it corresponds to -0.14/0.41 or -34% of the two-year expected gain (just over six school months).

Finally, another way to express the practical importance of the intervention effect is to convert the effect size into an improvement index using the properties of the normal distribution. In a normal distribution, a 1.0 standard deviation effect size is equivalent to approximately 34 percentile points. Therefore, an effect size of 0.13 equates to an improvement index of 4.42 (34 × 0.13) percentile points. This suggests (hypothetically) that if the students in the comparison districts were on average at the mean of a normed sample, the 50th percentile, then students in intervention districts would score on average at the 54th percentile. Again, the improvement index for year 3 should also be noted and that index is 4.76 (34 x 0.14) percentile points in favor of the comparison group.

Impact Study: Change in Effects Over Time

The findings of this study suggest that the effects of NACL became positive after several years of participation. The longitudinal pattern of cross-sectional effects suggests that there was an initial negative effect of the NACL with an eventual recovery by year 6, after which the effect became positive. This pattern is consistent with observations from the systemic change literature (Fixsen et al., 2005; Fullan, 2001) suggesting that education reform efforts often include an implementation "dip." In this context, such

Table 6. Intervention Effects Over Time

Year	Treatment effect (unstandardized coefficient, B_1)	Standard error	t	p	Hedges' *g* (95% CI)	Improvement index (%)
1	0.10	2.01	0.05	.96	.01 (–.54, .56)	0.24
2	0.22	1.77	0.13	.90	.02 (–.54, .57)	0.61
3	–1.95	2.19	–0.89	.38	–.14 (–.69, .40)	–4.83
4	–1.39	2.79	–0.50	.62	–.09 (–.64, .45)	–3.17
5	–1.30	2.58	–0.50	.62	–.07 (–.62, .48)	–2.49
6	–0.56	2.85	–0.20	.85	–.03 (–.58, .52)	–0.99
7	2.55	2.46	1.02	.31	.13 (–.42, .68)	4.54

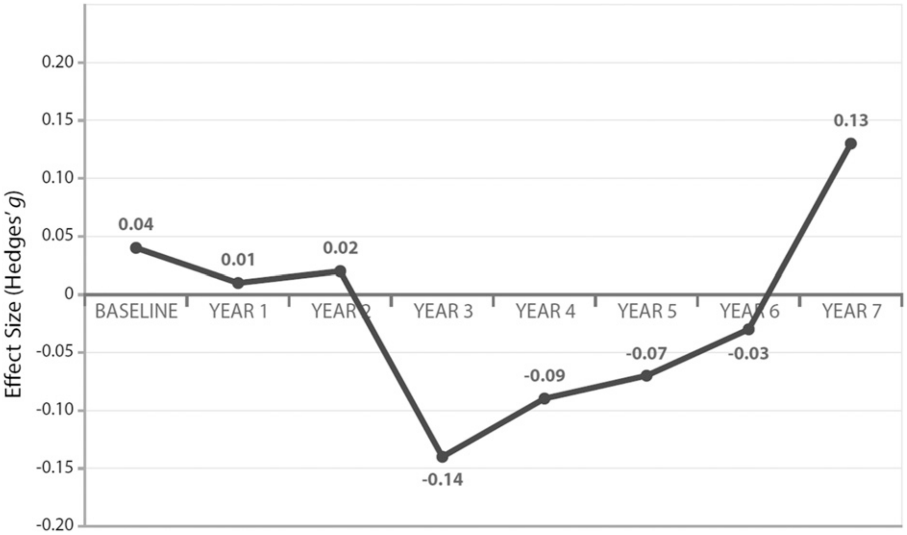

Figure 1. Intervention effects (Hedges' *g*) from baseline through eight timepoints after onset of the NACL.

a dip means that leader practice, teacher practice, and student learning are likely to show an initial decline as leaders and teachers first implement an unfamiliar practice before a rebound when student gains are realized (Fixsen et al., 2005; Fullan, 2001; Fullan, 2014). Few implementation studies have documented the length of time required for implementers to become comfortable and effective with new programs or practices, but some estimate that at least two to four years of use are required before participants move from "awkward" implementation to the full and sustained implementation that can lead to improved outcomes (Fixsen et al., 2005). Furthermore, there is some evidence from the implementation study (described more fully below) that the timing of the dip and recovery in student outcomes could be linked to a time-corresponding dip and recovery in the implementation of instructional materials.

Implementation study

Professional learning communities. The strong relationship observed in this study between the implementation of PLCs and student achievement is consistent with the results of meta-analytic work (see Vescio, Ross, & Adams, 2008) as well as a small but growing set of more recent empirical studies (e.g., Gallimore, Ermeling, Saunders, & Goldenberg, 2009; Saunders, Goldenberg, & Gallimore, 2009). For example, Vescio, Ross, and Adams (2008) reviewed research on the impacts of PLCs on student outcomes. This synthesis of 11 empirical studies drew the overall conclusion that a positive relationship occurred between the existence of well-developed PLCs and student achievement.

Table 7. Descriptive Statistics for Intermediate Outcomes (NACL Districts Only)

Intermediate outcome	Maximum possible score, frequency, or percentage	Mean Score, Frequency, or Percentage	Standard Deviation	Number of Districts
Team mean frequency of leadership practices	20.00	13.42	2.28	19
Team mean frequency of PLC practices	30.00	19.54	2.60	19
Team mean organizational support rating	72.00	51.71	7.78	19
Percent use of research-based instructional materials	100.00%	4.80	3.51	24
Percent team remaining	100.00%	5.42 (50-60%)	3.22	24

Leadership team sustainability. Although we found no other studies that linked turnover within a district team structure (i.e., like that of the NACL) to student learning, the positive relationship between team member retention (a proxy for sustainability) and student achievement observed in this study is consistent with extant literature on superintendent and principal turnover. For example, Waters and Marzano (2006) conducted a meta-analysis that observed a summary correlation of 0.19 ($p < .05$) between superintendent tenure and student achievement. Similarly, Burkhauser, Gates, Hamilton, and Ikemoto (2012) studied first-year principals from 40 urban schools across the United States and found the loss of a principal was a strong predictor of decline in student scores the following year. We suggest that some of the same mechanisms that make superintendent and principal turnover influential on student achievement (e.g., program coherence and support) were likely at play in the relationship between district team sustainability and achievement that we observed in this study.

Research-based instructional materials. The NACL encouraged the use of instructional materials that were based on extant research on science teaching and learning. Many of the instructional materials supported by the NACL had been developed with funding from the NSF (Bintz & Martin, 2007). While all the instructional materials promoted by the NACL were designed to be consistent with research, only the program described by Taylor and colleagues (2015), at the time this article was published, had been supported by an impact study that met the evidence requirements of the What Works Clearinghouse or other exclusive syntheses such as that of Cheung and colleagues (2015). In light of the sparse literature on the effectiveness of the instructional materials supported by the NACL, we find the mixed effects of instructional materials exposure (negative in year 3, positive in year 7)

Table 8. Implementation Study: Effects of NACL Intermediate Outcomes on Student Achievement (n = 19 districts with complete data)

Intermediate outcome	Student outcome year	Unstandardized coefficient	Standard error	t	p	Standardized coefficient
Leadership	Year 3	–0.67	0.87	–0.77	.46	–0.10
	Year 7	–0.06	1.27	–0.04	.97	–0.01
Organizational support	Year 3	–0.05	0.27	–0.16	.87	–0.02
	Year 7	–0.18	0.52	–0.35	.73	–0.07
PLC	Year 3	2.05	0.88	2.33	.04	0.33
	Year 7	2.09	1.15	1.82	.10	0.25
RBIMs	Year 3	–0.58	0.75	–0.77	.46	–0.12
	Year 7	0.63	0.93	0.68	.51	0.10
Team remaining	Year 3	–	–	–	–	–
	Year 7	2.33	1.19	1.96	.08	0.37
Baseline achievement	Year 3	0.70	0.16	4.39	.001	0.78
	Year 7	1.04	0.22	4.78	.001	0.85

Note: For both models (Year 3 and Year 7), the F-test of overall model significance was statistically significant (Year 3: p <.001, Year 7: p = .001) with corresponding R^2 values of .85 for Year 3 and .84 for Year 7.

consistent with the current state of ambiguity in the knowledge base. Again, although the magnitude of the effects was mixed, the timing of the effects was more interesting. That is, the fact the effect of instructional materials exposure was negative in year 3 and then positive in year 7 (see Table 7) supports our theory-based assertion that the dip and recovery in student outcomes was synced with a dip and recovery in the implementation of new instructional materials.

Leadership and Organizational Support. We hypothesized a positive relationship between achievement and levels of organizational support and between achievement and the frequency with which team members engaged in leadership practices. Although not statistically significant, our results are suggestive of a negative relationship. These results are difficult to interpret at this time as the extant literature does not appear to have any comparable studies that have examined in a direct way the size and direction of these relationships.

Limitations and Validity Threats

Impact study. A notable internal validity threat is the lack of information about potential confounds to the intervention effect. Although we were able to gather some information through the retrospective study of implementation about competing initiatives or other factors that could feasibly influence outcomes for intervention districts, we have virtually none of this contextual information for the comparison districts. Further, it is important to acknowledge that the intervention districts' desire and readiness for curriculum-based reform might have resulted in a type of unmeasured volunteer effect that affected outcomes. Therefore, although we are not aware of a specific confound that differentially influenced one or both groups, we cannot say conclusively that such a factor did not exist. This limitation is accentuated in a retrospective study like the one reported here where the long window of time not only allows for helpful longitudinal analyses but also increases the likelihood that potentially confounding initiatives could emerge in the interim.

Finally, a formal test of the extent to which the intended district-level intermediate outcomes mediated the effect of the NACL on student outcomes could not be pursued in this study. Formal mediation analyses would have required us to collect evidence about those intermediate outcomes in the comparison group. The comparison districts had not been identified at the beginning of the study (i.e., were not known until after the matching process) and thus had not consented to the research.

Implementation study. In this study, our knowledge about use of research-based instructional materials was limited to the proportion of students in a given district who were exposed to those materials in each of the two years leading up to the outcome comparison year. Beyond that, intervention districts varied in the nature of materials implementation in 9th and 10th grade. For example, the instructional materials selected and implemented varied. Some districts implemented biology-focused RBIMS, others implemented physical science programs or multidisciplinary programs, and others implemented a combination of these programs across the 9th and 10th grades. Further, we did not have access to individual teacher outcomes including fidelity of implementation measures or to student-level achievement (scale) scores. Both additional measures might have been valuable in interpreting the intervention effects or in refining our assessment of student impacts. Future studies including these measures may be warranted. Finally, some implementation data were self-report in nature, based on survey items that were highly customized to NACL participants (i.e., limiting the generalizability of the results) and/or based in the perceptions and memory of a single individual.

Implications for Future Research

Impact. It will be important to track the longitudinal pattern of cross-sectional effects to determine whether the current achievement advantage experienced by intervention districts is sustained. Over time, we plan to develop relationships with comparison districts that would allow for data collection in those districts. For example, contextual information from comparison districts could support future moderator analyses of the intervention effects. Similarly, collecting information in comparison districts related to the NACL intermediate outcomes would support the formal mediation analyses mentioned above and would have the potential to contribute to theory building and revision of the NACL program.

Clearly, replicating this intervention and subsequent impact study would contribute to the robustness of our findings. However, a strategy that would provide reproducibility evidence in a more timely way would be to examine the retrospective impacts of other leadership development programs that share many of the same intervention components as the NACL. These include the Institute for Learning (Learning Research and Development Center, 2017) and the K-12 Alliance NGSS Science Institutes (WestEd, 2017). Comparing outcomes across the three programs would constitute one of the few (if any) attempts at the comparative research Borko (2004) identified as critical to well-informed policy decisions.

Implementation. A limitation of the implementation study was that we used measures of leadership practices unique to this intervention and measures of organizational support that have not been widely used, if at all, in other impact studies of leadership programs. This made it impossible to situate our findings for these implementation factors in the context of similar studies. In future research, we will continue to use survey items that are customized to the NACL goals but will embed these items within instruments that also contain items from leadership measures that are more widely used by researchers in the field. This will facilitate a greater contribution from future NACL research toward advancing the field's knowledge in these areas.

Implications for Funders of Impact Research

Researchers who anticipate delayed onset of effects should plan longer impact studies where outcomes are not compared until sufficient time has elapsed for effects to manifest. Of course, the ability to design such studies is influenced by the maximum duration of awards from funders. For example, the maximum duration for an efficacy study in the Education Research Grants program at the Institute of Education Sciences (IES) had been just four years (without extensions) for some time. Recent progress has been made in this regard; beginning with its 2016 Education Research Grants solicitation, IES allowed researchers to propose five-year efficacy trials. This is clearly a step in the right direction. However, even a five-year study of the NACL would have drawn a conclusion of negative program impact. Fortunately, when it was funded in 2013, the necessary retrospective design of the current study was consistent with the funding priorities of the NSF Discovery Research K-12 program. Support for such studies is now clearly in place at IES as well. In another recent change that first affected the 2014 solicitation, IES initiated a dedicated funding stream for supporting retrospective efficacy trials that analyze historical outcome data to estimate the effects of interventions implemented in the past.

We make this recommendation to funders acknowledging that longer, retrospective impact studies, where outcomes are compared several years after the formal portions of interventions end, invite specific threats to validity. These include *history* effects (Shadish, Cook, & Campbell, 2002), where other district initiatives are influencing student outcomes as much as, if not more than, the residual impacts of the intervention. Furthermore, with extended timelines, the opportunity for district leader turnover is increased, and turmoil at the district level can clearly compromise sustained systemic efforts to increase student outcomes.

References

August, D., Branum-Martin, L., Cárdenas-Hagan, E., Francis, D. J., Powell, J., Moore, S., & Haynes, E. F. (2014). Helping ELLs meet the Common Core State Standards for literacy in science: The impact of an instructional intervention focused on academic language. *Journal of Research on Educational Effectiveness*, 7(1), 54-82.

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is—or might be—the role of curriculum materials in teacher learning and instructional reform? *Educational Researcher*, *25*(9), 6-14.

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

Bintz, J., & Martin, G. (2007). A model for high school science reform: The BSCS National Academy for Curriculum Leadership. *The Natural Selection, vol. Spring*, pp. 13-23. Colorado Springs, CO: BSCS.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289-328.

Bolam, R., McMahon, A., Stoll, L., Thomas, S., & Wallace, M. (2005). *Creating and sustaining professional learning communities*. Research Brief No. RB637. London, England: Department for Education and Skills, General Teaching Council for England.

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

Boyd, V. (1992). Creating a context for change. *Issues… about change*, *2*(2).

Boyd, V., & Hord, S. M. (1994, April). *Principals and the new paradigm: Schools as learning communities*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Burkhauser, S., Gates, S. M., Hamilton, L. S., & Ikemoto, G. S. (2012). *First-year principals in urban school districts: How actions and working conditions relate to outcomes* [Technical report]. Rand Corporation.

Bryk, A. S., & Schneider, B. (2003). Trust in schools: A core resource for school reform. *Educational Leadership, 60*(6), 40-45.

BSCS. (2011). *BSCS National Academy for Curriculum Leadership Washington State, Cohort 2: Final Evaluation Report.* BSCS Evaluation Report (ER 2011-04), Colorado Springs, CO: BSCS.

Carroll, T. G., Fulton, K., & Doerr, H. (2010). *Team up for 21st century teaching and learning: What research and practice reveal about professional learning.* Condensed Excerpts. National Commission on Teaching and America's Future.

Center for Comprehensive School Reform and Improvement (2005). *Research brief: What does the research tell us about teacher leadership*? Retrieved from http://www.centerforcsri.org/files/Center_RB_sept05.pdf

Cervetti, G. N., Kulikowich, J. M., & Bravo, M. A. (2015). The effects of educative curriculum materials on teachers' use of instructional strategies for English language learners in science and on student learning. *Contemporary Educational Psychology*, *40*, 86-98.

Cheung, A., Slavin, R. E., Kim, E., & Lake, C. (2015, June). *Effective secondary science programs: A best-evidence synthesis*. Baltimore, MD: Johns Hopkins University, Center for Research and Reform in Education.

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher*, *32*(6), 3-12.

Cochran-Smith, M., & Lytle, S. L. (1999). Relationships of knowledge and practice: Teacher learning in communities. *Review of Research in Education, 24*(2), 249-305.

CORS. (2016) *Center on organization and restructuring of schools*. Retrieved from: http://archive.wceruw.org/cors/

Davis, E. A., & Krajcik, J. S. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher*, *34*(3), 3-14.

Desimone, L. M., & Hill, K. L. (2017). Inside the Black Box: Examining Mediators and Moderators of a Middle School Science Intervention. *Educational Evaluation and Policy Analysis*, 0162373717697842.

DuFour, R. & Eaker, R, (1998). *Professional learning communities at work: Best practices for enhancing student achievement*. Bloomington, IN: Solution Tree Press.

DuFour, R., DuFour, R., Eaker, R. & Karhanek, G. (2004). *Whatever it takes: How*

professional learning communities respond when kids don't learn (1st ed.). Bloomington, IN: National Educational Service.

Fishman, B. J., Marx, R. W., Best, S., & Tal, R. T. (2003). Linking teacher and student learning to improve professional development in systemic reform. *Teaching and Teacher Education*, *19*(6), 643-658.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: Synthesis of the literature.* Tampa, FL: University of South Florida, Louis de Ia Parte Florida Mental Health Institute. The National Implementation Research Network (FMHI Publication# 231).

Fullan, M. (2001). *Leading in a culture of change*. San Francisco, CA: Jossey-Bass.

Fullan, M. (2014). *Leading in a culture of change personal action guide and workbook*. Hoboken, NJ: John Wiley.

Fulton, K., & Britton, T. (2011). *STEM Teachers in Professional Learning Communities: From Good Teachers to Great Teaching*. Retrieved September 18, 2017, from http://nctaf.org/wp-content/uploads/2012/01/NCTAFreportSTEMTeachersinPLCsFromGoodTeacherstoGreatTeaching.pdf

Gallimore, R., Ermeling, B. A., Saunders, W. M., & Goldenberg, C. (2009). Moving the learning of teaching closer to practice: Teacher education implications of school-based inquiry teams. *The Elementary School Journal, 109*(5), 537-553.

Gamoran, A., Anderson, C.W., Quiroz, P.A., Secada, W.G., Williams, T., & Ashman, S., (2003). *Transforming teaching in math and science: How schools and districts can support change*. New York: Teachers College Press.

Garmston, R. J., & Wellman, B. M. (2009). *The adaptive school: Developing and facilitating collaborative groups*. NY: Christopher-Gordon.

Guskey, T. R. (2000). *Evaluating professional development*. Thousand Oaks, CA: Corwin Press.

Hall, G. E., & Hord, S. M. (2011). Learning builds the bridge between research and practice. *Standards for Professional Learning, 32*(4), 52-57.

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

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1-28.

Holdren, J. P., Lander, E. S., & Varmus, H. (2010). *Prepare and inspire: K-12 education in science, technology, engineering, and math (STEM) for America's future*. Executive Report. Washington, DC: President's Council of Advisors on Science and Technology.

Hord, S.M. & Boyd, V. (1995). Professional development fuels a culture of continuous improvement. *Journal of Staff Development, 18*(1), 10-15.

Hord, S. M. (2007). Learn in community with others. *Journal of Staff Development*, *28*(3), 39-40.

Hord, S. M. & Sommers, W. A. (2008). *Leading professional learning communities: voices from research and practice*. Thousand Oaks, CA: Corwin Press; Reston, VA: National Association of Secondary School Principals; & Oxford, OH: National Staff Development Council.

Institute of Education Sciences (IES) (2016). *What Works Clearinghouse procedures and standards 3.0* Retrieved from: https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf

Jacob, R., Goddard, R., Kim, M., Miller, R., & Goddard, Y. (2015). Exploring the causal impact of the McREL Balanced Leadership Program on leadership, principal efficacy, instructional climate, educator turnover, and student achievement. *Educational Evaluation and Policy Analysis*, doi: 0162373714549620.

Kelly, D., Nord, C. W., Jenkins, F., Chan, J. Y., & Kastberg, D. (2013). *Performance of US 15-year-old students in mathematics, science, and reading literacy in an international context*. First Look at PISA 2012. NCES 2014-024. National Bureau of Economic Research.

Kennedy, A., Slavit, D. & Nelson, T. H. (2009). Supporting collaborative teacher inquiry. In Slavit, D., Nelson, T. H., & Kennedy, A. (Eds.), *Perspectives on supported collaborative teacher inquiry* (pp. 166-180). NY: Routledge.

Kruse, S., Louis, K. S., & Bryk, A. (1994). Building professional community in schools. *Issues in restructuring schools*, *6*(3), 67-71.

Kruse S. & Louis K.S. (2009). *Building strong school cultures: A guide to leading change*. Thousand Oaks, CA: Corwin Press.

Landes, N. M., Powell, J. C., & Short, J. B. (2004). AIM for Professional Development. *Science and Children*, *41*(5), 40.

Lara-Alecio, R., Tong, F., Irby, B. J., Guerrero, C., Huerta, M., & Fan, Y. (2012). An experimental study of science intervention among middle school English learners: Findings from first year implementation. *Journal of Research in Science Teaching*, *49*(8), 987-1011.

Lawton, M., Berns, B., & Sandler, J. (2009). Putting the curriculum at the center of science education reform. In Berns, B. & Sandler, J. (Eds.), *Making science curriculum matter: Wisdom for the reform road ahead* (pp. 7-21). Thousand Oaks, CA: Corwin Press.

Learning Research and Development Center. (2017). *What we do*. Retrieved from http://ifl.pitt.edu/index.php/what_we_do

Lee, O., Deaktor, R. A., Hart, J. E., Cuevas, P., & Enders, C. (2005). An instructional intervention's impact on the science and literacy achievement of culturally and linguistically diverse elementary students. *Journal of Research in Science Teaching*, *42*(8), 857-887.

Leithwood, K., Louis, K.S., Anderson, S., & Wahlstrom, K. (2004*). How leadership influences student learning*. New York: Wallace Foundation. Retrieved September 14, 2008.

Lieberman, A., & Friedrich, L. D. (2010). *How Teachers Become Leaders: Learning from Practice and Research. Series on School Reform.* New York: Teachers College Press.

Louis, K. S. (2006). *Organizing for school change*. New York: Routledge.

Louis, K. S., Leithwood, K., Wahlstrom, K. L., & Anderson, S. E. (2010). *Investigating the links to improved student learning: Final report of research findings.* St. Paul, MN: University of Minnesota.

Louis, K.S. & Marks, H.M. (1998). Does professional learning community affect the classroom: Teachers' work and student experience in restructuring school. *American Journal of Education, 106*(4), 532-575.

Loucks-Horsley, S., Love, N., Stiles, K. E., & Mundry, & S. Hewson. PW (2003). *Designing professional development for teachers of science and mathematics*. New York: Corwin Press.

Loucks-Horsley, S., Stiles, K. E., Mundry, S., Love, N., & Hewson, P. W. (2009). *Designing professional development for teachers of science and mathematics*. New York: Corwin Press.

Lynch, S., Kuipers, J., Pyke, C., & Szesze, M. (2005). Examining the effects of a highly rated science curriculum unit on diverse students: Results from a planning grant. *Journal of Research in Science Teaching*, *42*(8), 912-946.

McLaughlin M.W. & Talbert, J.E. (2006) *Building school-based teacher learning communities: Professional strategies to improve student achievement*. New York: Teachers College Press.

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

National Research Council (NRC). (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

National Research Council (NRC). (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: National Academy Press.

National Research Council (NRC). (2002). *Scientific research in education*. Washington, DC: The National Academies Press.

National Research Council (NRC). (2005). *How students learn: Science in the classroom*. Committee on How People Learn, A Targeted Report for Teachers, M. S. Donovan and J. D. Bransford, (Eds.) Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council (NRC). (2010). *Rising above the gathering storm, revisited: Rapidly approaching category 5*. Washington, DC: The National Academies Press.

National Research Council (NRC). (2014). *Exploring Opportunities for STEM Teacher Leadership: Summary of a Convocation*. S. Olson and J. Labov, Rapporteurs. Planning Committee on Exploring Opportunities for STEM Teacher Leadership: Summary of a Convocation, Teacher Advisory Council, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. p. 36.

Office of Superintendent of Public Instruction (OSPI). (2016). Retrieved from http://www.k12.wa.us/

Pellicer, L. O., & Anderson, L. W. (2001). *Teacher leadership: A promising paradigm for improving instruction in science and mathematics*. Retrieved from https://eric.ed.gov/?id=ED465586

Roth, K. J., Anderson, C. W., & Smith, E. L. (1987). Curriculum materials, teacher talk and student learning: Case studies in fifth grade science teaching. *Journal of Curriculum Studies*, *19*(6), 527-548.

Roy, P., & Hord, S. M. (2007). It's everywhere, but what is it? Professional learning communities. *Journal of School Leadership*, *16*(5), 490-501.

Saunders, W. M., Goldenberg, C. N., & Gallimore, R. (2009). Increasing achievement by focusing grade-level teams on improving classroom learning: A prospective, quasi-experimental study of Title I schools. *American Educational Research Journal, 46*(4), 1006-1033.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *Splintered vision: An investigation of US mathematics and science education*. Norwel, MA: Kluwer Academic Publishers.

Schneider, R. M., & Krajcik, J. (2002). Supporting science teacher learning: The role of educative curriculum materials. *Journal of Science Teacher Education*, *13*(3), 221-245.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton, Mifflin.

St. John, M., Hirabyashi, J., Helms, J.V., Tambe, P., (2005). *A Summative report of contributions and impacts of the SCI Center*. Inverness, CA: Inverness Research Associates.

Stoll, L., Bolam, R. McMahon A, Wallace, M., & Thomas, S. (2006). Professional learning communities: A review of the literature. *Journal of Educational Change, 7*(4), 221-258.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Supovitz, J.A. (2002). Developing communities of instructional practice. *Teachers College Record, 104*(8), 1591-1626

Supovitz, J.A. & Christman, J.B. (2003). *Developing communities of instructional practice: Lessons from Cincinnati and Philadelphia* (CPRE No. RB-39). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

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.

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open*, 2(1), 2332858415625975.

Taylor, J., Gardner, A., & Bybee, R. (2009). The role of curriculum materials in reform. In Gess-Newsome, J., Luft, J.A., & Bell, R.L. (Eds.), *Reforming Secondary Science Instruction* (pp. 27-38). Washington, DC: NSTA Press.

Taylor, J. A., Getty, S. R., Kowalski, S. M., Wilson, C. D., Carlson, J., & Van Scotter, P. (2015). An Efficacy Trial of Research-Based Curriculum Materials with Curriculum-Based Professional Development. *American Educational Research Journal*, 52 (5), doi:0002831215585962

Usiskin, Z. (1985). We need another revolution in secondary school mathematics. *The secondary school mathematics curriculum*, pp. 1-21.

Vescio, V., Ross, D., & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. *Teaching and teacher education*, *24*(1), 80-91.

Waters, T. J., & Marzano, R. J. (2006). *School district leadership that works: The effect of superintendent leadership on student achievement*. A Working Paper. Mid-Continent Research for Education and Learning (McREL).

WestEd (2017). *K-12 alliance NGSS science institutes: Promoting change and fostering Excellence*. Retrieved from https://www.wested.org/service/k-12-alliance-ngss-science-institutes/

Westine, C., Spybrook, J., & Taylor, J. (2013). An Empirical Investigation of Variance Design Parameters for Planning Cluster Randomized Trials of Science Achievement. *Evaluation Review* 37(6), 490-519.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*. Issues & Answers. REL 2007-No. 033. Regional Educational Laboratory Southwest (NJ1).

Corresponding Author: Joseph A. Taylor, BSCS Science Learning/University of Colorado, Colorado Springs:
jtaylor@bscs.org/jtaylo18@uccs.edu

Acknowledgements. The development and implementation of the NACL in WA was jointly funded by Agilent Technologies Foundation, Battelle, Intel Foundation, and districts with major support from the Pacific Northwest National Laboratory and the Pacific Science Center. The original development of the NACL was funded by the National Science Foundation under grant [DRL 9911615]. This retrospective impact study was funded by the National Science Foundation under grant [DRL 1316202]. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The authors would also like to acknowledge Lisa Carey and Valerie Maltese for their assistance with data collection, cleaning, and management, and Mark Bloom, Connie Hvidsten, Brooke Gorman and Paul Numedahl or their assistance with data collection and analysis.