American Educational Research Journal
http://aerj.aera.net

Increasing Achievement by Focusing Grade-Level Teams on Improving Classroom Learning: A Prospective, Quasi-Experimental Study of Title I Schools

William M. Saunders, Claude N. Goldenberg and Ronald Gallimore

Am Educ Res J 2009 46: 1006 originally published online 20 March 2009
DOI: 10.3102/0002831209333185

The online version of this article can be found at:
http://aer.sagepub.com/content/46/4/1006

Published on behalf of

American Educational Research Association

and

http://www.sagepublications.com

American Educational Research Journal
December 2009, Vol. 46, No. 4, pp. 1006–1033
DOI: 10.3102/0002831209333185

Increasing Achievement by Focusing Grade-Level Teams on Improving Classroom Learning: A Prospective, Quasi-Experimental Study of Title I Schools

William M. Saunders
Pearson Education
Claude N. Goldenberg
Stanford University
Ronald Gallimore
University of California, Los Angeles

The authors conducted a quasi-experimental investigation of effects on achievement by grade-level teams focused on improving learning. For 2 years (Phase 1), principals-only training was provided. During the final 3 years (Phase 2), school-based training was provided for principals and teacher leaders on stabilizing team settings and using explicit protocols for grade-level meetings. Phase 1 produced no differences in achievement between experimental and comparable schools. During Phase 2, experimental group scores improved at a faster rate than at comparable schools and exhibited greater achievement growth over 3 years on state-mandated tests and an achievement index. Stable school-based settings, distributed leadership, and explicit protocols are key to effective teacher teams. The long-term sustainability of teacher teams depends on coherent and aligned district policies and practices.

KEYWORDS: professional development, school/teacher effectiveness, educational reform, longitudinal studies, elementary schools, organization theory/change

School-based teacher collaboration, inquiry, and learning are a focus of many efforts to improve student learning. Such efforts are variously described as professional learning teams or communities, and their central goals include deprivatizing teaching through school-based collaboration and reflective dialogue to improve classroom instruction and student learning. In spite of their intuitive appeal and compelling logic, and the approbation of many educators, the limited evidence base is reason for pause. Scarce resources and urgent needs in public schools make funding learning teams and communities a fraught choice for districts because they require culture changes

that have been historically difficult to achieve and sustain (Elmore, 1999–2000). The reasonable demand for more evidence is a challenge to the research community (Smylie, 1994; Visscher & Witziers, 2004; Whitehurst, 2002).

In this article, we present a brief review and critique of the research base and describe a 5-year quasi-experimental investigation comparing achievement gains in nine Title I schools relative to six matched schools. The nine experimental schools organized and trained grade-level teams to increase tested achievement by focusing on the improvement of students' classroom learning. The results indicated that significant achievement gains were achieved when grade-level teams were provided with consistent meeting times, schoolwide instructional leadership, and explicit protocols that focused meeting time on students' academic needs and how they might be instructionally addressed. This might be one of the first quasi-experimental investigations demonstrating increased average achievement over time in schools that implemented teacher teams focused on improving student learning.

Literature Review

Reflecting on the failure of the 1960s "new math" reform, Sarason (1971) concluded that schools need to be places of learning for teachers if we are to improve classroom instruction and student achievement. Inspired partly by previous theory, research, and experience in the commercial sector (e.g., Schon, 1983; Senge, 1990), attention turned to creating "working conditions in schools that enhance the commitment and expertise of teachers" (Rowan, 1990, p. 353). Then and since, many have argued that teachers left working in isolated classrooms with little school-based opportunities for collaboration and learning are unlikely on their own to improve instruction (e.g., Elmore, 1999–2000, 2002; Goldenberg, 2004b; Little, 1982).

Breaking down barriers to greater collaborative learning opportunities is hampered by tradition and school culture. Willingly or not, most teachers spend the overwhelming majority of their work time in the classroom, with

---

WILLIAM M. SAUNDERS, PhD, is vice president, Learning Teams Group Pearson, 2701 Ocean Park Blvd. Suite 220, Santa Monica, CA 90405; e-mail: *bill.saunders@Pearson.com*. He is also a research associate at UCLA. He investigates school and teaching improvement and literacy development, and he recently contributed to a synthesis of research on English language learners. Saunders directs R&D and nationwide implementation for the Pearson's learning teams program.

CLAUDE N. GOLDENBERG, PhD, is a professor of education in the School of Education, Stanford University, Stanford, CA 94305; e-mail: *cgoldenberg@stanford.edu*. His research has focused on family, classroom, and school influences to improve educational outcomes for Spanish-speaking students.

RONALD GALLIMORE, PhD, is Distinguished Professor Emeritus in the Department of Psychiatry & Biobehavioral Sciences at UCLA, Los Angeles, CA 90095; home page: *http://ronaldg.bol.ucla.edu*. He investigates teaching and its improvement and is coauthor of a monograph on the teaching practices of basketball coach John Wooden.

limited paid time to interact with other professionals. Occasionally, the principal or another teacher might walk in for brief periods, but they are not functional parts of the classroom. As Lortie (1975) documented, teachers' fundamental orientation is to the students in their classrooms. Given the history and culture of schools, this is expectable.

Many ideas and reforms emerged in the 1980s and 1990s for addressing the isolation problem, moving teacher collaboration and learning to national attention (see, e.g., Bird & Little, 1986; Brandt, 1987; Fullan, 1991, Hord, 1997; Lieberman, 1988a, 1988b, 1995; Little, 1982; Little & McLaughlin, 1993; Murphy & Hallinger, 1993; Rosenholtz, 1989, 1991; Rowan, 1990; Tharp & Gallimore, 1988). Essential characteristics of school learning teams and communities emerged, including breaking down barriers to increase collaboration among teachers, nurturing shared values and commitments, focusing on student learning instead of teaching strategies, reflective dialogue, and joint analysis of assessments and planning of instruction (e.g., Hord, 1997; Louis, Kruse, & Marks, 1996; Newmann & Associates, 1996; Schmoker, 1996).

Further impetus came in the 1990s era of whole school reform, as national models struggled to achieve their aims when they relied solely on the traditional "loosely coupled" structure of teachers' working autonomously to implement improved instruction (Elmore, 1999–2000, 2002). Such experiences magnified interest in learning communities or teams as vehicles of school and instructional improvement. With apologies to those not cited, following is an incomplete list of some representative contributors to this development: Annenberg Institute for School Reform (2005); Cochran-Smith and Lytle (1999); Darling-Hammond and Sykes (1999); Elmore and McLaughlin (1988); Elmore, Peterson, and McCarthey (1996); Franke, Carpenter, Levi, and Fennema (2001); Hord (1997); Kruse, Louis, and Bryk (1994); Lieberman (1988a, 1988b, 1995); Murphy and Hallinger (1993); Newmann and Wehlage (1995); Peterson, McCarthey, and Elmore (1996); O'Neil (1995); and Tharp and Gallimore (1988). The long-term trend is apparent in the National Staff Development Council's (2001) standards for professional development, which include organizing educators into learning communities and meeting regularly to collaboratively learn, jointly plan lessons, and solve problems, and in the programs of a national commission that is promoting the practice and publicizing successful teacher learning teams (National Commission on Teaching and America's Future, 2008).

Research Critique

There are at least three problems with the existing research base. First, the intuitive appeal of learning teams and communities must be measured against a sobering reality: Few studies have investigated their impact on student achievement (Goldenberg, 2004b; Stoll, Bolam, McMahon, Wallace, & Thomas, 2006; Vescio, Ross, & Adams, 2008; Whitehurst, 2002). Vescio et al. (2008) identified 55 books, papers, and articles in the U.S. literature from 1990 to 2005 using various search terms, for example, *professional*

learning communities (PLCs), *teachers and learning communities, critical friends, communities of practice*, and so on. Eleven of the 55 were empirical studies that considered the impact of school-based PLCs on teaching and student learning. Of the 11 empirical studies, 8 reported student achievement data of some kind. Of these 8 studies, 5 were case studies, and 3 used self-report surveys to assess PLC practice. Two survey studies reported significant correlations between school reports of professional community practices and student achievement (Bolam, McMahon, Stoll, Thomas, & Wallace, 2005; Supovitz, 2002). Apparently, no study of PLCs or similar approaches has used an experimental or quasi-experimental design to investigate the effects of grade-level teams' working collaboratively to improve student learning. This limitation has been noted for investigations of teacher work redesign in general, with few longitudinal or methodologically rigorous comparisons of participating and nonparticipating schools or teachers (Smylie, 1994).

Second, Fullan (2000) noted that many collaborative teams and communities were identified for investigation after they were successful. Correlations between the prevalence of community or team features and higher student achievement have usually been interpreted to mean that the former causes the latter. However, it is plausible that a faculty struggling over time to improve student achievement develops the psychosocial qualities associated with PLCs only after its efforts begin to show success. One case study documented such a trajectory: Led by a principal who pressured as well as supported them and with significant resistance from some members, a highly conflicted staff implemented grade-level meetings focused on improving student learning and evolved into a collegial group with shared values and a group ethos as their school's achievement rose over several years from lowest to highest in its district (Goldenberg, 2004b). Certainly in other social group contexts, it is a cliché that struggle followed by success often positively improves a team's internal dynamics. Such outcomes are consistent with theories predicting that positive psychosocial relations often grow out of sustained, joint productive activity focused on shared challenges and problems (Tharp, Estrada, Dalton, & Yamauchi, 2000; Tharp & Gallimore, 1988). To clarify direction of effects, Fullan argues that more studies are needed that implement an explicit, repeatable framework in schools not yet deploying learning teams, and document improvements in student achievement "before the beginning" (Sarason, 1972, p. 4). If a group of underperforming schools introduce collaborative learning teams, will objectively measured student achievement increase relative to achievement at comparable schools that do not? The evidence base is too limited to be certain, in spite of numerous promising self-reports, case studies, and anecdotal accounts (Vescio et al., 2008) and the intuitively appealing logic that underlies PLC and learning team programs.

Third, there is substantial variation in definitions and practices of teacher learning teams or communities in the literature (Dufour, 2004; Vescio et al., 2008). Some teams focus on the social adaptation of students, others choose topics to study or read about using a seminar-like structure, some use the

time for more conventional professional development, and some use meeting time to address specific student academic needs and jointly plan instruction (Vescio et al., 2008). Vescio et al. (2008) concluded that the only common feature in studies of PLCs reporting improved achievement was a "persistent focus on student learning and achievement by the teachers in learning communities" (p. 87). Focusing on what students are learning, and not learning, might shift attention more to improving instruction and correspondingly less to teacher relations and satisfaction within the team or PLC structure (Goldenberg, 2004b; Saunders & Goldenberg, 2005; Smylie, 1994; Visscher & Witziers, 2004). Likewise, groups focused on increasing teacher knowledge but not its direct application to student learning difficulties might have a diluted effect on achievement. This is an example of the contrast we intend to draw: a PLC that learns to use digital cameras for general application to classroom teaching compared with a team focused on its school's persistently low performance on items measuring fraction concepts or reading comprehension. Both can benefit students, but a specific student academic need might be a more productive and sustaining team focus than an opportunity to learn something of more general value. Such a focus is consistent with national survey evidence that teachers prefer school-based, long-term, active learning that relates directly to classroom instruction (Garet, Porter, Desimone, Birman, & Yoon, 2001). In addition, Lortie (1975) concluded that given a choice in how to use extra work time, teachers overwhelmingly preferred to work in their classrooms or on classroom-related matters. Congruence with teacher preferences might further account for the finding that PLCs and teams focused on improving classroom learning have been reported to be more successful at increasing achievement. Such a focus seems a promising choice for designing a prospective test of the achievement effects of grade-level teams.

Study Goals and Questions

Our goals for this study were limited to testing the effects on student achievement in schools that introduced grade-level teacher teams focused exclusively on improving students' classroom learning. For this study, we defined learning teams as grade-level teams in elementary schools that meet two or three times a month. We hypothesized that significant gains in achievement might result if grade-level teams simply focused less during meeting time on noninstructional issues and more on their students' academic struggles. This derived in part from a not very novel observation that grade-level groups, departmental meetings, faculty meetings, and other familiar settings rarely focused on students' academic needs and how to instructionally address them.

The theory of action on which the intervention was based emerged from reviewing the effective schools research (e.g., Bliss, Firestone, & Richards, 1991) and a 6-year case study in a single Title I school that succeeded in significantly improving student achievement compared with other schools in

the district (Goldenberg, 2004b; Goldenberg & Sullivan, 1994). Grade-level teams in the case study school examined student assessments, set and shared specific academic goals, jointly developed instruction to address these goals, and reviewed student work products resulting from jointly developed instruction. This focus is little more than spending more time on the familiar teaching cycle of plan-teach-assess, is a common feature of many PLC and teacher team programs (see reviews by Stoll et al., 2006; Vescio et al., 2008), and is consistent with previously cited standards (National Staff Development Council, 2001). One difference worth noting was a heavy emphasis on providing grade-level teams and schoolwide instructional leadership groups with predictable, consistent settings during which they could carry out this work (Goldenberg, 2004b; Saunders & Goldenberg, 2005).

To stand up and support teams in the experimental schools, the initial 2-year implementation relied solely on training principals. When this proved too weak to achieve consistent team meetings and focus, a more robust implementation design was used in the final 3 project years. Both implementation designs had the goal of focusing grade-level and schoolwide instructional leadership teams (ILTs) on improving student learning.

We report the quantitative comparison of treatment and comparison schools on standardized achievement tests and a statewide academic performance index, using an analytic technique that took into account the midproject change in implementation design. In the final year of the project, a separate formative evaluation of the intervention and its effects was conducted by an independent investigator (McDougall, Saunders, & Goldenberg, 2007); in addition, project researchers conducted a case study of a single experimental school using materials collected over 5 years. Both of these formative evaluations are summarized in a section under "Methods" below. The purpose here, however, is to provide what might be the first evidence from a comparison of treatment and comparison groups that introducing teacher teams focused on improving student learning in low-performing schools will over time significantly increase achievement.

Methods

Sample Recruitment and Description

The study was conducted in one administrative region of a large urban district in Southern California. At the request of the regional superintendent, the project was called Getting Results (GR). Nine of the 11 Title I schools in this region participated in the 5-year scale-up study reported here and constituted the GR treatment group. A comparison group of 6 elementary schools was recruited from regions within the same district and adjacent to the one that included the 9 treatment schools. Twelve candidate comparison schools were identified on the basis of their similarity to the treatment schools. Six of the 12 candidate Title I schools were selected as best matches on demographics and achievement. All 6 agreed to participate.

Table 1

Demographic Information on Getting Results and Comparison Schools

School	Number of Students	Percentage Receiving Free or Reduced-Price Lunch	Percentage Hispanic	Percentage ELLs in Grades 2 and 5	Percentage Enrolled in School ≥2 Years, Grades 2 and 5[a]
Getting Results schools					
1	586	71	51	45	70
2	660	96	76	83	69
3	953	89	55	64	69
4	1,297	77	76	78	71
5	1,016	96	81	84	70
6	570	73	46	63	76
7	584	85	51	66	71
8	1,040	97	90	78	77
9	1,300	92	93	75	75
Mean	890	86	69	71	72
Comparison schools					
1	1,356	84	98	73	80
2	764	72	51	58	76
3	832	98	97	95	68
4	1,284	98	99	79	85
5	1,041	99	98	77	81
6	654	81	57	68	77
Mean	988	89	83	75	78

a. Estimate of school population stability on the basis of the percentages of second and fifth graders who had been enrolled in the school for at least 2 academic years.

Because both treatment (n = 9) and comparison (n = 6) groups were part of the same larger urban school district, all were obligated to use the same guidelines and policies related to curriculum, instruction, assessment, class size, English learner programming, and school calendar.

The two groups of schools were statistically similar in terms of enrollment and percentages receiving free or reduced-price lunch, Hispanic, and English-language learners. These comparisons are presented in Table 1. The last column in Table 1 indicates population stability, estimated as the percentage of students in Grades 2 and 5 who had been enrolled in the school for at least 2 academic years; higher numbers in this column indicate more stability among the school's student population. Altogether, the 15 schools served nearly 14,000 students, 73% of whom were English-language learners.

To establish the initial achievement comparability of the treatment and comparison schools, we used Stanford 9 Achievement Test (SAT-9) normal curve equivalent (NCE) scores from 1997 (the spring preceding the start of the GR intervention). For each school, NCE scores were available for four SAT-9 subtests by grade level only: reading, mathematics, language, and spelling (scaled scores and school-level means and standard deviations were

Table 2
Baseline Stanford 9 Achievement Test (SAT-9) Results for Getting Results and Comparison Schools in 1997a

	Getting Results Schools (n = 9)		Comparison Schools (n = 6)	
Grade[b]	NCE *M*	*SD*	NCE *M*	*SD*
2	19.11	7.22	22.29	11.07
3	22.00	6.93	18.25	8.801
4	22.33	8.01	22.63	10.86
5	22.64	8.13	20.17	7.39

a. End-of-school-year testing in the year prior to Phase 1 intervention.
b. No scores were available for Grade 1 because the SAT-9 was not administered by district at this time. No school-level means and standard deviations were available for the SAT-9 for 1997 and earlier.

not available in California for 1997 and earlier years). Average performance on the four subtests was computed across schools for each grade level (see Table 2), separately for the GR (n = 9) and comparison (n = 6) schools. A series of independent-group t tests were conducted to evaluate average differences between the treatment and comparison schools at each grade level. None of these tests yielded significant differences between the two groups of schools.

Implementation Goals

To support grade-level groups focused on improving classroom instruction and student learning, an ILT was established at each of the nine GR schools.

ILTs. An ILT was formed at each school, composed of at least one representative from each grade level, the principal, and other appropriate school administrators, coaches, or coordinators. ILTs met monthly for approximately 2 hours (a total of 20 hours across the school year). The broad purpose of the ILTs was to collectively set direction for and lead schoolwide efforts to improve instruction and student achievement. This involved: (a) leading the process of transforming academic standards (state or district) into explicit instructional goals; (b) identifying assessments and indicators to assess the goals; (c) regularly evaluating schoolwide achievement and determining next steps (action plans); (d) identifying common instructional challenges teachers face and organizing assistance to augment faculty content and pedagogical knowledge, such as building and district specialists and formal professional development experiences; (e) working to ensure future professional development aligned with teachers' instructional challenges; and, most important, (f) leading, facilitating, and planning weekly (or bimonthly) grade-level team meetings, wherein teachers focused explicitly on addressing identified student academic needs.

Grade-level meetings. Facilitated by the representative who served on the ILT, each grade level was to meet two or three times per month for approximately 45 to 50 minutes. These meetings were to take place during the school day, with each grade level meeting on a different day of the week when a physical education instructor (described as a psychomotor specialist by the district) was available for student supervision. Grade-level teams with the support of the principal and grade-level leader were to collaborate to develop instruction to address student academic needs, check indicators of progress, revise instruction as necessary, and once an academic need had been adequately addressed choose another and repeat the process. Grade-level meetings were to focus on areas of continuing student academic needs identified jointly with the principal and the school ILT, for example, specific areas of literacy need as identified by high-stakes, periodic, and teacher assessments. Researchers were sometimes consulted, but choosing an area of focus or an instructional strategy was not part of the experimental intervention; by design, the intervention was agnostic with respect to what grade-level teams worked on or what instructional strategies were deployed. These choices were left to the staffs of each school.

Implementation Phases

The original study design specified an implementation based on training the principals of the nine experimental schools to stand up and facilitate ILTs and job-a-like grade-level teams at each of their respective schools. After 2 years, this plan produced limited implementation of the goals described in the preceding section, and a revised implementation plan was introduced for the final 3 years of the project. Although unexpected, this resulted in the opportunity to test two forms of implementation, which we label Phase 1 and Phase 2.

Phase 1. During Phase 1 (2 years), only building principals were trained to implement grade-level teams and ILTs. Two-hour trainings were set aside by the regional superintendent during her monthly meeting with principals for the first 2 years of the project. The superintendent attended all training sessions. During these monthly trainings, GR staff members (the authors) made presentations and provided simulation opportunities for principals to practice what they were to do at their respective schools. Principals were trained to establish monthly ILT and weekly grade-level team meetings, facilitated by the ILT representative for each grade. They were provided with protocols or procedures their ILTs and grade-level teams could use, for example, how to examine student work to identify academic problems and indicators of progress. When requested during Phase 1, project staff members met with individual principals to discuss their implementation efforts and problems and responded to phone calls in addition to the regularly scheduled monthly trainings.

However, despite continuing support by the regional superintendent and favorable principal responses to the GR intervention, the Phase 1 yielded limited implementation, minimal implementation or impact of any kind, and no appreciable gains in student achievement. Competing demands for their time and attention were typically cited as reasons for the lack of progress in implementation. Principals expressed uncertainty about the content or structure of ILT meetings and how they should lead or guide ILT representatives, who were in turn expected to lead their colleagues at grade-level team meetings. It turned out to be very difficult for ILT representatives to function effectively as grade-level leaders for instructional improvement, and principals were challenged to provide them with the necessary guidance. It became clear that a "train the principal" approach yielded little implementation, ineffective teacher teams, or no gains in student achievement.

As the project team weighed the implications of this lack of progress, changes in the larger district intervened that further suggested that a revised implementation plan was needed. At the end of Phase 1, the larger district of the participating region began to require high-need schools to adopt a school reform design. These changes led to a change in the project that was mutually agreed on by the region superintendent, school principals, and the project team. This change led to Phase 2 of the project.

Phase 2. At the beginning of Phase 2 implementation, the nine treatment schools were granted permission by the district to choose GR as the school reform model to meet the district requirement. The six comparison group schools were also required by the district to adopt school reform designs and were allowed to choose from a district-developed site-based management program or a variety of nationally available designs available from for-profit and nonprofit vendors. A variety of models were chosen by the six comparison schools, making a contrast of GR and another model not statistically possible.

During Phase 2, the monthly GR principal meetings used in Phase 1 were continued with the principals of the nine GR treatment schools, but the focus and agenda were changed. During Phase 2 principal meetings, project advisors and principals met monthly for approximately 2 hours to discuss the progress of ILTs and grade-level teams and share strategies across schools.

In addition, project advisors met monthly with each principal at his or her school to set school-specific goals, analyze and interpret achievement data, debrief and discuss GR implementation, and plan monthly ILT meetings. They discussed the successes and challenges of each of grade-level team and about how best to support and, where needed, apply pressure to sustain their work. Project advisors also attended the monthly ILT meetings at each of the nine schools to support the principal who led the meeting, and some grade-level meetings at the request of a principal or ILT member. The on-site support for principals and ILTs plus the more focused monthly principals meetings represent one of several significant augmentations to the implementation design introduced in Phase 2 (see Figure 1 for a graphic representation of the support and assistance provided in Phase 1 and Phase 2).

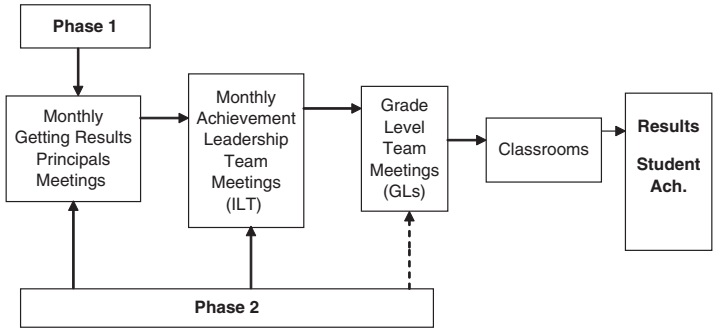

Figure 1. **Getting Results training and external assistance design.**
Note. The dotted arrow to grade-level team meetings indicates as-needed rather than ongoing external assistance.

Another significant augmentation in Phase 2 was summer and winter institutes organized and conducted by the GR staff. To initiate GR program implementation, each school sent its principal and leadership team to a 2.5-day summer institute held prior to the start of the academic year. The winter institute included all nine experimental schools and was 1 day long, during which selective training was provided and issues were discussed that had arisen at individual schools during the first semester of grade-level team meetings. Attendance at the institutes included administrators, content coaches, and grade-level team representatives. Planning of the institutes was done during the monthly principals' meetings. Participants were trained in the definition and theory of action for grade-level learning teams and in the use of a published manual of protocols for administrators and instructional leaders (grade-level team facilitators). On the basis of experiences and observations during Phase 1, the following protocols were included in the manual: analyzing standardized and periodic assessments, unit and instructional planning, and focusing on and addressing common student needs. The latter protocol was used by grade-level teams to focus their meeting time and included these steps:

1. Identify and clarify specific and common student needs to work on together.
2. Formulate a clear objective for each common need and analyze related student work.
3. Identify and adopt a promising instructional focus to address each common need.
4. Plan and complete necessary preparation to try the instructional focus in the classroom.
5. Try the team's instructional focus in the classroom.
6. Analyze student work to see if the objective is being met and evaluate the instruction.
7. Reassess: Continue and repeat cycle or move on to another area of need.

In summary, major augmentations to the implementation design in Phase 2 included summer and midyear institutes, a published manual of protocols, and assistance by GR staff members to principals and ILTs at each of the nine GR treatment schools.

Measures

SAT-9. During both Phases 1 and 2, all schools in the study (treatment and control) were required by the state to administer the SAT-9 to all students in Grades 2 through 5. The SAT-9 is a standardized, norm-referenced achievement test with subtests in reading, language, spelling, and mathematics. Tests were administered each spring by school personnel following state and district guidelines. The SAT-9 meets customary standards for reliability and validity of achievement tests (Berk, 1998).

State academic performance index (API). Devised by the state department of education and first reported in 1999, the API provides a single-numeric, composite index of school-level achievement. On the basis of a weighted algorithm that takes into account the numbers and proportions of students scoring in different quintiles across all SAT-9 subtests, the API is used to measure school growth toward designated improvement targets. The API ranges from 200 to 1,000. The state has determined that 800 represents acceptable performance. Schools not scoring at or above 800 are expected to meet annual growth targets toward reaching a score of 800. APIs are reported for the school overall and for specific subgroups as well: Whites, Hispanics, African Americans, socioeconomically disadvantaged, and special needs. In addition, each year, the state ranks all schools throughout the state on the basis of API scores and assigns each school a ranking ranging from 1 to 10 (Decile 1 = 1st to 10th percentiles, Decile 2 = 11th to 20th percentiles, etc.). API scores are included in this article for two reasons. First, disaggregating APIs by subgroup allows us to report results specifically for Hispanics, the predominant population in treatment and control schools. Second, the annual API rankings provide an additional indicator of school progress relative to state averages. Because each school is reranked every year, to maintain its ranking, it must match the rate of growth in the state; to improve its rank, it must exceed the rate of growth in the state.

Data Analysis

The research question specified school-level effects of the implementation of grade-level learning teams. However, the only SAT-9 data available from the district and state were in a form that did not permit averaging grade-level results into school-level estimates of achievement. The SAT-9 reports student outcomes through several types of scores, which have different mathematical properties and thus enable different kinds of analyses, among them national percentile ranks, NCEs, and scaled scores. *Z* scores were used in the analyses of SAT-9 scores because of validity issues that arise when

averaging national percentile ranks, NCEs, or scale scores. A summary *z* score was computed for each school for each year to compare change in SAT-9 scores over time between GR and comparison schools. These *z* scores describe, in a scale-free way, the average performance of each school relative to the state mean. The *z* scores were computed as follows:

1. For each grade level and subtest (reading, math, language, and spelling) the average scale score for each school, minus the state average, was divided by the state standard deviation. This yielded 16 *z* scores for each school for each year for which appropriate data were available (1998, 1999, 2000, 2001, and 2002). Because scale scores were not available in California for 1997 and earlier, it was not possible to compute a *z* score for the baseline year of 1997.
2. For each year tested, the 16 *z* scores for each school were averaged to produce an overall index of each school's performance relative to the state average.

With comparability established on the basis of t tests of NCE means at baseline, 1997 (see Table 2), analyses of SAT-9 scores focused on testing for differences within Phase 1 and Phase 2 of the GR intervention. A repeated-measures analysis of variance (ANOVA) was conducted on scores from 1998 and 1999 (Phase 1) and then again from 2000, 2001, and 2002 (Phase 2) to determine the effects, if any, produced in each of the two intervention phases.

This analysis of treatment and comparison schools during the years of intervention did not control for baseline scores, ruling out a longitudinal mixed model (hierarchical linear model) that would take in account the schools' starting status. However, neither the state nor the district could provide SAT-9 data in a form that permitted such an analysis. The 1997 "baseline" data were not available in a comparable form from the state database. The prior performance of schools was taken into account in each of the analyses by using a repeated-measures design. Thus, the effect of interest here is the interaction between preintervention and postintervention (the repeated measures) and group (treatment vs. control). The baseline performance of the two groups of schools is reported because it was the only basis for evaluating the comparability of samples at the time of their recruitment.

To compare change in API results over time between GR and comparison schools, data were analyzed using a repeated-measures ANOVA with one variable (treatment vs. control) and one outcome analyzed over the 4 years in which APIs were reported by the state: 1999, 2000, 2001, and 2002. No API scores were available prior to 1999. We conducted the same repeated-measures ANOVA for APIs reported specifically for the Hispanic subgroup. In addition, we compiled frequency distributions for each year on the basis of statewide rankings, that is, the number of GR and comparison schools each year that were ranked in each decile (1 to 10).

Implementation Fidelity

During the last year of the project (Phase 2), a formative evaluation was conducted by an external evaluator. Full details on the method, analysis, and

results of the external evaluation are available in McDougall et al. (2007). Observations, interviews, and focus groups were data sources used to assess the extent of GR implementation in four of the nine GR schools, along with the presence of comparable or parallel processes in three of the comparison schools. McDougall et al. concluded from blind ratings of all data sources that implementation was fairly strong in three GR schools and slightly weaker at a fourth. In summary, a combination of factors distinguished ILT and grade-level team meetings at GR schools relative to comparison schools, including more focus in meetings on student academics, systematic and joint planning, the purposeful use of assessment data (of all kinds), and agreements to implement and evaluate goal-directed instruction. Teacher meetings at GR schools had more consistent focus, planning, and time for academic topics, goals, and indicators and spent less time on nonacademic topics and tasks. GR school meetings spent more time discussing the relations between instruction and student outcomes and worked more on instructional improvements. At comparison schools, meetings were more loosely structured and were frequently canceled, curtailed, or rescheduled at the last moment. Comparison schools were more focused on shared or site-based governance than on improving teaching and learning. At GR schools relative to comparison schools, teachers had higher expectations for student achievement and were more likely to attribute student academic gains to their joint efforts to improve instruction. On the basis of this evaluation, it appears likely that many of intended effects of the GR intervention were present in at least four of the experimental schools.

In addition to McDougall et al.'s (2007) formative evaluation, two authors of this article conducted a case study of one GR experimental school (Saunders & Goldenberg, 2005). This study used annual focus groups and interviews over the 5 years of the project to obtain teacher reports on their experiences with the GR intervention. During the 1st year of the project (Phase 1), when the intervention was first introduced on the basis of principal training only, one teacher at Pine Elementary described how meetings at the school sometimes contributed to a personal sense of incoherence:

> Speaking only for myself, sometimes, . . . [with all the things that come down to our school from above] it feels very overwhelming because I'm the kind of person who needs to focus on one thing and work it through and know that I understand what's going on. And when I am bombarded, [or] it feels like I am bombarded—this is very personal—then I lose my focus. (Teacher focus group)

Five years later, once the intervention was well implemented, a Pine Elementary teacher reported that her second grade team regularly followed the GR protocol during grade-level meetings:

> **Teacher:** [We] formulate an objective. Assess for that objective. [Plan a lesson together and then we all teach it]. Look at the result. Did we meet the objective? No . . . let's go ahead and, you know, do it again. We all know this process.

Teacher: Very focused.
Teacher: We all know what we're doing at this meeting. We all know what we're doing at next week's meeting. We have an idea of what we will be doing, you know, 2 months from now.
Interviewer: Is that schoolwide? Is not just something at one grade level only?
Teacher: Schoolwide. (Teacher focus group)

When asked specifically about striving for high levels of academic achievement, teachers participating in one focus group remarked that the staff's expectations had increased after the GR intervention, that they had seen improving results each year, and that grade-level teams worked closely together to meet those expectations:

Teacher: I think it's something that we pride ourselves in, you know, knowing that we're gonna meet with our grade level and together we're fighting for this common goal, you know, and it all comes together as a school.
Teacher: I know my grade level, when I was in second grade, we were really tight and in first grade we are really tight. Very cohesive.
Interviewer: So if you had to characterize the academic climate for the school.
Teacher: Very academic. Very high expectations. (Teacher focus group)

A final excerpt indicates that grade-level meetings at Pine had improved, specifically because the ILT was taking time during its meetings to plan and prepare written agendas for each meeting:

Teacher: Grade-level meetings are [now] very well planned and organized. And they have agendas. And the agendas are reviewed and checked at the ILT. And suggestions are made. And revisions are made.
Teacher: Our classrooms are much more focused now than they have been.
Teacher: For sure. [All laugh]
Teacher: Oh yeah.
Interviewer: What is this a result of?
Teacher: A combination of things.
Teacher: I think the ILT members were kind of forced [by the principal]—[someone laughs] which helped though. I mean it was a big help to keep us focused and to keep a continued focus throughout every week—to keep our mind on a certain aspect of what we need to work on.
Teacher: And setting [instructional] goals every week. Besides all the big school goals that we created in grade levels and as a school at the beginning of the year, every week we're making weekly goals at each grade level. Agreeing on them, writing them down, adhering to them the following week, following up on them [in the classroom]—all based on student needs. (Teacher focus group)

In sum, two formative evaluations suggested that GR school meetings were more stable and focused on instruction than in comparison schools,

Table 3
_Z_-Score Means, Standard Deviations, and Effect Sizes (_d_) for Stanford 9 Achievement Test Scores: Getting Results and Comparison Schools, 1998 to 2002

Variable	1998	1999	2000	2001	2002
Getting Results schools: mean *z* score (*SD*)	–0.4599 (0.1993)	–0.4461 (0.2442)	–0.3557 (0.2339)	–0.2760 (0.1922)	–0.1865 (0.1977)
Comparison schools: mean *z* score (*SD*)	–0.5053 (0.2668)	–0.5051 (0.2665)	–0.5172 (0.2552)	–0.4513 (0.2718)	–0.4278 (0.2734)
z-score difference, Getting Results vs. comparison schools	0.0454	0.0590	0.1615	0.1753	0.2413
Pooled *SD*	0.2355	0.2556	0.2448	0.2354	0.2384
d based on pooled *SD*	0.1928	0.2308	0.6597	0.7447	1.0121
d based on comparison group *SD*	0.1702	0.2214	0.6328	0.6450	0.8834

Note. Because scale scores for the Stanford 9 Achievement Test were not available in California for 1997 and earlier, it was not possible to compute *z* scores and effect sizes for the baseline year.

but these results do not indicate whether they mediated changes in instruction or learning and achievement. They do suggest a significant degree of implementation of the intended changes in grade-level focus over time in some of the nine experimental schools.

Results

Effects of Phase 1 and Phase 2 Interventions on SAT-9 Scores

The averages across schools of *z* scores separated by condition are plotted in Figure 2 for each of the 5 test years. Mean *z* scores and standard deviations are presented in Table 3. A few features of Figure 2 are worth noting. First, there was a general improvement in the district average over the 5 years of the study, relative to state results. Second, the GR schools, which started out well below the district average, appeared to surpass the comparison schools and even the district average by the end of the 5 years. Finally, no impact of the GR intervention appeared during the first 2 years (Phase 1), but an impact did appear in the subsequent 3 years (Phase 2).

Because the nature of the intervention changed after the 2nd year, when Phase 2 began, two separate analyses of the data were performed. In the first analysis, only the 1998 and 1999 scores (Phase 1 outcomes) were included in a 2 (treatment condition) × 2 (year of testing) ANOVA in which year of testing was treated as a repeated measure. Confirming what Figure 2 presents, the analysis produced no main effect of treatment condition, $F(1, 13) = 0.176$,

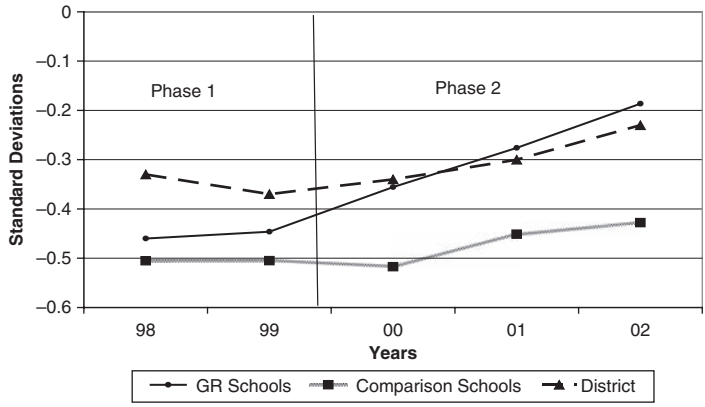

Figure 2. **Academic achievement on the Stanford 9 Achievement Test for Getting Results schools, comparison schools, and the district relative to statewide results.**

$p = .682$, or of year, $F(1, 13) = 0.085$, $p = .775$, and no significant interaction of treatment by year, $F(1, 13) = 0.075$, $p = .783$.

A separate repeated-measures ANOVA was performed that included the 3 school years of the Phase 2 intervention: 1999–2000, 2000–2001, and 2001–2002. Again, the analysis confirmed what is apparent in Figure 2. A significant interaction of condition by year was found, $F(3, 39) = 7.34$, $p < .01$, indicating that the difference between the GR and comparison schools increased over time during the Phase 2 intervention when external assistance was provided.

Table 3 also presents effect sizes (d) for the final year of Phase 1 intervention and for each of the 3 years of Phase 2. The z scores for SAT-9 results used in the ANOVAs were used to calculate effect sizes. We calculated two sets of effect sizes: one on the basis of pooled standard deviations and one on the basis of comparison group standard deviations. Comparison group standard deviations tend to yield a smaller and therefore more conservative effect size. Using comparison group standard deviations yielded effect sizes that were modest during the last year of Phase 1 (.22) and substantially higher during each year of Phase 2 intervention (0.63, 0.64, and 0.88). By the last year of Phase 2, the effect size had quadrupled over the effect size of the last year of Phase 1, increasing from 0.22 to 0.88.

The apparent difference in impact during the first 2 years (Phase 1) compared with the latter years (Phase 2) of the study makes sense. During Phase 1, the intervention consisted of working with the school principals: The principals were supposed to implement the GR program in their own schools, independent of the research team. As the results confirm, the observation at the end of the 2nd year (end of Phase 1) was correct that this approach was not producing implementation of the GR program or

Table 4
Academic Performance Index (API) for All Students: Getting Results and Comparison Schools, 1999 to 2002

Variable	API Mean for Each Year				
	1999	2000	2001	2002	Gain
Getting Results schools					
M	474.6	538.8	596.6	647.2	172.6
SD	74.4	72.7	50.0	48.4	
Comparison schools					
M	460.2	485.8	536.8	582.3	122.1
SD	78.4	80.2	79.8	66.2	
Difference between Getting Results and comparison schools	14.4	53.0	59.8	64.9	+50.5
Pooled *SD*	73.5	77.7	68.0	63.2	

improvements in instruction or learning at school sites. It was at this point that the augmented Phase 2 intervention began, which included summer and winter institutes, explicit protocols to support grade-level team and ILT meetings, and external assistance to schools, including project advisors meeting each month with each building principal and attending the monthly ILT meetings and occasional support of the weekly grade-level team meetings.

Achievement: API

API scores, available for 1999 to 2002, provide another estimate of Phase 2 effects. Table 4 includes API means and standard deviations for GR program and comparison schools. On average, API scores at GR schools increased by 172.6 points from 1999 to 2002. During the same time, API scores at comparison schools rose by 122.1 points. A repeated-measures ANOVA testing GR and comparison school means across Phase 2 (1999–2002) produced a significant time-by-group interaction, $F(3, 39) = 5.015$, $p < .01$, indicating that GR schools' improvement surpassed that of the comparison schools. The main effect for group was not significant, $F(1, 13) = 1.90$, $p > .15$.

The most conservative values on the basis of comparison school standard deviations (rather than pooled standard deviations) yielded increasing effect sizes from 1999 to 2002: 0.18, 0.66, 0.75, and 0.98, respectively. Not surprisingly, because APIs are based on SAT-9 results, the pattern is nearly identical to the results presented in Figure 2: similar performance for GR experimental and comparison schools in 1999 and strong divergence beginning in 2000, once Phase 2 intervention had begun.

However, the advantage of the API scores is the availability of results disaggregated for major subgroups, including Latino or Hispanic students,

Table 5
Academic Performance Index (API) for Hispanic Students: Getting Results and Comparison Schools, 1999 to 2002

	API Mean for Each Year				
Variable	1999	2000	2001	2002	Gain
Getting Results schools					
M	437.0	515.6	580.9	626.7	189.7
SD	57.7	59.6	45.8	46.4	
Comparison schools					
M	450.3	477.7	533.3	562.0	111.7
SD	61.7	66.7	71.7	68.6	
Difference between Getting Results and comparison schools	–13.3	37.9	47.6	64.7	+78.0
Pooled *SD*	57.7	63.2	60.1	57.8	

who constituted the largest percentage of students in both the GR and comparison schools. Table 5 presents API results for Hispanic students only, who constituted, on average, 69% of enrollment at GR schools and 83% of enrollment at the comparison schools. The same general pattern observed for all students was evident in the data for Hispanic students, except that Hispanic students at the GR schools were slightly lower in their API scores than Hispanic students at the comparison schools near the beginning of the study. Between 1999 and 2002, APIs for Hispanic students at GR schools grew by 189.7 points, in contrast to 111.7 points for Hispanic students at the comparison schools. A repeated-measures ANOVA testing GR and comparison schools means across 1999 to 2002 again produced a significant interaction, $F(3, 39) = 11.51$, $p < .001$. The main effect for group was not significant, $F(1, 13) = 1.27$, $p > .25$. The effect size was 1.03, using the pooled standard deviations for 2002 (57.8). Effect sizes were computed using comparison school standard deviations and were, from 1999 to 2002, respectively, –0.22, 0.57, 0.66, and 0.94.

Finally, Table 6 reports statewide rankings for each GR and comparison school from 1999 to 2002, as well as year-by-year group averages. Each year, the state reranks schools according to their APIs and their standing in relation to all other elementary schools in the state. After ranking all schools from top to bottom, the state assigns decile designations, with 1 corresponding to the lowest achieving 10% of schools in the state and 10 to the highest achieving 10%. In 1999, at the conclusion of the Phase 1 intervention, average ranks were virtually identical between GR and comparison schools, 2.1 and 2.0, respectively. Initially, five of the nine GR schools and three of the six comparison schools were in Decile 1, or the lowest 10% in the state. The other schools in both groups were distributed across Deciles 2 through 5. By 2002, the average ranking for GR schools had increased by 1.7 points to 3.8, and the average ranking for comparison schools had increased by .3 points

Table 6
Academic Performance Index Statewide Rankings for Getting Results and Comparison Schools, 1999 to 2002

	Statewide Ranking for Each Year (1–10)				
School	1999	2000	2001	2002	Gain
Getting Results schools					
1	5	5	5	6	1
2	1	1	2	3	2
3	2	3	4	6	4
4	1	2	2	2	1
5	1	2	2	2	1
6	4	4	4	5	1
7	3	4	4	5	2
8	1	1	2	2	1
9	1	1	2	3	2
Mean	2.1	2.6	3.0	3.8	1.7
Comparison schools					
1	3	3	3	4	1
2	4	4	4	4	0
3	1	1	1	1	0
4	1	1	1	2	1
5	2	2	2	2	0
6	1	1	1	1	0
Mean	2.0	2.0	2.0	2.3	.3

to 2.3. Between 1999 and 2002, every GR school increased its statewide rank: Five schools increased their rank by 1 decile, three schools increased by 2 deciles, and one school increased by 4 deciles. During that same period, only two comparison schools increased their ranking, both by 1 decile. The other four comparison schools showed no change in their statewide ranking at all.

The findings that emerged from the statewide ranking results correspond to the findings that emerged from the SAT-9 comparison. Both analyses showed GR schools producing significantly greater gains than comparison schools during the Phase 2 intervention (1999–2002). Moreover, both analyses showed GR schools improving during that time at a rate that exceeded the rate of growth in the rest of the state. The z scores calculated in comparison with year-by-year state averages (Figure 2) showed GR schools closing the gap between achievement levels at GR schools and state averages, and API rankings showed GR schools increasing in their standing relative to other schools in the rest of the state.

Discussion

The results were consistent with expectations and similar to those obtained in a preceding longitudinal study of a single Title I school

(Goldenberg, 2004b). The quasi-experimental trial reported here provided evidence that grade-level teams focused on improving student learning can produce school-level effects of both statistical and practical significance. However, these effects were obtained only during Phase 2, when three significant augmentations were added to the implementation design.

During Phase 1, when only principals were provided monthly training sessions, there was minimal implementation of the intended changes, and there were no detectable effects on achievement measures. After summer and winter institutes, protocols, and external assistance were introduced at the beginning of Phase 2, significant achievement gains were realized. There is no basis for assessing the relative contributions of the three augmentations to the implementation design. However, the results are consistent with evidence that more fully specified interventions that include external assistance typically produce better results than programs that do not provide effective external assistance and are not well specified (Bodilly, 1998; Borman, Hewes, Overman, & Brown, 2003; Datnow, Borman, & Stringfield, 2000; Datnow, Borman, Stringfield, Overman, & Castellano, 2003; Desimone, 2002; Goldenberg, 2004b).

Alternatively, it is possible that the appearance of significant achievement gains in the 3rd year (the 1st year of Phase 2), following the 2 years of the Phase 1 principals-only intervention, represented a delayed effect. Perhaps it took 2 full years of principals-only intervention to stabilize the program, and it was a coincidence that gains appeared at the end of the 3rd year, when the augmented intervention was introduced. There is no satisfactory basis for ruling out this plausible alternative explanation except to note that during the 2nd year of Phase 1, principals in the GR group were consistently reporting difficulties implementing the program. Observations by the authors confirmed the lack of implementation reported by principals. For example, in the school of a principal who was strongly committed to the program, there was an evident lack of implementation. A primary grade-level team was observed in November of the 2nd year of Phase 1 to be confused about what they were to do with the time set aside. The same team in spring was observed to be still struggling with how to effectively make use of grade-level meeting time. It was the consensus of the nine principals in the GR school group that they were all struggling, which resulted in the augmented program introduced in Phase 2. At the end of the project, there was also consensus among the GR principals that the augmentations were critical to the achievement gains.

Limitations

The generalizability of the results is limited in at least three respects. First, the nine schools in the experimental group were volunteer participants, and their experiences with grade-level teams and ILTs during Phase 1 likely increased their willingness to accept the augmented implementation design of Phase 2. During Phase 2, principals, ILTs, and the majority of teachers at

all GR schools generally understood program components and requirements and voluntarily decided each year of the study to continue their involvement. This is consistent with previous research: Implementation is typically more successful when school personnel understand program components and reach consensus, if not unanimity, to implement the program (Bodilly, 1998; Borman et al., 2003; Datnow et al., 2003; Hamilton et al., 2003). However, when school-based personnel are required by district or state authorities to implement PLCs or learning teams, staff response is likely to vary depending on the circumstances. Subsequent experiences implementing learning teams suggest that the adoption of the approach as a policy mandate is a solvable barrier to successful implementation, if it is perceived as giving teachers greater participation in addressing instructional challenges. However, we do not yet have sufficient data to report on the relative achievement gains in volunteer versus nonvolunteer samples (Gallimore, Goldenberg, Ermeling, & Saunders, in press).

Second, only elementary schools participated in the study (Grades K–5). Work on a secondary version of the GR program has demonstrated the feasibility of teams organized as subject specific, for example Algebra I teachers, but no outcomes data are yet available (Ermeling, 2005, in press). This pilot work does suggest that cross-subject or cross-course groups struggle to maintain a focus on improving instruction and learning in comparison with groups that are teaching the same subject matter, for example, ninth grade algebra or English.

Third, during Phase 2, all nine GR experimental schools were serviced by the developers (the authors). Work carried out by developers or as part of research and development programs generally produces larger effect sizes (Borman et al., 2005; Cronbach et al., 1980; Lipsey, 2003). Even a fully standardized and manualized intervention ready for implementation by nondevelopers is likely to achieve less robust effect sizes than obtained in the study reported here. But the magnitudes were of sufficient size that even a halving of effect would still rank the GR intervention in good company. Effect sizes less than 0.40 have been the norm in a number of meta-analyses of intervention or school reform programs, (e.g., Borman et al., 2003, 2005), with effect sizes of 0.20 to 0.25 being the average expectable outcome (Borman et al., 2005). Even if the developer-implemented nature of this study inflated the effect sizes obtained, the results suggest that the GR intervention is of sufficient promise to warrant further examination in other contexts. The intervention's agnosticism with respect to curricula and materials suggests a potential to work with a wide range of programs. Although many of the schools and grade-level teams focused on teaching reading and writing, other subjects could be and were accommodated by the protocol.

Implications and Conclusions

The approach tested in this study rested on the prosaic observation that students' academic needs and how to instructionally address them are

seldom discussed in the settings common in American schools: grade-level, department, and faculty meetings. This study provides some evidence that providing teachers with structural opportunities and skills to more often and consistently focus on improving classroom instruction and student learning might produce significant achievement gains.

Teachers' lack of focus on curriculum and instruction during meeting times was suggested by Miller and Rowan (2006) as a possible explanation for the absence of a relationship they and others have found between student achievement and teacher reports of staff collaboration, teacher empowerment, and supportive administration (also known as organic management). This hypothesis is consistent with the findings of this study, as well as Vescio et al.'s (2008) conclusion that a common feature of PLCs that reported achievement gains was teacher teams focusing on improving instruction and achievement. Time for collaboration by itself, even when administratively supported, is unlikely to improve achievement unless additional conditions are in place that structure its use.

Left unspecified by this study is what the teachers in GR schools were doing with more time in grade-level meetings focused on instruction and learning, and what instructional changes resulted from the collaborations, leaving another "black box" challenge for the research community. Did teachers simply become more focused, for example, developing better lesson plans addressed to specific student academic struggles? Did they continue using the same instructional practices as before, only more organized and focused? Or did the collaborative opportunities result in incremental adjustments or even more significant instructional changes? In possibly related findings, a recent study of the effects of standards-based instruction and accountability pressures suggested that they have galvanized increased reports of instructional changes (Hamilton, Stecher, Russell, Marsh, & Miles, 2008; Sawchuk, 2008). However, most of the variance in reported innovation is within schools, suggesting that a majority of teachers are responding individually rather than collaborating to converge on effective practices. This study suggests that provided the right conditions, leadership, and protocol, teachers will make use of collaborative time in ways that improve achievement. Providing time for collaboration and supportive administration alone appears to be insufficient to secure the desired outcomes.

Despite the fact that every school in the GR group moved up at least 1 decile in its state ranking, five of the nine schools still remained in 2002 in Decile 2 or 3, in other words, in the bottom third of schools in the state. This is meaningful improvement for schools that 3 years earlier ranked in the bottom 10th of the state. Using these statewide rankings as a measuring stick, there is clearly much more work to be done to understand and demonstrate how to improve student achievement at schools in poor and working-class communities that tend to score well below state and national averages. It remains an empirical question whether and to what extent well-implemented learning teams can continue to move such schools closer to or beyond state and national norms. The achievement gap is wide, and more than refocused

grade-level teams and ILTs will be needed to close it. Researchers and analysts from Comer (1980) and Edmonds (1979) to Darling-Hammond (2007) and Neuman (2008) have all noted the multiple and complex sources of the achievement gap and the corresponding multilayered and comprehensive responses that are needed if we are to attack it meaningfully: attracting and retaining top-flight teachers and administrators, creating schoolwide contexts that promote high levels of teaching and learning, providing economic resources for families, comprehensive health care and social services, among other measures (Goldenberg, 2004a; Rothstein, 2004). These measures require galvanizing political will and policy action to become reality, as Edmonds noted 30 years ago and no one has successfully challenged.

Finally, if evidence mounts that learning teams can help if not fully close the achievement gap, and adoption of the practice spreads, an old challenge is likely to arise to threaten their sustainability, and that concerns coherence. As Fullan, Bennett, and Rolheiser-Bennett (1990) noted almost two decades ago, the "greatest problem faced by school districts is not resistance to innovation, but the fragmentation, overload, and incoherence resulting from the uncritical acceptance of too many different innovations which are not coordinated" (p. 19). Decades later, the problem remains common (Allen, Osthoff, White, & Swanson, 2005; Childress, Elmore, Grossman, & King, 2007; Olson, 2007): "Districts wind up with a host of unrelated programs piled on each other, each with its own funding stream. . . . This lack of coherence is rampant" (Olson, 2007).

While interest in and excitement about learning teams and communities grow, we anticipate two potential risks as a result of this enduring incoherence. First, the critical time that teachers need to engage in structured inquiry as a learning team will compete with other state- and district-mandated and conventional professional development. Our experience suggests that those two engagements, learning teams and conventional professional development, can support each other in important ways if planned and managed carefully rather than simply piled on top of each other. Second, the critical time that teachers need to engage in structured inquiry as a learning team will compete with time needed to carry out countless other initiatives and mandates (implementing new curricula, changes in grouping practices, etc.). Again, our experience suggests that learning teams can be used productively to work through the actual practices associated with curricular or instructional initiatives and mandates if aligned and managed well.

In our experience, the potential of learning teams for improving classroom learning can easily become a casualty of well-intended district policies and initiatives that attempt to do too much within the actual time schools have for professional learning and development. Effective implementation of learning teams will require district leadership to improve coherence and alignment of professional development initiatives, perhaps by employing some of the same collaborative principles and practices school-level personnel are asked to adopt.

Note

This research was supported by the Spencer Foundation, the Office of Educational Research and Improvement, the Center for Culture and Health at the University of California, Los Angeles, and the LessonLab Research Institute. Grateful thanks to Nicole Kersting and James Stigler for their contributions to the statistical analysis of the results. Subsequent to the investigation summarized in this article, the program LT Learning Teams, based on the results, has been offered by Pearson Education, which employs two of the authors (Saunders and Gallimore). The views expressed are those of the authors and do not necessarily reflect those of the funding institutions.

References

Allen, L. E., Osthoff, E., White, P., & Swanson, J. (2005). *A delicate balance: District policies and classroom practice*. Chicago: Cross City Campaign for Urban School Reform. Retrieved March 23, 2008, from http://www.crosscity.org

Annenberg Institute for School Reform. (2005). *Professional learning communities: Professional development strategies that improve instruction*. Retrieved March 7, 2006, from http://www.annenberginstitute.org/images/ProfLearning.pdf

Berk, R. A. (1998). Review of the Stanford Achievement Test, Ninth Edition. In J. C. Impara & B. S. Plake, (Eds.), *The thirteenth mental measurements yearbook* (pp. 925–928). Lincoln, NE: Buros Institute of Mental Measurements.

Bird, T., & Little, J. W. (1986). How schools organize the teaching profession. *Elementary School Journal*, *86*(4), 493–512.

Bliss, J., Firestone, W., & Richards, C. (Eds.). (1991). *Rethinking effective schools: Research and practice*. Englewood Cliffs, NJ: Prentice Hall.

Bodilly, S. J. (1998). *Lessons from the New American Schools' scale-up phase: Prospects for bringing designs to multiple schools*. Santa Monica, CA: RAND.

Bolam, R., McMahon, A., Stoll, L., Thomas, S., & Wallace, M. (2005). *Creating and sustaining professional learning communities* (Research Report Number 637). London: General Teaching Council for England, Department for Education and Skills.

Borman, G., Slavin, R. E., Cheung, A., Chamberlain, A., Madden, N. A., & Chambers, B. (2005). Success for all: First year results from the national randomized field trial. *Educational Evaluation and Policy Analysis*, *27*(1), 1–22.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, *73*(2), 125–230.

Brandt, R. (Ed.). (1987). Collegial learning [Entire issue]. *Educational Leadership*, *45*(3).

Childress, S., Elmore, R., Grossman, A. S., & King, C. (2007, January 31). Note on the PELP Coherence Framework (PEL-010). Cambridge, MA: Public Education Leadership Project at Harvard University.

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

Comer, J. (1980). *School power: Implications of an intervention project*. New York: Free Press.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., et al. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements*. San Francisco, CA: Jossey-Bass.

Darling-Hammond, L. (2007, January 9). A Marshall Plan for teaching. *Education Week*. Retrieved October 12, 2008, from http://www.edweek.org.

Darling-Hammond, L., & Sykes, G. (Eds.). (1999). *Teaching as the learning profession: Handbook of policy and practice*. San Francisco, CA: Jossey-Bass.

Datnow, A., Borman, G., & Stringfield, S. (2000). School reform through a highly specified curriculum: A study of the implementation and effects of the Core Knowledge sequence. *Elementary School Journal*, *101*(2), 167–192.

Datnow, A., Borman, G. D., Stringfield, S., Overman, L. T., & Castellano, M. (2003). Comprehensive school reform in culturally and linguistically diverse contexts: Implementation and outcomes from a four-year study. *Education Evaluation and Policy Analysis*, *25*(2), 143–170.

Desimone, L. (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research*, *72*, 433–479.

Dufour, R. (2004). What is a "professional learning community"? *Educational Leadership*, *61*(8), 6–11.

Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership*, *37*(1), 15–27.

Elmore, R. (2002). *Bridging the gap between standards and achievement: The imperative for professional development in education*. Washington, DC: Albert Shanker Institute.

Elmore, R. F. (1999–2000). *Building a new structure for school leadership*. Washington, DC: Albert Shanker Institute.

Elmore, R. F., & McLaughlin, M. W. (1988). *Steady work: Policy, practice, and the reform of American education*. Santa Monica, CA: RAND.

Elmore, R. F., Peterson, P. L., & McCarthey, S. J. (1996). *Restructuring in the classroom: Teaching, learning, and school organization*. San Francisco, CA: Jossey-Bass.

Ermeling, B. A. (2005). *Transforming professional development for an American high school: A lesson study inspired technology powered system for teacher learning*. Unpublished doctoral dissertation, University of California, Los Angeles.

Ermeling, B. A. (in press). Tracing the effects of teacher inquiry on classroom practice. *Teaching and Teacher Education*.

Franke, M. L., Carpenter, T. P., Levi, L., & Fennema, E. (2001). Capturing teachers' generative change: A follow-up study of professional development in mathematics. *American Educational Research Journal*, *38*, 653–689.

Fullan, M. (1991). *The new meaning of educational change*. New York: Teachers College Press.

Fullan, M. (2000). The three stories of educational reform. *Phi Delta Kappan*, *81*, 581–584.

Fullan, M. G., Bennett, B., & Rolheiser-Bennett, C. (1990). Linking classroom and school improvement. *Educational Leadership*, *47*(8), 13–19.

Gallimore, R., Ermeling, B. A., Saunders, W. M., & Goldenberg, C. (in press). Moving the learning of teaching closer to practice: Teacher education implications of school-based inquiry teams. *Elementary School Journal* (special issue).

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*, 915–945.

Goldenberg, C. (2004a). Literacy for low-income children in the 21st century. In N. Unrau & R. Ruddell (Eds.), *Theoretical models and processes of reading* (5th ed., pp. 1636–1666). Newark, DE: International Reading Association.

Goldenberg, C. (2004b). *Successful school change: Creating settings to improve teaching and learning*. New York: Teachers College Press.

Goldenberg, C., & Sullivan, J. (1994). *Making change happen in a language minority school: A search for coherence* (Educational Practice Report No. 13). Washington, DC: National Center for Research on Cultural Diversity and Second Language Learning.

Hamilton, L. S., McCaffrey, D. F., Stecher, B. M., Klein, S. P., Robyn, A., & Bugliari, D. (2003). Studying large-scale reforms of instructional practice: An example from

mathematics and science. *Education Evaluation and Policy Analysis*, *25*(1), 1–29.
Hamilton, L. S., Stecher, B. M., Russell, J. L., Marsh, J. A., & Miles, J. (2008). Accountability and teaching practices: School-level actions and teacher responses. *Research in Sociology of Education*, *16*, 31–66.
Hord, S. M. (1997). *Professional learning communities: Communities of continuous inquiry and improvement*. Austin, TX: Southwest Educational Development Laboratory.
Kruse, S., Louis, K., & Bryk, A. (1994). *Building professional community in schools*. Madison, WI: Center on Organization and Restructuring Schools.
Lieberman, A. (1988a). Expanding the leadership team. *Educational Leadership*, *45*(5), 4–8.
Lieberman, A. (1988b). Teachers and principals: Turf, tension, and new tasks. *Phi Delta Kappan*, *69*, 648–653.
Lieberman, A. (Ed.). (1995). *The work of restructuring schools: Building from the ground up*. New York: Teachers College Press.
Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *Annals of the American Academy of Political and Social Science*, *587*, 69–81.
Little, J. (1982). Norms for collegiality and experimentation: Workplace conditions of school success. *American Educational Research Journal*, *19*, 325–340.
Little, J., & McLaughlin, M. (Eds.). (1993). *Teachers' work: Individuals, colleagues, and contexts*. New York: Teachers College Press.
Lortie, D. (1975). *School teacher: A sociological study*. Chicago: University of Chicago Press.
Louis, K. S., Kruse, S. D., & Marks, H. M. (1996). School wide professional community. In E M. Newman & Associates (Eds.), *Authentic achievement: Restructuring schools for intellectual quality* (pp. 179–203). San Francisco, CA: Jossey-Bass.
McDougall, D., Saunders, W. M., & Goldenberg, C. (2007). Inside the black box of school reform: explaining the how and why of change at Getting Results schools. *International Journal of Disability, Development and Education*, *54*, 51–89.
Miller, R. J., & Rowan, B. (2006). Effects of organic management on student achievement. *American Educational Research Journal*, 43(2), 219–253.
Murphy, J., & Hallinger, P. (1993). *Restructuring schooling: Learning from ongoing efforts*. Newbury Park, CA: Corwin.
National Commission on Teaching and America's Future. (2008). *NCTAF's Learning Teams Initiatives*. Retrieved March 19th, 2008, from http://www.nctaf.org/resources/demonstration_projects/LTInitiatives.htm
National Staff Development Council. (2001). *NSDC standards for staff development* (revised). Retrieved December 4, 2006, from http://www.nsdc.org/standards/index.cfm
Neuman, S. (Ed.). (2008). *Educating the other America: Top experts tackle poverty, literacy, and achievement in our schools*. Baltimore, MD: Paul H. Brookes.
Newmann, F., & Wehlage, G. (1995). *Successful school restructuring: A report to the public and educators*. Madison, WI: Center on Organization and Restructuring of Schools.
Newmann, F. M., & Associates. (1996). *Authentic achievement: Restructuring schools for intellectual quality*. San Francisco, CA: Jossey-Bass.
Olson, L. (2007, July 18). Project distills lessons of "coherent" district-level reforms. *Education Week*, *26*(43), 12–13.
O'Neil, J. (1995, April). On schools as learning organizations: A conversation with Peter Senge. *Educational Leadership*, *52*(7), 20–23.

Peterson, P., McCarthey, S., & Elmore, R. (1996). Learning from school restructuring. *American Educational Research Journal*, *33*, 119–153.

Rosenholtz, S. (1991). *Teachers' workplace: The social organization of schools*. New York: Teachers College Press.

Rosenholtz, S. J. (1989). Workplace conditions that affect teacher quality and commitment: implications for teacher induction programs. *Elementary School Journal*, *89*(4), 421–439.

Rothstein, R. (2004). *Class and schools*. Washington, DC: Economic Policy Institute.

Rowan, B. (1990). Commitment and control: Alternative strategies for the organizational design of schools. *Review of Research in Education*, *16*, 353–389.

Sarason, S. B. (1971). *The culture of the school and the problem of change*. Boston: Allyn & Bacon.

Sarason, S. B. (1972). *The creation of settings and the future societies*. San Francisco, CA: Jossey-Bass.

Saunders, W., & Goldenberg, C. (2005). The contribution of settings to school improvement and school change: A case study. In C. O'Donnell & L. Yamauchi (Eds.), *Culture and context in human behavior change: Theory, research, and applications* (pp. 127–150). New York: Peter Lang.

Sawchuk, S. (2008). Leadership gap seen in post-NCLB changes in U.S. teachers. *Education Week*, *28*(3), 1, 16.

Schmoker, M. (1996). *Results: The key to continuous school improvement*. Alexandria, VA: Association for Supervision and Curriculum Development.

Schon, D. A. (1983). *The reflective practitioner*. San Francisco, CA: Jossey-Bass.

Senge, P. (1990). *The fifth discipline: The art and practice of the learning organization*. New York: Currency Doubleday.

Smylie, M. A. (1994). Redesigning teachers' work: Connections to the classroom. *Review of Research in Education*, *20*, 129–177.

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.

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

Tharp, R. G., Estrada, P., Dalton, S. S., & Yamauchi, L. A. (2000). *Teaching transformed: Achieving excellence, fairness, inclusion, and harmony*. Boulder, CO: Westview.

Tharp, R. G., & Gallimore, R. (1988). *Rousing minds to life: Teaching, learning, and schooling in social context*. Cambridge, UK: Cambridge University Press.

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*, 80–91.

Visscher, A. J., & Witziers, B. (2004). Subject departments as professional communities. *British Educational Research Journal*, *30*(6), 785–800.

Whitehurst, G. J. (2002, March). *Research on teacher preparation and professional development*. Paper presented at the White House Conference on Preparing Tomorrow's Teachers, Washington, DC. Retrieved August 22, 2002, from http://www.ed.gov/admins/tchrqual/learn/preparingteachersconference/whitehurst.html

Manuscript received April 28, 2008
Revision received October 14, 2008
Accepted January 13, 2009