THE IMPACT OF ELEMENTARY MATHEMATICS COACHES ON STUDENT ACHIEVEMENT

ABSTRACT
Elementary mathematics coaches are placed in schools to construct leadership roles and to provide on-site, collaborative professional development addressing mathematical content, pedagogy, and curriculum in an effort to enhance instruction and improve student achievement. This 3-year randomized control study found that over time coaches positively affected student achievement in grades 3, 4, and 5. In these grades, this significant positive effect on student achievement was not evident at the conclusion of the first year of placement of a coach in a school but emerged as knowledgeable coaches gained experience and as a school's instructional and administrative staffs learned and worked together. The coaches in this study engaged in a high degree of professional coursework addressing mathematics content, pedagogy, and coaching prior to and during at least their first year of placement. Findings should not be generalized to coaches with less expertise.

Patricia F. Campbell
Nathaniel N. Malkus
UNIVERSITY OF MARYLAND

IN an effort to enhance student performance and achievement, schools and school districts across the nation are searching for mechanisms to provide for school-wide models of improvement in mathematics teaching and learning. Cognizant of the fact that traditional one-stop workshops and go-away professional conferences are ineffective routes for sustained growth (Ball & Cohen, 1999), many locales are embracing coaching as a model of professional development for teachers. Underlying this policy is the recognition that addressing the challenge of instructional

 0013-5984/2011/11103-0004 $10.00

change requires schools to become places where teachers can learn (Hawley & Valli, 1999). Indeed, recent reports have suggested that school-based mathematics specialists or coaches may be a vehicle to support the improvement of mathematics teaching and learning in elementary schools by targeting teachers' understanding and action (e.g., National Research Council, 2001). The intent is for a knowledgeable colleague with a deep understanding of mathematics and of how students learn, as well as pedagogical expertise, to serve as an on-site resource and leader for teachers. The function of the mathematics coach is to break the culture of teacher isolation whereby teachers work in private without observation or feedback and to collaborate with other professional development efforts in order to increase a school's instructional capacity (Neufeld & Roper, 2003). In this model, the mathematics coach or specialist catalyzes and sustains the implementation of content-focused work addressing mathematics curriculum, instruction, and assessment, supporting the emergence of collective professional habits that advance schoolwide growth and change as well as student learning and achievement (Campbell & White, 1997; Marzano, Walters, & McNulty, 2005; York-Barr & Duke, 2004). In practice, many schools are using their Title I funds to finance mathematics coaches, many rural areas are turning to on-site teacher leaders as a means of offering leadership to small populations of teachers spread over large geographical areas, and a number of urban districts are positioning mathematics coaches within their schools in an effort to advance test scores (Keller, 2007).

Background and Rationale

Whole-School Coaching

The rationale for the use of mathematics coaches as a vehicle for instructional change and teacher learning is rooted in research on learning and on effective models of professional development. In particular, Bransford, Brown, and Cocking (2000) cited three established principles regarding learning: (1) learners have prior knowledge, and if that knowledge is not accessed during instruction, learners may have difficulty learning or fail to learn new material; (2) those who learn—that is, those who retain and access or utilize what they have learned—actually understand; and (3) successful learners actively monitor their learning, reflecting on what they do and do not understand, and use strategies such as asking questions and explaining to oneself and to others to increase their knowledge. Ideally, coaching positions this vision of learning in the realm of what Desimone (2009) described as the "core conceptual framework" of professional development (p. 183). The core features of this framework are

- Content focus, whereby the coach facilitates activities in which teachers address mathematics content and pedagogy, as well as how students learn mathematics;
- Active learning, whereby the coach not only models instruction and coteaches but also engages with teachers in the work of teaching via coplanning, assessment design, observation, debriefing reflections addressing pedagogy and learning, and data-driven decision making;
- Coherence, whereby a coach supports teachers' efforts to understand, to examine ideas and relationships, and to connect prior knowledge and beliefs with new learning as well as teachers' efforts to reconcile state, district, and school policy demands;

- Duration, whereby a coach is consistently present to provoke and sustain attention toward addressing problems of practice; and
- Collective participation, whereby a coach facilitates inquiry, reflection, and experimentation within a community of practice focused on curriculum, instructional approaches, and interpretation of student meaning.

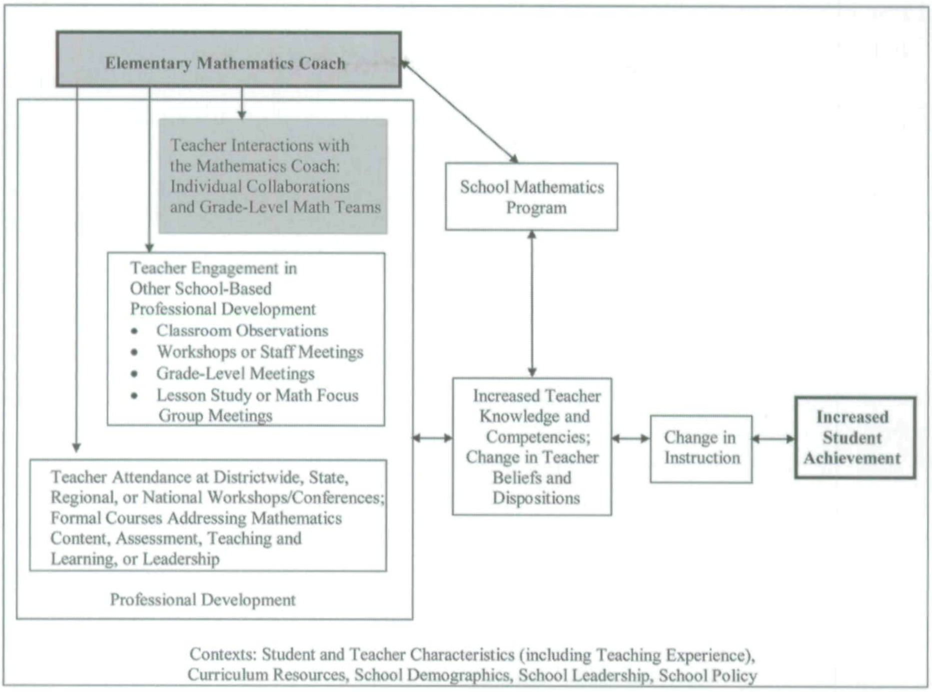
There is no single model of coaching; both past studies and current implementation efforts embody a variety of approaches. Joyce and Showers (1980) coined the term *peer coaching* to describe pairs of teachers providing reciprocal feedback and support to each other in an effort to improve their knowledge and skills. Seven years later, Loucks-Horsley and colleagues used the term *helping teachers* to describe those teachers who served to enhance the teaching of others through mentoring and professional dialogue (Loucks-Horsley et al., 1987, p. 83). Whether termed a specialist, coach, support teacher, or teacher leader, in many school districts today the intent is to place a highly knowledgeable teacher, who frequently does not have responsibility for the instruction of a classroom of students, in a school in order to advance instructional and programmatic change across the whole school.

There is a small body of research addressing the work and influence of specialists or coaches who work with multiple teachers (not peer coaches), generally addressing the intended practices of coaches plus teachers' perceptions of a coach's impact in terms of teachers' self-reports of changed instructional behavior, frequently within reading or writing instruction (Ai & Rivera, 2004; Dempsey, 2007; Rodgers & Rodgers, 2007; West & Staub, 2003). There is an emerging body of work that characterizes the challenges that whole-school coaches or specialists initially experience, such as understanding the curriculum across grades or courses, employing a variety of coaching modalities skillfully (e.g., joint lesson planning, coteaching, debriefing), understanding and addressing the growth of teachers, dealing with principals, transitioning from teaching students to coaching teachers, balancing multiple responsibilities and ambiguity, understanding and negotiating school culture, and setting priorities within time constraints (Neufeld & Roper, 2003; Poglinco & Bach, 2004; West & Staub, 2003).

This literature review identified only one publication reporting a positive relationship between mathematics student learning and professional development that highlighted coaching, but this posttest-only, intact-group design did not have a randomized control group nor did it address possible initial differences between participating schools (Foster & Noyce, 2004).

Conceptual Model

Desimone (2009) proposed a model representing critical components of the relationship between professional development, teacher knowledge and beliefs, classroom practice, and student learning. Figure 1 presents a modification of this model that incorporates coaches and characterizes forms of professional development distinct from coaching. The specification noted in Figure 1 identifies variables that may explain the effect of elementary mathematics coaches or interact to influence the effect of elementary mathematics coaches. Because improvements in student learning are ultimately tied to instructional change, the presumption may be that coaches focus solely on coaching for content instruction targeting individual teachers and



Note.—Highlighted cells are only a component in the treatment schools. This study only examined the bold components in this model.

Figure 1. Conceptual framework for studying the impact of elementary mathematics coaches on teachers and students.

grade-level teams via strategies such as coplanning, coteaching, observation, demonstration teaching, debriefing, and mentoring. But that is not always the case for elementary mathematics coaches, as they may also be called on to provide programmatic leadership as they assume the role of "community organizer" for mathematics in their schools (Neufeld & Roper, 2003).

For example, in their specification of the work that mathematics coaches may engage in within a school setting, a collaborative group of mathematicians, mathematics educators, and school district administrators in Virginia did specify "collaborate with individual teachers through coplanning, coteaching, and coaching" as an expectation, but they also included

- Assist administrative and instructional staff in interpreting data and designing approaches to improve student achievement and instruction;
- Ensure that the school curriculum is aligned with state and national standards and their school division's mathematics curriculum;
- Promote teachers' delivery and understanding of the school curriculum through collaborative long-range and short-range planning;
- Facilitate teachers' use of successful, research-based instructional strategies including differentiated instruction for diverse learners such as those with limited English proficiency or disabilities;
- Work with parents/guardians and community leaders to foster continuing home/school/community partnerships focused on students' learning of mathematics; and
- Collaborate with administrators to provide leadership and vision for a schoolwide mathematics program (Virginia Mathematics and Science Coalition, 2008, p. 1).

These programmatic efforts that are ancillary to content-focused coaching may influence teachers' beliefs and dispositions, if not knowledge. Coaches may also influence the degree to which teachers access other avenues for professional development. Each of these variables, along with instructionally focused mathematics coaching targeted to individual teachers or grade-level teams, may then affect teacher knowledge, competencies, beliefs, and dispositions, potentially yielding instructional change that influences student achievement. It is recognized that other elements influence the quality of instruction, such as teachers' attention to and management of students, how students make sense of and engage in instructional tasks, the quality of available resources, teachers' professional identity, the intended curriculum, provisions for opportunity to learn, teaching experience, and the nature of student-teacher interactions, as well as contextual factors in the classroom, school, and district.

This investigation addressed only some aspects of the conceptual model depicted in Figure 1. This report does not directly assess teacher knowledge, beliefs, or disposition, nor does it measure instructional practice; however, these variables do explain and potentially interact to influence effects on student achievement. Further, this work does not include a measure of the quantity or quality of other forms of professional development that teachers participating in this study may have engaged in, nor does it distinguish between school or district mathematics programs. This report broadly addresses the impact of coaches on student achievement revealed in a 3-year randomized control-treatment design, controlling for teacher experience, prior school academic tradition in mathematics, school size, and student demographics. Student achievement data were measured by the high-stakes, standardized assessment administered in Virginia in grades 3–5 as required by No Child Left Behind federal regulations.

Methodology

Coaches

Five school districts in Virginia, coded by the National Center for Education Statistics as representing urban and urban-edge/rural-fringe communities, participated in this study. Each district identified triples of schools with comparable student demographics and comparable traditions of student performance on state mathematics assessments. One large urban district identified two triples of schools, while two other midsize urban districts identified two triples and four triples of schools, respectively. One of the urban-edge/rural-fringe districts identified three triples of schools, while the other urban-edge/rural-fringe district identified a single triple of three schools. Triples of schools, rather than pairs, were identified in order to yield comparable school placement sites for two differing cohorts of coaches while maintaining corresponding control schools. This study accessed two cohorts of coaches who were participating in a funded teacher-enhancement effort addressing the development and refinement of mathematics content, pedagogy, and leadership courses for the coaches.

These 36 schools were each assigned a unique two-digit number; using a random number table, one school was randomly selected from each of the 12 triples by the first author as the site for placement of a Cohort 1 coach. School district personnel subsequently assigned the coaches from the ranks of experienced elementary teacher applicants to the identified treatment schools within districts. The first cohort of 12 coaches completed five mathematics content courses and one leadership-coaching course during 2004 and 2005 prior to placement, as well as a second leadership-coaching course during their first year of service as a coach. These mathematics coaches were placed in the selected schools at the beginning of the 2005–2006 school year. The school districts and coaches participating in the project agreed that for at least the next 2 years, these coaches and schools would participate in the research study, with these coaches remaining in their positions within their assigned schools during that time. Of the 12 coaches in the first cohort, 10 remained in their original treatment schools for 3 school years (August 2005 through June 2008). One treatment school closed due to redistricting after the 2006–2007 school year, and one coach in this cohort retired at that time, accepting a position as half-time supervisor of coaches across that school district. While the coach displaced by the school closing was reassigned to the school formerly supported by the newly retired coach, because the core features of duration and collective participation (Desimone, 2009) were affected by this change in coaching assignment, the analysis of the third year of data in this report includes only 10 Cohort 1 treatment schools.

Districts subsequently identified a second cohort of 12 prospective coaches, and these individuals completed an updated offering of the five mathematics content courses and the first leadership-coaching course during 2006 and 2007. Using the same random-selection procedure described previously, the first author identified one of the two control schools in each of the original triples as the site for the placement of a Cohort 2 elementary mathematics coach. School district administrators then assigned the Cohort 2 coaches from their districts to an identified school, with placement occurring during August 2007. As with the first cohort, the Cohort 2 coaches agreed to participate in the project for at least the next 2 years, with the school district agreeing to maintain their coaching assignments during that time. The analysis of data in this report includes the first year of treatment data as collected from this cohort in addition to the 2 prior years of control data drawn from these same schools. The third school in each triple retained its control status throughout this 3-year period.

This design permitted a controlled, 3-year data collection addressing the impact of coaches. As would be the case in practice, the coaches were not randomly assigned to a school; however, the schools that were identified to receive a coach were randomly selected from the triples of matched participating schools. It is presumed that, from the differing applicant pools that were available in Spring 2004 and again in Spring 2006, the school districts hired those that they perceived to be the best candidates for the Cohort 1 and the Cohort 2 coaching positions.

School districts were paid an allotment of $25,000 per coach per year in order to offset the cost of replacement classroom teachers. At the onset of the project, the school districts expected this subsidy to be in place only for the first 2 years of placement for the coaches in each cohort. During the second year of Cohort 1 placement, when Cohort 2 coaches were enrolled in their fourth preparation course, additional resources became available to provide the Cohort 1 subsidy for a third year. While the school districts received funding to assist in meeting the costs of 3 years of Cohort 1 placement and 2 years of Cohort 2 placement, this financial distinction was not known prior to the selection of a coach for either cohort.

Table 1. Coaches' Prior Professional Experience and Background

	Cohort 1			Cohort 2		
Source	*M*	*SD*	Range	*M*	*SD*	Range
Master's degree recipient (0/1)	.5	.52	0 to 1	.42	.51	0 to 1
No. of graduate credits (no master's degree)[a]	21.17	19.26	0 to 57	2.57	3.64	0 to 9
No. of graduate credits beyond master's degree[b]	4.17	7.05	0 to 17	5.4	8.85	0 to 21
No. of credits in mathematics content	10.42	7.44	3 to 24	6.0	3.38	3 to 12
No. of credits in mathematics education	7.50	6.72	3 to 27	4.0	2.26	0 to 9
Years of teaching experience (elementary school)	14.83	9.60	5 to 31	12.92	8.63	4 to 29
Years of teaching experience (middle school)	4.67	4.73	0 to 10	.42	2.24	0 to 5

[a] These statistics are calculated on the data submitted by the subset of coaches who did not have a master's degree at the beginning of their coursework for the coaching program.

[b] These statistics are calculated on the data submitted by the subset of coaches who did have a master's degree at the beginning of their coursework for the coaching program.

Coaches were paid an annual stipend of $2,500 for participating in the data collection phase of the study. All 24 coaches were female. Eight of the coaches were African American, one coach was Asian, and the remaining coaches were Caucasian. A summary of the prior professional experience and backgrounds of the 24 coaches as of the time they began their first course is noted in Table 1. As indicated, Cohort 1 was a somewhat more seasoned group of teachers than were the individuals in Cohort 2.

Professional development of coaches. The 24 coaches completed five mathematics courses designed for them by a course-development team consisting of college mathematics and mathematics education faculty, experienced school-district mathematics coordinators, and experienced classroom teachers from Virginia. The Numbers and Operations, Geometry and Measurement, Algebra and Functions, and Probability and Statistics courses used relevant modules of the Developing Mathematical Ideas professional development series (e.g., Schifter, Bastable, & Russell, 1999), together with other materials created or culled by the development teams. The Rational Numbers and Proportional Reasoning course accessed differing case-based materials (Fosnot & Dolk, 2002; Lamon, 1999). Coaches completed these courses at one of three locations, each with different instructors. All courses were team taught, with the team typically including both a mathematician and a mathematics educator. There was variation in the emphasis given to the goals of increasing teachers' content and pedagogical knowledge. For example, the Numbers and Operations course emphasized pedagogical issues, with numerous activities that required teachers to examine children's thinking. In contrast, much of the Geometry and Measurement course involved teachers grappling with mathematical concepts as students, focusing more on the mathematics content and less on the pedagogical implications. The first educational leadership course accessed standards documents from the National Council of Teachers of Mathematics (1991, 2000) as well as *Adding It Up* (National Research Council, 2001). Coaches were placed in schools following completion of this first leadership course and the mathematics courses. Subsequently, during their

Table 2. Percent of Mean Contract-Day Time over Coaching Activities by Cohort and Year

Activity	Cohort 1 2005–2006 (Year 1)	Cohort 1 2006–2007 (Year 2)	Cohort 1 2007–2008 (Year 3)	Cohort 2 2007–2008 (Year 1)
Coaching teachers (individual and grade-level teams)	21.9	13.1	12.9	10.2
Preparing for teaching/coaching	11.8	12.4	12.5	11.8
Supporting assessment	10.6	13.5	13.7	12.5
Teaching or supporting students (not demonstration or coteaching)	3.0	4.4	4.5	3.6
Supporting the school mathematics program	5.0	4.2	5.1	5.1
Performing school-based duties	6.5	9.2	10.4	9.8
Materials management/communication tasks	9.7	11.0	11.8	11.4
Attending meetings	9.2	6.8	6.7	9.5
Engaging in personal professional activity	13.2	14.7	10.9	14.4
Noneducational activities (lunch, travel, all-school event)	9.0	10.8	11.3	11.8

first year of placement, they completed the second educational leadership course focused on coaching, accessing a variety of published references (e.g., Wood, Nelson, & Warfield, 2001) as well as accompanying video segments and cases (Miller, Moon, & Elko, 2000; West & Staub, 2003).

Work activities of coaches. To account for their changing actions in school within and across the school year, coaches detailed their daily activities using a data-collection-transmittal program operating on a personal digital assistant (Dell Axim X50; PDA). Instructional Specialist Activity Manager (ISAM) is a menu-oriented entry interface that allows coaches to log the duration and category of their daily activities.

Within the Daily Activity Log option of ISAM, coaches chronologically indicate the duration of an activity and then click the primary identification of that activity. Based on a branching network, activities of interest trigger the presentation of more detailed subchoices, which coaches again select by clicking on the button of interest. After the activities of a complete day are entered, coaches may review the day's entries and, if necessary, modify the listing prior to confirmation. Daily confirmed data are subsequently transmitted over the Internet onto a comprehensive data-management platform housed on a server at the authors' university.

The ISAM data from the PDAs characterize the duration and nature of coaches' activity across up to 3 years of placement in a school. These daily data are not being utilized in the statistical analyses of student achievement data addressed in this report because the control schools have no parallel data to include in the analyses. Future analyses will investigate the relationships between coaching activities and student achievement using only data from coached schools. However, a descriptive summary of these data is included in Table 2 as a means of quantifying the activity of the coaches.

The contract days for the 24 coaches ranged from 7 hours to 8 hours, depending on the school district, with a mean length of 7 hours, 22 minutes. Thus, on average, the coaches were paid to spend 36 hours, 50 minutes at school each week with a 40-week school calendar. In terms of hours per day, the values in Table 2 may be interpolated according to the formula that 13.6% is equivalent to 5 hours per week (1 hour per day).

The amount of time that each of the two cohorts spent coaching teachers (observation of teaching, demonstration teaching, coteaching, coplanning, debriefing, meeting with grade-level teams) was more consistent when the year of work was constant (both cohorts in 2007–2008) than when the extent of experience was constant (Cohort 2 in 2007–2008 and Cohort 1 in 2005–2006). This may mean that during 2007–2008 there were common outside influences affecting coaches' decisions as to how much available time they had to spend working with individual teachers, with the amount of time that Cohort 1 coaches spent coaching decreasing over time. The shifting of coaches' time to assessment responsibilities is probably a local school response to concerns associated with district assessment demands, with frequency of time devoted to assessment responsibilities evident in all districts and a consistently modal feature in the urban districts. In contrast, the time coaches spent teaching or supporting students without an observing teacher present (thus not coaching) varied by individual coaches, not districts. Thus, this was most likely a reflection of a principal's request and not a consistent response to district policy or pressure. The amount of time that coaches spent addressing communication, such as e-mail correspondence, was more comparable by academic year than by year of expertise. All of the participating school districts provided e-mail addresses and access to their instructional and administrative staffs. The increase in communication time evidenced between 2005–2006 and 2007–2008 is most likely a reflection of changes in school culture and not the project. For the most part, time allocated to noneducational activities was at midday and reflected a break for lunch.

While these patterns of coaching activity were not likely unique to this implementation, the dominant school-based duty is most likely a project-related artifact. The coaches advised each other to "volunteer for bus duty" as a way to build entrée into their school placements, noting that this was a time when few, if any, teachers would be available to meet with a coach. The time devoted to personal professional development is also influenced by the research design, as all coaches completed the second leadership-coaching course during their first year of placement. Further, many of the coaches in each cohort completed an additional graduate course or two during their first year of placement as they completed requirements for a master's degree within the following summer or fall semester.

Data Sources

Student-level data. All students in grades 3 through 8 in Virginia are expected to complete a statewide standardized achievement test in mathematics termed the Standards of Learning Assessment (SOL) annually. Through administration of this high-stakes measure and the aligned collection of student demographic data, Virginia meets the expectations for assessment as required under the No Child Left Behind federal regulation. SOL data include a total scale score for mathematics (possible scores ranging from 200 to 600), as well as five subscale scores (possible scores ranging from 0 to 50) addressing number and number sense; computation and estimation; measurement and geometry; probability and statistics; and patterns, functions, and algebra. For purposes of labeling performance, total scale scores at or above 500 are deemed advanced passing, total scale scores from 400 to 499 are

Table 3. Descriptive Statistics for SOL Mathematics Scores by Grade and Cohort across Years

	2005–2006		2006–2007		2007–2008		
	Cohort 1 12 Schools	Control 24 Schools	Cohort 1 12 Schools	Control 24 Schools	Cohort 1 10 Schools	Cohort 2 12 Schools	Control 12 Schools
Grade 3:							
Mean	494.43	484.25	500.44	489.13	504.48	478.79	492.65
SD	73.27	76.82	66.03	70.96	73.55	80.58	80.89
Students (*n*)	912	1,927	825	1,823	808	922	990
Teachers (*n*)	46	99	43	96	42	46	46
Grade 4:							
Mean	459.25	449.77	488.31	477.32	475.99	462.71	469.96
SD	74.13	79.94	68.31	66.85	78.26	78.90	79.73
Students (*n*)	945	2,014	914	1,733	877	838	948
Teachers (*n*)	43	99	44	85	38	43	38
Grade 5:							
Mean	482.93	478.64	515.88	504.93	503.89	481.03	493.69
SD	84.04	85.53	68.65	72.58	81.83	86.35	81.75
Students (*n*)	951	2,018	914	1,735	878	839	948
Teachers (*n*)	40	93	38	82	34	34	40

deemed passing proficient, and total scale scores below 400 are deemed failing. The SOLs are administered annually, typically during the last half of May.

While the SOLs in grades 3 and 5 have been administered since the 2001–2002 school year, the grade 4 SOLs were administered for the first time during the 2005–2006 school year, the first year of placement of coaches in this study. Further, while the grade 4 and grade 5 SOLs only assess content associated within the grade-level standards of that single grade, the grade 3 SOL assessment measures content from kindergarten through grade 3. Thus, the analysis that follows separately considers the third-, fourth-, and fifth-grade students' scores across 3 years (2005–2008).

For each grade level, the primary dependent variable was the overall SOL mathematics scale score across 3 years. This dependent variable posed two challenges for these analyses. First, in each grade level, the distribution of test scores shifted in a nonlinear fashion as the difficulty of the SOL mathematics assessment varied from year to year. Second, while the range for the SOL scale scores was 200 – 600, there was a substantial but varying number of students in each year and in each grade achieving a score of 600. This ceiling was problematic because it varied with the difficulty of the test and because it increased the type II error rate, making detection of significant treatment effects less likely. Because of these two challenges, standardizing the dependent variable for each grade (by standardizing scores within each year, and then standardizing those scores across years) was not a feasible option. Therefore, in order to control for differences in the testing year, this analysis used the scale scores in the original metric and included binary indicators for each testing year in our models. Table 3 presents selected descriptive statistics for the analyzed mathematics SOL scores by grade and cohort. Note that there was shifting student enrollment in all of the participating schools over the 3 years, marked by a reduction in two districts. While this loss of student enrollment occurred in all of the participating schools in these two districts, it was more pronounced in the control schools.

Including prior student-level scale scores in the model was also problematic, in part because of the ceiling effect and the inability to standardize scores. Further, there were no prior-year SOL scores for grade 3 students in any school year or for grade 5 students in the 2005–2006 school year. In addition, when prior-year SOL assessments were administered, missing data due to student mobility were present and not evenly distributed across schools as coded by Title I status and minority composition. Including a student-level prior achievement variable in the analysis would remove approximately one-fifth of the students from the sample and restrict the analysis to the achievement scores of nonmobile students, biasing the sample.

While controlling for student-level prior achievement was not possible, two school-level measures of the prior academic tradition were included. Low Academic Tradition and High Academic Tradition identified those schools whose mean 2004–2005 SOL mathematics scale scores in both grades 3 and 5 were one standard deviation below or one standard deviation above the 2004–2005 sample mean for all 36 schools.[1] By identifying schools at the tails of the distribution of the 2004–2005 SOL scores, these variables provide a measure of school tradition prior to the initiation of treatment. While the three schools within each of the 12 triples of schools were comparable in terms of their demographics and prior student performance on the state mathematics assessments, there were differences between the triples of schools, reflecting the differing communities and districts participating in the study. The variables of Low Academic Tradition and High Academic Tradition serve as a control for prior school-level differences between the triples of schools, while randomization serves to balance prior differences within each triple of comparable schools.

Teacher- and classroom-level data. Over the course of the 3 years, there were 1,593 teachers of kindergarten through grade 5 mathematics in the 36 cooperating schools who agreed to participate in the project. As indicated in Table 4, the teachers in the three cohorts of schools did not differ substantively in terms of their professional experience or demographics. In Table 4, Cohort 1 schools are those randomly assigned to treatment status for 3 years (2005–2008); Control 1/Cohort 2 schools are those randomly assigned to control status for 2 years (2005–2007) and then treatment status for 1 year (2007–2008); the remaining schools randomly assigned to control status for 3 years (2005–2008) are labeled as Control Throughout. Note that while the coaches were responsible for building the capacity of teachers across grades K–5 in their schools in order to enhance instruction and increase student learning, the analysis of student achievement only addresses mathematics achievement data from the assessed grades 3–5. Thus, the number of teachers in Tables 3 and 4 differ.

The teachers of mathematics in the control and treatment schools completed demographic surveys upon entry into the study. These demographic surveys also included information regarding years of teaching experience, educational background, and certification. No measures of teachers' mathematical content knowledge were available.

School-level data. The primary school-level variable of interest in this study was whether a school was randomly assigned to receive a coach. While the ISAM data provide detailed information on the activities of the coaches, these data are not included in this control-treatment analysis because control schools have no parallel data. An investigation of the relationship between the activities of coaches and student achievement will be addressed in future analyses limited to treatment schools. Other school-level measures, in addition to academic tradition, were school size and Title I status.

Table 4. Grade K–5 Teachers' Professional Demographics (2005–2008)

	Cohort 1[a]			Control 1/Cohort 2			Control Throughout		
	2005–2006	2006–2007	2007–2008	2005–2006	2006–2007	2007–2008	2005–2006	2006–2007	2007–2008
Master's degree (%)	36.4	38.2	35	40.5	46.7	45.9	35.1	40.6	41.8
Years of teaching experience (%):									
1–2 years	16.3	18.7	17.9	24.8	15.6	14.8	14.5	11.3	13.1
3–4 years	10.1	9.8	6.8	11.8	18.0	17.2	8.4	14.3	14.8
5–9 years	34.9	32.5	33.3	27.0	28.7	29.5	19.8	21.1	16.4
10+	38.8	39.0	41.9	35.9	37.7	38.5	57.3	53.4	55.7
Certified teachers (%)	98.4	95.9	96.6	94.1	94.3	95.9	93.9	94	96.7
Female (%)	93.8	93.5	92.3	83.7	86.1	83.6	87.8	86.5	88.5
Race/ethnicity (%):									
American Indian/Alaskan Native	0	0	0	.7	.8	.8	0	0	.8
Black/African American	25.6	22.0	20.5	30.7	30.3	27.0	23.7	27.1	27.9
White	72.1	74.0	76.1	65.4	68.0	70.5	73.3	68.4	64.8
Asian/Asian American	0	0	0	1.3	0	0	1.5	2.3	1.6
Hispanic/Latino	1.6	2.4	1.7	.7	.8	.8	0	0	0
More than one race	.8	1.6	1.7	.7	0	0	.8	1.5	3.3
Teachers (*n*)	333	316	309	393	333	332	356	323	319
School attrition (%)	. . .	29.4	23.7	. . .	36.4	30	. . .	32.6	22.9

[a] 12 schools in 2005–2007; 11 schools in 2007–2008.

Models and Results

To determine whether coaches affect student mathematics achievement scores as measured by standardized assessments, this analysis accessed data on 24,759 student test scores drawn from grades 3, 4, and 5 of 36 treatment and control schools over 3 years. To account for the nested structure of the data (students nested within classrooms, nested within schools), hierarchical linear modeling (HLM) was used to analyze the data.[2] Across these 3 years, this sample included 1,169 teachers/classrooms of students in grades 3, 4, and 5, of which 368 were in Cohort 1 schools, 406 were in Control 1/Cohort 2 schools, and 395 were in control schools throughout. Students and teachers were linked to their schools in the sample in each year of the study.

The student data were analyzed separately by grade ensuring complete independence between measures at the student level from one year to the next. However, at the classroom level, teachers could have taught in a school in the first, second, or third year of the study, remaining in a single grade in a participating school for up to 3 years. This violates the assumption of independence of teachers across years. Likewise, there is not independence between schools because schools were entered into the analytic sample by year. This violation of the assumption of school-level independence is problematic because it increases the type I error rate, raising the likelihood of identifying significant treatment effects by an indeterminate amount. Nevertheless, the analyses that follow are offered with this limitation because no practical analytic alternative is available for analyzing the data and because there is much to be learned from this unique data set. Any conclusions drawn from these analyses should be made with this limitation in mind.

Some portion of the variance in students' scores can be attributed to the class to which a student belongs and some to the school that a student attends, rather than to individual or treatment differences. By estimating a baseline model for each grade, with no predictors at the student, class, or school level, it is possible to determine the interclass correlation coefficient (ICC) for the class level (level 2) and for the school level (level 3). These measures indicate the proportion of the total variance of students' scores associated with the classes and schools attended by students. The ICC for level 2 indicated that 11.0% of the variance in third-grade scores, 16.3% of the variance in fourth-grade scores, and 15.8% in fifth-grade scores were associated with classes. The ICC for level 3 showed that 9.5% of the variation in grade 3 students' scores was associated with the school the student attended, with school accounting for 9.3% of the variance in grade 4 and 7.4% in grade 5. Thus, on average, 14.4% of the difference in student scores in this sample is associated with class grouping and 8.8% with school grouping, while 76.8% is attributed to individual student differences. The level 2 reliabilities of the SOL mathematics scale score all exceeded .70 with a mean of .76, indicating ample reliability to identify class differences in mathematics scores within the same school. The level 3 reliabilities were slightly lower (.69 in grade 3, .60 in grade 4, .54 in grade 5) with a mean of .61, indicating sufficient reliability of the school means. Power analysis indicated that sample sizes and ICCs would permit detection of a minimum standardized effect size of .20 with a power of .80 (Raudenbush, Liu, Spybrook, Martinez, & Congdon, 2006; Raudenbush, Martinez, & Spybrook, 2007).

Treatment versus Control Analysis

Models. For each grade-level analysis, the base student-level model (shown below) included controls for student age at time of testing (AgeTest), gender (Female), limited English proficiency status (LEP), special education status (SpecEd), free- and/or reduced-meal status (FARM), and minority status (Minority). The model also includes two binary indicators for the tests in the second and third year of the study (2007 Test and 2008 Test, respectively) to capture the differences in the mathematics achievement score (Y) of student i, in class j, in school k. The reference category for the year of test was the first year of the study, the 2005–2006 school year. All of the variables were centered on the grand mean and controlled for these student characteristics across all three levels (Raudenbush & Bryk, 2002).

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}\text{AgeTest}_{ijk} + \pi_{2jk}\text{Female}_{ijk} + \pi_{3jk}\text{LEP}_{ijk} + \pi_{4jk}\text{SpecEd}_{ijk} + \pi_{5jk}\text{FARM}_{ijk} + \pi_{6jk}\text{Minority}_{ijk} + \pi_{7jk}\text{Test 2007}_{ijk} + \pi_{8jk}\text{Test 2008}_{ijk} + e_{ijk}.$$

The random effects for special education and FARM status were significant at the classroom and school levels in some grade-level analyses. In each grade-level analysis, an analysis was completed to determine whether these relationships varied across groups at both the classroom and school level. In models where these random effects were significant ($p < .05$), the error terms are random. However, these random slopes are not modeled to preserve similarity across all of our models and because these random effects are not the focus of this study.

Classroom-level variables included measures of teacher experience and education. Because prior research has indicated that students in classrooms with novice teachers may have significantly lower achievement on standardized mathematics assessments and because urban schools have higher teacher mobility with teachers who have less than 5 years of teaching experience, this analysis controlled for years of teaching experience at the teacher or classroom level. This analysis compared the SOL achievement scores of students whose teachers had 5 to 9 years of teaching experience to the scores of students whose teachers had 2 or fewer years (1–2 Years Experience), 3 or 4 years (3–4 Years Experience), or 10 or more years of teaching experience (10+ Years Experience). The model also specified a binary indicator to indicate teachers who held a master's degree. Measures of classroom composition, such as percentage of minority students in a class, are controlled for with the grand mean centered, student-level variables and are not included at the classroom level.

For all analyses, the level 2 model shown below included the binary indicator for teachers with master's degrees and the measures of teacher experience on the intercept, π_{0jk}. All variables were grand mean centered with fixed effects:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}\text{Masters}_{jk} + \beta_{02k}\text{1–2YrsExp}_{jk} + \beta_{03k}\text{3–4YrsExp}_{jk} + \beta_{04k}\text{10+YrsExp}_{jk} + r_{0jk}.$$

With two exceptions (free and/or reduced meals and special education status), all level 1 predictors had fixed effects at the class and school levels such that the effect of each predictor was equal to the average group effect of that predictor.

The primary independent variables in these analyses were school-level variables that indicated whether a school had the services of an elementary mathematics coach. Since Cohort 1 schools had coaches in place as of 2005–2006 and Cohort 2 schools only had coaches in the third year of the study (2007–2008), separate indicators were used for each cohort (Cohort 1 and Cohort 2 Year 3). Additional school-level variables provided controls for Title I services, school size, and the academic tradition of the school. In particular, Title I School served as a binary variable indicating federal funding for schoolwide Title I services, when 40% or more of a student body qualified for Title I support. Minority composition at the school level was not included in the final models because, after controlling for Title I School status, the proportion of minority students in a school was not a significant variable, did not improve model fit, and did not affect the estimation of other effects. This model also included a standardized measure of school size (School Size) with a mean of 0 and a standard deviation of 1.

The level 3 intercept model shown below includes an indicator of whether the school was eligible for schoolwide Title I services (Title I School), the measure of school size (School Size), and the indicators for academic tradition (Low Academic Tradition and High Academic Tradition). This model addresses differences between treatment schools and control schools by cohort using two binary indicators for each treatment cohort. The first (Cohort 1) indicated the 10 schools that had the same Cohort 1 coach across all 3 years of the study, as well as the two schools that had the same Cohort 1 coach across the first 2 years of the study. The second, (Cohort 2 Year 3) indicated the 12 schools that had a Cohort 2 coach placed during the third year of the study. These 12 schools are included in the control group for the first 2 years of the study:

$$\beta_{00k} = \gamma_{000} + \gamma_{001}\text{Title I School}_k + \gamma_{002}\text{High Academic Tradition}_k \\ + \gamma_{003}\text{Low Academic Tradition}_k + \gamma_{004}\text{School Size}_k + \gamma_{005}\text{Cohort 1}_k \\ + \gamma_{006}\text{Cohort 2 Year 3}_k + u_{00k}.$$

The combined three-level model for the control versus treatment analyses for each grade is specified below.

$$Y_{ijk} = \gamma_{000} + \gamma_{001}\text{Title I School}_k + \gamma_{002}\text{High Academic Tradition}_k \\ + \gamma_{003}\text{Low Academic Tradition}_k + \gamma_{004}\text{School Size}_k + \gamma_{005}\text{Cohort 1}_k \\ + \gamma_{006}\text{Cohort 2 Year 3}_k + \gamma_{010}\text{Masters}_{jk} + \gamma_{020}\text{1–2YrsExp}_{jk} + \gamma_{030}\text{3–4YrsExp}_{jk} \\ + \gamma_{040}\text{10+YrsExp}_{jk} + \gamma_{100}\text{AgeTest}_{ijk} + \gamma_{200}\text{Female}_{ijk} + \gamma_{300}\text{LEP}_{ijk} + \gamma_{400}\text{SpecEd}_{ijk} \\ + \gamma_{500}\text{FARM}_{ijk} + \gamma_{600}\text{Minority}_{ijk} + \gamma_{700}\text{Test 2007}_{ijk} + \gamma_{800}\text{Test 2008}_{ijk} + u_{00k} \\ + r_{0jk} + e_{ijk}.$$

Results. Findings from analyses of grades 3, 4, and 5 student mathematics achievement data are presented in Table 5, with the statistics for differing independent variables presented in each row and the grouped columns specifying the grade. In all three grades the Cohort 1 coefficients were positive and significant. In grade 3, students in Cohort 1 schools averaged 10.7 points, or 14% of the grade 3 pooled standard deviation,[3] higher than the mean on the SOL mathematics scaled score ($p = .040$). In grades 4 and 5, students in Cohort 1 schools scored 13.7 ($p = .0095$) and 15.3 ($p = .004$)

Table 5. REML Parameter Estimates and Standard Errors for Coaching Effects on Student SOL Mathematics Overall Scale Score by Grade

	Grade 3		Grade 4		Grade 5	
Scale Score	Coefficient	*SE*	Coefficient	*SE*	Coefficient	*SE*
Intercept	493.95***	2.34	470.63***	2.45	496.95***	2.46
Student variables:						
Age at test	−8.49***	1.61	−14.37***	1.43	−13.29***	1.56
Female	−1.86	1.26	−7.72***	1.55	−.69	1.64
LEP	−8.02	7.25	−18.71**	6.33	−21.89**	6.74
Special education	−41.84*** a, b	3.45	−40.17*** a, b	3.30	−51.53*** a, b	4.36
Free or reduced meal	−17.10***	2.20	−17.80*** a	2.41	−18.24*** a	2.78
Minority	−35.76***	2.16	−33.74***	2.21	−27.68***	2.08
2007 test	3.71	5.26	18.52**	5.20	20.32***	5.37
2008 test	7.32	5.96	12.12	7.31	11.30	6.09
Teacher variables:						
Master's degree	2.12	2.42	.21	2.94	−5.14	3.39
1–2 years experience	−7.87	5.15	−4.96	5.24	−13.67*	5.98
3–4 years experience	−6.54	4.27	−7.39	5.02	−6.01	5.38
10+ years experience	.76	2.95	10.45*	4.11	10.60**	3.75
School variables:						
Title I school	7.04	5.57	4.70	5.67	12.99*	5.34
High academic tradition	35.53***	8.97	39.51***	9.71	51.45***	6.47
Low academic tradition	−13.26	7.65	−17.06*	7.88	−2.08	9.65
School size	−2.78	2.48	−8.43**	2.69	−7.98**	2.44
Cohort 1	10.71*	5.05	13.68**	5.17	15.25**	5.08
Cohort 2 Year 3	−3.89	7.43	9.08	8.31	−3.16	8.74
	Grade 3		**Grade 4**		**Grade 5**	
	Variance	χ^2	Variance	χ^2	Variance	χ^2
Variance estimates:						
Student-level variance (σ_2)	3,927.03	. . .	3,776.08	. . .	4,319.54	. . .
Class-level variance ($\tau_{00\pi}$)	382.88***	676.93	523.16***	518.86	554.62***	569.58
School-level variance ($\tau_{00\beta}$)	307.98***	288.36	277.73***	238.65	289.61***	207.43

[a] Unmodeled level 2 random effect.
[b] Unmodeled level 3 random effect.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

points above the mean, respectively, which corresponds to 18% *SD* on the grade 4 tests and 19% *SD* on the grade 5 tests.

In contrast, the Cohort 2 Year 3 variable, representing the placement of a first-year coach during the third year of the study, was not significant in any of the analyses. With only this single year of treatment data for these 12 schools, the variance for the Cohort 2 measure was, as expected, larger than the variance of the Cohort 1 variable, which had 3 years of data for 10 schools and 2 years of data in 2 schools. These disparate findings between the 3 years of Cohort 1 treatment data and the single year of Cohort 2 treatment data are addressed below.

At the classroom level, students whose teachers had a master's degree did not have significantly different SOL scores than their peers who were taught by teachers without a graduate degree. The effects of teacher experience were not consistently significant across the grade-level analyses but were in the expected direction, with students with early-career teachers having somewhat lower SOL scores than did students of teachers with 5 to 9 years of teaching experience. In the grade 5 analyses, on average, students of novice teachers scored 13.7 points (17% *SD*, $p = .023$) lower on the SOL assessment than did their peers in classrooms with more experienced teachers. The mean scale scores of grade 4 and grade 5 students whose teachers had 10 or more years of teaching experience were approximately 10.5 points (14% *SD*, $p = .012$, in grade 4; 13% *SD*, $p = .005$, in grade 5) greater than the mean scale scores of students in these grades whose teachers had 5 to 9 years of teaching experience. Thus the magnitude and significance of student achievement differences associated with teacher experience generally increased by grade.

Across all three grades, the individual effects of age, poverty, race/ethnicity, and special education status had consistently significant negative effects on total SOL mathematics scores ($p < .001$). The effects of gender and LEP status were negative but not consistently significant. The only significant impact of gender was in grade 4 for females; in grades 4 and 5, LEP students had significantly lower scale scores. In 2007 the average SOL mathematics scale score was significantly higher for grades 4 and 5 (25% *SD*, $p < .01$). In 2008 the year-of-test effects were not significant; however, the magnitude of the coefficients underscores the importance of including these controls in the model.

This model explained similar amounts of variance across all three grade-level analyses (see bottom of Table 5). Compared to the fully unconditional baseline model with no predictors at any level, the final models explained between 13% and 15% of the individual or level 1 variance in student mathematics achievement scores. The models explained between 41% and 47% of the variance at the classroom level, and between 41% and 50% of the variance at the school level.

Cohort-by-Year Analysis

Models. As reported previously, this analysis indicated a significant effect associated with the Cohort 1 variable, but the treatment effect was not evident in the Cohort 2 Year 3 variable. This difference between findings may reflect the differing amounts of time the coaches in the two cohorts had to work with teachers and the school mathematics program. To further address possible differences between treatment cohorts, a second set of analyses were completed to examine the effect of each treatment cohort by year. In these analyses the Cohort 1 variable was replaced with three treatment variables, one for Cohort 1 schools in each year of the study, and the Cohort 2 Year 3 variable remained in the model unchanged. Examining cohorts by year is a more conservative analytic approach because each cohort-by-year variable is based on a smaller set of data (12 Cohort 1 schools in Years 1 and 2; 10 Cohort 1 schools in Year 3). In contrast, over the 3 years of data collection, 34 sets of school data were included in the Cohort 1 indicator in the prior analyses. While it is recognized that this is likely to increase the standard error for these coefficients, as was the case with the Cohort 2 Year 3 coefficient in the previous model, this conservative analysis was completed in order to determine whether a pattern was evident across the cohort-by-year coefficients.

To complete this second analysis examining differences between control schools and treatment schools by cohort and year, the Cohort 1 variable in the model was replaced with three variables for Cohort 1 schools, noting the first (Cohort 1 Year 1), second (Cohort 1 Year 2), and third (Cohort 1 Year 3) years of the study. Since Cohort

2 schools only had a coach in the third year of the study, the Cohort 2 Year 3 indicator remained the same. The student- and school-level models remained constant in all analyses. The level 3 and combined models for this secondary analysis at each grade level are noted below:

$$\beta_{00k} = \gamma_{000} + \gamma_{001}\text{Title I School}_k + \gamma_{002}\text{High Academic Tradition}_k + \gamma_{003}\text{Low Academic Tradition}_k + \gamma_{004}\text{School Size}_k + \gamma_{005}\text{Cohort 1 Year 1}_k + \gamma_{006}\text{Cohort 1 Year 2}_k + \gamma_{007}\text{Cohort 1 Year 3}_k + \gamma_{008}\text{Cohort 2 Year 3}_k + u_{00k}.$$

$$Y_{ijk} = \gamma_{000} + \gamma_{001}\text{Title I School}_k + \gamma_{002}\text{High Academic Tradition}_k + \gamma_{003}\text{Low Academic Tradition}_k + \gamma_{004}\text{School Size}_k + \gamma_{005}\text{Cohort 1 Year 1}_k + \gamma_{006}\text{Cohort 1 Year 2}_k + \gamma_{007}\text{Cohort 1 Year 3}_k + \gamma_{008}\text{Cohort 2 Year 3}_k + \gamma_{010}\text{Masters}_{jk} + \gamma_{020}\text{1–2YrsExp}_{jk} + \gamma_{030}\text{3–4YrsExp}_{jk} + \gamma_{040}\text{10+YrsExp}_{jk} + \gamma_{100}\text{AgeTest}_{ijk} + \gamma_{200}\text{Female}_{ijk} + \gamma_{300}\text{LEP}_{ijk} + \gamma_{400}\text{SpecEd}_{ijk} + \gamma_{500}\text{FARM}_{ijk} + \gamma_{600}\text{Minority}_{ijk} + \gamma_{700}\text{Test 2007}_{ijk} + \gamma_{800}\text{Test 2008}_{ijk} + u_{00k} + r_{0jk} + e_{ijk}.$$

Results. Findings from these cohort-by-year analyses of grades 3, 4, and 5 student mathematics achievement data are noted in Table 6. The estimates at the student and classroom levels are remarkably stable, with no substantive changes in any coefficients except for the year-of-test control variables (2007 Test; 2008 Test). In all grades, the test-year variable coefficients are smaller when Cohort 1 is entered into the model by year, as compared to these estimates when the Cohort 1 treatment effect was examined across all 3 years. These differences in the test-year-variable estimates reflect the higher average scores in Cohort 1 in each of these years. When Cohort 1 is entered into the model as a single variable, the 2007 Test and 2008 Test estimates are the average difference between the average SOL scale score in the first year of the study and the average scores in 2007 and 2008 tests for all students. With Cohort 1 by Year in the model, the differences between Cohort 1 and control students' mean scale scores each year is removed from the 2007 Test and 2008 Test estimates and attributed to the year-specific Cohort 1 estimate. Thus the 2007 Test and 2008 Test coefficients are reduced as the Cohort 1 coefficients in those years increase. This pattern is particularly evident in the grade 4 analysis. Unlike the grade 3 and grade 5 analyses, the Cohort 1 Year 2 at grade 4 coefficient is slightly larger than the Cohort 1 Year 3 coefficient; concurrently the Test 2007 estimate is greater than the Test 2008 estimate. A similar adjustment occurs in the Cohort 2 Year 3 estimate when the Cohort 1 Year 3 variable enters the model.

The Cohort 1-by-year variables reveal a consistent pattern of results over time, although, as expected, the increased variance of the estimates reduced the number of significant coefficients. In grade 3 none of the Cohort 1-by-year variables were significant. In the first year of the study, the SOL mathematics scores of the Cohort 1 students were, on average, 6.8 points (9% *SD*, $p = .25$) higher than those of the students in the control schools; in Year 2 the coefficient increases to 10.4 points (14% *SD*, $p = .24$), and in Year 3 it increases to 16.5 points (22% *SD*, $p = .14$). While increasing coefficients are apparent in the second and third years of placement of an elementary mathematics coach, the increasing Cohort 1-by-year coefficients in this

Table 6. REML Parameter Estimates and Standard Errors for Coaching by Year Effects on Student SOL Mathematics Overall Scale Score by Grade

	Grade 3		Grade 4		Grade 5	
Scale Score	Coefficient	*SE*	Coefficient	*SE*	Coefficient	*SE*
Intercept	493.91***	2.31	470.66***	2.50	497.05***	2.44
Student variables:						
Age at test	−8.50***	1.61	−14.37***	1.43	−13.28***	1.56
Female	−1.86	1.26	−7.72***	1.55	−.71	1.64
LEP	−8.00	7.25	−18.72**	6.33	−21.85**	6.71
Special education	−41.82***[a, b]	3.45	−40.17***[a, b]	3.30	−51.62***[a, b]	4.36
Free or reduced meal	−17.11***	2.19	−17.79***[b]	2.41	−18.25***[b]	2.78
Minority	−35.76***	2.16	−33.73***	2.21	−27.64***	2.09
2007 test	2.69	6.91	17.48**	6.56	16.21*	7.13
2008 test	3.41	7.45	11.87	11.22	6.17	8.13
Teacher variables:						
Master's degree	2.15	2.43	.21	2.94	−5.03	3.37
1–2 years experience	−7.67	5.09	−5.00	5.27	−13.15*	6.02
3–4 years experience	−6.56	4.27	−7.41	5.01	−5.98	5.31
10+ years experience	.82	2.95	10.45*	4.10	10.66**	3.73
School variables:						
Title I school	7.23	5.71	4.57	5.66	13.49*	5.42
High academic tradition	35.44***	8.82	39.46***	9.65	51.80***	6.10
Low academic tradition	−13.88	7.89	−17.03*	7.88	−2.77	9.53
School size	−2.81	2.49	−8.45**	2.70	−8.09**	2.36
Cohort 1 Year 1	6.81	5.83	12.27	7.26	6.34	7.67
Cohort 1 Year 2	10.38	8.83	15.35*	7.62	19.61*	7.82
Cohort 1 Year 3	16.48	11.03	13.25	11.98	20.31*	9.24
Cohort 2 Year 3	−1.11	8.32	8.86	10.90	−.63	9.44
	Grade 3		**Grade 4**		**Grade 5**	
	Variance	χ^2	Variance	χ^2	Variance	χ^2
Variance estimates:						
Student-level variance (σ_2)	3,926.95	. . .	3,776.06	. . .	4,319.37	. . .
Class-level variance ($\tau_{00\pi}$)	382.60***	677.02	523.33***	518.83	554.37***	569.50
School-level variance ($\tau_{00\beta}$)	305.20***	287.61	276.90***	238.26	276.01***	203.25

[a] Unmodeled level 2 random effect.
[b] Unmodeled level 3 random effect.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

analysis are not significant, due in large part to the increased standard errors associated with this more conservative analysis.

In grade 4 there is a similar pattern. In the first year, on average the Cohort 1 students scored 12.3 points higher (17% *SD*, $p = .09$) than the control schools on the SOL mathematics assessment, though this coefficient is not significant. In the second year, the coefficient was significant, as on average Cohort 1 students scored 15.4 points higher (21% *SD*, $p = .046$) than grade 4 students in the control schools. In the third year of the study, the coefficient for Cohort 1 students fell somewhat to 13.3 points (18% *SD*, $p = .27$), with a substantially larger standard error.

In the grade 5 analysis, the pattern of growth is more compelling, with larger and significant differences in both the 2007 and 2008 testing years. The Cohort 1 Year 1 coefficient for grade 5 was small and nonsignificant at 6.3 points (8% *SD*, $p = .41$).

However, during the second and third year of the placement of a coach, on average the Cohort 1 students scored 19.6 (25% *SD*, $p = .01$) and 20.3 (25% *SD*, $p = .03$) points higher, respectively, than the students in the control group; both of these estimates are statistically significant.

Across all three grades, the Cohort 2 Year 3 variable had smaller coefficients in the year-by-treatment analysis as compared to the coefficients in the analysis of treatment versus control over 3 years. The reductions in these coefficients were due to the entry of the Cohort 1-by-year variables. Specifically, the Cohort 1 Year 3 estimate captures the average benefit for students in Cohort 1 schools over and above the average difference for control schools in Year 3. With the Cohort 1 Year 3 variable in the model, the Test 2008 variable is reduced to the average difference between the Year 1 and Year 3 SOL scores for control schools. The Cohort 2 Year 3 estimate is then compared to the average score for control schools. Specifically, in grades 3 and 5, these Cohort 2 coefficients change from small negative coefficients to negligible estimates. In grade 4, the Cohort 1 Year 3 estimate is essentially equal to the average Cohort 1 effect in the previous grade 4 model, and thus the Cohort 2 Year 3 estimate remains essentially the same (8.9 points compared to 9.1). In sum, the Cohort 2 Year 3 estimates in the second set of models better represent the true differences between Cohort 2 treatment schools and the control schools. Further, these more accurate estimates are consistent with the pattern of Cohort 1 coefficients, which indicated no statistically significant improvements in student scores in the first year of coaching, with larger increases evident in following years.

The cohort-by-year models explain the same amount of variance as the control-versus-treatment models. The similar amounts of variance explained are not surprising since the by-year models are nested in the control-versus-treatment models. Thus the same variance is explained but attached to the by-year estimates differently. Similarly, across all grades the deviance statistics of the two sets of models had negligible differences, favoring the more parsimonious set of models by any measures of model fit (χ^2, AIC, or BIC; McCoach & Black, 2008).

Discussion

Mathematics coaches are placed in elementary schools to construct leadership roles and provide professional development addressing mathematical content, pedagogy, and curriculum. Theoretically, these leaders support collective collaborative professional development, providing knowledgeable "critical collegiality" (Lord, 1994). But substantive change is neither rapid nor consistent. Coaches are called upon to navigate not only the complexity of teaching and student learning as it emerges in the classrooms of multiple teachers, but to do so while provoking the development of those teachers by advocating for their change, nurturing their performance, advancing their thinking, increasing their mathematical understanding, and saluting their attempts (Campbell, 1996). This is a demanding role, and a role that the profession does not understand and is only beginning to examine.

This study was designed to address the fundamental question that educational policy makers, district administrators, and school leaders ask: Does the placement of an elementary mathematics coach affect student achievement across a school? As such it employed a control-treatment design with triples of like schools randomly assigned to either 3 years of coaching placement, 3 years of control status, or 2 years of control status and 1 year of coaching placement. Each of the matched schools in a triple were within a single district, with no school having benefited from the placement of an elementary mathematics coach in the past, thereby limiting confounding district variation in this group-randomized trial. All of the cooperating districts were in a single state, establishing consistency in terms of intended curriculum standards and assessment objectives in mathematics. As an efficacy study, the schools for this project are all in urban or urban-edge districts that each employed a district-level mathematics supervisor and are located within commuting distance of a university, wherein the coaches could complete their mathematics content, pedagogy and leadership/coaching courses. The combination of the 12 triples in this study provide a sample of matched schools that together represent differing academic traditions in mathematics and enroll students drawn from a variety of urban and urban-edge settings reflecting both poorly resourced and adequately resourced communities. This study then relied upon random school assignment to treatment or control status within the triples to limit confounding school differences.

Ideally, baseline individual student achievement data would have been available for most students in all grades, student mobility would have been minimal and not correlated with other school demographics, school enrollment would have been stable, and the standardized student assessments would have year-to-year consistency, always permitting room for growth. However, this was not the case, reflecting the reality of public schooling. Thus, this analysis relied upon randomization of the matched schools. While prior achievement was not available at the student level, two measures of prior school-level academic tradition in mathematics (High Academic Tradition and Low Academic Tradition) were used to control for large differences in prior school performance. Nevertheless, the lack of an individual prior student-achievement covariate and the fact that standardizing student achievement scores was not an option in the analysis are acknowledged research limitations.

In addition, two countervailing challenges influence the type I and type II error rates in these analyses. The ceiling effects present in the SOL data—effects that were more prevalent in the data from the schools with coaches as compared to the control schools—decreased the likelihood of finding significant treatment effects by an indeterminate amount. At the same time, the violation of the assumption of independence at the classroom and school levels increased the likelihood of finding significant treatment effects by an indeterminate amount. This lack of independence is a serious limitation. While the pattern of growth over time in the treatment effect supports and further explains the positive impact of elementary mathematics coaches on student achievement, the results of the analyses presented in this report should be interpreted in light of these admitted limitations. As such, further research is needed to support or refute the conclusions that follow.

Over a 3-year period, the students in this study who were enrolled in schools with an elementary mathematics coach had significantly higher scores on their state's high-stakes standardized mathematics achievement tests (grades 3–5) than did students in the control schools. While significant at all three grades, this positive impact was stronger in grades 4 and 5. It may be that this reflects the increased rigor and abstraction of the upper-elementary mathematics curriculum, evidencing more challenge for teachers and students and the resulting potential for coaches to have more entrée or influence because of perceived need. Or it may simply reflect the confounding measurement demands that a paper-and-pencil, timed assessment places on third graders.

The subsequent conservative, year-by-year analyses revealed that the effect of mathematics coaches on student achievement was not significant in the first year of coach placement at any grade. While these more conservative analyses did not consistently find significant effects for treatment groups consisting of only 12 schools per year, it did reveal a consistent pattern of results across all three grades. Student achievement in the treatment schools was consistently greater than the achievement of students in the control schools during the second year of coach placement (a significant difference in grades 4 and 5), and this difference in student achievement either increased (grades 3 and 5) or was comparably maintained in the third year of the placement of a coach. This finding aligns with the core features of Desimone's (2009) professional development framework. It may be that coaches do not have a positive impact on student achievement until the school-based professional interaction between coaches and teachers is of sufficient duration to permit emergence of coherent collective efforts marked by active learning and focused on mathematics content and pedagogy, as well as on student understanding. The pragmatic implication of this finding is the caution that a coach's positive effect on student achievement develops over time as a knowledgeable coach and the instructional and administrative staffs in the assigned school learn and work together. There is no evidence that elementary mathematics coaches will yield increased student achievement in their first year of placement.

Relying on quantitative design, this project did not study the following: how coaching practices were implemented or influenced practice in these schools; how these coaches varied in their focus, organization, priorities, coaching knowledge, and skills; how these coaches interacted with their teachers; how teachers' existing instructional practices meshed or conflicted with the instructional ideal of their coaches; or how these local school administrators and teachers perceived the role and value of a coach. These types of field investigations are needed if we are to understand how to maximize the potential of coaching as site-based professional development supporting student learning and teacher enhancement.

Similarly, as implied in Figure 1, this quantitative analysis did not address many of the variables of interest in the coaching, professional development, or teacher education literature. In particular, this analysis did not include controls or measures addressing the status or growth of teachers' mathematical content knowledge, pedagogical content knowledge, knowledge of mathematics for teaching, or beliefs about mathematics teaching and learning. There are no measures controlling for the degree to which teachers accessed other avenues of professional development, such as formal workshops or conferences, graduate courses, collegial networks, or school-based collaborative peer groups. These teacher-level data are potential control measures or dependent measures that should be addressed in future investigations.

While Desimone's (2009) "core conceptual framework" (p. 183) frames an ideal interpretation of how coaching may target the professional development of individual teachers, the distribution of coaches' activity as revealed in this study's PDA data suggests that many of the coaches in this study had limited time to coach teachers, as on average coaches spent over twice as much time addressing assessment, teaching students, managing materials, and attending meetings than they did coaching. This suggests that the potential for coaching's impact on student achievement may be even greater than reported herein. However, that would require teachers and principals to understand the role and responsibilities of the coach as an agent and catalyst who establishes and maintains a safe environment for instructional improvement in mathematics and would require supportive principals and knowledgeable coaches to work together as instructional leaders in their schools.

Finally, the coaches who are the subjects of this study engaged in substantive academic coursework that was designed to foster and support their transition to the position of whole-school elementary mathematics coach. As such, the results herein should not be generalized to other settings where an experienced teacher is simply named as the school-based mathematics coach with little or no prior professional development addressing the responsibilities and expertise presumed of coaches.

Notes

This study was developed with the support of the National Science Foundation, grant no. ESI-0353360. The statements and findings herein reflect the opinions of the authors and not necessarily those of the foundation.

1. These academic tradition measures accessed the 2004–2005 SOL data of grade 5 students who were not subsequently the source of any analyzed student-level data in this report, as well as the 2004–2005 SOL data of grade 3 students. Those 2004–2005 third graders who subsequently moved away from a participating school were then not the source of any analyzed student-level data in this report. But the nonmobile grade 3 students in 2004–2005 were the source of analyzed student-level grade 4 SOL for 2006 (first year of the study) and analyzed student-level grade 5 SOL data for 2007 (second year of the study). However, the source of the grade 4 SOL data and grade 5 SOL data analyzed in this study was not limited to these students. The analyzed fourth-grade data also included the SOL data from all mobile grade 4 students tested in 2006 as well as from all grade 4 students tested in 2007 and 2008. Similarly, the analyzed fifth-grade data also included the SOL data from all grade 5 students tested in 2006 and in 2008 and from all mobile grade 5 students tested in 2007.

2. All HLM analyses used restricted maximum likelihood estimation using the EM Algorithm. Models were run using HLM 6.0 (Raudenbush, Bryk, & Congdon, 2004). No problems were encountered during model estimation.

3. Effect sizes were not computed because standardizing the data was not a viable option. To provide perspective on the magnitude of coefficients, the percentage of the pooled standard deviation is presented across all 3 years within each grade, indicated by *SD* throughout.

References

Ai, X., & Rivera, N. (2004). Linking ideas to practice: Effectiveness of coaching upon teacher practice. In J. Gangel & E. Guyton (Eds.), *Teacher education yearbook XIII: Research on alternative and non-traditional teacher education* (pp. 191–204). Lanham, MD: Rowman and Littlefield.

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

Bransford, J. D., Brown, A., & Cocking, R. E. (Eds.). (2000). *How people learn: Brain, mind, experience and school.* Washington, DC: National Academy Press.

Campbell, P. F. (1996). Empowering children and teachers in the elementary mathematics classrooms of urban schools. *Urban Education,* **30**, 449–475.

Campbell, P. F., & White, D. Y. (1997). Project IMPACT: Influencing and supporting teacher change in predominantly minority schools. In E. Fennema & B. S. Nelson (Eds.), *Mathematics teachers in transition* (pp. 309–355). Mahwah, NJ: Erlbaum.

Dempsey, N. (2007). Five elements combine in a formula for coaching: South Carolina initiative carves out time for science and math coaches in schools. *Journal of Staff Development,* **28**(2), 10–13.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher,* **38**(3), 181–199.

Fosnot, C. T., & Dolk, M. (2002). *Young mathematicians at work: Constructing fractions, decimals and percents.* Portsmouth, NH: Heinemann.

Foster, D., & Noyce, P. (2004). The Mathematics Assessment Collaborative: Performance testing to improve instruction. *Phi Delta Kappan,* **85**(5), 367–374.

Hawley, W. D., & Valli, L. (1999). The essentials of effective professional development: A new consensus. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 127–150). San Francisco: Jossey-Bass.

Joyce, B., & Showers, B. (1980). Improving inservice training: The messages from research. *Educational Leadership,* **51**(2), 57–61.

Keller, B. (2007, December 12). Coaching teachers to help students learn. *Education Week,* **27**(15), 22–24.

Lamon, S. J. (1999). *Teaching fractions and ratios for understanding: Essential content knowledge and instructional strategies for teachers.* Mahwah, NJ: Erlbaum.

Lord, B. (1994). Teachers' professional development: Critical colleagueship and the role of professional communities. In N. Cobb (Ed.), *The future of education: Perspectives on national standards in education* (pp. 175–204). New York: College Entrance Examination Board.

Loucks-Horsley, S., Harding, C. K., Arbuckle, M. A., Murray, L. B., Dubea, C., & Williams, M. K. (1987). *Continuing to learn: A guidebook for teacher development.* Andover, MA: Regional Laboratory for Educational Improvement of the Northeast and Islands and the National Staff Development Council.

Marzano, R., Walters, T., & McNulty, B. (2005). *School leadership that works: From research to results.* Alexandria, VA: Association for Supervision and Curriculum Development.

McCoach, D. B., & Black, A. C. (2008). Evaluation of model fit and adequacy. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 245–272). Charlotte, NC: Information Age.

Miller, B., Moon, J., & Elko, S. (2000). *Teacher leadership in mathematics and science: Casebook and facilitator's guide.* Portsmouth, NH: Heinemann.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

National Research Council. (2001). *Adding it up: Helping children learn mathematics.* Washington, DC: National Academy Press.

Neufeld, B., & Roper, D. (2003). *Coaching: A strategy for developing instructional capacity.* Cambridge, MA: Education Matters. Retrieved from http://www.annenberginstitute.org/Products/Coaching.php

Poglinco, S. M., & Bach, A. J. (2004). The heart of the matter: Coaching as a vehicle for professional development. *Phi Delta Kappan,* **85**, 398–400.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Raudenbush, S., Bryk, A., & Congdon, R. (2004). HLM: Hierarchical and nonlinear modeling (Version 6.0) [Computer software]. Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W., Liu, X.-F., Spybrook, J., Martinez, A., & Congdon, R. (2006). Optimal design software for multi-level and longitudinal research (Version 1.77) [Computer software]. Available at http://sitemaker.umich.edu/group-based

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis,* **29**, 5–29.

Rodgers, A., & Rodgers, E. M. (2007). *The effective literacy coach: Using inquiry to support teaching and learning.* New York: Teachers College Press.

Schifter, D., Bastable, V., & Russell, S. J. (1999). *Number and operations, Part 1: Building a system of tens.* Parsippany, NJ: Dale Seymour.

Virginia Mathematics and Science Coalition. (2008). *DR K–12 support materials.* Retrieved from http://www.vamsc.org/

West, L., & Staub, F. (2003). *Content focused coaching: Transforming mathematics lessons.* Portsmouth, NH: Heinemann.

Wood, T., Nelson, B. S., & Warfield, J. (2001). *Beyond classical pedagogy: Teaching elementary school mathematics.* Mahwah, NJ: Erlbaum.

York-Barr, J., & Duke, K. (2004). What do we know about teacher leadership? Findings from two decades of scholarship. *Review of Educational Research,* **74**, 255–317.