Dividing by Zero:
Exploring Null Results in a Mathematics Professional Development Program

Heather C. Hill
Harvard Graduate School of Education
Douglas Corey
Brigham Young University
Robin T. Jacob
University of Michigan

The research reported here was supported by the National Science Foundation through Grant # 108051-5021742 and DRL 0918383 to Harvard University. The opinions expressed are those of the authors and do not represent views of the National Science Foundation.
Abstract

Since 2002, U.S. federal funding for educational research has favored the development and rigorous testing of interventions designed to improve student outcomes. However, a large proportion of the programs developed and rigorously tested in the past decade have shown null results on both student outcomes and, often, intermediate variables. Scholars reporting on null results often explain such results by reporting on factors identified informally as they either delivered or observed the program. In this paper, we argue for a more systematic approach to examining null results and illustrate this approach via an examination of one program’s failure to impact teaching and learning. Although identifying the causes of a null result is a little like dividing zero by zero—results are always indeterminate owing to the large number of possible contributing factors—gaining purchase on potential explanations would allow programs to adopt more effective designs.
Dividing by Zero: Exploring Null Results in a Mathematics Professional Development Program

Since 2002, U.S. federal funding for educational research has favored both the development of interventions designed to improve student outcomes and the use of randomized trials to evaluate these interventions. To the surprise of many, a large proportion of the programs developed and tested in the past decade have shown null results on both student outcomes and, often, intermediate variables. A recent report (Coalition for Evidence-Based Policy, 2013) found that of roughly 90 Institute for Education Sciences studies funded since 2002, 88% produced weak or null results; this included many highly regarded and widely implemented programs evaluated in well-designed studies (e.g., Bos et al., 2012; Garet et al., 2008; Garet et al., 2011).

These findings have met with a fair amount of dismay among scholars and educators but, to date, little systematic inquiry regarding why such results might occur. Some have speculated that this pattern of results is consistent with those from other fields, where no-impact findings are common (Coalition for Evidence-Based Policy, 2013). It is also possible that methodological issues, such as weak statistical power for detecting effects (Schochet, 2008; Spybrook & Raudenbush, 2009), may have caused at least some of these null results. In this paper, however, we argue for a more substantive analysis of the reasons programs fail. Although identifying the causes of a null result is a little like dividing zero by zero—results are always indeterminate owing to the large number of possible contributing factors—gaining purchase on potential explanations would allow programs to adapt to features of schools and classrooms that may hamper their effectiveness. Similarly, identifying weaknesses in interventions’ theory of action and implementation might lead to better program design.

In this vein, this paper uses a theory-driven analysis of data to understand why a mathematics professional development program failed to show an effect on teaching quality and student learning. To do so, we drew upon prior research in the field of implementation—from
both inside and outside education—to structure hypotheses for testing. We then tested these hypotheses using a rich data corpus, including teacher interviews, lesson observations, post-lesson teacher logs, records from professional development sessions, and interviews with professional development providers. These data also suggested additional hypotheses for exploration and testing, and we describe these as well. This is an unusual method but, we think, one worth investing in given the large number of null-result studies.

In what follows, we review the policy and program implementation literature from which we generate our hypotheses, and we describe our data and analytic plan. We then evaluate our formal hypotheses and describe two more generated from working with the data. We conclude with thoughts about the replicability of this method and about implications for policy and intervention design.

**Background**

Professional development is simultaneously the major method for instructional improvement in the United States and also a major reason instructional improvement has been so difficult to achieve. Although almost every teacher receives some professional development each year (Choy, Chen, & Bugarin, 2006), the majority of teachers attending both content- and teaching methods-focused professional development report spending one day or less in such an activity (Scotchmer, McGrath, & Coder, 2005). This typically available professional development may vary dramatically in quality (Author, 2004). Thus it is no surprise that in administrative datasets, teachers’ attendance at professional development is, on average, often unrelated to their effectiveness, at least as measured by student test score outcomes (Jacob & Lefgren, 2004; Harris & Sass, 2011). Results of single-program evaluations have been quite mixed, with early efforts showing positive and often lasting results (e.g., in mathematics,
Carpenter, Fennema, Peterson, Chiang, & Loef, 1989; Franke, Carpenter, Levi, & Fennema, 2001; McGill-Franzen, Allington, Yokoi, & Brooks, 1999) but more recently completed evaluations showing mainly null results (Buysse, Castro, & Peisner-Feinberg, 2010; Cabalo, Ma, & Jaciw, 2007; Garet et al., 2008; Garet et al., 2011; Gersten, Dimino, Jayanthi, Kim, & Santoro, 2010; Santagata, Kersting, Givvin, & Stigler, 2011; for exceptions, see Landesman Ramey et al., 2011; Penuel, Gallagher, & Moorthy, 2011).

System-level analyses quite sensibly point to the superficial, diffuse, and incoherent offerings as a reason that typical professional development does not result in improved instruction (Wilson & Berne, 1999; Wilson, Rozelle, & Mikeska, 2011). However, evaluators have yet to apply disciplined inquiry to examine the reasons for a particular program’s lack of success. Instead, evaluators often follow a disclosure of null results with a list of several potential contributing factors, often gleaned from their experience providing or observing the professional development rather than from a rigorous examination of evidence (see, e.g., Santagata et al., 2011; Gersten et al., 2010). Yet as more evaluations gather data on factors mediating the relationship between program inputs and student learning, and as evaluators collect more information on contextual factors surrounding programs more generally, disciplined inquiry should replace ad hoc explanations for failure.

To structure such inquiry, we argue the field should turn toward the policy implementation literature, both within and outside education, to identify and test common reasons for program failure. Such an approach would take advantage of forty years of research on why policies and programs do and do not work, yielding strong hypotheses regarding possible reasons for failure. Although alternative literatures are available in which to ground such work (e.g., principal-agent theory, program fidelity), the implementation literature is preferable...
because it is wide-ranging and well-established, yielding findings that generalize past the particulars of specific programs and policies. We review the implementation literature that structures our analyses next.

**Lessons from Policy Implementation**

In the early 1970s, the failure of many Johnson-era Great Society programs to deliver relief to impoverished communities spurred academics to inquire about the reasons for such disappointments; the result was a thriving literature on the perils of policy implementation (e.g., Sabatier & Mazmanian, 1979; Pressman & Wildavsky, 1973). By the mid-1980s, implementation scholars had identified over 300 variables (O’Toole, 1986) responsible for the adaptation, transformation, or collapse of policy at the local level. Many of these variables related to a newly recognized class of agents—termed street-level bureaucrats (Lipsky, 1980)—responsible for implementing social policy in schools, police stations, public defenders’ offices, welfare-to-work programs, and elsewhere. These agents interact directly with policy “recipients,” tend to have wide discretion in the construction of policy and practice, and often work in situations with few resources and high demand for services. Early research showed that street-level bureaucrats’ attitudes, capacity, and practice profoundly shape the implementation of policy (Berman & McLaughlin, 1978). Over the next three decades, scholars elaborated on these findings, arriving at several broad reasons policies and programs fail.

Street-level bureaucrats’ *lack of willingness* to implement policies and reforms constitutes one such reason. Such unwillingness can take several forms: selectively attending to policy by screening out unwanted messages; symbolic acceptance of policy and reform by superimposing superficial changes over typical practice; and outright rejection of policy altogether (Berman, 1978; Meyer & Rowan, 1977). Education scholars examining teachers’ reaction to reform
initiatives have documented all such forms, for instance selective attention to data resulting from accountability policies (Terhart, 2013), superficial change in response to standards-based reforms (Spillane & Zeuli, 1999), and outright rejection of school reform, particularly among those veteran teachers who have experienced multiple waves of such reform (Cuban, 1993; Olsen & Sexton, 2009). Such responses could in fact be rational, as policies may not address important teacher and student needs, may ask teachers to perform new activities with inadequate resources, and/or may require actions that teachers view as in conflict with students’ best interests.

Implementers’ sense-making around policy constitutes another reason for policy failure. Logically, street-level bureaucrats must understand what policy implies for practice before enacting that practice. Scholars have observed that street-level bureaucrats’ understanding of policy can vary and that the “meaning” of policy is created not only from the actual words of legislation but also from the knowledge and values implementers bring to their jobs and from the milieu in which implementation occurs (Yanow, 1996). In education, studies of the 1990s standards-based reforms suggested that teachers’ interpretation of policy differed in significant ways from policy intent (Author, 2001; Spillane & Zeuli, 1999). Moreover, they differed in predictable ways; in a review of the role cognition plays in implementation, Spillane, Reiser, and Reimer (2002) noted that teachers were likely to implement reforms in light of existing frames of reference, leading many to interpret policies in conventional ways; they also noted that reforms requiring the fundamental reorganization of teacher knowledge and beliefs were less likely to succeed. Finally, recent work in this field has focused on the role colleagues play in influencing implementers’ understanding of policy, particularly when teachers have collaborative opportunities in which they interpret and respond to policy (e.g., Coburn, 2001).
Another finding from this literature focuses on street-level bureaucrats’ insufficient resources to implement policy. In the wider implementation literature, resources are often construed as financial inputs and support from key stakeholders (Derthick, 1972; Sabatier & Mazmanian, 1979). In education, some scholars have argued that the most relevant resources are those closest to classroom practice—the supports teachers draw upon when enacting instruction. One such resource is thought to be teachers’ knowledge (Ball, 1990; Cohen, 1990; Heaton, 1992; Putnam, Heaton, Prawat, & Remillard, 1992). Teachers must have, obviously, basic knowledge of the policy or program they are to enact. In addition, they are likely to need deep knowledge of the content they are to teach, as well as knowledge of how students are likely to learn that content. A second resource for classroom practice is thought to be curriculum materials, particularly in the case where policies and programs call for very different forms of instruction from that typical in U.S. schools (Remillard, 2005; Stein, Remillard, & Smith, 2007). Such materials can support changes in day-to-day classroom activities, providing novel tasks and problems for students to work on, allowing for more conceptually based presentations of content, and supporting new forms of assessment and student learning.

A fourth finding from this literature focuses on organizations and the broader environmental milieu in which street-level bureaucrats work. Sandfort and Moulton (2015), for instance, describe organizations as the “critical middle position” between policy and street-level bureaucrats; organizations interpret, translate, and operationalize the signals received from policies and programs and direct resources toward implementation. These organizations often exist in multi-level, multi-layered networks, creating problems of coordination between agencies tasked with solving problems (Sandfort & Moulton, 2015). In education, Berman (1978) noted that policy implementation required changes in local organizational routines, rules, and processes.
and that congruence between organizational goals and policy aims affected the probability of full implementation. Cuban (1993) argued that norms, such as long-held traditions about the role of the teacher and the organization of classrooms and schools, affect the fate of curricular and pedagogical reforms. Factors outside of implementers’ organizations—rival policies, political upheavals, and unexpected events—can distract organizations and individuals from intended reforms (McLaughlin, 1987, 1990). Finally, simple things matter: For instance, McLaughlin (1990) noted that supportive district administrators increase the likelihood of instructional improvements.

A fifth and more recently emergent explanation for implementation failure in education focuses on the relative difficulty of enacting ambitious instructional practice. Kennedy (2005) notes that what reformers typically want—instruction that yields more student intellectual engagement—may be unattainable because such practice is both highly demanding and exhausting for teachers. Cohen (2011) elaborates this further, analyzing the demands that ambitious, rather than conventional, teaching places on teachers and their students. The former calls for students to engage intellectually, often around difficult problems and tasks; this puts the onus on teachers to motivate learners to take on these more complex tasks and assignments. Once students are willing to engage, ambitious instruction calls for more real-time instructional responsiveness on the part of the teacher. Such work requires deep and flexible knowledge of content, a willingness and capacity to heed learners’ ideas and to make use of them in instruction, skill in leveraging classroom discussions and novel tasks to attain learning goals, and decision-making expertise responsive to both the content and students’ needs (Cohen, 2011; Cohen & Barnes, 1993; Heaton, 2000; Lampert, 2003).
A final lesson from this research literature is that policy itself may be poorly designed, lack an adequate theory of action, and/or have failed to attract “fixers” or other interested sovereigns to oversee its implementation (Bardach, 1977, Pressman & Wildavsky, 1973; Sabatier & Mazmanian, 1980). In education, this corresponds to some observers’ notions that policies and educational interventions are often weak treatments, incapable of securing long-term change in organizational and individual behavior (McLaughlin, 1987).

These observations from the policy implementation literature form the basis for most of the analysis we undertake below. Although we cannot evaluate all potential explanations for the null findings, by using the strong structure provided by the implementation literature and checking these hypotheses against our data, we argue that we can contribute to both theory and practice regarding the design of professional development and educational interventions more broadly.

**Methods**

**The Math Solutions Professional Development Program**

Math Solutions (www.mathsolutions.com) supplied the professional development for teachers in this study. Math Solutions is one of a handful of mathematics professional development programs with a national reach, using hand-selected, rigorously trained professional developers based in locations across the country. Founded by Marilyn Burns in 1984, at the beginning of the study (2010), Math Solutions had provided professional development to over 600 districts in 48 states, reaching over 100,000 teachers.

According to Math Solutions materials, four goals are central to the professional development—helping teachers (a) to learn more mathematics, (b) to understand how children learn math, (c) to use formative assessment to develop insight into what specific students know
and do not know, and (d) to develop effective classroom instructional strategies. Specifically, our observations of the program suggest that it emphasizes instructional strategies that involve engaging, high-cognitive-demand tasks that allow students to develop their own solutions to mathematical problems (often in collaboration with other students), to participate in mathematical discussions, and to deepen their understanding of mathematical concepts. Math Solutions sessions included teachers collaboratively exploring and solving the mathematics tasks they could choose to use in classrooms, discussions about best practices in instruction, videotapes of students solving challenging mathematics problems via novel strategies, examples of and teacher practice in interview-based assessment techniques, and planning sessions aimed at allowing teachers to integrate Math Solutions ideas with district instructional guidance. The program also provided teachers with supplemental curriculum materials, most often in the form of books with detailed lesson plans and written explanations of both the mathematics and how students might learn the mathematics of the lesson.

Setting and Participants

The professional development and research study were conducted in one mid-size school district, Eastern, serving a racially and socio-economically diverse population of over 30,000 students across 46 school locations. Approximately 60% of the district’s elementary schools received Title 1 funds and all experienced relatively high rates of student and teacher turnover due to a nearby military presence; Eastern’s schools also have performed consistently below state averages in elementary mathematics. Eastern was nominated to participate in the study by Math Solutions, which had worked with teachers as well as many of the individuals in the district’s mathematics office roughly five years prior to the current project. Recruitment efforts focused on schools and teachers not already trained by Math Solutions.
The study enrolled 105 fourth- and fifth-grade teachers over two separate cohorts, with 88 teachers beginning the study in fall 2010 and an additional 17 teachers added in fall 2011 to replace 29 departing teachers. All teachers who left the study did so for reasons unrelated to the study; they left study schools, left teaching, were no longer teaching fourth or fifth grade, or stopped teaching mathematics. Teachers in our sample were 72% white and 9% male, and 55% held a graduate degree. They had on average almost nine years of teaching experience. The baseline mathematical knowledge for teaching (MKT) scores of the teachers in the study were slightly below the national average for elementary teachers of all grades, and moderately (0.20-0.30 SD) below teachers of the same grades in a nationally representative sample.

Teachers were randomly assigned within schools to either the treatment or control group. Teachers in the treatment group began their Math Solutions experience with a four-day summer institute in August 2010, with similar summer institutes held in two subsequent summers through August 2012. Attendance was relatively high during the first year, when 35 out of 42 treatment teachers attended the summer session (83%), but dropped in Years 2 and 3 for reasons described below. In addition to the summer institutes, Math Solutions staff taught either four (Year 3) or six (Years 1 and 2) one-day in-person sessions during the school year. Almost all of the teachers in the treatment group participated in these sessions during the first year of the study, but participation dropped during Years 2 and 3, again for reasons described below.

Data Collection and Measures

A complete description of data collected for this study exists in (blinded for review). Here we describe only the data we draw upon for our analyses below.

Observations of professional development activities. Two authors of this study attended 19 days of professional development led by Math Solutions staff. These days were split
roughly equally between the summer and the school year. Observers took notes on the tasks completed by teachers, the discussions teachers had with one another, and the whole-group discussions between the providers and teachers. The observers also took photographs of displays created for or by teachers within the professional development (e.g., posters with a mathematics problem and teachers’ solution methods). We used these notes to both characterize the program and to evaluate the strength of the Math Solutions program relative to the task of helping teachers change their practice.

**Teacher surveys and logs.** Teachers completed surveys at the beginning of the study (the pretest) and each spring for the following three academic years (Year 1, Year 2, Year 3). These surveys provided two sources of information for this analysis. During the spring, teachers in both the treatment and control groups reported on the quality of their professional development during the prior year, including whether they learned content and pedagogical strategies that were relevant to their practice and whether they changed their teaching as a result. As this set of items was given to both treatment and control teachers, the questions did not name Math Solutions specifically, and thus teachers reported on all professional development experiences in the year prior to the survey. Reliabilities on the professional development quality metric were 0.92 for Year 1, 0.93 for Year 2; and 0.98 for Year 3. We used these reports to assess the hypothesis regarding lack of will.

Teacher surveys also included MKT items covering late-elementary number/operations and geometry (Hill, Schilling & Ball, 2004). These multiple-choice items both measure teachers’ “common” content knowledge, or the mathematical knowledge common to educated adults, and “specialized” content knowledge, or the mathematical knowledge and skills uniquely needed by teachers. For instance, a common content knowledge item may ask a teacher to select
the correct answer to the problem $35 \times 25$; a specialized version of this item asks teachers to assess the generalizability of several nonstandard methods for solving this problem. Specialized content knowledge items also asked teachers to identify mathematical explanations for common rules and procedures, and to assess the usability of content representations. Off-the-shelf MKT forms supplied by the Learning Mathematics for Teaching project were used, with reliabilities ranging between 0.62–0.78 across different years and test forms. We used teacher MKT scores to evaluate the hypothesis regarding whether and how teacher content knowledge affects changes in instruction.

Teachers also completed a short log following each of their video-recorded lessons, described below. This log asked teachers to reflect on lesson goals, what went well, and what did not go well. We used this data to assess the hypothesis associated with sense-making by comparing teachers’ reports to observers’ accounts of lesson activities. The log also asked teachers to report on the curriculum materials used in the lesson—whether it was the district-adopted text (*Math Expressions*), a supplemental set of materials used in the district (Investigations), materials from Math Solutions, or district-created materials. Teachers not using one of these resources could write in the name and source of the lesson materials. We used this data to assess whether use of Math Solutions materials helps foster the implementation of Math Solutions instructional techniques, as suggested by resource theory.

**Video-recorded lessons.** During the first two years of the study, lessons taught by teachers in the treatment group were video recorded; in the third and final year of the study, lessons taught by both treatment and control group teachers were captured. At each time point, participating teachers recorded six lessons averaging roughly one hour each in length. Lessons

---

1 At the time the study was proposed to the National Science Foundation, cost associated with video recording prohibited capturing lessons from all teachers in all years.
were sampled in three blocks of two back-to-back taping days. To avoid repetition of content, blocks were scheduled at least two weeks apart. Teachers chose the day and time to be video recorded, although we did request that teachers not record on days on which they would be giving a test.

We scored these videos using the Mathematical Quality of Instruction (MQI) observation instrument, which captures lesson quality along four dimensions. MQI’s richness dimension captures the depth of the mathematics provided to students, including the extent to which the lesson features mathematical practices and emphasizes the meaning behind facts and procedures. Student participation in meaning-making and reasoning (SPMMR) captures evidence of students’ involvement in cognitively demanding mathematical activities, such as providing explanations, engaging in reasoning, posing mathematically motivated claims or questions, or working on novel/complex tasks. Teacher errors reports on the presence of mathematical errors introduced by the teacher during instruction, and working with students captures teacher use of students’ mathematical ideas during the course of instruction and teachers’ remediation of incorrect student thinking. Each dimension consists of between two and five items that identify specific behaviors, and each item was scored on a scale of 1 (not present) to 3 (present and high quality).

Following scoring at the item level, raters were also asked to complete an overall assessment of the lesson (Overall MQI) on a 1–5 scale, with scores of 1–2 reserved for lessons that are mathematically problematic and scores of 4–5 reserved for lessons with both strong richness and student participation. Two trained raters were randomly assigned to each lesson and were blind not only to treatment/control condition, but also to the fact that this data stemmed from a study of professional development, as lessons from this study was scored alongside those
from a much larger study of instruction. The adjusted intraclass correlations of the within-year dimension-level scores ranged from 0.39 (errors) to 0.71 (SPMMR), low by conventional standards for reliability but typical for observational metrics (Bell et al., 2012; Kane & Staiger, 2012). MQI scores were used to check hypotheses regarding the role of teachers’ knowledge and curriculum materials in the implementation of Math Solutions instructional techniques.

For the analyses reported here, we also used lesson videos to construct case studies of treatment teachers, noting instructional changes between the first and third year of videotaping and examining those changes (or lack thereof) in light of our hypotheses about insufficient resources and the difficulty of enacting ambitious instruction.

**Interviews.** At the conclusion of the project, school researchers conducted interviews with 31 of the 33 Eastern teachers who remained in the treatment group at the end of the project. Interviews averaged about 30 minutes in length and covered topics such as teachers’ beliefs about mathematics teaching, experiences in the Math Solutions professional development, and reported changes in practice. Although self-reports are not always reliable indicators of actual practice, they can provide insight into the ways in which teachers both perceived changes in their practice and the role of Math Solutions in fostering those changes. We used these interviews to assess hypotheses associated with lack of will and organizational barriers to implementation of the Math Solutions program.

We also interviewed the Math Solutions providers who worked with teachers in Eastern. The interview was loosely structured around trying to understand the evaluation results presented below. Providers discussed their own explanations for the null results and commented on some of the hypotheses generated from our own observations of the program.

**Summary of Evaluation Results**
Because we have reported in depth on our formal evaluation of the program elsewhere (see Authors, under review), we only briefly summarize those results here. We expected that teachers’ participation in the Math Solutions professional development program would lead to an increase in teachers’ mathematical knowledge for teaching, classroom instruction that features more student participation in mathematical thinking and reasoning, and stronger student test score outcomes as compared to a control group. Results did not bear these hypotheses out. Although there were some effects on teachers’ MKT favoring the treatment group (a significant +0.40 SDs on the numbers and operations assessment at the end of Y2; +0.32 SD and insignificant at the end of Y3), there were no statistically significant impacts on teachers’ instructional practice as measured by teachers’ MQI scores. Examining treatment teachers’ MQI scores between Y1 and Y3 of the study actually suggested they remained relatively static (Y1 = 2.76, SD = 0.46; Y3 = 2.87, SD = 0.25). However, this masks a slight improvement in the rate of teacher errors over this time period and a 1 SD drop in scores on the MQI dimension student participation in meaning-making and reasoning (Y1 = 1.31, SD = 0.20; Y3 = 1.12, SD = 0.08). Finally, there was no impact of the professional development on either a state standardized assessment or a project-administered assessment. This last result is unsurprising given the lack of effects on instruction, although it is also possible that student achievement could have improved as a result of factors other than those detected by the MQI.

**Current Analyses**

The primary purpose of this paper is to assess potential explanations for these mostly null results. To do so, we specified hypotheses relating to the factors identified by the implementation literature, then drew on our quantitative, observational, and interview data to evaluate each. In reading through our data corpus, we also remained alert for other, emergent
explanations. Once such an explanation surfaced, we recorded it and subjected it to more formal analysis. We organize our results by first discussing a priori hypotheses and then reviewing emergent explanations. Although our analyses cannot definitively identify the reason(s) for program failure, we argue that this highly structured search is an improvement over the ad hoc commentaries typically provided in the evaluation literature.

We were aided in this work by case memos written about each of the 24 teachers who completed both Year 1 and Year 3 of the study (nine additional teachers completed Year 3 but not Year 1). Observers watched up to six videos per teacher at each time point and also examined these teachers’ postobservation logs \( n = 133 \). Observers recorded notes on each lesson’s mathematical strengths and weaknesses, including the degree of student participation in mathematical meaning-making and reasoning, teacher use of student ideas, and teacher problems with mathematical content. This enabled us to understand the degree to which instruction overall matched Math Solutions ideals. Then, comparing across years, observers answered specific questions regarding change in teachers’ practice. Although teachers had already experienced roughly 40 hours of professional development at the time the Year 1 videos were collected, we expected by Year 3 to see growth resulting from the additional two years of summer institutes and school-year activities. Together, these analyses initially served as checks on whether the quantitative analysis of instructional change using teachers’ MQI scores had obscured important ways in which teachers had implemented Math Solutions ideals. We subsequently used this data to evaluate hypotheses about teacher resources and the difficulty of enacting ambitious instructional practice.

In testing hypotheses, we omitted those associated with research design and instrumentation for several reasons. First, we had sufficient power to detect substantively
meaningful effects. With our final Year 3 sample \((n = 57)\), we had statistical power to detect effect sizes of 0.35 SD for the MKT measures, of 0.20 SD for the MQI measures, and of roughly 0.10 SD for student achievement outcomes. Back-of-the-envelope estimates suggest that for MKT, 0.35 SD correspond to answering about three more items correct out of 28 total, probably the smallest MKT gain likely to lead to gains in instruction and student outcomes. For MQI, 0.20 SD correspond to one or two more instances of student reasoning or rich mathematics per lesson. Impacts smaller than this are not likely to be educationally significant. Second, the study was well-instrumented, including measures that captured many possible outcomes from the professional development. Finally, there was a sufficient contrast between the amount and type of professional development that the treatment and control groups received as demonstrated by teacher responses to questions asking about the hours of professional development attended each year (see Authors, 2015). Therefore, we do not believe factors related to the design of the study played a significant role in the largely null results that were obtained.

Finally, as we worked to test the hypotheses drawn from the implementation literature, we noted a logical problem: All failure can be attributed to both program recipients and their environments (e.g., contextual conditions limited full-scale enactment of the program’s ideas) and the program itself (e.g., the intervention was not strong enough to overcome contextual conditions). This two-sides-of-the-same-coin issue led us to omit a direct evaluation of the weak treatment hypothesis, instead weaving observations regarding this topic into the discussions of other hypotheses below.

**Results**

As summarized above, the professional development had a modest effect on teachers’ mathematical knowledge but failed to realize an effect on instructional practice and student
outcomes. Our qualitative analysis of the 24 teachers with Year 1 and Year 3 video data further illuminated these findings. Across the sample and years, observers found instruction to be variable both across and within teachers but on average contained few extended instances of student participation in meaning-making and reasoning, and few episodes in which teachers substantially used students’ thinking. Observers also noted that about one third of the teachers had moderate to serious problems with the clarity and accuracy of the mathematical content provided to students.

Observers’ memos also suggested that there was a lack of substantial change in teachers’ practice between the first and third years of video data collection. Although seven teachers appeared to change their instruction toward the Math Solutions ideal, observers noted that in five of these cases, changes were either minor and/or superficial and did not result in an overall improvement in the quality of the lesson. For another 15 teachers, observers reported no differences between Year 1 and Year 3 videos. Finally, we noted two teachers for whom observers described Year 3 lessons as, in the aggregate, lower quality than Year 1 lessons.

This evidence, combined with that from our quantitative hypothesis testing above, led us to focus on the relationship between the intervention and instruction as the main cause of program failure. We consider both a priori and emergent arguments that might explain the lack of change in instruction.

A Priori Hypotheses

Lack of will. It is possible that teachers may have participated in the professional development for reasons other than wishing to change their practice. They may have felt pressure from the district math coordinator or their school principal—both of whom often attended recruiting sessions—to enroll, or they may have agreed to join the study because they
were paid up to $550 per year for their participation. Likewise, the professional development itself may not have convinced teachers that improvements in their practice were necessary, or teachers may have initially wished to improve their practice then changed their mind.

We tested for these possibilities in several ways. First, we examined teachers’ self-reports of learning and change in response to their professional development on spring surveys. Second, we analyzed the teachers’ interviews to understand their views of the program’s influence on their instruction. Finally, we used survey data to rule out differences in attitudes between treatment teachers who finished Year 3 and those who dropped out.

The results of these analyses indicate that teachers were overwhelmingly positive about the program, making lack of will an unlikely explanation for the failure of the program to improve instruction and student outcomes. Survey results show that teachers in the treatment group were significantly more likely than control-group teachers in all three years to report that they had learned a great deal from their professional development experiences. For example, at the end of Year 3, 75% of the treatment group but only 20% of the control group indicated strong agreement with the statement “I learned things that were helpful to my everyday practice” in their math professional development—despite the fact that most control-group teachers also participated in math professional development. Similarly, 63% of teachers in the treatment group strongly agreed that “the professional development was useful to me” as compared to only 15% of the control group.

In interviews, every treatment teacher indicated making at least one change to their practice in response to their Math Solutions experience, and approximately 60% indicated making three or more such changes. Table 1 lists the changes teachers report; many of these were themes we heard Math Solutions staff emphasize during professional development sessions.
In discussing the program, teachers also volunteered statements such as “great program” and “I love it. The kids love it.” There were no teachers who made negative comments about the program. This analysis of interview data was corroborated by survey findings, in which 83% of the treatment group teachers agreed or strongly agreed that they had made changes to the way they taught because of what they learned in program.

This enthusiasm did not appear to result from differential attrition. On the Year 3 spring survey, there was no statistically significant difference on the above items between teachers who left the study at the end of Year 1 or 2 and those who stayed for all three years.

This evidence suggests that teachers’ lack of will to change was not an issue in this study. However, teachers’ reports of having made substantial changes in their practice does raise a new set of questions about how such perceptions could coexist with observational evidence of more conventional and static practice. To explore this further, we turn to sense-making.

**Sense-making.** As Spillane and colleagues (2002) show, individuals typically interpret new ideas through existing cognitive frameworks, creating a gap between what programs or policies desire and what implementers understand must be done. Comments made by Math Solutions’ on-site professional developers raised this as a possibility:

> If we just went off of [teachers’ comments] in the sessions, we would be thrilled with what they were saying, but then every time we watched somebody teach or saw stuff in the classroom, it was a reality check about what was really going on in classrooms. . . . The wait time was very minimal, and for a lot of teachers when they did ask their questions, it was the traditional “I’m waiting to hear the right answer” strategy, and when they didn’t hear the right answer, they moved on to the next question, instead of really assessing what percentage of the class really understands it and how deep they understand it. It was really a race against time of “move, move, move” and not really being comfortable with students’ wrong answers.

Our own notes corroborate this trend: during professional development sessions, teachers’ comments suggested they had significantly reformed their practice, yet recorded lessons
uncovered few that contained evidence of significant student mathematical thinking. On this evidence, we might have concluded that teachers’ sense-making was a factor in instruction remaining static.

However, further analysis suggested otherwise. We reasoned that if there was such a gap between teachers’ and professional developers’ understandings of reform, we should see a mismatch between how teachers and observers described specific lesson. Teachers may remark that the lesson featured student reasoning and problem solving whereas observers would characterize the lesson as procedural and recall-oriented. To this end, we examined Year 3 responses to the open-ended postlesson log questions that invited teachers to comment on what went well and did not go well during their lessons. For each, we coded responses for mentions of student thinking, reasoning, discussion, multiple solution strategies, and other terms associated with the Math Solutions professional development. To start, we found only very few mentions of these terms; out of all \((n = 187)\) Year 3 treatment-group logs, we noted only 12 where the teachers named such an activity as something that went well during the lesson. For the inverse question, what did not go well, only three logs recorded a wish for more student reasoning or communication. This suggests that, at a minimum, teachers were not attending to Math Solutions principles in reflecting on their lessons.

Next, we compared the 12 cases where teachers used a Math Solutions-type term to describe their instruction to observers’ written characterizations of the lessons. In most cases, we did find a correspondence between the two descriptions of the lesson; when teachers mentioned asking why questions or fostering student discussions or multiple solutions strategies, for instance, observers’ notes often explicitly mentioned or alluded to the same. Observers noted, however, that these activities were often enacted with lower levels of quality than providers
would wish, an issue Math Solutions providers also hinted at in their comments about lessons they observed (for example: “I think there was definitely more math talk, but it wasn’t necessarily more productive”).

Together, this evidence suggests that sense-making was not a major explanatory factor for static instruction. The 14 teachers who referenced Math Solutions practices related to their instruction seemed to understand and interpret the meaning of those practices in the same way as the observers. It is more likely that inattention to Math Solutions goals in the day-to-day conduct of instruction and the difficulty of enacting high-quality practice might have been. We return to these issues below.

**Insufficient resources.** A third potential reason for the lack of change in instructional practice relates to the resources teachers use when instructing students. As identified in prior research (e.g., Ball, 1990; Cohen, 2011; Lampert, 1990), teachers’ knowledge of content is one such resource. In this line of thinking, deeper knowledge of mathematics would ease the development of practice that features more rigorous mathematics and more student participation. If this were true in our data, we would see that instructional change may correlate with initial MKT status, with teachers who had stronger mathematical knowledge for teaching exhibiting more change and weaker teachers remaining the same. Our data, however, suggests that baseline MKT was not a significant predictor of change in MQI scores ($r = -0.18, p = 0.39$).

Curriculum materials can also serve as a support for the enactment of reformed instructional practice (Davis & Krajcik, 2005; Ball & Cohen, 1996). Teachers who undertake ambitious instruction while using conventional materials face steep costs—designing or unearthing more challenging tasks, crafting open-ended questions for students, anticipating students’ conceptual misunderstandings and how to remediate those misunderstandings. At the
time the study began, Eastern was using *Math Expressions* (Fuson, 2006). An inspection of fourth- and fifth-grade lessons suggests that there was considerable overlap between *Math Expressions* and Math Solutions in the mathematical content covered. However, an inspection of Math Solutions materials, which were presented to the teachers in the form of both professional development activities and published books, suggests that the lessons written by Math Solutions authors tended to follow a quite different format than those in *Math Expressions*. Where *Math Expressions* lessons support teachers’ careful scaffolding concepts and skills for students to learn in a whole-group setting, Math Solutions materials begin with teachers posing a complex question or problem for students to solve, often collaboratively. Thus a lack of alignment between the district’s assigned textbook and Math Solutions materials may have suppressed the implementation of the ideas transmitted through professional development.

However, an examination of teachers’ Year 3\(^2\) reports of curriculum use against observers’ notes suggests that lesson quality—in particular the presence of student participation in mathematical explanation and reasoning—did not substantially vary by the materials used by teachers. To confirm this, we calculated overall MQI for each major category of curriculum material reported by teachers (see Table 2). If anything, the use of Math Solutions materials was associated with the lowest MQI scores. This suggests that it was not a lack of appropriately aligned materials that contributed to the null results.

**Organizational barriers.** Similar to the policy implementation literature, our analyses identified a set of district- and school-level factors that appeared to hinder the implementation of Math Solutions practices in classrooms. Some of these factors relate to what McLaughlin (1987) termed environmental stability. Between teacher recruitment and the end of the study, state and

\(^2\) We conducted this test for Year 3 lesson data only because the implementation of Math Solutions lessons in Year 1 could have been complicated by the fact that teachers were new to the materials.
district mathematics standards and pacing guides were updated, the state assessment changed from paper and pencil to computerized and was reportedly upgraded to include more cognitively demanding problems, Eastern had three different superintendents, and the district’s math coordinator (a “fixer” of problems related to the program’s implementation; see Bardach, 1977) retired. As a result of these changes, district support for the professional development waned over the years, and teachers often could not attend scheduled professional development sessions because the district failed to prioritize time for them to leave their classrooms, particularly in the second year of the program. Limited teacher access to appropriate technology also prevented the implementation of an online learning community, intended to provide support to teachers between formal professional development sessions. In interviews for the study, Math Solutions providers spoke of these as barriers to their work in the district.

Another district-level factor appeared to be competing instructional guidance. Eastern, like many other urban districts struggling to make progress on state assessments and accountability metrics, maintained a series of instructional policies regulating teachers’ content coverage, instructional strategies, and time use in the classroom. In Eastern, this took the form of specific grade-level objectives, frequent benchmark assessments, and pacing guides. District leaders rewrote some of these documents during the second year of the project, asking teachers to reconfigure their lesson sequencing and also to locate new lessons to fill content gaps. Eastern also required the use of specific student grouping structures and, in Year 3 of the study, instituted a mandatory “calendar math” session for daily review of place value and basic mathematical skills.

This district instructional policy was a frequent topic in our Year 3 interviews with Math Solutions teachers. In these interviews, we asked whether any aspects of Math Solutions were
difficult to implement in the classroom—a relatively broad, open-ended question meant to elicit any considerations teachers considered relevant. Over half—17 out of the 31 teachers interviewed—identified Eastern’s instructional guidance as a significant reason for nonimplementation of Math Solutions. Some teachers specifically mentioned the district curriculum and pacing guides. For example:

You know and even like the materials that we get from Math Solutions . . . because of our curriculum and our pacing and all of those things, you’re not always able to, you know, use it, incorporate it.

A second set of teachers identified lesson length—either their prescribed mathematics time blocks or the time demands of Math Solutions lessons—as a major factor:

I think the only thing is a lot of [Math Solutions] is time consuming. Like I have an hour for math. Like I don’t have all this time. And there’s a pacing guide, like I’ve got to get to the next thing. And ideally it’s not Math Solutions who’s wrong; it’s more of the way that the system is set up in the district and the state by expecting all these things to be done in a short amount of time.

Finally, a third set of teachers also noted a lack of fit between Math Solutions and their school or district requirements regarding the use of ability-sorted small groups in mathematics. For instance, a teacher commented:

There are a lot of [Math Solutions] activities and a lot of approaches and strategies and things that I would have liked to do in my classroom. However, what principals want to see and what we’re allowed to do is a different story. So I would have liked to do a lot more of heterogeneous groups and a lot more of, you know, problem-solving-based learning as the way we did it in Math Solutions. . . . And we’re not able to do that as often as we probably should because of what Eastern thinks is the best way to teach math and what my principal thinks is the best way to teach math.

As a result, many teachers saw the Math Solutions material and approach as separate from their “regular instruction.”

This suggests that the district environment may have played a significant role in producing the results observed in our evaluation. In our experience, it is not uncommon for...
districts to increase efforts to control instruction in light of state testing and accountability; it is also not uncommon to see multiple individuals and offices within the district add programs and additional guidance on top of extant decisions about curriculum and instruction. In this case, such guidance seems to have hampered the implementation of the instructional practices recommended by Math Solutions staff.

The difficulty of enacting ambitious practice. Kennedy (2005) and Cohen (2011) suggest that one reason for the nonimplementation of ambitious mathematics reforms may be the difficulty of the practice itself. We tested this idea by examining observers’ write-ups and corresponding interviews for the seven teachers who appeared to be making some changes toward Math Solutions practices between Year 1 and Year 3. As noted above, five of those teachers enacted the instructional questioning and tasks called for by the Math Solutions program superficially, suggesting that although the teacher was trying to make changes in the direction of the program, such changes were difficult to put in practice. For instance, one teacher had adopted Math Solutions-recommended questioning strategies, in Year 3 asking students, “Why?” or to explain their thinking more often than in Year 1. However, the observer noted that students’ responses to these “why” questions were more procedural than conceptual, a report on the series of steps the student went through to solve a problem rather than an explanation of why a solution was true or a particular method worked. Asking why, as the program recommends, is easy; however, knowing when to ask why, discerning and encouraging meaning-focused student responses, and responding to those students’ ideas is a far more difficult teaching task (Cohen, 2011).
The same pattern was observed in teachers’ use of tasks aligned with those found in the Math Solutions program. Though cognitively complex tasks were used, they were rarely used to their potential. For instance, one observer wrote:

[Year 3] lessons featured a number of tasks that had potential to be inquiry/exploratory in nature but were enacted in a way that either removed the possibility of students finding their own way through the task or just devolved the cognitive demand. . . . If anything, I might argue that the teacher-designed tasks in [Year 3] reflect a misguided attempt at meaning-making. . . . It does seem that she has the idea to do more “engaging” longer tasks, but they never go deep and suffer from the same problems as her other lessons.

Similarly, other implementations of open-ended tasks featured slow-paced mathematics without the richer discussions or the student mathematical thinking and reasoning intended by Math Solutions developers. In one lesson, the class spent 47 minutes estimating how many beans would fit into a cup, then how many cups of beans would fit into a larger container. Students spent a majority of the lesson recording numbers, with no connections to larger ideas about estimation, volume, or three-dimensional shapes. An observer summarized what she saw for another teacher:

Although the [Year 3] lessons were clearly more open-ended and more student-centered, they were not appreciably higher in quality. The pacing was very slow and very little mathematics got done over fairly large chunks of time (for example, students were asked to find two numbers that multiply to 360 in groups and worked on this task for over a half-hour). Ultimately, the math worked on in these “investigations” was somewhat superficial (listing the factors of a product, dividing three pizzas among six people) and connections to big ideas were rarely surfaced.

To successfully enact such high-cognitive-demand tasks, these teachers would have had to either possess or learn several skills: how to launch a mathematical task without devolving its demands on student thinking; how to thread mathematical ideas through an entire investigation, drawing students’ attention to and engagement with those ideas; and how to maintain the mathematical density and pace of the lesson while allowing for depth in discussion.
We argue that these observations, along with those above regarding the quality of lessons that use Math Solutions materials, provide evidence for the “difficulty of ambitious practice” hypothesis. Teachers who reported making sincere efforts to change, and whose classrooms did feature superficial differences between the first and third year of the study, did not improve along the major dimensions targeted by the intervention, likely because the new form of instruction required skills they had not yet developed. Another way of looking at this problem, however, is that the treatment was too weak to support teachers’ transition to high-quality, ambitious practice. In fact, a read of our professional development field notes suggested that while the program itself prized ambitious instruction, its activities and features stayed at a superficial level rather than building core technical skills to enable high-quality implementation. For example, discussion was valued, and mathematically meaningful discussions were often on display in videos and in the modeling by program staff; however, teachers themselves neither identified nor practiced the specific techniques involved in leading mathematically productive discussions (see, e.g., Chazan & Ball, 1999). Teachers were cautioned not to devolve the cognitive demand of tasks but did not work on how to launch tasks in a way that both allow students to work productively but does not directly give students solution methods (Jackson, Shahan, Gibbons & Cobb, 2012). One reason for the lack of work on such technical skills may have been a district and program culture that pulled away from difficult public work; we discuss this next.

**Emergent Hypotheses**

**Weak instructional press.** Our own observations of professional development, as well as discussions with professional developers after the conclusion of the study, suggested another potential reason for null effects: that factors both inside and outside the professional development program led to a relatively weak “instructional press” (Firestone & Rosenblum,
on teachers to improve their craft. This was most visible in work around improving teachers’ mathematical knowledge and in assisting them in developing effective classroom instructional strategies.

On the first issue, the professional development was filled with what we judged to be mathematical learning opportunities for teachers: teachers solving problems their students would solve, with attention to meaning; videos of students using inventive and correct solution methods to complex mathematical problems; examples of ways to link numeric and visual solution methods. However, as teachers carried out these tasks, the developers often stopped short of pressing hard on the underlying mathematics. For example, one task asked teachers to determine how many possible pentominoes exist (a pentomino is a shape made with five squares with edges and vertices touching). Teachers worked in groups, finding all 12 possible pentominoes, then shared each pentomino at length with the whole group. But rather than delving deeper into the mathematics (How do we know we have them all? What strategies could you use to make sure you had all sextominoes without actually making the arrangements of shapes? What is the mathematical point of this activity? What mathematical practices does this activity involve?), the subsequent conversation focused on group dynamics and other instructional matters. In other cases, the providers overlooked teachers’ difficulties solving mathematics tasks. For instance, when a group of teachers failed to successfully model problems such as $1/2 + 2/3$ using manipulatives, the provider simply worked one example publicly and moved on without further comment, despite her disclosure later that she knew many of the teachers struggled with the task (which used mathematics they were expected to model at their grade level).

On the second issue, a lack of instructional press was also evident during demonstration lesson debriefings, according to Math Solutions developers. For both lessons taught by Math
Solutions providers and by Eastern teachers themselves, few public critiques of the instruction surfaced. Math Solutions providers reported that they were hesitant to press critiques, for they needed to ensure future teachers would feel safe participating in model lessons.

These two areas provide examples of ways Math Solutions staff failed to press on the skills and knowledge necessary for ambitious instruction. Although our study was not designed to examine this issue in detail, we hypothesize that the lack of press relates to the “culture of nice” (Mangin & Stoelinga, 2011) and privatization of instruction that exists within schools and under which teachers are protected from questions about their professional judgment and expertise. As well, Math Solutions staff raised concerns about retaining teachers in the study, noting that if teachers did not view the professional development as a positive experience, they would not return to the professional development and thus likely count against the program in the impact evaluation. Whatever the cause, conditions in this study and district did not allow close work on topics that could have supported improved instruction.

**Fit of the intervention to teachers’ needs.** Finally, our analyses uncovered one further potential reason for the null results: the fit of the intervention to teachers’ needs. We make this claim based on evidence present in observers’ notes that suggested that teachers’ instructional quality varied widely after the first year of the intervention. For one group of teachers, Math Solutions was in fact a good fit; their Year 1 lessons featured little student thinking and reasoning, and the mathematics conveyed seldom went beyond procedures. Professional development focused on changing teachers’ beliefs about good instruction, providing models for such instruction, and exposing teachers to mathematics and classroom mathematics activities seems like a reasonable starting place for teachers in this category.
However, a second group of teachers, whether due to the Year 1 program itself or to pre-existing experiences and learning opportunities, already displayed a fair number of the practices the program desired. Although these practices were not necessarily enacted consistently throughout each Year 1 lesson or at the highest level of sophistication, roughly one third of the sample showed some productive work in these domains. The presence of these teachers in the professional development may make providers’ work more complex as these teachers have already taken a first step to change instruction and may have needed help fine-tuning those practices and encouragement to make them more prevalent in classrooms.

Our analysis also uncovered multiple areas of need not related to the cognitive demand of lessons. One was teachers’ mathematical content knowledge. Although observers’ notes did not record outright content errors in teachers’ lessons (e.g., solving problems incorrectly, misdefining terms), teachers’ communication of the mathematics to students often suffered in Year 1 lessons. In some cases, teachers’ presentations and launches of lesson tasks were muddled, with students sometimes unsure of how to engage in the mathematical task laid out. In other cases, observers noted instances in which the teacher likely had a correct idea of the mathematical content, but the language used to communicate the idea was not precise (for instance defining perimeter as “distance around” a shape and telling students that finding the perimeter consisted of “counting the squares around” rather than measuring the length of the shape’s edge).

Another area of need identified in observers’ Year 1 notes was around responding to students’ thinking and misconceptions in productive ways. In some cases, teachers chose not to engage incorrect student answers; in other cases, students made correct and sometimes elegant contributions to the mathematical discussion, but teachers struggled to use those thoughts to
develop a mathematical point. The thinking entailed in parsing student thoughts, connecting them to the mathematical point, and deciding how to move instruction forward using those thoughts has been shown to be a quite complex (Lampert, 2003).

Our analysis also showed that some Year 1 lessons meant to be open-ended and exploratory also suffered from a lack of a clear mathematical point. For example, in a game in which students could have developed sophisticated strategies for multiplying successive rolls of dice by a variable to arrive at a target number, a teacher was unable to describe the mathematical content beyond twice stating, “We’re using probability, and we’re computing.” Student written work during the lesson consisted only of computation, and by the end of the lesson, there had been no discussion of using principles from probability to better reach the target value.

It is worth noting that the use of Math Solutions materials may have exacerbated this issue. For instance, one observer described the enactment of a Math Solutions fractions game:

Teacher devolves cognitive demand a bit by extensively modeling not only the game but the thinking required to play it. [Teacher] takes about 28 minutes to describe game, model it, and clarify directions. Students don’t begin game until 28:00.

For another Math Solutions lesson, the observer noted that “[the teacher’s] directions were jumbled up. She seemed to not know the mathematical point herself.” In these and other lessons, students were often left engaging in the task-specific procedures demonstrated by the teacher (e.g., roll dice, record result), but without a clear mathematical purpose and connection. It may have been that such a purpose was not made clear within the Math Solutions professional development and materials themselves; alternatively, making sense of the mathematics in this way may be a relatively difficult task in the enactment of ambitious instruction. Whatever the reason, observers often noted that the lessons that specifically relied on Math Solutions materials
were not mathematically dense, in terms of containing important mathematical ideas and/or opportunities for student thinking and reasoning about those ideas.

A final issue observed in Year 1 lessons relates to pacing and the resulting density of the mathematics in instruction. For half of the teachers in the sample, observers noted that the time spent on some activities was quite long relative to the amount of mathematical content covered. In some cases, this occurred because of lengthy preparation of or passing out of materials (e.g., cutting out triangles or copying definitions onto a poster), transitions between lesson activities, and off-task discussion and chatter. In other cases, teachers and students simply solved problems very slowly, sometimes covering only two or three in a given lesson without compensating features such as in-depth discussion or careful attention to meaning. Math Solutions materials may have exacerbated this problem for some teachers, especially when the enacted activity lacked a mathematical point and students engaged mainly in computational tasks or clerical activities, again decreasing the mathematical density of the lesson. For many of these activities, mathematical closure was seldom evident.

Thus teachers’ needs differed from what Math Solutions provided. The Math Solutions program, at the time we studied it, focused more on creating a new vision for instruction than providing the technical skills to support that vision. But these technical skills—launching tasks so that students can profitably work, responding to student ideas—may be necessary before high-quality implementation can proceed.

Discussion

In this paper, we evaluated potential reasons for the failure of a professional development program to affect practice and student outcomes. We found little or no support for three phenomena common in other implementation and evaluation studies: a lack of will to change on
the part of teachers, insufficient resources, and sense-making. These factors may play important roles in other program and policy failures, but we think the probability that they were at work here is low.

We identified evidence that organizational barriers, particularly in the form of competing district instructional guidance, likely played a significant role in program failure. While the effects of conflicting district guidance has been noted (e.g., Spillane, 1998), the amount and specificity of the guidance around instructional activities in Eastern (e.g., calendar math, homogenous grouping) seemed to us different than what has been observed in prior literature. The culture of teacher autonomy—long described in the literature as “close the door and teach how you want” (Lortie, 1975)—appears to have eroded, at least in this district (see also Spillane, Parise, & Sherer, 2011). This has implications for professional development, especially those programs not tightly aligned to the mix of instructional guidance available in districts.

We also evaluated Kennedy’s (2005) hypothesis that ambitious practice is simply difficult to enact. From one perspective, the data appears to support this assertion—there was little strongly reformed practice among the lessons we viewed generally, and even among teachers whose lessons showed some sign of change between Year 1 and Year 3, those changes were typically modest. We also observed teachers struggle with difficult aspects of enacting such ambitious practice, such as responding to student thinking and making the mathematical point of a lesson clear. However, another way of viewing this relates to the weak treatment hypothesis raised in the implementation literature: Our analysis of the contents of Math Solutions suggests that many of the skills necessary to implement ambitious practice were not addressed by the program. Likely, these are two sides of the same coin; the difficulty of the new practice desired by a program must be matched by the supports that program offers for enactment.
Our first emergent hypotheses also suggested that several factors may have constrained providers from pressing on teachers’ mathematical knowledge and instruction. One such factor is the “culture of nice.” Although this is often used in descriptions of reform efforts (Mangin & Stoelinga, 2011; MacDonald, 2011), we know of only one prior study that has empirically identified this phenomena as a potential barrier to the efficacy of professional development (Wilson, Lubienski, & Mattson, 1996). Yet the professional development staff appeared familiar with this dilemma, suggesting that this may be a widely faced problem in this sector. In our view, it is one that must be overcome for real change to occur. Thinking through the design of programs that allow heightened instructional press seems important if professional development is likely to survive as a viable forum for teacher change.

Our second emergent hypothesis—the lack of fit between the professional development and teachers’ needs—is also one we have seldom seen in the literature. This is curious and surprising in an era in which one-size-fits-all professional development has long dominated. Just as teachers themselves are increasingly being asked to understand and meet the needs of individual students, the system must do so for teachers; it makes no sense to us for teachers with negative classroom climates to be in math-focused professional development, or for very accomplished teachers to be in routine forms of professional development at all. With the advent of better data from teacher observations and video analysis of practice (e.g., Kane & Staiger, 2012), it seems likely to us that districts can—and should quickly—do better.

Finally, we ask what both teacher professional development and teaching would have to look like to enable ambitious instructional practice—in other words, how “strong” the treatment would need to be to see effects. Such questions are typically answered in terms of hours of professional development or by pointing to the formats and topics contained in effective
professional development (see, e.g., Landesman Ramey et al., 2011; Desimone, 2009). What seems missing from such formulations, however, is a backwards mapping from instructional goals to specific teacher competencies to the components of professional development that might develop such competencies. For instance, if ambitious instruction is the goal, teachers might need to build skill in reasoning-in-practice—what Rowland, Huckstep and Thwaites (2005) would call “contingent knowledge” or the types of thinking Lampert (2003) elegantly describes. To develop such thinking, teachers would surely need opportunities to face complex instructional situations, try out instructional moves and questions, receive feedback on their performance, and to reflect on the experience. Teachers may also have to build content knowledge—not just the actual content being taught, but also the structure of the discipline (N. Joglar Prieto, personal communication, July 12, 2015) and how particular classroom activities and student ideas relate to that structure. Finally, professional development may need to more resemble comprehensive school reform in carrying components meant to organize environments that support instructional improvement (Peurach, 2011).

In addition to backward mapping, we also argue that the individualization of professional development would maximize learning gains for both teachers and students. As we move past the era of one-size-fits-all workshops, districts might invest in diagnostic systems, perhaps linked to teacher evaluation systems, that would deliver the right tools at the right time. Such systems appear possible and efficacious with students (Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007), and it should be tried with teachers. We imagine that this work would more often take place in classrooms themselves, through coaching and individualized support designed to help teachers work on the specific aspects of practice required for improvement; working in classrooms with individual teachers would also allow more nimble responses to implementation
barriers. Math Solutions itself has moved away from workshop-style professional development and toward such coaching.

Finally, we conclude with thoughts about the replicability of this method. We argue that for professional development to become more effective, more analyses of this kind should be undertaken. Doing so requires enhanced forward planning and data collection: Analysts must anticipate likely reasons for program failure, perhaps using a similar approach to ours, which relied upon the implementation literature; then data collected during and after the study must allow for evaluation of these hypotheses and others. We argue that this is an advance over the ad hoc explanations sometimes seen at the end of null result reports and worth the additional expense. Over time, as more analysts use such a method, we may triangulate among reports and identify means for both strengthening professional development design and better matching specific programs to environments likely to allow them to thrive.
References

Author (2001).

Author (2004).

Authors (2015).

Authors (under review).


Table 1

*Reported Changes from Participant Interviews Attributed to the Math Solutions Professional Development*

<table>
<thead>
<tr>
<th>Reported Change</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher asked more questions/better questions</td>
<td>41</td>
</tr>
<tr>
<td>Teacher focused on deeper thinking, not just answers</td>
<td>34</td>
</tr>
<tr>
<td>Teacher used more exploration activities</td>
<td>31</td>
</tr>
<tr>
<td>Teacher used or allowed multiple ways to solve a problem</td>
<td>31</td>
</tr>
<tr>
<td>Teacher taught the “why” and developed understanding</td>
<td>25</td>
</tr>
<tr>
<td>Teacher increased use of technology or manipulatives</td>
<td>22</td>
</tr>
<tr>
<td>Teacher listened for and used student thinking</td>
<td>22</td>
</tr>
<tr>
<td>Teacher talked less</td>
<td>13</td>
</tr>
<tr>
<td>Teacher presented multiple ways</td>
<td>13</td>
</tr>
<tr>
<td>Teacher acted more like a facilitator</td>
<td>13</td>
</tr>
<tr>
<td>Teacher engaged in more “math talk”</td>
<td>13</td>
</tr>
<tr>
<td>Teacher made more mathematical connections</td>
<td>9</td>
</tr>
</tbody>
</table>
Table 2

*Overall MQI Lesson Score by Curriculum Materials, Year 3*

<table>
<thead>
<tr>
<th>Curriculum Materials</th>
<th>Overall MQI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math Solutions ($n = 15$)</td>
<td>2.74</td>
</tr>
<tr>
<td><em>Math Expressions</em> ($n = 22$)</td>
<td>2.71</td>
</tr>
<tr>
<td>Norfolk online materials ($n = 21$)</td>
<td>2.88</td>
</tr>
<tr>
<td>Other ($n = 75$)</td>
<td>2.96</td>
</tr>
</tbody>
</table>