A Lesson for the Common Core Standards Era from the NCTM Standards Era: The Importance of Considering School-Level Buy-in When Implementing and Evaluating Standards-Based Instructional Materials

Steven Kramer, Jinfa Cai, and F. Joseph Merlino

Those who cannot remember the past are condemned to repeat it.

George Santayana

In June 2010, the Council of Chief State School Officers and National Governor's Association promulgated the Common Core State Standards (CCSS) for mathematics and literacy (CCSSO/NGA, 2010). The new standards are expected to "stimulate significant and immediate revisions…in classroom curriculum materials (Council of Chief State School Officers, Brookhill, & Texas Instruments, 2011)." Similarly, the new *Next Generation Science Standards* may require educators to develop and implement new instructional materials (NSTA, 2012).

Today's new standards build on previous efforts, including earlier standards promulgated by the National Council of Teachers of Mathematics (NCTM, 1989, 2000, 2009a, 2009b). Soon after the publication of the first *Standards* document (NCTM, 1989), the National Science Foundation (NSF) funded development of a number of elementary, middle, and high school mathematics curricula (hereafter referred to as "NSF-funded curricula") designed to implement the *Standards*. Studies evaluating

Research reported in this paper was supported by funding from the National Science Foundation Award Numbers: 9731483 and 0314806. Any opinions expressed herein are those of the authors and do not necessarily represent the views of the National Science Foundation.

S. Kramer (✉) • F.J. Merlino
The 21st Century Partnership for STEM Education, Conshohocken, PA, USA
e-mail: skramer@21pstem.org

J. Cai
University of Delaware, Ewing Hall 523, Newark, DE 19716, USA
e-mail: jcai@udel.edu

J.A. Middleton et al. (eds.), *Large-Scale Studies in Mathematics Education*, Research in Mathematics Education, DOI 10.1007/978-3-319-07716-1_2

the effectiveness of the NSF-funded curricula can provide important lessons for today's new curriculum reform efforts.

The current study investigates the effectiveness of NSF-funded middle school mathematics curricula implemented with the assistance of the Greater Philadelphia Secondary Mathematics Project (GPSMP). The GPSMP, operating between 1998 and 2003, was an NSF-funded Local Systemic Change (LSC) initiative. This study differs from previous studies by evaluating the mediating effects of a school-level measure of stakeholder buy-in. We found that the degree of principal and teacher buy-in had a large impact on curriculum effectiveness. These results have potentially important implications for today's efforts to implement new instructional materials, providing insights both about how to support implementation and about how to evaluate the effectiveness of those materials.

Background

The original NCTM *Standards* emphasized student reasoning as being central to learning mathematics. Mathematics curriculum materials that had been in widespread use prior to promulgation of the *Standards* were perceived as placing too much emphasis on procedural fluency at the cost of ignoring conceptual understanding and applications (Hiebert, 1999). Based on early field trials of new curricula designed to implement the *Standards*, developers cautioned that teachers could find it difficult to change their practice (Cai, Nie, & Moyer, 2010). The NSF attempted to address this issue by establishing the LSC Initiative. The LSC theory of action argued that providing teachers with extensive professional development in the context of implementing the new NSF-funded curricula would result in teachers having both the inclination and capacity to implement the curricula. Between 1995 and 2002, NSF funded 88 multi-year LSC mathematics and/or science projects in Grades K-12 (Banilower, Boyd, Pasley, & Weiss, 2006).

The passage of the No Child Left Behind Act in 2001 and the establishment of the What Works Clearinghouse in 2002 heralded a new wave of educational reform focusing on student assessment and "scientifically based research" to investigate the effects of educational innovations (Slavin, 2002). Researchers began investigating the effectiveness of NSF-funded mathematics curricula. Syntheses of this early research tended to report positive achievement effects on researcher-administered tests using open-ended problems, but near-zero achievement effects on standardized tests measuring basic mathematical skills (Cai, 2003; Kilpatrick, 2003; Slavin, Lake, & Groff, 2008; U.S. Department of Education, 2007a). These early studies of NSF-funded curricula generally used a quasi-experimental *intent-to-treat* analysis, comparing achievement growth in schools and/or classrooms implementing new curricula with achievement growth in matched comparison schools/classrooms that implemented business-as-usual curricula. As shown in Fig. 1, intent-to-treat views curriculum implementation as a black box, comparing the achievement of students assigned to Treatment classrooms to the achievement of students assigned to Comparison classrooms without regard to actual classroom instruction (see, e.g., Riordan & Noyce, 2001).

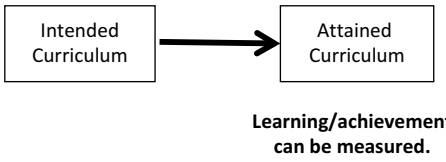

Fig. 1 Intent-to-treat evaluation model

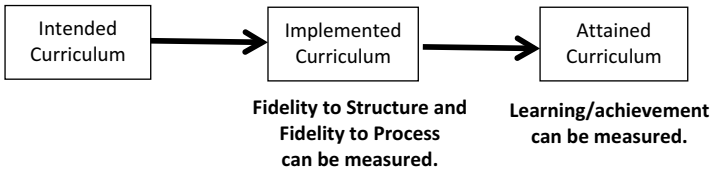

Fig. 2 Evaluation model with implementation fidelity

Recently, social science researchers have become concerned about information that is obscured by intent-to-treat studies. Evaluators have emphasized the importance of focusing not only on average effects, but also on mediating factors which can affect program outcomes when delivered under naturalistic conditions (Vuchinich, Flay, Aber, & Bickman, 2012). Among the potential mediating factors, evaluators have put particular emphasis on *fidelity of implementation* (Flay et al., 2005; U.S. Department of Education, 2007b). In an extensive review of the literature, O'Donnell (2008) defined fidelity of implementation as the degree to which an intervention is implemented as originally intended in the program design. Without a measure of fidelity of implementation, researchers may not be able to determine whether unsuccessful outcomes are due to an ineffective program or are due to failure to implement the program as intended (Dobson & Cook, 1980; Forgatch, Patterson, & DeGarmo, 2005; Hohmann & Shear, 2002; O'Donnell, 2008). As shown in Fig. 2, researchers focusing on fidelity of implementation differentiate between the intended curriculum embodied in curriculum materials, the implemented curriculum as seen in the classroom, and the attained curriculum as reflected in tests and other measures of student achievement (Cai, 2010). O'Donnell (2008) extended the concept of fidelity to include both *fidelity to structure* and *fidelity to process*. Fidelity to program structure means actually using the program materials as intended—and will be seen only in Treatment groups. Fidelity to process, in contrast, involves implementing processes congruent with the underlying program theory of action and might be seen in both Treatment and Control/Comparison groups.

Recent effectiveness studies have indeed confirmed an interaction between fidelity and treatment effects. In an evaluation of a supplemental elementary school math intervention aimed at increasing computational fluency, VanDerHeyden, McLaughlin, Algina, and Snyder (2012) found that a measure of fidelity to structure in Treatment classrooms predicted higher achievement on statewide test scores. Four recent studies evaluated inquiry-based middle school mathematics or science curricula while investigating fidelity to process (Cai, Wang, Moyer, Wang, & Nie, 2011; O'Donnell & Lynch, 2008; Romberg, Folgert, & Shafer, 2005; Tarr et al., 2008). All four found that Treatment classrooms with high fidelity to process

showed more achievement growth than either Control classrooms or Treatment classrooms with low fidelity to process.

While many researchers and funders investigating program effectiveness have focused on fidelity of implementation, other researchers have taken a contrasting *mutual adaptation* or *co-construction* perspective that views fidelity of implementation itself as too simplistic a construct (e.g., Cho, 1998; Remillard, 2005). The mutual adaptation perspective emphasizes that any curriculum implementation necessarily involves teachers transforming the written curriculum, working with those materials to co-construct the enacted curriculum.

Researchers working on design experiments have used an evolutionary metaphor to describe this view. Some program changes are "lethal mutations" which decrease quality of learning, whereas other changes are "productive adaptations" which increase quality of learning (Brown & Campione, 1996; Design-Based Research Collective, 2003). Brown and Edelson (2001) described one such "productive adaptation" to the Global Warming Project (GWP), an inquiry-based middle school science curriculum that Brown had helped develop. They described how one teacher, Janet, implemented *The Sun's Rays*, an investigation that occurs approximately midway through the GWP.

> Rather than have her students follow the "recipe" for doing the lab, she decided to turn the activity into an opportunity for them to engage in experimental design. Instead of providing them with a set list of materials, she gave them access to a host of supplies which she gathered from her own supply closet and borrowed from other teachers. And rather than just connect the elements of the lab model to the actual phenomena they represented, she relied on the model throughout the lesson as a means to stimulate deep reflection and analysis of the results (p. 15).

The authors viewed Janet's extensive adaptations as consistent with the program philosophy and, if anything, an improvement on the original lesson materials.

The current study extends the "implementation fidelity" model in Fig. 2 by introducing *buy-in*, a concept taken from research into Comprehensive School Reform (Cross, 2004; Glennan, Bodilly, Galegher, & Kerr, 2004; Schwartzbeck, 2002). While buy-in is often discussed in the Comprehensive School Reform literature, the buy-in concept has seldom been formally defined. One exception is Turnbull (2002), who used a five-part operational definition for buy-in to a school reform model. Teachers bought in to the model if they understood the model, believed they could make the model work, were personally motivated to do so, and believed the model was good for their school and would help them become better teachers. Our definition of buy-in is consistent with Turnbull's, but more general, applying not only to comprehensive school reform, but to any school program, and to principals and other stake-holders as well as to teachers. Our definition is also influenced by a co-construction view of program implementation. We define buy-in as *the degree to which stakeholders understand the underlying program theory and embrace that theory*. Stakeholders who buy in to a program are less likely to introduce lethal mutations—and, to the degree their ability and situation allows, they are more likely to introduce productive adaptations. When applied to curriculum materials, buy-in reflects stakeholders' attitudes toward, beliefs about, and understandings of those

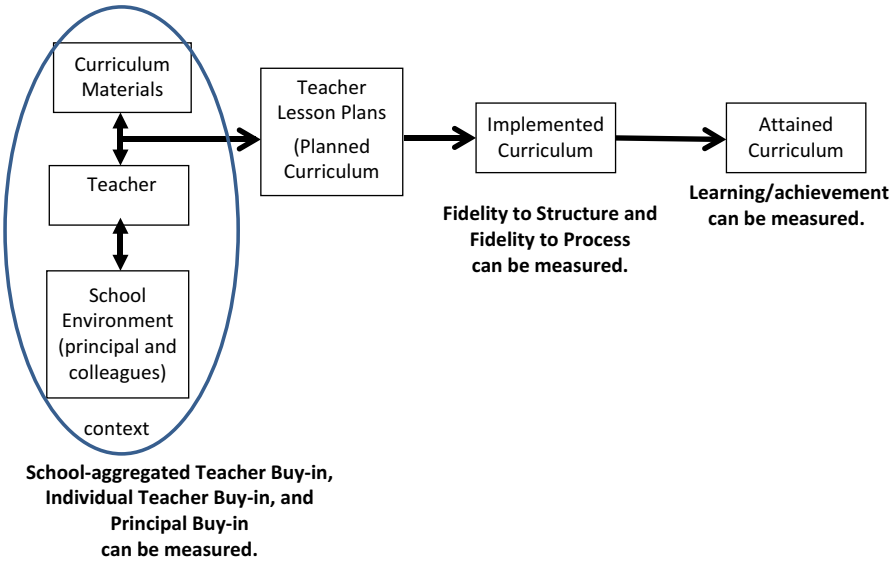

Fig. 3 Evaluation model with buy-in and implementation fidelity

materials. It is hypothesized that stakeholders with strong buy-in will be more likely to implement a program with fidelity to both structure and process.

Buy-in might be easier to measure than fidelity, especially fidelity to process. Measures of buy-in might be obtained via interviews or surveys of teachers and other stakeholders, or even indirectly via interviews or surveys of teacher leaders and mentors, whereas measures of fidelity to process might require teacher logs or classroom observations. Furthermore, professional development and other efforts to support program implementation might be more successful if such efforts seek to secure buy-in and help teachers and other stakeholders become active co-constructors of the curriculum, rather than seeking only to help them implement the program with fidelity to its structure and processes. Figure 3 (modified from Remillard, 2005) represents a curriculum evaluation model incorporating both buy-in and fidelity of implementation.

Buy-in has not often been addressed by curriculum effectiveness studies. In their evaluation of four elementary school math curricula, Agodini et al. (2010) asked teachers to state their interest in using their assigned curriculum again, if they were given a choice. The authors reported mean responses by curriculum, but did not attempt to analyze whether choosing "yes" correlated with higher achievement, perhaps because the dichotomous variable may have been too weak a measure of buy-in to achieve valid results. Similarly, in their evaluation of an intervention implementing supplemental math curriculum materials, VanDerHeyden et al. (2012) measured teacher-rated "acceptability" of the intervention by computing the school-level mean of teacher responses to 15 Likert-scale items measuring the program's perceived effectiveness, practicality, ease of implementation, potential risks,

etc. The authors did not report whether or not schools with high average teacher buy-in achieved better results than did schools with low average teacher buy-in. An unpublished supplemental analysis did not detect any correlation between buy-in and achievement, but because buy-in variation among the schools was not large and the sample consisted of only seven Treatment schools, the data available would have very low power to detect any such correlation (A. VanDerHeyden, personal communication, 2013). Thus, the current study breaks new ground by quantitatively analyzing the relationship between buy-in and program effectiveness.

The current study investigates the effectiveness of one particular LSC, the GPSMP. The GPSMP operated between 1998 and 2003. The study was commissioned by NSF in 2003 to use retrospective data to analyze GPSMP effects at twenty middle schools that implemented one of two NSF-funded middle school curricula, either *Mathematics in Context* (MiC) or *Connected Mathematics* (CMP). This study differs from previous studies of NSF-funded curricula in that it investigates the effectiveness not only of the curricula themselves, but of the LSC theory of action, which combined curriculum adoption with extensive professional development for teachers. The study also differs from previous studies by using a measure of buy-in as a mediating variable.

Over its 5 years of operation, the GPSMP provided an average of 59 h of professional development to each of 249 middle school teachers at the 20 middle schools that participated in the retrospective study. GPSMP differed from some other LSCs in that mentors working for the project supplemented professional development by providing extensive assistance to mathematics teachers implementing NSF-funded curricula.

The LSC theory of action predicted that, at most middle schools, implementing an NSF-funded mathematics curriculum combined with extensive teacher professional development would lead to sufficiently high fidelity curriculum implementation so that positive impacts on student achievement might be expected. In contrast, anecdotal reports from GPSMP mentors indicated that this was not the case. The mentors reported that even when all schools participated in extensive professional development activities, there remained systematic differences among middle schools in quality of implementation. These differences appeared to be a function of initial teacher and principal assent or buy-in and district-level support for the new curriculum, a factor we have named "Will to Reform."

The mentors' focus on the importance of Will to Reform was supported by previous research investigating conditions that facilitate or inhibit the process of implementing a new curriculum. Important factors identified included the teachers' buy-in, as well as support from school principals and district superintendents (e.g., Fullan & Pomfret, 1977; Krainer & Peter-Koop, 2003; Little, 1993). The LSC capstone report (Banilower et al., 2006) also noted that, over time, LSC Principal Investigators came to feel that school principals had a larger impact on the quality and impact of program implementation than had been recognized in the original theory of action.

This paper has two goals. First, it evaluates the effects of a moderately large scale (20 middle schools) implementation of the LSC model (adopting a problem-based mathematics curriculum combined with extensive professional development) on

student achievement, as measured by high-stakes, state-administered, standardized mathematics tests. Second, it focuses on Will to Reform, a school-level measure of buy-in to implementing the problem-based curriculum. It investigates how Will to Reform interacts with the LSC model to predict student achievement.

As shown in Fig. 3, teacher/curriculum interactions take place within a wider school context. While researchers from the 1980s (e.g., Goodlad, 1983) emphasized that teachers tend to work in isolation, more recent research has found that teachers see themselves as part of a larger coordinated system of instruction. For example, Kennedy (2004) reported that, when planning and implementing lessons, teachers often focused on their obligation to make sure students mastered the particular content the teachers who received their students the following year would expect them to have learned. Congruent with this "coordinated system" view, the GPSMP mentors described a school-wide gestalt Will to Reform. Consequently, it was reasonable to believe that this school-level measure of buy-in might mediate the effect of curriculum on student achievement. While it would have been worthwhile to evaluate the effects of buy-in measured at the teacher-level in addition to the effects of buy-in measured at the school level, the retrospective nature of our data made doing so impossible. Also due to the retrospective nature of the data, we do not have available any direct measure of implementation fidelity. Thus, the current study investigates how school-level buy-in variables interact with curriculum materials to predict student achievement, without considering the effects of any intervening variables. Nonetheless, establishing whether or not such school-level buy-in variables predict student achievement is an important first step towards studying the more complete model displayed in Fig. 3.

Method

Achievement Measures

Student achievement was measured using eighth-grade state mathematics tests: the New Jersey Grade Eight Proficiency Assessment (GEPA) and the Pennsylvania System of State Assessment (PSSA) test. There are several advantages to using these two tests as the measure of student achievement. High stakes state tests measure what Confrey et al. (2004) called "curriculum alignment with systemic factors"—i.e., the degree to which students achieved the learning goals laid out for them by local authorities. Between 1999 and 2004, the GEPA was designed to assess student mastery of the *New Jersey Core Curriculum Content Standards for Mathematics* (New Jersey Department of Education, 1996). Subsequent to 1998, the Grade 8 PSSA assessed mastery of the Pennsylvania *Academic Standards for Mathematics* (Pennsylvania Department of Education, 1999). By using students' scores on both GEPA and PSSA, we assessed mathematics content that was both important to local stakeholders and well-aligned with the state curriculum goals.

Will-to-Reform Scale

An independent educational research firm located in Philadelphia, PA was employed to gather information about program implementation within each of the districts participating in the GPSMP follow-on. Two qualitative researchers were assigned to the project. They interviewed school administrators, teachers, and math supervisors at each of the schools, talked to the mentors, and reviewed hundreds of pages of notes mentors had made about their interactions with GPSMP schools. They also collected quantitative data on the amount of professional development attended by teachers at the various districts.

At the same time, the GPSMP mentors were asked to rate teachers and school administrators on their Will to Reform. They did so blind to the work of the two independent qualitative investigators. In February, 2006, the two qualitative researchers, the mathematics mentor(s) for each school/district, and the GPSMP Principal Investigator held a meeting to compare ratings of each school on Will to Reform.[1] Will to Reform was determined as the sum of a school's gestalt rating on each of four subscales: Teacher Buy-in, Principal Support, District Coherence, and Superintendent Support. In advance of the meeting, the scales were defined in a fairly cursory manner, with a more detailed definition developed by group consensus at the meeting.

Teacher Buy-in:

1 = Low Buy-in. Some teachers avoid professional development, taking personal leave days since they would "rather miss a workshop day than a school day." Most teachers who do attend professional development go through the motions: they tend not to ask questions and tend not to be responsive. There is a subset of teachers who are vocal in criticizing the new program, including complaining to parents about it. In schools with adequate test scores, some teachers express the attitude, "If it ain't broke, don't fix it." In schools with lower test scores, teachers tend to see the problem as being located in "kids today," "parents today," "society today," or in poor elementary school teaching. No matter what the test scores, there is a tendency for teachers to believe that the curriculum and pedagogy "won't work with our type of kids."

3 = Medium Buy-in. Teachers tend to view themselves as professionals. They are willing to be team players and implement what the school asks of them. They make an effort to attend professional development, as long as they receive a stipend for doing so. They tend to believe that their old ways of teaching might be improved, and to be willing to give something new a shot.

[1] We attempted to keep raters as blind as possible to student test achievement data. Before they started their research, the qualitative researchers were instructed not to review such data. Further, we asked mentors not to discuss or compare standardized test data across districts when making their ratings. However, it is possible that during the time when mentors were working with the districts they had learned whether test scores were improving, and perhaps even developed some sense about which schools were seeing relatively weaker or relatively stronger improvement.

5 = High Buy-in. Teachers tend to be excited about the new program and are eager to attend related professional development regardless of the pay. There is already a school culture supporting learning for the staff as well as for the students. Teachers tend to participate vocally in professional development and in math meetings and are willing to have others come in and observe them and provide feedback. As a group, teachers are proactive in dealing with the community and the school board, organizing parent meetings and similar activities. They aren't just walking into curriculum change, but have been building up to it: looking at student data to diagnose problems, trying out new teaching techniques like cooperative learning, etc. In general, the curriculum fits with where the majority had already been going.

Although only three levels (low, medium, and high) were described, each rater could rate Teacher Buy-in as "2" (between low and medium) or "4" (between medium and high).

Principal Support:

1 = Individual does not support the program. He/she may give lip service to it in front of district officials, but in private will criticize it.
2 = Neutral and disengaged. Often, mentors had never met these individuals even after spending many hours working with teachers in the building.
3 = Generally supportive of the program, but not an advocate; allows it to happen, takes an interest, but not willing to go out and fight for it. If there is any flak from the community, the principal defers to math teachers, who are expected to be the experts and defend what they are doing.
4 = Not only supportive, but also an advocate. Talks about the program in public meetings, and runs "interference" defending teachers from any community criticism. Lets teachers know that he/she is strongly behind the new curriculum.
5 = Supportive and an advocate, and a mathematics instructional leader. The principal understands the mathematics and learning theory behind the curriculum. He/she uses this knowledge to inform discussions with teachers about classroom practice, to inform teacher observations, to decide what types of professional development activities are appropriate for the staff, etc.

District Coherence:

1 = Program is incoherent. There is a lot of conflict and/or disagreement among the school board, superintendent, principals, teachers, and community about exactly where the program should go or what should be done.
3 = Medium coherence. There is no overt or obvious conflict about mathematics among the school board, superintendent, principals, and teachers. Community disagreements tend to be dealt with in a spirit of communication, not conflict.
5 = High coherence. Everyone is "pulling in the same direction." Programs like ongoing professional development for new teachers and advanced professional development are in place. District support staffs take an active interest in the math program, in collecting data about mathematics achievement, etc.

Similar to Teacher Buy-in, although only three levels (incoherent, medium coherent, and high coherent) were described for district coherence, each rater could rate District Coherence as "2" (between incoherence and medium coherence) or "4" (between medium coherence and high coherence).

Superintendent Support:

1 = Individual does not support the program. In cases where this happened, it tended to be a superintendent who inherited a predecessor's program, and who was interested in setting a new direction.
2 = Neutral and disengaged. Again, these tended to be superintendents who inherited the math program, but who were not actively hostile to it.
3 = Generally supportive of the program, but not an advocate; allows it to happen, takes an interest, but not willing to go out and fight for it.
4 = Not only supportive, but also an advocate. Talks about the program in public meetings, and runs "interference" defending principals and teachers from any community criticism. Lets principals and teachers know that he/she is strongly behind the new curriculum.
5 = Supportive and an advocate, and a mathematics instructional leader. The superintendent understands the mathematics and learning theory behind the curriculum and can use this knowledge in explaining what the district is doing, and in making plans with principals and other instructional leaders.

To rate a school on each subscale, each member in the group first described any information and experiences relevant to that school. The description was intended to cover the period from initial implementation through the spring of 2004, so individuals were asked to describe information relative to the overall tenor of teacher buy-in, district coherence, and support of principals and superintendents about the implementation during this period. Then, the group as a whole developed a consensus rating for each factor for each particular school. Although the rating scales were developed using retrospective data, they were based on the input of independent observers who had interviewed relevant stakeholders and reviewed detailed field notes taken by the mentors, plus the observations of the mentors themselves, who had acted as participant-observers. (Recall that the average math teacher at these middle schools participated in 59 h of professional development, much of it either one-on-one with the mentor or in group sessions taught by the mentor.) Additional observations and information were provided by the GPSMP Principal Investigator who had worked closely with district administrators and principals throughout the 5-year GPSMP project.

Each Treatment school was assigned a composite Will-to-Reform score by summing the subscale scores. Since each of the four subscales was scored from 1 to 5, the composite Will-to-Reform score could theoretically vary from a minimum of 4 to a maximum of 20. In practice, there was wide variation among Treatment schools in Will to Reform, with an observed minimum score of 5 and an observed maximum score of 19. Across the 20 Treatment schools, the mean Will-to-Reform score was 11.5 and the standard deviation was 3.62. Figure 4 displays a dot-plot of the observed

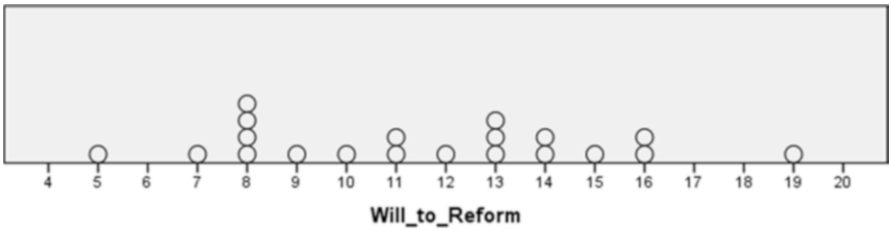

Fig. 4 Dot plot of observed Will-to-Reform scores

Will-to-Reform scores at the 20 Treatment schools. While necessarily imperfect due to the retrospective nature of the data, Will to Reform was a reasonable proxy measure for school-wide buy-in to the NSF-funded math curriculum that each Treatment school had implemented with GPSMP support.

Comparison Schools

School districts in suburban Pennsylvania and New Jersey are usually small, composed of one or two high schools and their feeder elementary and middle schools. For this reason, the participating GPSMP school districts each contained only one to four middle schools. Within each participating district, all middle schools adopted the chosen reform curriculum. Thus, we matched each GPSMP middle school to similar Comparison schools that were located in other, similar districts in the same state (either Pennsylvania or New Jersey).

Each GPSMP middle school was matched to a unique set of Comparison schools according to similar demographics (as reported by the National Center for Education Statistics 2004 data base) and test scores prior to GPSMP implementation. We chose to match using pre-determined "calipers" (maximum distance) on a set of covariates. While a number of studies have used propensity scores to match Treatment and Comparison groups (e.g., Stuart, 2007), for our study calipers had two advantages over propensity scores. Calipers allowed us to prioritize among covariates so that we matched most closely on baseline math scores and second most closely on baseline reading scores—the two covariates that were the best predictor of later-year math scores. Second, we found that while propensity scores match the entire set of Treatment schools so that on average they are similar to Comparison schools on each covariate, propensity scores might match an individual Treatment school to Comparison schools that are very different on specific covariates. The statistical models we used assumed that each Treatment school was individually matched to similar Comparison schools. Using calipers enabled us to accomplish this goal.

To select Comparison schools, we required a match within ±0.2 standard deviations on baseline scores in eighth grade mathematics and reading scores. We chose

0.2 standard deviations following the rule-of-thumb described by Rubin (2001) for propensity scores. For other variables, we aimed at finding schools within roughly the same quintile. Experience with data sets from several states (Arkansas, Illinois, Massachusetts, Michigan, Minnesota, Pennsylvania, New York, and Washington) led us to estimate that on average this could be accomplished by accepting schools within approximately ±17 % on Free and Reduced lunch and ±27 % on Percent White. We set the calipers for acceptable distance in Percent other races to be the same as calipers for Percent White. The actual matching proceeded along the following steps:

First, we first identified "Priority One" matches, defined as follows:

1. School-level Grade 8 math and reading scores in the "baseline year," which we defined as the school year prior to beginning GPSMP-supported professional development and/or curriculum implementation (1998 for all but one treatment school in Pennsylvania, and 1999 for New Jersey schools and the remaining Pennsylvania school), were within ±0.2 school-level standard deviations.
2. Within ±17 % Free and Reduced Lunch.
3. Within ±27 % for EACH of the following races: White, Black, Hispanic, Asian, and Native American.
4. Greater than 40 students enrolled in eighth grade in 2004.
5. School organization: Schools where students attended grades 6–8 but not earlier were matched to similar schools (either 6–8 or 6–12). Schools where students attended grades 5–8 were matched to similar schools (either 5–8 or K-8). Junior High schools (grades 7–8) were matched to other Junior High schools.

After identifying the set of Priority One Comparison schools, we sorted by three variables, in order: closeness of baseline math scores, closeness of baseline reading scores, and percent free/reduced lunch. For each Treatment school, we selected the top ten Priority One matches. (We used a predetermined algorithm to assign each Comparison school matching more than one Treatment school to one unique Treatment school.) If fewer than ten Priority One matches existed, we accepted all Priority One matches. In the few cases where this process yielded fewer than three Comparison schools, we used a predetermined algorithm to relax our criteria until we identified three acceptable Comparison schools. Table 1 compares Treatment to matched Comparison schools on baseline math and reading scores, as well as demographic variables. Because each Treatment school was paired with 3–10 Comparison schools, depending on how many good matches were available, for each variable we computed the average reported in Table 1 by first computing the mean for Comparison schools within each school-group, and then averaging across the 20 school-groups. Table 2 lists for each Treatment school the school's Will-to-Reform score, its math achievement growth between baseline year and 2004, the average math achievement growth of its matched comparison schools, and the number of Comparison schools identified by our matching algorithm. For each Treatment and Comparison school, math achievement growth was computed as within-state school level z-score in 2004 minus within-state school level z-score in baseline year.

Table 1 Average baseline ability and demographic characteristics of schools in the study

Variable	Treatment schools ($n=20$)	Comparison schools ($n=118$)
Baseline Math (in school-level standard deviations from state mean)	0.09	0.11
Baseline Reading (in school-level standard deviations from state mean)	0.18	0.17
Percent free/reduced lunch (%)	28	29
Percent White (%)	78	86
Percent Black (%)	9	8
Percent Hispanic (%)	10	4
Percent Asian (%)	3	2
Percent Native American (%)	<1	<1

Table 2 Mean growth for treatment and comparison schools

	Will-to-Reform score	Growth for treatment school	Mean growth comparison schools	Number of comparison schools
School 1	8	–0.03	–0.05	3
School 2	11	–0.52	0.27	10
School 3	7	–0.80	–0.23	10
School 4	19	0.54	–0.31	3
School 5	14	–0.63	0.03	10
School 6	15	0.11	0.11	10
School 7	16	1.12	–0.28	3
School 8	9	–0.82	0.27	3
School 9	8	–0.25	0.06	3
School 10	5	–0.51	0.46	3
School 11	8	–0.37	–0.20	10
School 12	16	0.05	–0.01	4
School 13	14	0.98	0.68	3
School 14	8	–1.09	0.18	4
School 15	12	0.32	0.37	3
School 16	11	–0.29	–0.20	10
School 17	13	–0.10	–0.09	10
School 18	13	0.12	–0.08	7
School 19	13	–0.18	–0.25	6
School 20	10	0.00	0.06	3
Mean		–0.12	0.04	

Statistical Model

Each school and all its Comparison schools were assigned to the same unique "school group." Further, all schools in each district plus all their Comparison schools were assigned to the same unique "district group." We used a growth model identical in form to Hierarchical Linear Models (HLMs) that track growth in individual achievement over time—but in our case, schools served as the "individuals" whose growth we were analyzing. Thus, the 4-level model measured observations, nested within schools, nested within school-groups, nested within district-groups. There were either six or seven observations per school, one for the mean math test score in the spring of each school-year from 1998 through 2004. (As noted above, for a few school groups the baseline year was 1999 instead of 1998.) We used an unstructured correlation matrix to model the six or seven observations within each school as being correlated with each other. We allowed the effects of year, of treatment, and of treatment-by-year to vary randomly between school-groups and between district-groups. (Because each school-group consisted of a Treatment school and all its matched Comparison schools and each district-group consisted of the Treatment schools within a district and all their matched Comparison schools, groups were defined by underlying similar characteristics that might lead to correlated results.) We treated State (Pennsylvania or New Jersey) as a fixed effect and allowed the fixed effect of "year" to vary between the two states. All statistical tests were run using SAS Proc Mixed. The "Satterthwaite" formula was used to estimate degrees of freedom.

Investigating the main effect of Treatment. To investigate the "main effects" of adopting an NSF-funded curriculum (either CMP or MiC) with GPSMP support, we used the model in Eq. 1:

$$\begin{aligned}\text{Math Test Score} &= \text{Baseline Score} + \beta_1 * \text{New Jersey} + \beta_2 * \text{Year} \\ &+ \beta_3 * \text{New Jersey} * \text{Year} + \beta_4 * \text{Treatment} + \beta_5 * \text{Treatment} * \text{Year} \\ &+ (\text{Error Terms})\end{aligned} \tag{1}$$

Definitions. Math Test Score: Mean score on the eighth-grade state math test (PSSA in Pennsylvania or GEPA in New Jersey) at a particular school in a particular year. For each year, these scores were standardized to a school-level *z*-score by subtracting the statewide average of school mean test scores and dividing by the statewide standard deviation of school mean test scores. This is analogous to what other large scale program evaluations have done (e.g., Garet et al., 2008) when they recentered student achievement data on each state's distribution by creating standard scores, except that we used schools, instead of students, as the unit-of-analysis.

Baseline Score: Model-estimated 1998 mean score for the Pennsylvania Comparison schools.

β_1: Difference between model-estimated 1998 mean score for Pennsylvania Comparison schools and model-estimated 1998 mean score for New Jersey Comparison schools.

β_2: Yearly growth rate in *z*-score for Pennsylvania Comparison schools. Because the dependent variable was a within-year *z*-score, this parameter would be significantly different from zero only if over time math test scores at the Comparison schools were systematically getting better or worse than test scores at other schools in Pennsylvania—an unlikely prospect.

β_3: Difference between the yearly growth rate in *z*-score for Pennsylvania Comparison schools and yearly growth rate in *z*-score for New Jersey schools. Like β_2, this parameter would ordinarily be near zero.

β_4: Model-estimated difference between Math Test Score at Treatment schools and Math Test Score at Comparison schools in the baseline year, 1998. Because each Treatment School was matched to Comparison schools using baseline test scores, by design this parameter was near zero.

β_5: This is the parameter of primary interest in the main-effects model. It is the difference in yearly achievement growth rate between Treatment and Comparison schools. A positive value would indicate that on average implementing *Mathematics in Context* or *Connected Mathematics* under the LSC model had a positive effect on achievement growth. A negative value would indicate that on average the program had a negative effect on achievement growth.

Error Terms: These were the error terms computed by the 4-level HLM. The following error terms were used: random differences among school groups in baseline score, yearly growth rate, baseline treatment effect, and treatment-by-growth interaction; random differences among district groups in baseline score, yearly growth rate, baseline treatment effect, and treatment-by-growth interaction; and seven correlated error terms for each year-within-school.

Investigating the effect of Will to Reform. We theorized that strong school-wide Will to Reform might catalyze the impact of NSF-funded middle school curricula, whereas low school-wide Will to Reform might interfere with the impact of the curricula. To test this theory, a valid and intuitively appealing approach would be to add for each Treatment school a recentered "Will-to-Reform" variable, i.e., the original Will-to-Reform score recentered around the middle value of 12 (halfway between 4 and 20).[2] For each comparison school, the recentered Will-to-Reform variable would be entered as zero. The new model would then add Will-to-Reform and Will-to-Reform*Year as fixed effects. The Will-to-Reform*Year slope would test whether the Will-to-Reform variable predicted how much achievement grew at Treatment schools, relative to achievement growth at other Treatment and Comparison schools.

This intuitively appealing approach had one potential drawback. Perhaps schools tended to have higher or lower Will to Reform because of some underlying background characteristic that was also associated with achievement growth. For example, perhaps baseline year achievement scores might predict Will to Reform and

[2] We recentered Will-to-Reform because otherwise the main effects for Treatment in Eq. 2 (reported in Table 4) would have been misleading. Table 4 would have reported Treatment effects at implementation schools where the Will-to-Reform was 0, a score below the minimum possible actual score of 4.

also predict achievement growth. In that case, Will to Reform might be correlated with achievement growth, but the correlation would be due to underlying school characteristics, not to the interaction between Will to Reform and the Treatment. One way to control for this possibility was to take advantage of each Treatment school's similarity to its matched comparison schools. In this model, each Treatment school's recentered Will-to-Reform score would be assigned both to the Treatment school and to its matched comparison schools. Then, four additional variables would be added to Eq. 1: Will-to-Reform, Will-to-Reform*Year, Will-to-Reform*Treatment, and Will-to-Reform*Treatment*Year. A positive slope for the Will-to-Reform*Treatment*Year interaction term would indicate that Treatment schools with high Will to Reform had a larger growth rate, relative to their matched Comparison schools, than did Treatment schools with low Will to Reform. A negative slope would indicate the opposite.[3]

Because both of these models were defensible, we ran each separately. Results of the two models were nearly identical. We report results from the second model, since that model theoretically did a better job controlling for possible spurious results. The model we used is described in Eq. 2.

$$\begin{aligned}\text{Math Test Score} &= \text{Baseline Score} + \beta_1 * \text{New Jersey} + \beta_2 * \text{Year} \\ &+ \beta_3 * \text{New Jersey} * \text{Year} + \beta_4 * \text{Treatment} + \beta_5 * \text{Treatment} * \text{Year} \\ &+ \beta_6 * \text{Will-to-Reform} + \beta_7 * \text{Will to Reform} * \text{Treatment} \\ &+ \beta_8 * \text{Will-to-Reform} * \text{Year} + \beta_9 * \text{Will-to-Reform} * \text{Treatment} \\ &* \text{Year} + (\text{Error Terms})\end{aligned} \tag{2}$$

Definitions of new parameters

β_6: The effect of Will to Reform on predicted mean 1998 scores for Comparison schools. It is possible that initially high-achieving Treatment schools might have systematically lower or higher Will to Reform than low-achieving Treatment schools.

[3] It might appear that, by assigning each Treatment school's recentered Will-to-Reform score to its matched comparison schools, we are claiming that Will-to-Reform is a meaningful construct for the comparison schools, and further that the Will-to-Reform happens to be exactly the same at the matched comparisons as at the Treatment school. That is not what we have done. Will-to-Reform is our (retrospective and imperfect) measure of school-level buy-in at the Treatment school to their reform math curriculum. The matched comparison schools did not implement a reform math curriculum, so Will-to-Reform is not a meaningful concept for them. Within our HLM, by assigning the same value of Will-to-Reform to all members of a school-group we have made Will-to-Reform a variable that applies to school-groups, not to individual schools within a school-group. Conceptually, the HLM first estimates the growth over time at each school by computing slope for Year within that school. Then, the HLM estimates how Treatment affects the growth rate in each school-group by computing the slope for Treatment*Year within that school-group. Finally, the HLM estimates how Will-to-Reform impacts Treatment effects by computing *across school groups* the slope of Will-to-Reform*Treatment*Year. To be imprecise but conceptually correct, the model is treating Treatment*Year as a dependent variable with school-group as unit of analysis, and Will-to-Reform as the independent variable. In this way, parameter β_9 in Eq. 2 estimates whether the effect of Treatment in school-groups where the Treatment school had a high Will-to-Reform is different from the effect of Treatment in school-groups where the Treatment school had a low Will-to-Reform.

If this were the case, then our matching procedures would ensure each school's matched comparison schools would have similarly high or low baseline math scores.

β_7: The effect of Will to Reform on the difference in baseline math scores between a Treatment school and its matched Comparison schools. Our matching procedures were designed to ensure that this parameter would be close to zero, since in theory each Treatment school would have nearly the same baseline scores as its Comparison schools. Including this term in the model corrected for any remaining noise caused by imperfect matching.

β_8: The effect of a school-group's Will-to-Reform score on predicted growth rate at its Comparison schools. Some demographic characteristics were associated with a higher achievement growth rate. For example, between 1998 and 2004 in Pennsylvania, low-SES middle schools improved their eighth-grade math test scores more than did high-SES middle schools. If demographic characteristics also predicted the Will to Reform of a Treatment school, then it is possible that school-groups whose Treatment school had high Will to Reform might have systematically higher (or lower) achievement growth rates than school-groups whose Treatment school had low Will to Reform.

β_9: This is the parameter of primary interest in the Will-to-Reform model. It measures the degree to which Will to Reform was associated with an increased or decreased difference in growth rate between a Treatment school and its matched Comparison schools. A positive slope would indicate that implementing *Mathematics in Context* or *Connected Mathematics* under the LSC model had a more positive effect in high Will-to-Reform schools than in low Will-to-Reform schools. A negative slope would indicate the opposite.

Results

Overall Treatment Effects

None of the parameters in Eq. 1 differed significantly from zero. Most importantly, the slope of Treatment*Year was not significantly different from zero ($t = -0.44$, $p = 0.6714$), with a mean treatment effect of only −0.012 school-level standard deviations per year. Thus, on average, mathematics achievement growth at the 20 treatment schools was not statistically different from math achievement growth at matched similar schools.

A rough 95 % confidence interval indicates that each year the Treatment schools' growth rate differed from that at Comparison schools between −0.064 school-level standard deviations per year and +0.041 school-level standard deviations per year. Over 6 years, this difference in growth rate would predict a 95 % confidence that the total effect of Treatment by 2004 would be in the confidence interval (−0.38, +0.25) school-level standard deviations, i.e., very near zero. Table 3 reports all fixed effects of the Main Effects model.

Table 3 SAS proc mixed solution for fixed effects, modeling treatment effects on yearly growth

Effect	Standard STATE	Estimate	Error	DF	*t* Value	Pr>|*t*|
Intercept		0.238	0.253	4.23	0.94	0.399
STATE	NJ	–0.241	0.515	8.81	–0.47	0.652
STATE	PA	0				
YEAR*STATE	NJ	0.004	0.021	36.5	0.20	0.843
YEAR*STATE	PA	–0.006	0.007	17.8	–0.76	0.455
YEAR		0				
TREAT		–0.030	0.030	108	–0.99	0.326
YEAR*TREAT		–0.012	0.026	9.95	–0.44	0.671

Buy-in Effects

To investigate the interaction between Will to Reform and the effects of the two NSF-funded curricula, we used a HLM that in essence compared the value-added of high Will-to-Reform Treatment schools to the value-added of low Will-to-Reform Treatment schools. That is, we compared the degree to which growth in eighth grade math scores from the baseline year (1998 or 1999) through 2004 of high versus low Will-to-Reform schools exceeded growth at their matched Comparison schools.

Figure 5 provides a simplified visual display of this analysis. The Composite Will-to-Reform scale was a sum of four scales scored from 1 to 5, so possible scores ran from 4 to 20. As the figure shows, GPSMP Treatment schools scored over a wide range of possible Will-to-Reform scores, from a minimum of 5 to a maximum of 19 on the composite scale. The figure displays the "Value Added" at each Treatment school—i.e., how much math achievement growth from the baseline through 2004 at the Treatment school exceeded growth at its matched Comparison schools—as a function of Will to Reform. Figure 5 clearly shows that students in the treatment schools with high values of Will to Reform had higher growth from the baseline to 2004 on state test scores than those students in the treatment schools with low values of Will to Reform. By the end of 6 years of treatment, some high Will-to-Reform schools showed an increase in state test scores of more than 1 school-level standard deviation (about 0.4 student-level standard deviations) in comparison to their matched schools while some low Will-to-Reform schools showed a decrease of more than 1 school-level standard deviation.

The actual HLM, described in Eq. 2, calculated each school's growth rate based on data from all available years, providing a more accurate and stable estimate than would have been possible using just the baseline and 2004 test scores used to create Fig. 5. The analysis showed a statistically significant slope for only one parameter: β_9, the Will-to-Reform*Treatment*Year interaction ($t=4.51$, $p<.0001$), with a mean effect size of 0.021 standard deviations per Will-to-Reform point per year (95 % confidence interval between 0.012 and 0.030). That is, the higher a Treatment

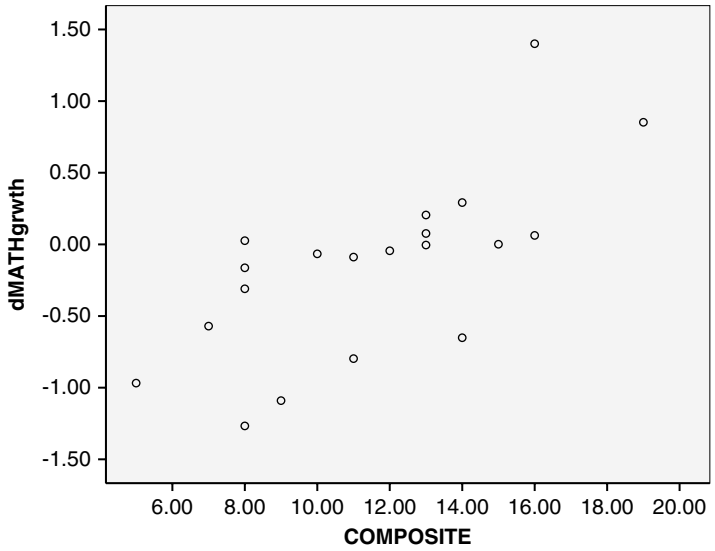

Fig. 5 Math achievement growth at treatment school from base year through 2004, minus math achievement growth at matched comparison schools (dMATHgrowth) as a function of composite Will-to-Reform score (COMPOSITE)

school's score on Will to Reform, the more achievement grew relative to that of matched Comparison schools. For detailed fixed effects from this analysis, see Table 4 in the appendix.

In practical terms, how much impact did the slope of 0.021 school-level standard deviations per Will-to-Reform point per year have on the relative effectiveness of Treatment schools? Our model's estimates for the three growth parameters in Eq. 2 were $\beta_5 = 0.002$, $\beta_8 = 0.000$, and $\beta_9 = 0.021$ where β_5 is the predicted growth rate difference between a Treatment school and its matched Comparison schools if the Treatment school had a middle value of 12 on the Will-to-Reform scale, β_8 is the (unsurprisingly zero) impact of a Treatment school's Will-to-Reform score on achievement growth at its matched Comparison schools, and β_9 is the impact of Will to Reform on achievement growth at the Treatment school. Thus, the predicted impact of the GPSMP Treatment at a school with the lowest observed Will-to-Reform score of 5 (7 less than 12) would be 0.002+ (−7*0.021)= −0.145 school-level standard deviations per year, or −0.87 school-level standard deviations over 6 years. That is, by 2004, a school that implemented a Reform curriculum but had the lowest Will to Reform would be expected to be performing about 0.87 school-level standard deviations *below* its matched comparison schools. Assuming a normal distribution, this would be enough to bring a school from the 50th percentile in math scores statewide in 1998 down to the 19th percentile in 2004. In contrast, the predicted impact of GPSMP Treatment at a school with the highest observed Will-to-Reform score of 19 (7 more than 12) was 0.002+(+7*0.021)=0.149 school-level

Table 4 SAS proc mixed solution for fixed effects, modeling composite Will-to-Reform effects on treatment-by-year slope

Effect	Standard STATE	Estimate	Error	DF	*t* Value	Pr > |*t*|
Intercept		0.214	0.293	4.55	0.73	0.500
STATE	NJ	–0.215	0.563	7.59	–0.38	0.712
STATE	PA	0				
YEAR*STATE	NJ	0.004	0.022	24.8	0.17	0.867
YEAR*STATE	PA	–0.004	0.009	4.12	–0.39	0.716
YEAR		0				
TREAT		–0.029	0.031	107	–0.94	0.348
YEAR*TREAT		0.002	0.016	125	0.10	0.920
WILL-TO-REF[a]		–0.044	0.054	17	–0.82	0.426
TREAT* WILL-TO-REF[a]		0.007	0.008	105	0.88	0.384
YEAR* WILL-TO-REF[a]		–0.0002	0.002	11.7	–0.09	0.930
YEAR*TREAT*WILL-TO-REF[a]		0.021	0.005	124	4.51	<.0001

[a]Recentered composite Will to Reform

standard deviations per year, or 0.89 school-level standard deviations over 6 years. That is, by 2004, a school that implemented a Reform curriculum but had the highest Will to Reform would be expected to be performing about 0.89 school-level standard deviations *above* its matched comparison schools. Assuming a normal distribution, this would be enough to bring a school from the 50th percentile in math scores statewide in 1998 up to the 81st percentile in 2004.

Movements of roughly this magnitude were in fact visible in the data set. For example, there were three middle schools in the data set (designated School 10, School 7, and School 4) that at the start of the GPSMP program in 1998 had the same PSSA score, at the 22nd percentile of all middle schools in Pennsylvania. School 10 (with the lowest observed Composite Will-to-Reform score of 5) moved from the 22nd percentile in 1998 down to the 18th percentile in 2004. In contrast, School 7 (tied for the second-highest observed Composite Will-to-Reform score of 16) moved from the 22nd percentile in 1998 up to the 69th percentile in 2004. School 4 (with the highest observed Composite Will-to-Reform score of 19) moved from the 22nd percentile in 1998 up to the 46th percentile in 2004.

Equation 2 did not control for reading achievement because both MiC and CMP incorporate extensive reading and thus might potentially improve eighth-grade reading as well as math scores. Nonetheless, we conducted a secondary analysis of Will-to-Reform effects on mathematics achievement while controlling for each school's reading score each year. After controlling for eighth-grade reading scores, the interaction between "Will to Reform" and the effects of GPSMP on mathematics achievement growth remained statistically significant ($t = 2.86$, $p < 0.005$). The point estimate for Will-to-Reform Effects was 0.011 school-level standard deviations per Will-to-Reform point per year (see Table 5). Thus, even after controlling for reading growth that might have been partly caused by the new

Table 5 SAS proc mixed solution for fixed effects, modeling composite Will-to-Reform effects on treatment-by-year slope, after controlling for school-level reading achievement

Effect	Standard STATE	Estimate	Error	DF	*t* Value	Pr>|*t*|
Intercept		0.073	0.157	4.86	0.47	0.659
YREAD[a]		0.610	0.024	842	25.74	<.0001
STATE	NJ	−0.031	0.297	7.48	−0.10	0.920
STATE	PA	0				
YEAR*STATE	NJ	−0.014	0.019	20.3	−0.76	0.454
YEAR*STATE	PA	0.001	0.008	5.11	0.13	0.905
YEAR		0				
TREAT		−0.050	0.035	109	−1.44	0.153
YEAR*TREAT		−0.00007	0.014	121	−0.00	0.996
COMPCTR[b]		−0.021	0.027	16.1	−0.78	0.447
YEAR*COMPCTR[b]		0.002136	0.002189	13.2	0.98	0.347
TREAT*COMPCTR[b]		0.006141	0.009673	105	0.63	0.527
YEAR*TREAT*COMPCTR[b]		0.01125	0.003929	121	2.86	0.005

[a]Current year mean eighth-grade reading score for the school
[b]Recentered composite Will to Reform

math program, math achievement at a school with the highest Will-to-Reform score would grow roughly 7*0.011 standard deviations faster than its comparison schools each year, or .462 standard deviations over 6 years—enough to bring a school from the 50th percentile in math scores statewide in 1998 up to the 68th percentile in 2004.

Effects of Will-to-Reform Subcomponents

While there was a significant interaction between composite Will To Reform and mathematics achievement growth, we were also interested in how each of the four Will-To-Reform subcomponents affected achievement growth. To that end, we ran four separate analyses and found that two of the components (Principal Support and Teacher Buy-in) were by themselves significant predictors of curriculum effectiveness. That is, when we replaced the Will-to-Reform variable in Eq. 2 with each of the individual subscale variables in turn, we could confirm the statistical significance of Principal-Support*Treatment*Year ($p=0.0007$) and Teacher-Buy-in*Treatment*Year ($p=0.0074$) (See Tables 6 and 7).

On their respective five-point scales, Principal-Support*Treatment*Year had a slope of 0.05 school-level standard deviations, and Teacher-Buy-in*Treatment*Year had a slope of 0.04 school-level standard deviations. Over 6 years, a school with principal buy-in of 5 would be expected to outperform a school with principal buy-in of 1 by $(5-1)*0.05*6=1.2$ school-level standard deviations. Over the same

Table 6 SAS proc mixed solution for fixed effects, modeling principal buy-in effects on treatment-by-year slope

Effect	Standard STATE	Estimate	Error	DF	*t* Value	Pr>|*t*|
Intercept		0.145	0.263	5.33	0.55	0.603
STATE	NJ	−0.093	0.521	9.58	−0.18	0.864
STATE	PA	0				
YEAR*STATE	NJ	0.002	0.021	31.9	0.07	0.943
YEAR*STATE	PA	−0.005	0.008	20.7	−0.63	0.534
YEAR		0				
TREAT		−0.026	0.032	107	−0.80	0.424
YEAR*TREAT		0.005	0.021	8.29	0.25	0.809
PRINCC[a]		−0.190	0.122	13.3	−1.55	0.145
TREAT*PRINCC[a]		0.018	0.027	109	0.67	0.504
YEAR*PRINCC[a]		−0.001	0.007	27.1	−0.13	0.901
YEAR*TREAT*PRINCC[a]		0.052	0.015	76.5	3.54	0.0007

[a]Zero-centered Principal buy-in

Table 7 SAS proc mixed solution for fixed effects, modeling teacher buy-in effects on treatment-by-year slope

Effect	Standard STATE	Estimate	Error	DF	*t* Value	Pr>|*t*|
Intercept		0.163	0.259	4.79	0.63	0.558
STATE	NJ	−0.067	0.527	9.49	−0.13	0.902
STATE	PA	0				
YEAR*STATE	NJ	−0.007	0.020	136	−0.32	0.747
YEAR*STATE	PA	0.0004	0.007	120	0.06	0.956
YEAR		0				
TREAT		−0.030	0.031	108	−0.94	0.351
YEAR*TREAT		−0.003	0.018	8.21	−0.16	0.876
TCHRBUYINC[a]		−0.145	0.109	14.2	−1.34	0.203
TREAT*TCHRBUYINC[a]		0.004	0.022	108	0.19	0.850
YEAR*TCHRBUYINC[a]		0.009	0.0057	121	1.84	0.068
YEAR*TREAT*TCHRBUYINC[a]		0.036	0.012	24.9	2.92	0.0074

[a]Zero-centered aggregate teacher buy-in

period, a school with teacher buy-in of 5 would be expected to outperform a school with teacher buy-in of 1 by $(5-1)*0.04*6=0.96$ school-level standard deviations.

It is important to note that the two school-level components of Will to Reform were not completely independent constructs. In fact, Principal Support and Teacher Buy-in were significantly correlated with each other ($r=0.686$, $p<0.01$). None of the other correlations among the four components of Will to Reform were statistically significant (See Table 8).

Table 8 Correlations between components of Will-to-Reform scale

	Teacher buy in	District coherence	Superintendent support
Principal support	0.686[a]	0.291	0.093
Teacher buy-in	0.286	0.094	
District coherence	0.329		

[a]Correlation is significant at the 0.01 level (2-tailed)

Neither of the two district-level variables was, by itself, a significant predictor for mathematics achievement growth (for Superintendent-Support*Treatment*Year, $p = 0.3184$, and for District-Coherence *Treatment*Year, $p = 0.0791$). The lack of statistical significance for district-level Will-to-Reform subcomponents may be an artifact of the small number of Treatment districts in the sample (only 9 district-groups, vs. 20 school-groups). Nonetheless, we cannot at this time confirm the independent importance of district-level Will-to-Reform subcomponents on the effectiveness of NSF-funded middle school mathematics curricula.

Discussion

In their comprehensive review of experimental and quasi-experimental studies that investigated the outcomes of mathematics programs for middle and high schools, Slavin et al. (2008) found a "lack of evidence that it matters very much which textbook schools choose (p. 42)." In particular, they reported a mean effect size of 0.00 standard deviations for 24 studies of NSF-funded curricula. At first blush, our findings appear to support the contention that choice of textbook doesn't matter. In our quasi-experimental study of 20 middle schools that adopted an NSF-funded math curriculum, the main effect was a statistically non-significant negative 0.012 school-level standard deviations per year.

However, when we added to our model Will to Reform, a measure of school-level buy-in to the new curriculum, we found that choice of textbook appears to have mattered very much indeed. Middle schools with very high scores on the Will-to-Reform scale saw dramatic improvements in mathematics achievement after adopting *Connected Mathematics* or *Mathematics in Context* with professional development support provided by the GPSMP. Middle schools with very low scores on the Will-to-Reform scale saw just as dramatic drops in mathematics achievement after adopting one of the new curricula—even though they too received significant professional development support from the GPSMP.

Our study also confirmed the importance of both the Teacher Buy-in and the Principal Support components of Will to Reform. The district-level components of Will to Reform—Superintendent Support and District Coherence—could not be confirmed as being independently important. It should be noted, however, that our sample consisted of only nine districts. A better test of district-level components would require a larger study incorporating a larger number of districts.

When considered in light of a co-construction view of program implementation (see Fig. 3), our results are consistent with a second finding reported by Slavin et al. (2008): reforms to instructional process strategies can have a strong positive effect on mathematics achievement. In our view, the implemented curriculum is a result not of the curriculum materials alone, but of an interaction between the teacher and the curriculum materials, as mediated by such factors as school context and teacher buy-in. That is, instructional processes, which actually affect learning, can only be predicted when curriculum materials and teachers' reactions to them are considered together.

This study is only a first step towards using the evaluation model displayed in Fig. 3 to study the effects of curriculum materials. Our study was limited by the retrospective nature of the data available. Will to Reform and its subscales were less than ideal measures of school-level buy-in. They were subject to potential limitations such as observer bias. Further, because we developed ratings by consensus, we did not have any measures of construct reliability. Moreover, our study did not have any teacher-level measure of buy-in, which might have been a more accurate predictor of program implementation than the school-level measures we used. Neither did we have available any direct measures of fidelity to implementation structure or fidelity to implementation process. To confirm the evaluation model and gain a deeper understanding of the interaction between buy-in, implementation fidelity, and student outcomes, future studies will need to correct these problems. Ideally, such studies would also include qualitative data documenting whether curriculum materials actually undergo lethal mutations in classrooms with low buy-in and productive adaptations in classrooms with high buy-in.

Future work to develop better measures of buy-in will need to consider trade-offs between the detail needed to obtain valid measures and the expense of collecting data. Would a Likert-type questionnaire for teachers, similar to that used by VanDerHeyden et al. (2012), have produced similar results to ours? Could a yes/no question about wanting to use the curriculum again, similar to that used by Agodini et al. (2010), have been sufficient?

In addition to replicating our findings using better measures of buy-in combined with measures of other variables in our model, it is also important to investigate whether our findings are applicable to other settings. Would a similar process occur in high school? In elementary school? Does buy-in predict results for subject matter other than mathematics? Compared to the reforms we implemented, many reforms (e.g., Saxon mathematics and Success for All language arts) are much more scripted. For such curricula, would strong buy-in lead to productive adaptations and positive results? Would weak buy-in lead to lethal mutations and negative results?

The retrospective nature of our study, in addition to limiting what independent variables we could study, limited us in several other ways. Only school-level, not student-level, data were available. A more detailed data set using student instead of school as unit-of-analysis would provide more precise estimates of program effects and would make it possible to investigate differential impacts on differing subgroups of students. Also, we had available only one measure of curriculum effect, the high-stakes eighth-grade mathematics tests administered by the local state

(Pennsylvania or New Jersey). A more diverse set of dependent variables would have been desirable. Past research has found that both *Math in Context* and *Connected Mathematics* tend to have a more positive impact on measures like the Balanced Assessment of Mathematics that are explicitly designed to test student problem-solving skill (Kilpatrick, 2003; Romberg et al., 2005; Tarr et al., 2008). Additionally, this was a quasi-experiment. While quasi-experiments can provide important and valid findings—especially when, as we did, they use large data bases and careful matching techniques—randomized control trials are less prone to error and provide more certain results.

Nonetheless, our results have potentially important implications for current and future implementations of instructional materials such as those designed to implement the newer *Common Core State Standards* or the *Next Generation Science Standards*. Researchers evaluating new instructional materials should strive to test the full model displayed in Fig. 3, including measures of principal buy-in, of school-wide teacher buy-in, and of individual teacher buy-in, as well as measures of structural fidelity and of process fidelity to the implementation. Further, quantitative data should be supplemented with qualitative data reporting how curriculum materials are adapted when those materials are actually used in the classroom.

Implementers of new instructional materials would be wise to attend to the role of principals and teachers as co-constructors of the planned and implemented curriculum—either by selecting materials that are a good match for local staff, or else by working closely with staff to ensure buy-in and minds-on implementation. Results of the current study support the hypothesis that doing so might encourage productive adaptations that improve student learning, while failing to do so might encourage lethal mutations that retard student learning.

References

Agodini, R., Harris, B., Thomas, M., Murphy, R., Gallagher, L., & Pendleton, A. (2010). *Achievement effects of four elementary school math curricula: Findings for first and second graders.* Washington, DC: Department of Education. Retrieved Jan 12, 2013, from http://ies.ed.gov/ncee/pubs/20094052/index.asp

Banilower, E. R., Boyd, S. E., Pasley, J. K., & Weiss, I. R. (2006). *Lessons from a decade of mathematics and science reform: A capstone report for the local systemic change through teacher enhancement initiative*. Chapel Hill, NC: Horizon Research, Inc.

Brown, A. L., & Campione, J. C. (1996). Psychological theory and the design of innovative learning environments: On procedures, principles, and systems. In R. Glaser (Ed.), *Innovations in learning: New environments for education* (pp. 289–325). Mahwah, NJ: Erlbaum.

Brown, M. W., & Edelson, D. C. (2001, April). *Teaching by design: Curriculum design as a lens on instructional practice.* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Cai, J. (2003). What research tells us about teaching mathematics through problem solving. In F. Lester (Ed.), *Research and issues in teaching mathematics through problem solving* (pp. 241–254). Reston, VA: National Council of Teachers of Mathematics.

Cai, J. (2010). Evaluation of mathematics education programs. *International Encyclopedia of Education, 3*, 653–659.

Cai, J., Nie, B., & Moyer, J. C. (2010). The teaching of equation solving: Approaches in *Standards-based* and traditional curricula in the United States. *Pedagogies: An International Journal, 5*(3), 170–186.

Cai, J., Wang, N., Moyer, J. C., Wang, C., & Nie, B. (2011). Longitudinal investigation of the curriculum effect: An analysis of student learning outcomes from the LieCal Project. *International Journal of Educational Research, 50*(2), 117–136.

Cho, J. (1998, April). *Rethinking curriculum implementation: Paradigms, models, and teachers' work*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Confrey, J., Castillo-Chavez, C., Grouws, D., Mahoney, C., Saari, D., Schmidt, W., et al. (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, DC: National Academies Press.

Council of Chief State School Officers and National Governors Association. (2010). *Common core state standards for mathematics*. Washington, DC: Council of Chief State School Officers and National Governors Association.

Council of Chief State School Officers, Brookhill Foundation, & Texas Instruments. (2011). *Common Core State Standards (CCSS) mathematics curriculum materials analysis project.* Washington, DC: Authors. Retrieved Jan 23, 2013, from https://www.k12.wa.us/CoreStandards/pubdocs/CCSSOMathAnalysisProj.pdf

Cross, C. T. (2004). *Putting the pieces together: Lessons from comprehensive school reform research*. Washington, DC: National Clearinghouse for Comprehensive School Reform.

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8.

Dobson, L. D., & Cook, T. J. (1980). Avoiding type III error in program evaluation: Results from a field experiment. *Evaluation and Program Planning, 3*, 269–276.

Flay, B., Biglan, A., Boruch, R., Castro, F., Gottfredson, D., Kellam, S., et al. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science, 6*(3), 151–175.

Forgatch, M. S., Patterson, G. R., & DeGarmo, D. S. (2005). Evaluating fidelity: Predictive validity for a measure of competent adherence to the Oregon model of parent management training. *Behavior Therapy, 36*, 3–13.

Fullan, M., & Pomfret, A. (1977). Research on curriculum and instruction implementation. *Review of Educational Research, 47*, 335–397.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., et al. (2008). *The impact of two professional development interventions on early reading instruction and achievement (NCEE 2008-4030)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Glennan, T. K., Bodilly, S. J., Galegher, J. R., & Kerr, K. A. (2004). *Expanding the reach of education reform: Perspectives from leaders in the scale-up of educational interventions.* Santa Monica, CA: Rand. Retrieved Jan 12, 2013, from http://www.rand.org/pubs/monographs/MG248.html

Goodlad, J. I. (1983). A study of schooling: Some implications for school improvement. *Phi Delta Kappan, 64*(8), 552–558.

Hiebert, J. (1999). Relationships between research and the NCTM Standards. *Journal for Research in Mathematics Education, 30*, 3–19.

Hohmann, A. A., & Shear, M. K. (2002). Community-based intervention research: Coping with the "noise" of real life in study design. *American Journal of Psychiatry, 159*, 201–207.

Kennedy, M. M. (2004). Reform ideals and teachers' practical intentions. *Education Policy Analysis Archives, 12*(13). Retrieved Jan 2, 2013, from http://epaa.asu.edu/epaaa/v12n13/

Kilpatrick, J. (2003). What works? In S. L. Senk & D. R. Thompson (Eds.), *NSF funded school mathematics curricula: What they are? What do students learn?* (pp. 471–488). Mahwah, NJ: Erlbaum.

Krainer, K., & Peter-Koop, A. (2003). The role of the principal in mathematics teacher development. In A. Peter-Koop et al. (Eds.), *Collaboration in teacher education* (pp. 169–190). Dordrecht: Kluwer Academic.

Little, J. W. (1993). Teachers' professional development in a climate of educational reform. *Educational Evaluation and Policy Analysis, 15*, 129–151.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

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

National Council of Teachers of Mathematics. (2009a). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*. Reston, VA: The Council.

National Council of Teachers of Mathematics. (2009b). *Focus in high school mathematics: fostering reasoning and sense making for all students*. Reston, VA: The Council.

National Science Teachers Association. (2012). *Recommendations on next generation science standards first public draft*. Arlington, VA: NSTA. Retrieved Feb 6, 2013, from http://www.nsta.org/about/standardsupdate/recommendations2.aspx

New Jersey Department of Education. (1996). *New Jersey core curriculum content standards for mathematics*. Trenton, NJ: Author. Retrieved Aug 7, 2007, from http://www.edsolution.org/CustomizedProducts/data/standards-frameworks/standards/09mathintro.html

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in KI-12 curriculum intervention research. *Review of Educational Research, 78*(1), 33–84.

O'Donnell, C. L. & Lynch, S. J. (2008, March). *Fidelity of implementation to instructional strategies as a moderator of science curriculum unit effectiveness*. Paper presented at the annual meeting of the American Educational Research Association, New York. Retrieved Jan 12, 2013, from http://www.gwu.edu/~scale-up/documents/AERA%20O%27Donnell%20Lynch%202008%20-%20Fidelity%20of%20Implementation%20as%20a%20Moderator.pdf

Pennsylvania Department of Education. (1999). *Academic standards for mathematics.* Harrisburg. PA: Author. Retrieved Aug 7, 2007, from http://www.pde.state.pa.us/k12/lib/k12/MathStan.doc

Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research, 75*(21), 1–246.

Riordan, J. E., & Noyce, P. E. (2001). The impact of two standards-based mathematics curricula on student achievement in Massachusetts. *Journal for Research in Mathematics Education, 32*(4), 368–398.

Romberg, T. A., Folgert, L., & Shafer, M. C. (2005).*Differences in student performances for three treatment groups.* (Mathematics in context longitudinal/cross-sectional study monograph 7). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research. Retrieved May 28, 2014, from http://micimpact.wceruw.org/working_papers/Monograph%207%20Final.pdf

Rubin, D. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology, 2*, 169–188.

Schwartzbeck, T. D. (2002). *Choosing a model and types of models: How to find what works for your school*. Washington, DC: National Clearinghouse for Comprehensive School Reform.

Slavin, R. E. (2002). Evidence-based educational policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15–21.

Slavin, R. E., Lake, C. & Groff, C. (2008). *Effective programs in middle and high school mathematics: A best-evidence synthesis*. Baltimore, MD: Johns Hopkins University Center for Data Driven Reform in Education (CDDRE) Best Evidence Encyclopedia. Retrieved May 29, 2013, from http://www.bestevidence.org/word/mhs_math_Sep_8_2008.pdf

Stuart, E. A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher, 36*, 187–198.

Tarr, J. E., Reys, R., Reys, B., Chávez, Ó., Shih, J., & Osterlind, S. (2008). The impact of middle school mathematics curricula on student achievement and the classroom learning environment. *Journal for Research in Mathematics Education, 39*(3), 247–280.

Turnbull, B. (2002). Teacher participation and buy-in: Implications for school reform initiatives. *Learning Environments Research, 5*(3), 235–252.

U.S. Department of Education. (2007a). *What works clearinghouse intervention report: connected mathematics project.* Washington, DC: Author. Retrieved Nov 1, 2007, from http://ies.ed.gov/ncee/wwc/pdf/WWC_CMP_040907.pdf

U.S. Department of Education. (2007b). *Mathematics and science specific 84.305A RFA*. Washington, DC: Author. Retrieved Dec 2, 2007, from http://ies.ed.gov/ncer/funding/math_science/index.asp

VanDerHeyden, A., McLaughlin, T., Algina, J., & Snyder, P. (2012). Randomized evaluation of a supplemental grade-wide mathematics intervention. *American Educational Research Journal, 49*(6), 1251–1284.

Vuchinich, S., Flay, B. R., Aber, L., & Bickman, L. (2012). Person mobility in the design and analysis of cluster-randomized cohort prevention trials. *Prevention Science, 13*(3), 300–313.