4-2013

The Linking Study: An Experiment to Strengthen Teachers' Engagement With Data on Teaching and Learning

Jonathan A. Supovitz
University of Pennsylvania, JONS@GSE.UPENN.EDU

The Linking Study: An Experiment to Strengthen Teachers' Engagement With Data on Teaching and Learning

Abstract
In this AERA 2013 paper, Dr. Jonathan Supovitz investigates what it means for *teachers* to fruitfully use data *to enhance the teaching and learning process*. Informed by research on the challenges teachers face to use data meaningfully, and clues from the rich literature on formative assessment, this paper reports on the design and effects of an intervention designed to help teachers *connect* data on their teaching with data on the learning of their students for the purpose of informing subsequent instruction which leads to better student outcomes. The hypothesis of this study, therefore, is that while examining data may be useful, the real value of data use is to examine the connection between data points – in this case the instructional choices that teachers make and the learning outcomes of students. Thus, 'data use' in this study means encouraging and facilitating teachers' analytical experiences of linking data on teaching to data on the learning of their students.

Disciplines
Educational Assessment, Evaluation, and Research | Teacher Education and Professional Development

Comments
View on the CPRE website.

The Linking Study:
An Experiment to Strengthen
Teachers' Engagement with Data on Teaching and Learning

Jonathan Supovitz
Consortium for Policy Research in Education
Graduate School of Education
University of Pennsylvania

Paper presented at the American Education Research Association Conference
San Francisco, CA.

April, 2013

INTRODUCTION

The allure of using data to improve performance is a source of tremendous activity in the education field today. "Data use" has spurred a wide variety of reforms at all different levels of the education system, ranging from infrastructure augmentation to state databases, to district dashboard systems that collect and display an array of indicators, to the formation of school data teams that conduct data-informed inquiries into subgroups of students, to specific formative assessment classroom techniques. From this cornucopia it is increasingly apparent that data use means different things to decision-makers at different levels of the education system, and that the type of data, frequency of the data, mode of inquiry, and decision-making processes look quite different from one another according to role, situation, and purpose (Supovitz & Klein, 2003; others). Thus, when we talk about the term 'data use', we must hone in on "for whom?" and "for what purpose?"

In this paper I am interested in what it means for ***teachers*** to fruitfully use data ***to enhance the teaching and learning process***. Informed by research on the challenges teachers face to use data meaningfully, and clues from the rich literature on formative assessment, this paper reports on the design and effects of an intervention designed to help teachers ***connect*** data on their teaching with data on the learning of their students for the purpose of informing subsequent instruction which leads to better student outcomes. The hypothesis of this study, therefore, is that while examining data may be useful, the real value of data use is to examine the connection between data points – in this case the instructional choices that teachers make and the learning outcomes of students. Thus, 'data use' in this study means encouraging and facilitating teachers' analytical experiences of linking data on teaching to data on the learning of their students.

Using a randomized control trial, the Linking Study tests the impacts of the intervention on teachers' perceptions of their fluency with data and their self-reported learning about their instructional practices and their students' thinking. Moreover, the study estimates effects on instruction caused by the intervention, based upon external trained raters' judgments of the quality of instructional practice. Finally, this research examines impacts of the intervention on student outcomes.

Overall, as a result of the linking intervention, we found substantive impacts on participating teachers' reports of learning about their instruction and gaining insights into the thinking of their students. Furthermore, there were statistically significant and educationally meaningful effects on external judgments of the quality of instruction associated with the intervention. Finally, there were small but statistically significant effects on student performance on end-of-unit assessments associated with the intervention. We found no impacts on teachers' perceptions of their data fluency from their experience.

The Linking Study was generously funded by the Spencer Foundation of Chicago, Illinois to explore teachers' use of data to inform and improve the teaching and learning process. This paper focuses on the experimental impacts of the study. Other Spencer funded papers include an examination of the role and moves of facilitators in guiding teachers' conversations in PLCs (Ebby & Oettinger, 2013), and several micro studies of the teacher learning process (Christman and Edmonds, forthcoming; Supovitz & Ebby, in imagination).

INFERENCES FROM THE LITERATURE

Over the past decade, a number of classroom-based data interventions have focused on providing both student test data and analytic schemas to teachers. Most of these data use approaches have focused on the organization of student test data by standards, learning objectives, etc. (Refs). While these data certainly provide teachers with some information about their students' levels of proficiency at the time, they are problematic for at least two reasons. First, they lack insight into how students misunderstand and therefore provide little guidance for subsequent actions (Supovitz, 2012). Second, they are solely lagging indicators because they ask teachers, absent of data, to infer back to what they did that produced these results (Supovitz, Foley & Mishook, 2012). Thus, rather than linking action to result, they focus only on result.

There is relatively little research, however, that explores the ways that teachers make sense of data and the ways they incorporate them into their practice. In one noteworthy study, Goertz, Nabors Olah and Riggan (2009) examined how a sample of 45 teachers of mathematics in nine elementary schools in two school districts used data from interim and classroom assessments. The researchers conducted three investigations to explore the quality of information contained in the assessments, teachers' ability to analyze the assessments, and the relationship between teacher capacity and their formative assessment practices. In one aspect of their study they presented teachers with student responses and common student misconceptions and asked the teachers to explain what they saw in order to understand teachers' interpretations of student errors. They found that teachers analyzed the assessment data in two ways. First, teachers *located* errors by examining whether or not students answered questions correctly. Second, mostly only after prompting, teachers *diagnosed* errors by focusing on why students answered questions incorrectly. Diagnoses ranged from procedural to conceptual explanations, with procedural explanations predominant.

In another part of their study, the researchers used classroom observation and teacher interview data to create teacher profiles to understand how a variety of assessments influenced instruction. From these analyses the researchers found that that the information that teachers gleaned from their assessment

data resulted in mostly what they called “organizational change strategies” (ie reteaching, identifying students for additional support, regrouping, when to move on to the next topic or concept). What the researchers called “instructional improvement strategies,” or occurrences in which teachers identified ways to adjust their teaching based on the assessment data, were much rarer. Interestingly, those teachers who gave conceptual interpretations of student errors were more likely to adjust the ways they taught. This study points to an important insight about using data to inform instructional practice. It suggests that teachers have trouble *getting underneath the numbers*, and understanding why students are responding the way that they are. Lacking more sophisticated diagnosis, teachers’ responses were largely organizational rather than more instructionally responsive.

Other studies have reiterated the challenges posed by the last finding of the Goertz, Nabors Olah and Riggan study, namely that one of the biggest challenges to teachers is the “Now what?” question of what actions to take as a response to information about student understanding that they have gained from the assessment. Heritage, Kim, Vendlinski & Herman (2009) conducted a study of what a sample of 118 sixth grade teachers would teach next based upon their interpretation of students responses to mathematics items that assessed the principle of the distributive property in algebra. Using a group of university mathematics experts and expert teachers, they rated teacher interpretation of student responses on a rubric that ranged from no explanation of the relevant concept to a procedural explanation of the concept, to a more sophisticated conceptual understanding of the concept. They found most teacher responses were empty or procedural. They also found that adjusting subsequent instruction based upon assessment information tended to be the most difficult task for teachers with subsequent choices narrowed by prior interpretation. This study suggests that teacher success in analyzing student understanding is an important precursor to subsequent instructional response.

Another fertile source of research that informs how teachers might use data to inform the improvement of teaching and learning comes from the formative assessment literature. The core theory of formative assessment based upon the theory of how instructors gain access to the current state of understanding of learners and move them towards a goal. According to Sadler (1989), “Formative assessment is concerned with how judgments about the quality of student responses (performances, pieces, or works) can be used to shape and improve the student's competence” (p.120). That is, an assessment becomes “formative” when its information is activated as feedback to the learner in order to reduce the distance between her present state of understanding and the desired state. A key element of teachers’ potential to use data is the extent to which they can gain insight into current student understanding to move them towards greater understanding. Thus, a key aspect of formative assessment is repeated efforts to connect action to improvements in performance to eventually reach a goal or level of mastery.

Several strands of the research related to formative assessment are relevant to the purpose of the Linking Study. First is attention to what kinds of data to examine. Several researchers have looked at the effects of different representations of past performance on subsequent performance. In educational research, this is most often represented by studies of the effects of grades on learners. For example, Butler (1988) compared the effects of grades only, grades and comments, and comments only on subsequent student performance. He found that both groups viewing grades declined in performance over time relative to the group with comments only. This study suggests that even when provided with

comments, grades get in the way of the processing associated with learning. Similarly, Schunk & Swartz (1993) compared providing feedback to 5th graders in writing and showed improved performance associated with process feedback as opposed to end product assessments. An in-depth qualitative study by Black, Harrison, Lee, Marshall & Wiliam (2007) reported that teachers substantially expanded their students' understanding by focusing on written feedback rather than grading student work.

Another source that informs our understanding of effective classroom data use is theory and research on inquiry cycles. Theory in this area ranges from the Plan-Do-Study-Act cycles of continuous improvement advocated by Edwards Deming (1986) to the cycle of question-investigation-action-evaluation of Smith & Ruff (1988) to the cycle of data examination advocated by Boudett, City, & Murnane (2005) in Data Wise. Much of this research stresses the iterative nature of data-informed inquiries and suggests that repeated cycles both reveal patterns from the data more readily and codify both the process and learning into practice.

I take several things from this short literature review. First, *regular feedback enhances learning*. Learners (whether they be teachers or students) need regular and repeated opportunities to examine their practice and apply these lessons to subsequent practice. Thus an experience that will impact practice must occur repeatedly in cycles, rather than as one experience. Second, *the form of the data, which are the source of feedback, are important*. They should be rich and nuanced, (ie qualitative or mixed data are better than numbers alone). Third, *data should seek to connect actions to outcomes*, rather than provide information on outcomes alone. Examinations of outcomes alone, ie lagging indicators, leave much room for speculation about what produced those outcomes, but linking data to actions (ie leading indicators) and exploring how they contribute to outcomes provides for a richer data experience.

STUDY BACKGROUND

Informed by this understanding of the literature, CPRE partnered with a school district to design an intervention and develop the Linking Study, a randomized experiment to test the hypothesis that timely feedback to teachers about their instruction, examined in conjunction with data on the performance of their students, can positively influence subsequent teaching and learning. More specifically, the research was designed to address the following four questions:

What is the impact of providing teachers of mathematics with feedback on both their teaching and their students' learning, in comparison to the usual condition of feedback on learning alone, on:

1. Teachers' views about the importance of teaching and learning data and their self-reported proficiency to use such data in their mathematics instructional practice;

2. Teachers' perceptions about their learning about mathematics instruction and their students thinking about mathematics;

3. Teachers' subsequent instructional practices in mathematics;

4. The subsequent mathematics learning of students.

To address these research questions, CPRE researchers worked with a moderate sized school district to conduct the Linking Study with teachers in grades 1-5 in mathematics. In this paper I describe the district context; the intervention that was co-constructed with the district to provide feedback to teachers under experimental conditions to test the research hypotheses; the process we used to recruit teachers to participate in the project; the data we collected to address the research questions; and the results of the experiment on teachers' perceptions, their practices, and the learning of their students. The paper concludes with a short discussion of the importance of the findings.

DISTRICT CONTEXT

The study was conducted in a mid-sized suburban district in southern New Jersey. The district has 20 schools, including 12 elementary schools, and serves approximately 12,000 students. Teachers in 10 of the 12 district elementary schools agreed to participate in the research. The research team and the district had a history of working together and collaboratively designed the intervention to fit into the district's efforts to encourage teachers to examine student data in professional learning communities (PLC). In the 2009-10 and 2010-11 school years, the district had invested in PLC training, providing teachers in the district with multiple day training on the DuFour model of professional learning communities (DuFour, Eaker & DuFour, 2008), delivered by Solution Tree In the 2009-10 school year a team of U. Penn graduate students working with CPRE observed a sample of PLCs in the district to understand how they used data to inform discussions of teaching and learning (Supovitz & Merrill, 2010).

The intervention was constructed in collaboration with the district's chief academic officer and mathematics supervisor. The district was using a combination of the *Investigations* curriculum, a reform-oriented mathematics curriculum, and the Scott Foresman mathematics book, which conveys mathematics more traditionally. The district also provided teachers at each grade level with common time each week to hold professional learning community (PLC) meetings. These PLC meetings were about 45 minutes each (the length of a class period), and focused on different subjects and topics each meeting. In some schools, the PLCs were facilitated by coaches or lead teachers. PLCs were expected to use their time to discuss curriculum, examine student work, develop assessments, and discuss students.

INTERVENTION DESIGN

The linking intervention consisted of providing a random sample of teachers with written feedback on an observed lesson of their teaching followed up by a facilitated discussion of both their teaching and their students' learning on that unit's end of unit assessment. The facilitated discussion occurred during the grade level PLC meeting that occurred shortly after the unit was completed.

The intervention occurred in three cycles across three different mathematics units during the 2011-12 school year. There were 8-11 units across the school year, depending on the grade, so that the intervention covered approximately a third of the school year. The units at each grade level were chosen by the district and research team to both focus on *Investigations* units and emphasize mathematics concepts that were revisited across the school year at that grade level (ie addition and subtraction in grades 1 and 2; number operations in grade 3; and multiplication and division in grades 4 and 5). This

was done to maximize the opportunity for the feedback in one lesson cycle to be used in a subsequent cycle.

Each intervention cycle followed a similar pattern. First, participating teachers (both treatment and control) within grade level teams were asked to identify a common lesson during the relevant unit to be observed. A common lesson was chosen to facilitate future conversation about the lesson. Observations took the form of videotaping the lesson, done by a substitute teacher from the district who was trained by the project as a videographer. Using substitute teachers had two advantages. First, they were familiar to both adults and children in the schools and therefore were minimally disruptive. Second, they had already gone through the background check required for adults to be in the school.

All videotaped lessons were reviewed by experienced mathematics teachers (either graduate students at the University of Pennsylvania or CPRE research team members), who were trained in identifying aspects of mathematics instruction based on the Instructional Quality Assessment (IQA), an established mathematics lesson observation tool. Written feedback to teachers and scoring the quality of the lesson on IQA rubrics focused on two dimensions of mathematics instruction: (1) the academic rigor of the lesson and (2) the accountable talk in the lesson, or teacher questioning and subsequent student-teacher interactions.

Treatment teachers received feedback from their lesson in two stages. First, they received private, emailed feedback within one week of the observed lesson provided by the trained observer. The feedback was written in prose, rather than providing numerical ratings, and was written to balance both positive things about the lesson and areas for improvement; in accordance with the literature on effective performance feedback (Kluger & DeNisi, 1996). The feedback focused on the academic rigor of the lesson and the teachers' interactions with students (accountable talk). The feedback was written up and sent privately via email to the teacher within one week of the observed lesson, regardless of where this took place in the timeline of the unit, so that the teacher received feedback on the lesson as close to the lesson itself as possible. We also wrote feedback for each lesson of the control group teachers and provided it to them at the end of the study.

The second component of feedback on the observed lesson came during a subsequent PLC meeting. At the end of the mathematics unit, the teachers met in their professional learning community and followed a structured routine that was facilitated by a trained facilitator. Each teacher brought with them their students' end of unit tests. In advance of the meeting, the facilitator chose 1-2 test items that (a) were central to the focus of the unit and (b) asked students to show their work, not just provide the answer.

The treatment group's 45 minute PLC meetings had two components. The first component, designed to take about 15 minutes, was to examine student test performance. Teachers were asked to group their student work by the strategy that students used to solve the problem, rather than by the correct answer. This allowed teachers to focus on students' solution strategies, which emphasizes how students are thinking about solving mathematical problems and the efficiency of their solution strategies, rather

than their ability to produce the correct answer. Facilitators were asked to facilitate the conversation using the following guiding questions:

1. What different strategies do you see in your students' work on the focus item?
2. In what way do these problem solving strategies give you insight into how the students understand the big idea(s) of the unit?
3. Using this understanding, how can you support students to move toward more sophisticated strategies?

The second component of the PLC meeting was used to examine 1-2 selected video clips of a teacher's interaction with students to discuss examples of student-teacher interactions, or accountable talk.[1] This component of the PLC session was intended to take about 25 minutes and facilitators were asked to facilitate this component of the conversation using these guiding questions:

1) How did the interaction begin?
2) What did the student(s) response(s) reveal about their understanding of the math?
3) What was the teachers' follow up?
4) Were there any missed opportunities here? How could you have changed this interaction to learn even more?

At the end of this instructional conversation, participants were asked to spend the final minutes making connections between the instruction in the unit, as exemplified in the instructional discussion, and end-of-unit student test performance.

Teachers in the control group were provided with a structured guide on how to examine student test data. The guide, which was developed in the pilot year of the study (2010-2011) was given all teachers in the district and used as a model for how to examine student work within PLCs. While we wrote up feedback on the three lessons for the treatment teachers, we did not return this feedback to them until after data collection was completed.

RECRUITMENT OF PARTICIPANTS AND RANDOM ASSIGNMENT

The Linking Study was conducted during the 2011-12 school year. In September 2011 we recruited teachers to participate. In preparation, the research team developed a series of recruitment materials, devised incentives, gained support from the teacher's union - whose president co-authored a letter of support with the district's superintendent that we included in our recruitment materials - and even scripted and produced a video of the district superintendent extolling the value of the study for teachers and the district. We also had active support of the district's elementary mathematics coach, who worked regularly with teachers and was well respected by teachers in the district, and who served as one of two treatment group PLC meeting facilitators.

[1] One of the teachers in the PLC agreed in advance to allow their video clip to be used. In the few cases where no one agreed, a video clip from another consenting teacher at that grade level from a different school was used.

With these resources we visited each of the 12 elementary schools in the district and presented collectively to teachers in grades 1-5 to explain the purpose of the study, which was to experimentally test the idea that cyclically analyzing feedback on instruction in connection to student learning within PLCs was more powerful than examining feedback in PLCs on student learning (ie test data) alone. We explained our research goals, the design of the study (an experiment), the commitment for teachers (observations, facilitated PLCs for treatment group, participation in data collection). We offered PLCs that agreed to participate (regardless of whether they ended up in the treatment group or the control group) a document camera which they could use during their professional learning community meetings or during instruction, as they wished. Initially, we asked that a majority (i.e 2 of 3 or 3 of 4 or 3 of 5) of a PLC agree to participate. But, as recruiting became more desperate, we relaxed this condition and allowed a few individual members of PLCs to participate (ie join in the random sampling process).

Despite these efforts, recruiting teachers to participate in the study was extremely challenging and we struggled to reach our goal of 80 teachers. Full treatment of the recruiting challenges and what they say about the climate of education today is a story for another article. That said, much of teacher reticence to participate focused on two issues. First, teachers were reluctant to be videotaped. Second, and more apparently, teachers worried that the data would be used for accountability purposes. We had anticipated this in our recruitment strategies and took great pains to create a firewall between observation for improvement purposes and observation for accountability purposes. We made it clear that all data were held by the researchers, not the district; that principals could not attend PLCs in which data from the study were being examined, and that our IRB agreement held individuals' information confidential at the risk of us losing our jobs and reputations. Nevertheless, we could not overcome this fear of many teachers. It makes me wonder if the omni-present pressure of accountability produces a closed and protective environment that is anathema to the openness required for the sharing of practice that is essential for learning and professional improvement.

During the recruiting process, we succeeded in enlisting 70 teachers in 28 PLCs in grades 1-5 to participate in the study. Since a component of the intervention was to facilitate conversations within professional learning community meetings, the unit of assignment to either the treatment or control condition was done at the grade level PLC. This did not mean that all teachers in a PLC had to participate in the study, but treatment and control teachers could not exist within a given PLC. Based upon this, we randomly assigned PLCs to treatment or control conditions: 36 teachers in 15 PLCs were assigned to receive the treatment, while 34 teachers in 13 PLCs were assigned to the control condition.

DATA COLLECTION

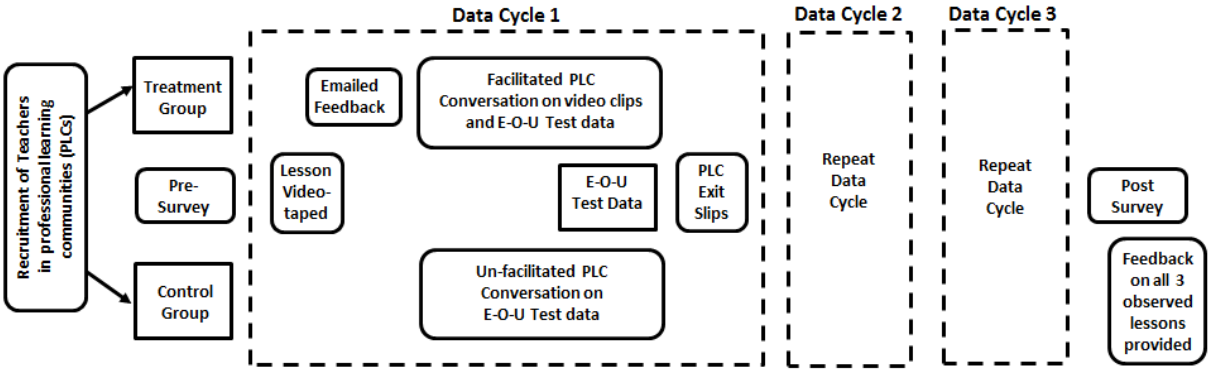
Data collection occurred before, during, and after the Linking Study intervention. The data included surveys, interviews, external ratings of lesson quality, and analysis of student test data.[2] The Linking Study data collection sequence and its alignment with the intervention is depicted in Figure 1. First, teachers in grade level professional learning communities were recruited to participate in the study.

[2] We also conducted interviews with participants and the PLC facilitators, but those data were not used in these analyses.

Even before teachers were assigned to treatment and control groups, all volunteers completed an online survey that asked about their beliefs about the value of different kinds of data, their experience analyzing data, their current mathematics practices, and demographic information.

During each of the three data cycles of the study we collected three forms of data from each participant, both treatment and control. First, based on the videotaped lesson, expert raters assessed the lesson quality of each teacher on two dimensions of mathematics instruction: academic rigor and accountable talk with an instrument called the Instructional Quality Assessment (IQA) (more in the "Measures" description below). Second, after the PLC meeting in which teachers discussed data on their teaching and the end-of-unit test data on their students, each teacher was asked to complete an 'exit slip' on which they rated the quality of the PLC and asked them to self-report on what they learned about instruction and their students. We also asked about their comfort examining data with their peers in the PLC. Third, we collected students' end-of-unit test performance for both treatment and control teachers. This protocol was followed for three cycles across the school year.

Figure 1. Linking Study Design and Data Collection



At the end of the school year, we re-administered the online survey to all teachers. Finally, we collected annual test data for all students whose teachers participated in study.

MEASURES

In this section I provide greater detail of each of the measures that were described briefly above. These include a pre-post online survey; expert ratings of videotaped mathematics lessons; short surveys conducted after each PLC in which teachers examined data associated with the project; and student achievement data. The specific survey items that were used to produce the scales described in this section are detailed in Appendix A.

Online survey

Both before and after the intervention, we conducted an online survey of participants. In the pre-survey we collected information about the background of the participants, including their education overall experience, and experience teaching their current grade. On both the pre- and post-surveys, we also measured four domains focusing on data use that we hypothesized might be impacted by the

intervention. These included two scales about the importance of teaching and learning data, and two scales about their proficiency using teaching and learning data. These are described briefly below, with the items that make up each scale enumerated in Appendix A:

1. *Importance of Instructional Data* – (alpha reliability = .78) was a four-item scale that measured teachers' agreement with statements about the importance of data on instruction and feedback.
2. *Importance of Student Test Data* – (alpha reliability = .76) was a three-item scale that measured teachers' agreement with statements about the importance of test data on instruction.
3. *Proficiency Using Teaching Data* – (alpha reliability = .93) was a four-item scale that measured teachers' perceived proficiency using teaching data to improve their instruction.
4. *Proficiency Using Test Data* – (alpha reliability = .77) was a three-item scale that measured teachers' perceived proficiency using test data to improve their instruction.

Ratings of observed lessons

The data from the videotaped lessons was used for both part of the treatment and part of the research. As part of the treatment, the videotaped lessons formed the basis for providing qualitative feedback to teachers in the treatment group about their instruction (see sequence of feedback in Figure 1). The lessons for all teachers were also numerically rated by trained raters for their instructional quality based upon the Instructional Quality Assessment (IQA), a mathematics classroom observation rubric developed and validated by the Learning Research and Development Center (LRDC) at the University of Pittsburgh. Based upon research of the types of mathematics instruction that lead to improved student achievement, the IQA produces individual teacher scores on two dimensions of mathematics instruction: Academic Rigor and Accountable Talk. We chose these dimensions because there was both research on their leverage to change instruction (Cobb, Boufi, McClain, & Whitenack, 1997; O'Connor & Michaels, 1996; Tharp & Gallimore, 1988) and because there were rubrics developed to assess them (Matsumura, Garnier, Pascal, & Valdés, 2002; Junker, Weisberg, Matsumura, Crosson, Wolf, Levison, 2005; Boston, & Wolf, 2006). Just as important, the developers note that the IQA can also be used to provide teachers with formative feedback about their instruction (Junker et al. 2005), which fit perfectly with our study design.

The IQA produced two scales:

1. *Academic Rigor* – A three item scale that assesses the rigor of the design and enactment of the lesson.
2. *Accountable Talk* – A four item scale that measures the quality of student-teacher interactions during the lesson.

In order to score the classroom observations using the IQA, we contracted with LRDC to come to Philadelphia and to provide two days of training to our coders. Next, we had coders practice using other videos of non-study mathematics lessons.

PLC Exit Slips

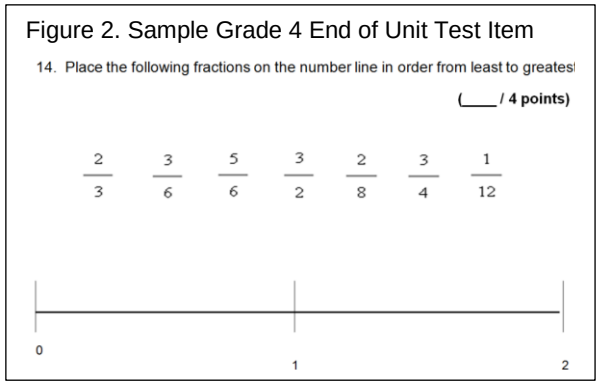
At the end of each of the three PLC meetings in which data were discussed as part of the project, both treatment and control teachers were asked to complete a short survey about the activities of that PLC. Using confirmatory factor analysis, we combined the survey items into three scales. These are listed below.

1. *Learning about Instruction* – (alpha reliability = .78) was a five-item scale that asked teachers about the extent to which they learned about their instruction in their PLC meeting examining data.
2. *Learning about Students* – (alpha reliability = .89) was a six-item scale that asked teachers about the extent to which they learned about their students in their PLC meeting examining data.
3. *PLC Group Interaction* – (alpha reliability = .72) was a four-item scale which asked teachers about their comfort discussing data in their PLC and the quality of the conversation.

End of Unit Test Data

At each grade, the district provided an end of unit assessment that aligned with the district's curriculum. The assessment varied by grade level, but generally consisted of a combination of multiple choice and word problems that tested both students' mastery of the content of the unit and asked them to show their thinking process. Each teacher was asked to administer the test to their students at the end of the unit, to score the work of their students, to enter it into a district database, and to use the resulting data to discuss student understanding in a subsequent professional learning community (PLC) meeting. This was an established district policy that had been going on for at least three years prior to the year in which the Linking Study occurred.

Figure 2. Sample Grade 4 End of Unit Test Item



The end of unit assessments served three purposes in the Linking Study. First, they were the standard data that every grade level was expected to discuss in PLCs across the school year; which formed the control condition of 'looking only at student test data.' Second, the PLC activity of teachers looking at student tests data were incorporated into the treatment condition and augmented with looking at instruction; hence the 'link' between teaching and learning. Third, they formed one of the common data sets to compare the performance of students of treatment and control teachers.

Figure 3. Sample Grade 5 End of Unit Test Item

10. The following table shows the times of four students who ran the 50-yard dash in their physical education class. Who took the longest to run the 50-yard dash? How do you know?

THE 50-YARD DASH	
Student	Time (in seconds)
Brian	7.225
Danjiro	7.713
Patricia	7.748
Fusako	7.204

End-of-unit tests consisted of a different number of items both across tests and across grades. For example, early grade tests had between 2-5 items, depending on the unit; whereas 4th and 5th grade

tests had between 12-15 items. To put all tests on the same scale, we converted each test into a percent correct on a 100-point scale.

For each student we identified four percent correct scores. The first score was the percent correct on the first mathematics unit of the year, which we used as a pre-test covariate in our models. We then utilized the student's performance in the unit following each curriculum unit upon which we intervened. The purest form of this strategy can be seen in the third grade in Table 1. Student performance (represented as the percentage of the items on which the student performed correctly) on *Investigations* Unit 1 was considered the pre-test. The first post-test was Unit 4, which was the unit following the first intervention unit, Unit 3. The second post-test was student performance in Unit 7, the unit following the second intervention unit, Unit 6. Unit 9 was the third post-test, as it following Unit 8, which was the third intervention unit at that grade.

Table 1. Sequence of intervention units and end-of unit test data used in student impact analyses.

Grade	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6	Unit 7	Unit 8	Unit 9	Unit 10	Unit 11
First	T1/Pre	Post1	T2	Post2		T3	Post3				
Second	T1/Pre	Post1	T2	Post2		T3	Post3				
Third	Pre		T1	Post1		T2	Post2	T3	Post3		
Fourth	Pre	T1	Post1	T2	T3/Post2	Post3					
Fifth	Pre	T1	Post1		T2	T3/Post2	Post3				

T1 = Linking Treatment Unit 1
T2 = Linking Treatment Unit 2
T3 = Linking Treatment Unit 3
Pre = Pretest
Post1 = Posttest 1
Post2 = Posttest 2
Post3 = Posttest 3

In the cases of grades 1 and 2, in which the first unit of the year was also the first of the three intervention units at that grade level, we also considered this the pretest. We considered this reasonable because, although teachers had received email feedback on their lesson during the unit, feedback to teachers in their PLC did not occur until after students had taken the end of unit test.

We also had to account for the fact that in grades 4 and 5 the treatment occurred in two concurrent units. In those cases, we used the second treatment unit also as the second post-test.

We plan to conduct additional analyses using state test data, but these are not yet completed.

ANALYSIS PLAN

Since a major part of the intervention occurred during PLC meetings, and therefore we could not mix treatment and control teachers within a PLC, the PLC became the unit of randomization in the study. Therefore, all analyses were conducted as multi-level models with students nested within teachers (where appropriate) and teachers nested within PLCs.

The analysis to address the first research question controlled for the pre-treatment measure and predicted the post-treatment measure, including a covariate for treatment. For this and subsequent models I report the fixed effects and covariance parameters (random effects). I also report the intraclass correlation (ICC) for each full model to show how the variation is distributed across the different levels. I

do not report the ICCs of unconditional models, as I am not primarily interested in the amount of pre-existing variation in outcome measures.

The analyses to answer the second and third research questions, which had three time points, employed two-level random intercept models in which teachers were nested within PLC and time was treated as a continuous fixed effect. Consequently, time was forced to be linear and its relationship with the treatment was also treated as linear. These models also included an interaction term between time and treatment, which allowed the relationship between the time and the outcome to differ by treatment group. An unstructured error covariance matrix was used for the mixed-effect model, which allows all elements to be freely estimated and makes the fewest assumptions about the error covariance structure. The results tables for these analyses also include least square group means and effect size calculations as Cohen's *D*, which were calculated based on differences between adjusted means and associated standard error.

The three level model of student outcomes had student performance nested within teacher, nested within PLC and allowed the effect for time (linear fixed effect) to vary randomly by teacher. This allowed the natural rate of student growth to vary by class around a population mean. Student achievement data was represented as the percent correct on the end-of-unit test.

STUDY SAMPLE

During the recruiting process, we succeeded in enlisting 70 teachers in 28 PLCs in grades 1-5 to participate in the study. Based upon this, we randomly assigned PLCs to treatment or control conditions: 15 PLCs with 36 participating teachers were assigned to receive the treatment, while 13 PLCs with 34 participating teachers were assigned to the control condition. Subsequently, six teachers dropped out of the project during the first round of data collection for the study. Two of these teachers, each from different PLCs, came from the treatment group. The other four teachers were from the control group. They comprised one teacher from a PLC and three teachers from another PLC (the entire PLC). We tried to convince these teachers to remain in the data collection portion of the study, but they refused.

The final sample, consisting of 64 teachers in 27 PLCs, is shown by grade level in Table 2.

Table 2. Final sample of teachers and PLCs by grade in Linking Study

Grade	Treatment Teachers	Treatment PLCs	Control Teachers	Control PLCs	Total Teachers	Total PLCs
First	3	1	9	3	12	4
Second	8	4	5	2	13	6
Third	6	3	8	3	14	6
Fourth	9	4	4	2	13	6
Fifth	8	3	4	2	12	5
Total	34	15	30	12	64	27

Several background characteristics of participants in the both the treatment and control groups are shown in Table 3 to give a sense of the experience of participants. Teachers in the study had an average of just over 12 years of teaching experience, which ranged from a minimum of two years of experience to a maximum of 33 years of experience. The variability in experience at their current grade level was large for teachers in both groups where, as shown in the standard deviation, some teachers were in their first year at that grade level, while others had taught their whole career at their current grade level.

Table 3. Background Characteristics of Study Participants

Characteristic	Treatment (n=34)	Control (n=30)
Experience Overall (mean and standard deviation)	12.32 (7.35)	12.29 (5.86)
Experience at Grade Level (mean and standard deviation)	6.77 (6.32)	6.71 (5.42)
Highest Degree (respondents and percentage)		
Bachelors	14 (41%)	12 (40%)
Masters	12 (35%)	9 (30%)
Masters Plus	8 (24%)	9 (30%)
Study Participants in a PLC (mean and standard deviation)	3.30 (.95)	3.59 (.75)

~ $p \leq .10$, *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$;

About 40 percent of the teachers in each group had bachelor's degrees, about a third of the teachers in each group had master's degrees, and about 25-30 percent of the teachers in each group had masters degrees plus coursework. The average number of study participants in the PLCs for each group was 3-4 teachers. Additionally, the lack of any significant differences in background characteristics between treatment and control group teachers provides substantiation of the effectiveness of randomization.

RESULTS

The results section is organized in alignment with the research questions. First, I examine the impacts of the intervention on teachers' views about data and their self-reported preparation to teach and facilitate student understanding. Second, I investigate the impact of the intervention on teachers' perceptions about their learning about instruction and their students. Third, I assess the impacts of the intervention on teachers' subsequent mathematics instructional practices. Finally, I examine the impact of the intervention on the learning of students on both end-of-unit tests and state assessments.

RQ1: What is the impact of providing teachers with feedback on teaching and learning on teachers' views about the importance of teaching and learning data and their self-reported proficiency to use such data in their practice?

The first research question focuses on teachers' views about data and their proficiency using teaching and learning data to inform their practice. The data to analyze these effects come from the pre-treatment and post-treatment survey that were administered immediately before teachers were assigned to treatment and control groups in September 2011 and re-administered again in May or June 2012.

The results of these adjusted post analyses are shown in Table 4. Each model looks for a difference in the post survey means for treatment and control groups after adjusting for the pre-treatment mean. While pre-treatment was a significant predictor of post-scores in almost every case, there is no treatment associated effect on any of these scales. In short, the intervention, which made substantial use of data on teaching and learning, did not significantly change teachers' perceptions of the importance of instructional data, the importance of student test data, nor their perceived proficiency to use either data on teaching or test data.

Table 4. Impact of Treatment Over Time on Teachers' Perceptions of Their Learning

	Importance of Instructional Data			Importance of Student Test Data			Proficiency using Teaching Data			Proficiency Using Test Data		
	ß	SE	ICC	ß	SE	ICC	ß	SE	ICC	ß	SE	ICC
Fixed Effects												
Intercept	2.035***	.404		3.053***	.446		2.677***	.300		1.799***	.320	
Treatment	-.042	.099		-.080	.118		.123	.148		.018	.101	
Pre-	.409***	.117		.181	.124		.200~	.105		.467***	.103	
Covariance Parameters												
PLC	.002	.001		.033	.032		.060	.049		.001	.001	
Residual	.146***	.027		.131***	.030		.210***	.050		.152***	.028	
Adjusted Post Means												
Treatment	3.451	.074	.014	3.715	.079	.201	3.280	.101	.222	3.204	.067	.007
Control	3.493	.066	.986	3.794	.089	.799	3.157	.113	.778	3.222	.075	.993

~ $p \leq .10$, *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$;

Interestingly, the survey also did not pick up any effects of teachers' perceived preparation to use either of the two instructional strategies – academic rigor or accountable talk – that were the main focus of the intervention. As we will see, several other measures in this study detected this effect.

RQ2: What is the impact of providing teachers with feedback on teaching and learning on teachers' perceptions about their learning about their instruction and their students?

At the end of each PLC in which participating teachers examined data, we administered an 'exit slip' which asked teachers to answer a series of questions about their perceptions of their experience. We

developed three scales from these survey questions: (1) the extent to which they learned about their mathematics instruction through their PLC experience; (2) the extent to which they learned about their students thinking about mathematics through their PLC experience; and (3) the extent to which they felt comfortable looking at data with their colleagues in a PLC. In the analyses that address this research question we compared the responses of treatment and control teachers in models that appropriately nest teachers within their PLCs.

Table 5. Impact of Treatment Over Time on Teachers' Perceptions of their Learning

	Learning About Instruction			Learning About Students			Comfort With PLC Group Interactions		
	ß	SE	ICC	ß	SE	ICC	ß	SE	ICC
Fixed Effects									
Intercept	2.507***	.133		2.637***	.127		3.453***	.142	
Treatment	.842***	.180		.589**	.172		.066	.192	
Time	.200***	.047		.092*	.044		.052	.053	
Treat*Time	-.152*	.064		-.053	.061		-.019	.072	
Covariance Parameters									
PLC	.085**	.029	.39	.067**	.029	.32	.077**	.031	.30
Teacher	.003	.012	.02	.031*	.017	.15	.014	.017	.06
Residual	.128***	.016	.59	.114***	.015	.53	.160***	.020	.64

~ $p \leq .10$, *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$;

Looking first at the covariance parameters, we see that, even after including treatment and time as predictors, there was significant variation across PLCs for all three outcomes, with between PLC variance explaining, respectively, 39 percent, 32 percent, and 30 percent of the variation in the outcomes across the three models. Significant variation between teachers within PLCs was only evident for the 'learning about students' outcome, which explained 15 percent of the total variation.

The fixed effects in table 5 for the outcome of teachers' perceptions of learning about their instruction shows a positive and statistically significant treatment effect, a positive and significant time effect, and a *negative* and significant interaction of treatment and time. This indicates that overall the treatment group significantly outperformed the control group, that there was significant growth over time of all participants in the study, but that the difference between treatment and control groups, was reduced across the three time points.

The fixed effects for the outcome of teachers' perceptions of their learning about their students showed a similar pattern. There was a positive and significant effect of the treatment, whereby the treatment group significantly outperformed the control group on this outcome, there was a significant and positive effect of time, whereby scores increased across all three time points, but there was a narrowing of the differences between the two groups over time, which in this case was not significant.

Finally, the fixed effects for teachers' feelings of comfort examining data with their colleagues within PLCs did not show any differences between treatment and control groups, nor any changes in responses over time.

The adjusted means presented in Table 6 show the patterns of effect on each exit slip survey scale for both treatment and control teachers. Looking first at the scale that measures teachers' perceptions of the extent to which their PLC experience helped them learn about their teaching, we can see that at all three time points there is a significant effect of the treatment; that is, teachers in the treatment group had significantly higher average scores on this scale than did teachers in the control group. The effect sizes were robust, ranging from two thirds of a standard deviation unit at time one to a third of a standard deviation unit at time three. While the treatment group mean grew slightly over time, the control group mean actually grew more (although the difference continued to be significant). This accounts for the negative treatment by time interaction shown in Table 5.

Table 6. Adjusted Means for treatment and control groups for Teacher Perception Outcomes

	Treat	Control	Difference	Standard Error	Cohen's D
Learning About Instruction					
Time 1	3.40	2.71	0.69***	0.16	0.616
Time 2	3.45	2.91	0.54***	0.15	0.521
Time 3	3.49	3.11	0.39*	0.16	0.342
Learning About Students					
Time 1	3.26	2.73	0.53***	0.14	0.569
Time 2	3.30	2.82	0.48***	0.12	0.569
Time 3	3.34	2.91	0.43**	0.14	0.453
Comfort With PLC Group Interactions					
Time 1	3.55	3.50	0.05	0.15	0.046
Time 2	3.58	3.56	0.03	0.13	0.031
Time 3	3.62	3.61	0.01	0.15	0.008

~ $p \leq .10$, *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$;

The scale of items that represent teachers' perceptions of the extent to which they learned about their students' understanding of mathematics during their PLC experience also showed that, at all three time points, there were significantly greater scores for treatment teachers than for control group teachers. At all there time points, the perceived learning about students from the treatment group was greater than that of the control group. Again, the standardized effect sizes were substantial, averaging about a half a standard deviation unit at each time point.

RQ3: What is the impact of providing teachers with feedback on teaching and learning on subsequent instructional practices in mathematics?

Next, I examined growth in the external ratings of the two measures of instructional practice, academic rigor and accountable talk, and found significantly greater growth in the treatment group in comparison to the control group on both outcomes. Table 7 shows the fixed and random effects (covariance parameters) for both academic rigor and accountable talk.

Table 7. Impact of treatment over time on Instruction

	Academic Rigor			Accountable Talk		
	ß	SE	ICC	ß	SE	ICC
Fixed Effects						
Intercept	2.853***	.151		2.751***	.178	
Time	-.039	.056		-.047	.063	
Treatment	.105	.204		-.176	.242	
Treat*Time	.149~	.077		.235**	.086	
Covariance Parameters						
PLC	.036~	.036	.14	.034	.119	.06
Teacher	.094**	.039	.28	.303***	.092	.54
Residual	.192***	.025	.57	.226***	.030	.40

~ $p \leq .10$, *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$;

The covariance parameter estimates in Table 7 indicate that there was significant variance at the PLC level for academic rigor, in which about 14 percent of the variance in the model was at PLC level (as measured by the intraclass correlation, or ICC) with 28 percent at the teacher level (ie between teachers), and 57 percent of the variance occurring within teacher. For accountable talk, very little of the variance, a non-significant 6 percent, was between PLCs; while the majority of the variance in the model, 54 percent, was between teachers. Despite the lack of difference across PLCs, we retained this level in the model due to its central role in the study design.

Looking at the fixed effects, we can see that there are neither significant main effects for time or treatment. However, for both academic rigor and accountable talk the treatment by time interaction was positive and statistically significant (although only at the .10 level for academic rigor); this indicates that there is a differing growth rate between treatment and control groups over time for both outcomes. I will explore this further through an examination of the adjusted means for each group.

Table 8 reveals an interesting story of the changes in time for teachers' mathematics instruction associated with the Linking treatment. The table shows the means for both the treatment and control group, adjusted for the nested relationship of teachers within PLCs, the differences between the means, the standard errors, and the standardized effect sizes associated with the differences.

The time 1 measures for both academic rigor and accountable talk show negligible and non-significant differences between the treatment and control groups. This is important because this assessment was conducted before the treatment occurred (ie the first videotape of a teachers' lesson was conducted before any intervention). At time 2, there was a statistically significant difference in the academic rigor

of the lessons of treatment teachers in comparison to control teachers, with a standardized effect size of .43. There was also a marginally significant difference between the accountable talk rating of teachers in the treatment and control groups at time 2, with a small effect size of .25.

Table 8. Adjusted Means for treatment and control groups for instructional outcomes

	Treat	Control	Difference	Standard Error	Cohen's D
Academic Rigor					
Time 1	3.07	2.81	0.25	0.16	0.236
Time 2	3.18	2.78	0.40**	0.13	0.434
Time 3	3.29	2.74	0.55**	0.16	0.511
Accountable Talk					
Time 1	2.76	2.70	0.06	0.19	0.045
Time 2	2.95	2.66	0.29~	0.17	0.248
Time 3	3.14	2.61	0.53**	0.19	0.398

~ $p \leq .10$, *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$;

At time 3, there continued to be statistically significant differences between treatment and control rating of both academic rigor and accountable talk, with an effect size of about a half a standard deviation unit.

RQ4. What is the impact of providing teachers with feedback on teaching and learning on student outcomes?

Table 9 presents the multi-level model of student end-of-unit test performance over time, appropriate adjusting for the nested relationship of students within teachers within PLCs. The fixed effects indicate,

Table 9. Impact of treatment over time on Instruction

	End of Unit Test Performance		
	ß	SE	ICC
Fixed Effects			
Intercept	.876***	.029	
Time	-.020***	.006	
Treatment	-.036	.039	
Treat*Time	.021**	.008	
Covariance Parameters			
PLC	.007***	.002	.292
Teacher	.001	.001	.042
Residual	.016***	.003	.667

~ $p \leq .10$, *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$;

most importantly, a significant and positive treatment by time interaction, which indicates that significantly different test performance trajectories for students of treatment and control group teachers over time.

To investigate this effect further, we produced the adjusted means for each group at each time point. These are shown in Table 10. The adjusted means show an increasing difference, albeit small, in the average test scores of students of teachers in the treatment and control groups across each of the time points. At time 1, before the treatment, there was a small negative difference between the performance of the two groups; at time 2 the difference is positive but negligible. Increasingly, the difference grows larger at each time point. While these differences are not statistically significant at any one time point, their cumulative difference is significant, as shown in the treatment by time interaction in table 9.

Table 10. Adjusted Means for treatment and control groups for student test outcomes

	Treatment Group Mean	Control Group Mean	Difference	Standard Error
Time 1	.841	.856	-.015	.036
Time 2	.842	.837	.006	.034
Time 3	.844	.817	.027	.035
Time 4	.845	.797	.048	.037

~ $p \leq .10$, *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$;

SUMMARY AND DISCUSSION

In this project, a CPRE research team worked with a New Jersey school district to develop an intervention that provided teachers with cyclical and facilitated conversations about data on their instruction examined in conjunction with data on the learning of their students (end-of-unit test data). The intervention was conducted within an experimental framework, with teachers in grade levels (PLCs) randomly assigned to participate in the experience or continue with their usual practice of examining only student end-of-unit test data in their PLCs.

The results of the experiment indicate large effects – on the order of about a third to a half standard deviation in magnitude – on what teachers felt they learned about their teaching and their students' understanding and, more importantly, on their subsequent instructional practice. The impacts on instructional practice are particularly notable because they are judgments of external raters, rather than teacher self-report. There were also small, but statistically significant, effects of the intervention on student learning over time. Notably, teachers did not report being better prepared to use data, nor did they perceive a greater importance for data as a result of their experience. Thus, even though this intervention was about using data, it was not framed nor perceived as such. Rather, it was more focused on looking at teaching and learning, and the mechanism to do so was data on practice and performance.

As a result of this research, what features of this intervention should we focus on as important clues about how to strengthen data-based experiences for teachers to provide opportunities to better hone

their craft and improve the learning of their students? Although the intervention featured data on teaching and learning, the data that teachers examined had several important and distinctive attributes. First, the intervention did not ask teachers to learn statistical or other numerical analysis techniques. Neither the data on teaching nor the data on student learning emphasized numerical information, but rather was designed to emphasize the substance represented by the data, rather than the data themselves. This helped teachers reflect on their instructional approaches and gain insight into the levels of understanding of their students, rather than to acquire new analytic skills to make sense of the data.

A second feature of the intervention was its cyclical nature; the treatment was designed to occur multiple times across the school year to increase teachers' chances to apply what they learned in subsequent teaching. This reinforces much of the research on the importance of embedded and sustained learning experiences.

A third feature of the intervention was that it linked what teachers do (teaching) with what it produces (student learning) and pressed teachers to both examine each individually, and to ask questions about the relationship between the two. A mean feature of the Linking Study was to facilitate teacher explorations of the connections between teaching and learning and to experimentally test the impacts of the experience. The results indicate this is a promising area for both further professional development and more precise research.

References

Black, P. P. J., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2007). *Assessment for learning*. New York: Open University Press.

Boston, M., & Wolf, M. K. (2006). *Assessing academic rigor in mathematics instruction: The development of the Instructional Quality Assessment toolkit* (CSE Technical Report No. 672). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Boudett, K.P., City, E.A., & Murnane, R.J. (2005). *Data Wise*. Cambridge MA: Harvard Education Press.

Butler, R. (1988) Enhancing and undermining intrinsic motivation; the effects of task-involving and ego-involving evaluation on interest and performance, British Journal of Educational Psychology, 58, pp. 1-14.

Christman and Edmonds (2013).

Cobb, P., Boufi, A., McClain, K., & Whitenack, J. (1997). Reflective discourse and collective reflection. *Journal for Research in Mathematics Education, 28*(3), 258-277.

Deming, W. 1986. Out of Crisis. Center for Advanced Engineering Study, Massachusetts Institute of Technology, Cambridge, Mass.

DuFour, R., Eaker, R., & DuFour, R. (2008) Revisiting Professional Learning Communities at Work: New Insights for Improving Schools. Bloomington, IN: Solution Tree Press.

Ebby,C.B. & Oettinger, A. (2013). Facilitating Productive Discussions in Professional Development Settings. Paper presented at the Research Presession of the Annual Meeting of the National Council of Teachers of Mathematics, Denver, CO.

Goertz, M.E., Nabors Oláh, L. & Riggan, M. (2009). From testing to teaching: The use of interim assessments in classroom instruction. CPRE Research Report #RR-65. Philadelphia, PA: Consortium for Policy Research in Education.

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. L. (2009). From evidence to action: A seamless process in formative assessment?, *Educational Measurement: Issues and Practice, 28*(3), 24–31.

Junker, B., Weisberg, Y., Matsumura, L. C., Crosson, A., Wolf, M., & Levison, A., (2005). *Overview of the instructional quality assessment* (CSE Technical Report No. 671). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.

Kluger, A.N. & DeNisi, A. (1996) The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory, Psychological Bulletin, 119, pp. 254-284.

Matsumura, L. C., Garnier, H., Pascal, J., & Valdés, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment, 8,* 207–229.

O'Connor, M. C., & Michaels, S. (1996). Shifting participant frameworks: Orchestrating thinking practices in group discussions. In D. Ghicks (Ed.), *Discourse, learning, and schooling* (pp. 63-103). New York: Cambridge University Press.

Sadler, R. (1989) Formative assessment and the design of instructional systems, Instructional Science, 18, pp. 119-144.

Schunk, D.H. & Swartz, C.W. (1993a) Goals and progress feedback: effects on self-efficacy and writing achievement, Contemporary Educational Psychology, 18, pp. 337-354.

Smith, D.R., & Ruff, D.J. (1988) Building a Culture of Inquiry. In Assessing Student Learning: From Grading to Understanding. New York: Teachers College

Supovitz, J. (2012). Getting at Student Understanding – The Key to Teachers' Use of Test Data. *Teachers College Record*, 114, 1-29.

Supovitz, J., Foley, E. & Mishook, J. (2012). In Search of Leading Indicators in Education. *Educational Policy Analysis Archives*, 20(19), 1-26.

Supovitz, J. & Klein, V. (2003). Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement. Philadelphia: Consortium for Policy Research in Education.

Supovitz, J. & Merrill, L. (2010). Teacher Use of Student Performance and Related Data in Professional Learning Communities. Paper presented at the American Educational Research Association Annual Conference, Denver, Colorado.

Appendix A – Survey Items and Scale Reliabilities

PRE-POST SURVEY SCALES

IMPORTANCE OF INSTRUCTIONAL DATA (Alpha =.78)
1. Classroom observation data are an important source of information to inform my classroom instruction.
2. Watching video of my teaching can help me become a better teacher.
3. I think it is important to have feedback on my classroom teaching to inform my educational practice.
4. Improving my ability to use feedback on my classroom instruction will help me to become a better teacher.

IMPORTANCE OF STUDENT DATA (Alpha =.76)
1. Data on my students' performance are an important source of information to inform classroom instruction.
2. I think it is important to have data on my students' performance to inform my educational practice.
3. Improving my ability to use my students' performance data will help me to become a better teacher.

PROFICIENCY USING TEACHING DATA (Alpha =.93)
1. Using feedback on my teaching to refine my instructional approaches.
2. Using feedback on my teaching to gauge student understanding.
3. Using feedback on my teaching to adjust how I engage student in class.

PROFICIENCY USING TESTING DATA (Alpha =.77)
1. Analyzing trends in student performance over time.
2. Translating student performance data into knowledge about student strengths and weaknesses.
3. Using student performance data to tailor my instruction to meet individual students' needs.
4. Targeting interventions for students based upon their student performance data.

Appendix A – Survey Items and Scale Reliabilities

EXIT SLIP SCALES

LEARNING ABOUT STUDENTS SCALE* (ALPHA =.89)
1. The data we examined today gave me useful insights into the performance of my students.
2. I learned something today about the mathematics content of the unit we discussed.
3. The conversation in today's meeting helped my PLC get on the same page about mathematics instruction.
4. The data we examined on student performance gave me useful insights into the understanding of my students.
5. I gained a better understanding of how to examine student test data for insights into student thinking.
6. I plan to make changes in my teaching as a result of things I learned from examining student performance data

LEARNING ABOUT INSTRUCTION SCALE* (ALPHA =.78)
1. I learned something today about designing challenging math lessons.
2. I learned about engaging students to explain their thinking about how they solve mathematics problems.
3. I learned something today about developing students' conceptual understanding of mathematics.
4. I learned new strategies to press students to explain their thinking.
5. I plan to make changes in my teaching as a result of things I learned in this PLC meeting.

PLC GROUP INTERACTION SCALE* (ALPHA =.72)
1. The conversation in today's meeting helped my PLC get on the same page about mathematics instruction.
2. I would have preferred to examine these data on my own instead of with my grade level team. (REVALENCED)
3. Examining data with colleagues made the meeting more meaningful than examining the data on my own.
4. Please rate the overall quality of the discussion in your PLC today (3 point scale of Lo, Medium, Hi Quality)

*All responses on a four point scale (strongly disagree, disagree, agree, strongly agree) unless otherwise specified.

Appendix A – Survey Items and Scale Reliabilities

CLASSROOM RATING SCALES

ACADEMIC RIGOR SCALE
1. Potential of the Task Did the task have the potential to engage students in exploring and understanding the nature of mathematical concepts, procedures, and/or relationships?
2. Implementation of the Task At what level did the teacher guide students to engage with the task in implementation?
3. Student Discussion Following the Task To what extent did students show their work and explain their thinking about the important mathematical content?

ACCOUNTABLE TALK SCALE
1. Participation Was there widespread participation (ie, a response to a mathematical question) in teacher-facilitated discussion?
2. Questioning Does the teacher ask academically relevant questions that provide opportunities for students to elaborate and explain their mathematical thinking?
3. Asking (Teacher Press) Were students pressed to support their contributions with evidence and/or reasoning?
4. Providing (Student Responses) Did students support their contributions with evidence and/or reasoning?