Using Data to Inform Decisions: How Teachers Use Data to Inform Practice and Improve Student Performance in Mathematics

Results from a Randomized Experiment of Program Efficacy

Linda Cavalluzzo • Thomas M. Geraghty • Jennifer L. Steele • Laura Holian • Frank Jenkins • Jane M. Alexander • Kelsey Y. Yamasaki

IRM-2013-U-006508
March 31, 2014

CNA
ANALYSIS & SOLUTIONS

Acknowledgements

The authors would like to thank Duval County Public Schools, school administrators, and teachers for participating in this study. We also thank members of our technical working group, Michael Puma, Jonathan Star, and Kathryn Boudett, as well as Ellen Mandinach, Jeffrey Wayman, Thomas Goode, Shira Solomon, Rebecca Thessin, Christine Mokher, and CNA reviewers who helped articulate the study plan; and seminar participants at the SREE 2013 Fall Conference (Society for Research on Educational Effectiveness) and the October 2013 CIERS: Causal Inference in Education Research Seminar (Ford School of Public Policy, University of Michigan). We also thank our IES program officers, Harold Himmelfarb and Wai-Ying Chow, for their advice. Any remaining errors are our own.

Author affiliations:

Linda Cavalluzzo, Ph.D.
Project Director, CNA

Thomas M. Geraghty, Ph.D.
CNA

Jennifer L. Steele, Ph.D.
RAND Corporation

Laura Holian, Ph.D.
Insight Policy Research

Frank Jenkins, Ph.D.
Westat

Jane M. Alexander
CNA

Kelsey Y. Yamasaki
CNA

Distribution unlimited. Specific authority contracting number: R305A100445.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A1000445 to CNA. The opinions expressed are those of the authors and do not necessarily represent views of IES or the U.S. Department of Education.

Contents

Executive Summary

The purpose of this study is to evaluate, using a randomized experimental design, the efficacy of TERC's *Using Data* program to change teacher behavior and improve student learning outcomes. The *Using Data* intervention provides professional development and technical assistance to teachers to help them use data collaboratively to identify and solve systemic student learning problems.

The intervention was implemented by school-based data teams composed of a designated data coach and four grade 4 and 5 math teachers in Duval County Public Schools, a large, urban school district serving Jacksonville, Florida, during school years 2011–12 and 2012–13. In the first year of the study, teachers in the treatment group participated in professional development events and technical assistance sessions that exposed them to *Using Data* and helped them implement its processes. In the second year the teachers received additional assistance with implementation.

Approach

Sixty (60) schools were recruited to participate in the study. Because the characteristics of the students, as well as the level of mathematics performance in these schools, differed widely, we grouped them into four blocks and conducted a block-randomized design, ensuring a balanced sample across the treatment and control-group schools.

Our confirmatory analyses measure teacher outcomes at the end of year 1, when all *Using Data* program materials have been introduced, and student outcomes at the end of year 2, when teachers had the benefit of a full year of UD training in preparation for year 2 instruction. In addition to measuring overall program effects, we conducted exploratory analyses to better understand when the *Using Data* program works, and for whom.

Measures

We evaluate the effects of the intervention on teachers by examining three outcomes:

- Teacher-reported frequency of use of individual and collaborative data-use practices
- Teachers' knowledge and skills pertaining to data use for instructional improvement (level of "data literacy")
- Teacher-reported attitudes and beliefs about the value of data to improve instruction

We evaluate the effects of the intervention on student outcomes based on performance on the state-administered end-of-grade mathematics assessment.

We supplement this quantitative analysis with extensive qualitative data to explore the contextual factors that may shape the implementation of the *Using Data* intervention.

Findings

Our evaluation finds a positive effect of the *Using Data* intervention on teacher behavior after year 1 of the study. Specifically, teachers in the treatment group at the end of year 1 reported using data more frequently, exhibited higher levels of data literacy compared with control teachers, and held attitudes and beliefs that were more favorable toward data use for instructional improvement.

We do not find an overall treatment effect of the UD program on schoolwide student performance on the annual state math assessment. Restricting the sample to the students of intent-to-treat (ITT) teachers at the end of year 2 (ie, limiting the sample to students of teachers who were in the randomized sample and offered the UD program from the beginning), also yields a null effect overall. However, students of ITT teachers from the block of lowest-performing schools at baseline show improvements in year 2 that are moderately sized and statistically significant (effect size = .40, p = .01).

Conclusions

The weight of the evidence presented here indicates that *Using Data* improves teachers' outcomes after one year, and improves the outcomes of their students in high-needs schools after two years. We conclude that further research and evaluation of the *Using Data* program is warranted.

1. Study Overview

1.1. Introduction

The last two decades have witnessed a vast expansion in the use of education data to improve classroom instruction and raise student achievement. Educators are making decisions using a wide variety of data about students, including state accountability test scores; interim progress test results; classroom tests, assignments, and homework; and attendance, mobility, and grade-level progression rates, as well as dropout and graduation rates (Allensworth & Easton 2007; Hamilton et al. 2009; Marsh et al. 2006).

For those schools and districts that use data effectively, significant gains in student achievement are possible. For example, Carlson et al. (2011), in a multi-state randomized study covering 500 schools in 59 districts, found that a data-driven reform initiative resulted in statistically significant improvement in student math achievement. Faria et al. (2012) reported that several research studies suggested that using a particular type of data—formative assessment—can result in student achievement gains.[1]

Schools and districts face important challenges in increasing data use for instructional improvement. One key challenge is the need for teachers and administrators to have "data literacy" (Halverson & Thomas 2007; Thessin 2007)—defined as the ability to work individually and collectively to examine different types of data (including summative assessments such as achievement data, as well as formative assessments of students' performance and work products) and to develop strategies for improvement based on these data. Few educators, however, are data literate or are prepared to use data effectively (Wayman & Stringfield 2006).

[1] See Black & Wiliam 1998a, 1998b; Black et al. 2004; Brookhart 2001; Christman et al. 2009; Hayward et al. 2004; Heritage 2007; Shepard 2005.

Staff data literacy is now widely recognized as a critical resource in improving the academic performance of schools (e.g., Duncan 2009; Fullan 2000; Haycock 2001; Johnson 1996; Love 2004; Schmoker 1999; Zalles 2005). Evidence suggests that teachers do tend to use multiple sources of data such as homework assignments, in-class tests, and classroom performance, as well as impressionistic, anecdotal, and experiential information, to shape their thinking about their students' strengths and weaknesses (Brunner et al. 2005; Honey et al. 2002; Light et al. 2004; Mandinach et al. 2005, 2006).

As the National Research Council (1996) noted, however, "far too often, more educational data are collected and analyzed than are used to make decisions or take action" (p. 90). Research suggests that teachers are more inclined to think on a case-by-case basis, rather than look for patterns in data at different levels of aggregation, such as classroom-wide or grade-level patterns. Systematically analyzing the relationship between students' classroom performance and their teachers' instructional strategies and materials is beyond the capacity of many teachers (Confrey & Makar 2002, 2005; Hammerman & Rubin 2002, 2003; Mandinach et al. 2005). Not only must practitioners be trained to use data, but they also must understand how to translate data into actionable instructional practice (Herman & Gribbons 2001; Mandinach & Honey 2008; Mandinach et al. 2006; Mason 2002).

The purpose of the current study is to evaluate, using a randomized experimental design, the efficacy of TERC's *Using Data* (UD) program—a professional learning experience designed to fill the gaps in the ways teachers use data to inform instructional decisions—to change teacher behavior and improve student learning outcomes. The program evaluated here rolled out over two school years (SY 11–12 and SY 12–13) in Duval County Public Schools (DCPS), a large, urban school district serving Jacksonville, Florida. In year 1, teachers participated in professional development (PD) events and technical assistance sessions that exposed them to *Using Data* and helped them implement its processes. In year 2 teachers received additional assistance with implementation.

With this timeline in mind, this two-year evaluation is most interested in rigorous examination of the efficacy of the UD program to affect teachers' data use practices and data literacy at the end of year 1, and

student outcomes at the end of year 2. However, it is also of interest to know, on an exploratory basis, whether student outcomes were affected in year 1 and whether teacher outcomes were sustained in year 2. This report examines these questions.

1.2. Study Background

1.2.1. The *Using Data* Professional Development Model

TERC's *Using Data* program provides professional development and technical assistance to help teachers work collaboratively, using data to identify and solve systemic student learning problems. The UD curriculum may be applied to any grade level or subject area, and can be delivered online or face-to-face. For the face-to-face program, the curriculum is introduced in three or four 2-day events in year 1, with additional support in year 2. The online course consists of three modules divided into 15 weekly sessions. The intervention in this study used only the face-to-face program.

The *Using Data* model was developed by TERC, with support from the National Science Foundation and the Eisenhower Regional Alliance for Mathematics and Science Education. The goal of the UD program is to help teachers use data more effectively to improve teaching and learning, and ultimately, through collaborative development of action plans to address systemic learning problems, to cause a cultural change in the classroom and school.

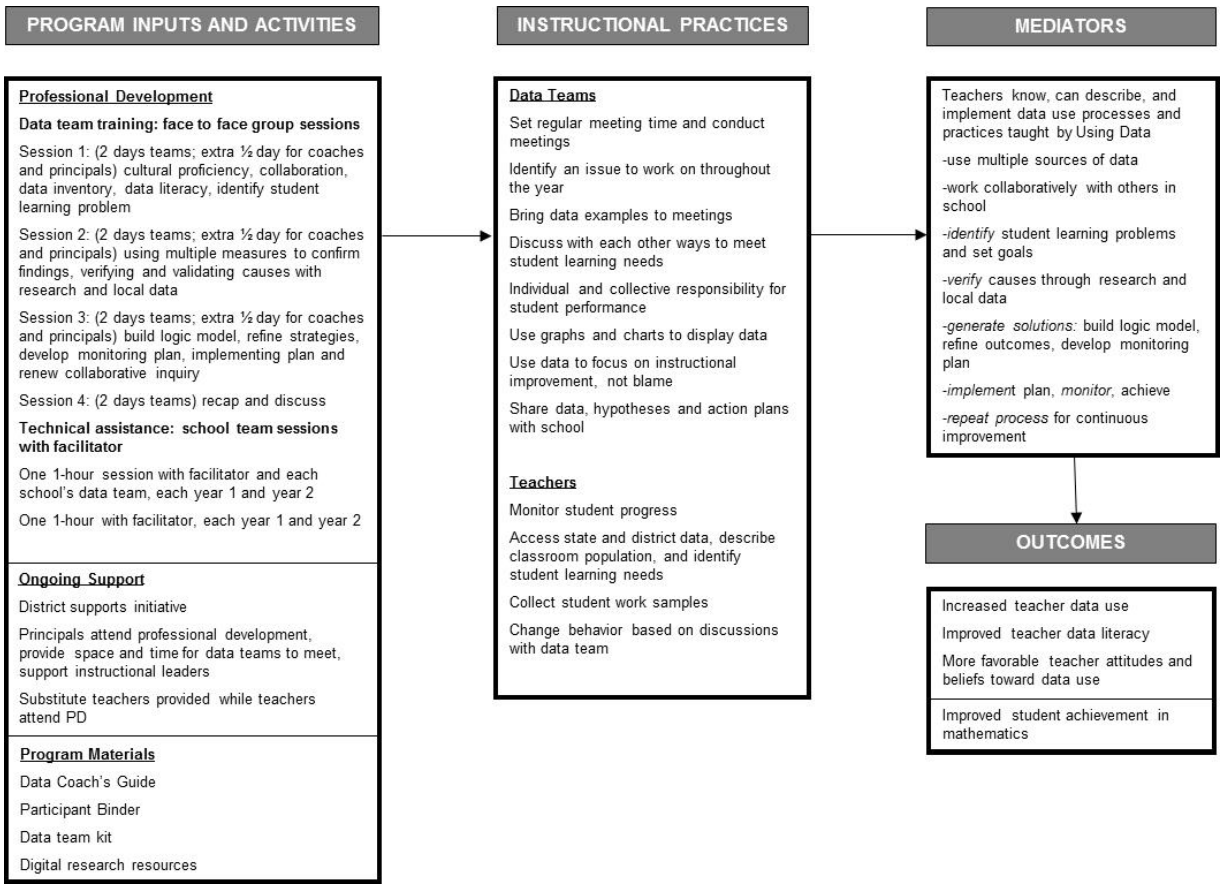
The program introduces teachers and school leaders to a process through which they learn to frame questions, collect data, formulate hypotheses, draw conclusions, take action, and monitor results (Love 2002, 2009b; Love et al. 2008). The UD process has five components: (1) building a foundation for data; (2) identifying student learning problems; (3) verifying causes of learning problems; (4) generating solutions; and (5) implementing, monitoring, and achieving results (for details, see Love 2002; Love et al. 2008). Program participants work in "data teams" typically composed of four teachers plus a "data coach," a leadership role designed to be occupied by a school-based instructional specialist who is not a classroom teacher.

Figure 1 provides a detailed logic model, summarizing how the different components of the program are intended to work together

to produce changes in teacher practices and improvement in student outcomes. The first column shows the program inputs and activities that take place in the professional development events and technical assistance sessions. The second column shows the instructional practices that teachers and data coaches are encouraged to engage in at their schools, along with the data literacy and collaborative behaviors that teachers are expected to practice, learn, and adopt.

These activities are intended to influence teachers' attitudes and beliefs about the value of data in helping to identify and resolve student learning problems, as well as to improve the data literacy skills teachers need in order to identify and address those problems. The expectation is that this leads, in turn, to changes in classroom practices and improvements in student achievement. Based on their experience, UD's developers anticipate that teachers require about a year of exposure to the UD process to be able to use data effectively in their practice.

Figure 1.1. *Using Data* Logic Model



1.2.2. Implementation of *Using Data* for This Study

The setting for the intervention, DCPS, contained 102 elementary schools at baseline, of which one-third were designated Title I. Roughly 50 percent of the students attending the schools were African American, and 68 percent of the students were eligible for free/reduced-price lunch (FRL). Test scores at baseline were below average for the state and were behind other large, urban districts in Florida. Prior to the intervention, DCPS had already implemented active professional development programs, including Leadership Academy, with a data use component.

A total of 60 schools participated in the study. Each school selected a data team. By design, principals identified a data coach and 4 teachers for the team who taught grade 4 or 5 math; but some participating teachers were English language arts or special education generalists at these grade levels, or math teachers at lower grade levels. Data coaches were most often instructional specialists in math, technology, or data in their school, but several were assistant principals, guidance counselors, or teachers. Coaches acted as team leaders and liaisons between data teams and principals. Coaches also were charged with organizing and planning school-based data team meetings and with gathering data for examination by the group.

Of the 60 study schools, 30 "treatment" schools received the *Using Data* professional development and 30 control schools did not. The treatment schools were split into three groups of 10 schools each for professional development sessions. Two TERC facilitators worked with each group.

In year 1, facilitators from TERC used grade 4 data for mathematics from DCPS to illustrate data-use skills and practices. In three 2-day professional development events, treatment data teams received training in the UD model and were encouraged to return to their school to share their findings and experiences, hypotheses about the sources of student learning problems, and action plans to solve them. In this way, data teams encouraged data use and collaborative solutions to systemic student learning problems among their colleagues throughout their school. In year 2, the program provided an additional 2-day professional development event, just prior to the start of the new school year. At this event, treatment data teams got a

refresher in UD processes and looked at aggregate data for their new group of students.

Each year, in winter and spring, the program offered two technical assistance sessions to each treatment school, one at a distance and one in person. A 1-day summative event was held at the completion of the program, in May of year 2. The *Using Data* project director gave a place at the data team table to treatment school principals. The program included them in the training and facilitated discussions about how they could support their school's team; however, they were not required to actively participate on the team. An additional half-day workshop for the coach and principal was attached to each 2-day professional development event to help keep the principal informed and supportive of the program. Although not part of the data team, the principal was expected to provide the time and space for the team to hold regular meetings.

Face-to-Face Group Events—The Curriculum

The first professional development event provided an overview of the *Using Data* program, covering its core beliefs; the data team's responsibilities; a structure for organizing data teams (different roles of facilitator, recorder, reporter, timekeeper, materials manager, and dialogue monitor; emphasis on the importance of switching roles); and norms of collaboration. These norms of collaboration are *pausing, paraphrasing, probing, putting ideas on the table, paying attention to self and others, presuming positive presuppositions,* and *pursuing a balance of advocacy.*

The event also introduced the four phases of "data-driven dialogue," UD's structured protocol for discussing data and drawing evidence-based inferences. The four phases in this process of collaborative inquiry are *predict, go visual, observe,* and *infer/question.*

At this first occasion, teams also began to inventory all of their sources of data and what questions different types of data (aggregate, disaggregate, student work, common assessments) can answer. The half-day for data coaches (and principals) explained how to lead data team meetings and focus on a particular student learning problem.

The second 2-day event continued to deepen the teachers' understanding of student learning by having them engage in data-

driven dialogues. Facilitators showed data teams how to use multiple measures to confirm (or challenge) initial findings. Multiple measures can include, but are not limited to, aggregate data, disaggregate data, and student work. Data team members were taught that when analyzing student work to deconstruct questions in four aspects: *knowledge, skills, big ideas or concepts,* and *level of cognitive demand*. Teams continued to practice using the four-phase dialogue with new data. Teams began to identify student learning problems.

Facilitators led data teams in organizing for causal analysis and using research and local data to verify causes. Teams prioritized causes by which had the biggest impact and what was changeable. When there were multiple student learning problems or multiple causes, the team practiced prioritizing and focusing. Teams identified one or multiple student learning problems. Common areas of student academic weakness that teams identified were fractions, geometry and measurement, and operations.

The third professional development event in year 1 focused on the teams building a logic model, identifying strategies to use on their identified student learning problem(s), and learning ways to monitor the impact of the strategies chosen to overcome the problem. Each team began by reviewing its target student learning problem and setting a student learning goal. The creation of a logic model helps show how particular strategies are linked to desired outcomes. Teams wrote action plans for monitoring their progress and also planned celebrations for success.

School Technical Assistance Sessions with Facilitator

The *Using Data* program's technical assistance (TA) sessions complement its formal professional development events. The purpose of these sessions is to check implementation progress, provide encouragement, and help facilitate getting data and ideas for implementing the program.

In this study, each treatment school's data team was invited to four TA sessions, two in year 1 and two in year 2. The TA sessions consisted of a one-hour "at a distance" session (via phone, Skype, etc.) between February and March of year 1 and between November and December of year 2, plus one face-to-face visit by the facilitators to each school each year. In April of year 1, the facilitators spent one to two hours at

each school, meeting face-to-face with its data team. Face-to-face visits were made in March of year 2. TERC's objectives for the March 2013 technical assistance session were to capture the data teams' work and achievements and to understand the challenges teams faced; and to provide assistance and share resources to help teams in their work.

The format of the TA sessions was an informal conversation between data team members and a TERC facilitator. The facilitators began a session by asking a number of questions to understand the team's progress and achievements. Throughout the conversation, the facilitators periodically suggested a strategy or resource that the data team could use to aid them in their work. The "at a distance" technical assistance sessions followed the same format as the in-person sessions: the facilitators asked questions to understand the team's progress and achievements, and suggested strategies and resources.

Ongoing Support

The program developers believe that district and administrator support is needed for the UD program to really succeed. Principals are invited to attend the four professional development events, and they are strongly encouraged to attend the half-day workshops with the data coaches that follow each. Administrators provide time and space for the data teams to meet during the year, in addition to shaping the school culture and getting teachers to buy into the program.

Program Materials

The UD program includes a variety of materials for teachers, administrators, and coaches. Every data coach, data team member, and building administrator receives a copy of *The Coach's Guide*. Every data coach and data team member receives a participant binder containing templates and activity organizers. Every data coach receives a Data Coach's Kit containing all of the materials needed to support the school in completing each step in the *Using Data* process. The kit includes presenter slides and handouts for all professional development events, "norms of collaboration" tent cards, and pocket guides to data-driven decision making. Every team receives access to online research resources to support its work in causal analysis, generating solutions, and implementing action plans.

The Data Coach's Kit and *Guide* and participant binders were furnished to attendees at the first professional development event. Online and other resources were provided as needed over the course of the program.

The *Using Data* materials provide the detail that professional development providers would need in order to offer the UD program with a high degree of fidelity in other schools or districts. As such, the program is positioned to be taken to scale.

Implementation Data

Tables 1.1 to 1.3 present some basic implementation data for the two years of the *Using Data* intervention. Table 1.1 provides attendance data for the UD professional development events for teachers: 6 days during year 1, and 3 days during year 2. The modal value for attendance was 6 out of 9 days in year 1, and 3 out of 4 days in year 2.[2]

Table 1.1. Attendance at *Using Data* Professional Development Events

Year 1		Year 2	
No. days attending	**No. of treatment teachers**	**No. days attending**	**No. of treatment teachers**
0	0	0	11
2	1	1	17
4	7	2	6
5	12	3	27
6	64	4	1
7	1		
9	3		
Avg. days	5.8	Avg. days	1.8
Att. rate	96%	Att. rate	61%

The distribution for year 2 is somewhat bimodal, as there were a relatively large number of teachers who attended none or just one of the four professional development days. The average attendance rate

[2] Data coaches could attend as many as 3 additional days of PD in year 1, and 1 additional day in year 2, for a total of 9 days and 4 days, respectively. Occasionally, a team whose coach could not make the extra sessions would send a teacher as an alternate. Thus, a few teachers in each year attended more than the standard 6 and 3 days of PD.

in year 1 was 96 percent (5.8 out of 6 days), and 61 percent (1.8 out of 3 days) in year 2. The typical *Using Data* participant thus attended 6 days of professional development in year 1, and 2 days in year 2.

Tables 1.2 and 1.3 provide some information about the school-level data team activities; the source is the data team logs kept by the teams. There were a total of 202 data team meetings in year 1, and 139 in year 2—although there may be some underreporting here, as not all teams turned in all of their logs, especially during year 2. Meeting length and attendance rates were fairly similar between year 1 and year 2 (table 1.2), as were the types of activities the teams engaged in (table 1.3).

Table 1.3 shows that data teams examined data in more than 80 percent of meetings for which we have logs, manipulated data and/or engaged in action planning in 40 to 50 percent of meetings, but engaged in the four-phase data-driven dialogue protocol only about 20 to 25 percent of the time.

Table 1.2. Attendance at *Using Data* School Data Team Meetings

			Attendance		
Study year	No. of data team meetings	Avg. length of meeting (hrs)	Avg. no. of grade 4/5 teachers	Avg. no. of data team members not attending	Attendance rate
Year 1	202	1.4	3.6	0.6	85%
Year 2	139	1.5	3.3	1.0	77%

Table 1.3. Activities at *Using Data* School Data Team Meetings

		Percentage of meetings engaging in ...			
Study year	No. of data team meetings	Examining data	Manipulating data	Action planning	Four-phase dialogue
Year 1	202	86%	50%	40%	26%
Year 2	139	84%	55%	45%	17%

1.2.3. The Control Condition

Control schools did not receive the *Using Data* program. There were, however, district-wide professional development programs already available that have a strong data-use component.

For the past several years DCPS has offered two professional development programs to K–12 teachers. The first program, lasting six to eight weeks, involves a subject-area coach working with a group of teachers at their school to develop and implement formative assessment. With the support of the coach, the teachers analyze student formative assessment results and develop a lesson plan based on that analysis.

The second program is a three-year professional development program that is infused with data work. Among other activities, teachers from different schools gather at the district's professional development center and learn how to continuously analyze student work and student results, and to reflect on their teaching practice. As part of this process, teachers receive training in how to build valid and reliable formative assessments. They are then expected to take these skills and apply them in their own school.

DCPS PD programs overlapped with *Using Data* during the two years of the study. Their existence illustrates that this was already a relatively intensive data-use district, even before implementation of the *Using Data* intervention.

This page intentionally left blank

2. Study Design

This section discusses the design of the study used in this evaluation. A total of 60 schools were recruited to participate in the study. Of those, 30 schools were assigned to treatment status and 30 to control status, using a school-level, block-randomization design.

2.1. Key Outcomes for Teachers and Students

2.1.1. Teacher Outcomes

The teacher impacts of the program are measured at the end of year 1, following the treatment group's exposure to all UD professional development materials and year 1 training. For this study, the research team designed surveys of teachers' data use, knowledge and skills pertaining to effective data use, and attitudes and beliefs about the value of data analysis as an effective tool for instructional improvement. The surveys were administered at the start of year 1 (baseline), the end of year 1, and the end of year 2. Impacts are evaluated based on posttest comparisons, controlling in the statistical model for pretest score.

Although the *Using Data* program lasts two years, evaluating teacher outcomes at the end of year 1 provides the best opportunity to capture direct impact on teachers' knowledge, beliefs, and practices because the data teams are exposed to all UD materials at that point, teacher turnover is limited, and *Using Data*'s professional development activities are most frequent and intensive during the first year.[3] However, given that the study is interested in documenting the solidification and persistence of program impacts over time, we also report on teachers' survey responses in year 2.

Three teacher outcomes are examined:

[3] The research team expected some teacher attrition from the data teams between years 1 and 2, given that about 15 percent of teachers in any given year do not return to the same school the following year (Keigher 2010; Leukens et al. 2004).

- Teacher-reported frequency of use of individual and collaborative data-use practices
- Teachers' knowledge and skills pertaining to data use for instructional improvement (level of "data literacy")
- Teacher-reported attitudes and beliefs about the value of data to improve instruction

2.1.2. Student Outcomes

One student outcome is examined:

- Students' performance on the state-administered end-of-grade mathematics assessment

The focus of this evaluation is on grade 4 and grade 5 students in SY 12–13, at the end of year 2, when students are assigned to schools and teachers who have been exposed to and practiced using all of the UD program materials.

Analysis of this year 2 student sample provides an estimate of the *Using Data* program's schoolwide impact on grade 4 and grade 5 mathematics achievement. Because only some teachers in treatment schools participated directly in the UD program, the schoolwide estimate captures the weighted average of the program effects on achievement for students who had a study teacher for math instruction in year 2 and students who did not.[4]

2.2. Sample Recruitment

2.2.1. Sample Recruitment Strategies

We sought a single large urban district to evaluate the intervention. This limited variation in district context, while providing us with an adequate sample size to rigorously evaluate the program. In addition, this choice allowed us to control costs. The recruited district, DCPS, strongly supported our work, and took the lead on recruiting schools

[4] For both teachers and students, data were collected at baseline and at the end of years 1 and 2. Exploratory analyses will examine year 1 data for students, providing a thorough analysis of the available data and giving additional insight into the conditions under which the *Using Data* program can provide meaningful effects.

within it to participate. Due to positive prior experiences using other of TERC's professional development services, as well as strong support for data use, the district was eager to be involved.

The superintendent of DCPS worked with the cluster chiefs (assistant superintendents) in January and February 2011 to gauge interest among the elementary schools in participating. The research team was assured that at least 80 of the 102 school were interested. After receiving IRB approval, the study team gave a presentation to school administrators on May 26, 2011. Representatives from 61 schools attended this meeting, where UD's director of professional development and the study's principal investigator described *Using Data*, the incentives participating schools would receive, and the responsibilities of the schools and data teams.

2.2.2. Respondent Incentives

Participating schools assigned to the treatment group received the face-to-face professional development intervention and all associated materials; those assigned to the control group were offered *Using Data*'s online course, to be accessed after the data collection for the impact study was completed. In each of the two years of the study, participating schools, principals, teachers, and data coaches received stipends to compensate them for the time required by the study and to encourage their continuation. Substitute teachers also were provided for treatment teachers attending year 1 professional development events during the school year.

2.2.3. Application Procedure

Principals of schools wishing to participate in the study filled out an application with their name and the names, grade levels, and positions of five educators from their school who agreed to participate in the study. Consent forms, signed by the principal, teachers, and designated data coach for each school, were required. Once all forms were obtained by the research team, the principal received a confirmation email that the school was eligible for inclusion in the study as either a treatment or control-group school. This process resulted in 60 valid and complete applications by the time of randomization.

2.2.4. Recruiting Challenges

Despite the large number of elementary schools in DCPS, many were small schools with only a single grade 4 or 5 teacher. Therefore, some interested schools could not be included in the study.

Somewhat larger schools were able to participate by substituting a teacher from another grade level or subject area. Of the 182 teachers in the year 1 intent-to-treat[5] (ITT) sample, 22 teachers in 19 schools had a primary job assignment as a grade 1, 2, or 3 teacher; another 5 teachers in 4 additional schools were primarily reading or science teachers. Of the 142 teachers in the year 2 intent-to-treat sample, 15 teachers in 12 schools had a grade 2 or 3 assignment; another 4 teachers in 3 additional schools were primarily reading teachers.

2.3. Sample Design

2.3.1. Process of Random Assignment

The study design is a block-randomized experiment with two levels of treatment—*Using Data* versus business-as-usual. Randomization took place at the school level, because within a school, data team members would need to take on the same assignment as their school (treatment or control). Moreover, the UD intervention is designed to alter the school culture by creating collaborative dialogue around the examination of data about systemic student learning problems and potential solutions. Treatment teachers were expected to collaborate with other, nonparticipating teachers in their school by spreading knowledge, strategies, and perspectives on data use throughout the school.

After the 60 participating schools sent in all consent forms (principal, data coach, and four teachers), we compared these volunteer schools with schools that were not participating. To make these comparisons, we downloaded from the 2010 state report cards selected information for all elementary schools in DCPS: state math assessment scores for school year 2009–10, percentage of students who were African American, and percentage who were FRL eligible.

[5] The "intent-to-treat" teacher sample for a given year consists of all of the teacher volunteers already in the study at the time of randomization for whom we have outcome measures at the end of that year.

The comparison, presented in table 2.1, shows the participating schools had fewer African American (47.4 percent compared with 54.4 percent) and fewer FRL-eligible (65.7 percent compared with 70.9 percent) students.

Table 2.1. Comparison of DCPS Schools, School Year 2009–10 Data

School characteristic	In study (n=60)	Not in study (n=42)	All schools (n=102)
% Title I (s.d.)	28.3 (.45)	38.1 (.49)	32.4 (.47)
% African American (s.d.)	47.4 (27.3)	54.4 (30.6)	50.5 (28.7)
% FRL eligible (s.d.)	65.7 (23.0)	70.9 (22.7)	68.4 (22.5)
Math scale score (s.d.)	1565.2 (102.8)	1558.0 (103.0)	1562.3 (102.4)

Table 2.1 also shows significant diversity among study schools. To ensure a balanced sample across the treatment and control groups for these characteristics, we used a block-randomized design, with four blocks based on an index that took into account the variables shown in table 2.1. These variables are intended to measure the extent of schools' "needs" at baseline. We ran a principal components analysis (PCA) on the 60 participant schools using the four variables. The first principal component extracted accounted for 77.45 percent of the variance.

Regression-based factor scores were saved, and schools were grouped by quartile on the factor score. Schools with the lowest scores were all non–Title I schools, had lower percentages of African American and FRL-eligible students, and had higher math scores. Descriptive statistics by factor score quartile are displayed below, in table 2.2.

Table 2.2. Block Averages for Key Variables Measuring School "Needs," Study Schools

School characteristic	Block 1 (n=14)	Block 2 (n=14)	Block 3 (n=16)	Block 4 (n=16)
% Title I (s.d.)	0.0 (0.0)	0.0 (0.0)	12.0 (34.2)	94.0 (25.0)
% African American (s.d.)	17.2 (6.8)	30.9 (9.5)	51.8 (12.7)	83.7 (10.9)
% FRL eligible (s.d.)	33.3 (11.9)	57.5 (9.6)	77.6 (7.6)	89.4 (5.2)
Math scale score (s.d.)	1694.3 (63.4)	1603.7 (51.6)	1500.2 (53.3)	1483.6 (65.9)

The schools' level of needs increases from Block 1 to Block 4. The table shows that Block 1 schools had, at baseline, no Title I members, the lowest percentages of students who are African American and FRL eligible, and the best schoolwide performance on the state math assessment. At the other end of the scale, Block 4 schools at baseline were 94 percent Title I, had the highest percentages of African American and FRL-eligible students, and had the lowest performance on the state math assessment. Block 1 schools, therefore, can be thought of as relatively the "lowest-needs" schools and Block 4 schools as the relatively "highest-needs" schools among those in the study.

The random assignment process was conducted at CNA under the direction of the principal investigator, in June 2011. Random numbers were assigned to all schools prior to assigning schools to blocks. Half the schools in each block were assigned to either the treatment or control group based on their random number, with the highest half of each block assigned to treatment.

2.3.2. Descriptive Statistics for the Randomized Sample of Schools

Descriptive statistics for the treatment and control-group schools on the four variables of interest are displayed in table 2.3, below. There are no statistically significant differences at baseline between the treatment and control groups for any of the four school-level characteristics.

Table 2.3. Descriptive Statistics: Study Schools by Treatment Status

School characteristic		Treatment group (n=30)	Control group (n=30)	Std. Mean Diff.
% Title I school	Mean	30.0	26.7	0.07
	Std. deviation	46.6	45.0	
	Std. error mean	0.1	0.1	
% African American	Mean	47.9	46.8	0.04
	Std. deviation	25.5	29.4	
	Std. error mean	4.7	5.4	
% FRL eligible	Mean	65.6	65.9	–0.01
	Std. deviation	21.0	25.2	
	Std. error mean	3.8	4.6	
Math scale score	Mean	1572.1	1558.3	0.13
	Std. deviation	87.8	117.0	
	Std. error mean	16.0	21.4	

2.4. Sample Attrition, Crossovers, and Noncompliance

2.4.1. School Attrition and Noncompliance

The intent-to-treat sample for schools consists of all study schools for which we had outcome data, regardless of whether the UD professional development occasions were attended (for treatment schools) or the data teams met as intended. Because we randomized the sample at the school level, we carefully considered the consequences of school attrition; however, we did not expect much, due to the high level of district support.

From the perspective of the student analysis, no schools were lost in year 1. However, one control school closed after year 1, reducing the number of schools in year 2 to 59. For the teacher analysis, one treatment school was lost from the ITT sample in year 1, because the principal created a new data team, by replacing after randomization all of the originally assigned teachers. These replacements took place prior to the start of the intervention at the beginning of the school year.

In year 2, one treatment school was noncompliant; the principal agreed to provide student and teacher outcome data, but said the school was too busy to continue the *Using Data* activities. The principal also noted that the school was continuing to use data for instructional improvement. Because we were able to collect data for this school, it remains in the ITT sample.

2.4.2. Teacher Attrition and Noncompliance

The year 1 intent-to-treat sample for teachers consists of all participant teachers who were applicants in the originally randomized set of study schools and for whom we have outcome measures at the end of year 1.

Principals were encouraged to select data team members who would likely remain in the school and be committed to the *Using Data* professional development. After randomization, six schools made changes to their data team because the original teacher participants changed grade levels or schools. In addition, four treatment schools changed their data coach because the original data coach was promoted and no longer available to take part in the study. (One was promoted to principal of a treatment school,)

The CONSORT table below (figure 2.1) displays the attrition levels for schools and teachers in the treatment and control groups during year 1. At the school level, overall attrition was 2 percent and differential attrition was 3 percentage points between the treatment (3 percent) and control (0 percent) groups. As figure 2.1 shows, 19 teachers were lost from the control group (18 left the study, for 1 we did not have outcome data), while 32 teachers were lost from the treatment group (30 left, for 2 we did not have outcome data).

At the teacher level, overall attrition was 22 percent, and differential attrition between control (16 percent) and treatment (27 percent) groups was 11 percentage points. This level of attrition exceeds the "optimistic" threshold level set in the What Works Clearinghouse's *WWC Procedures and Standards Handbook*.

Although our primary interest for teachers is in the ITT sample at the end of year 1, we continued to track the sample in year 2. Between year 1 and year 2, an additional 40 of the original teachers assigned at randomization left the study, so that by the end of the study the

intent-to-treat sample numbered 142 (62 treatment, 80 control). This corresponds to an overall attrition rate of 22 percent (30 percent treatment, 15 percent control) for year 2, the same overall rate as in year 1.

To maintain the integrity of the data teams, principals were asked to replace teachers who left the study. These replacements occurred at one of two times: during the summer of 2011, after randomization but before any professional development had been conducted (these "year 1 replacements" numbered 37; 27 treatment and 10 control teachers); and during the summer of 2012, between year 1 and year 2 of the study (these "year 2 replacements" numbered 33; 23 treatment and 10 control teachers).

The year 1 replacements received the full two years of *Using Data* training; whereas the year 2 replacements received only year 2 training. Neither year 1 replacements nor year 2 replacements are included in the teacher intent-to-treat samples.

Figure 2.1. CONSORT Table, Teacher Intent-to-Treat Sample, Year 1

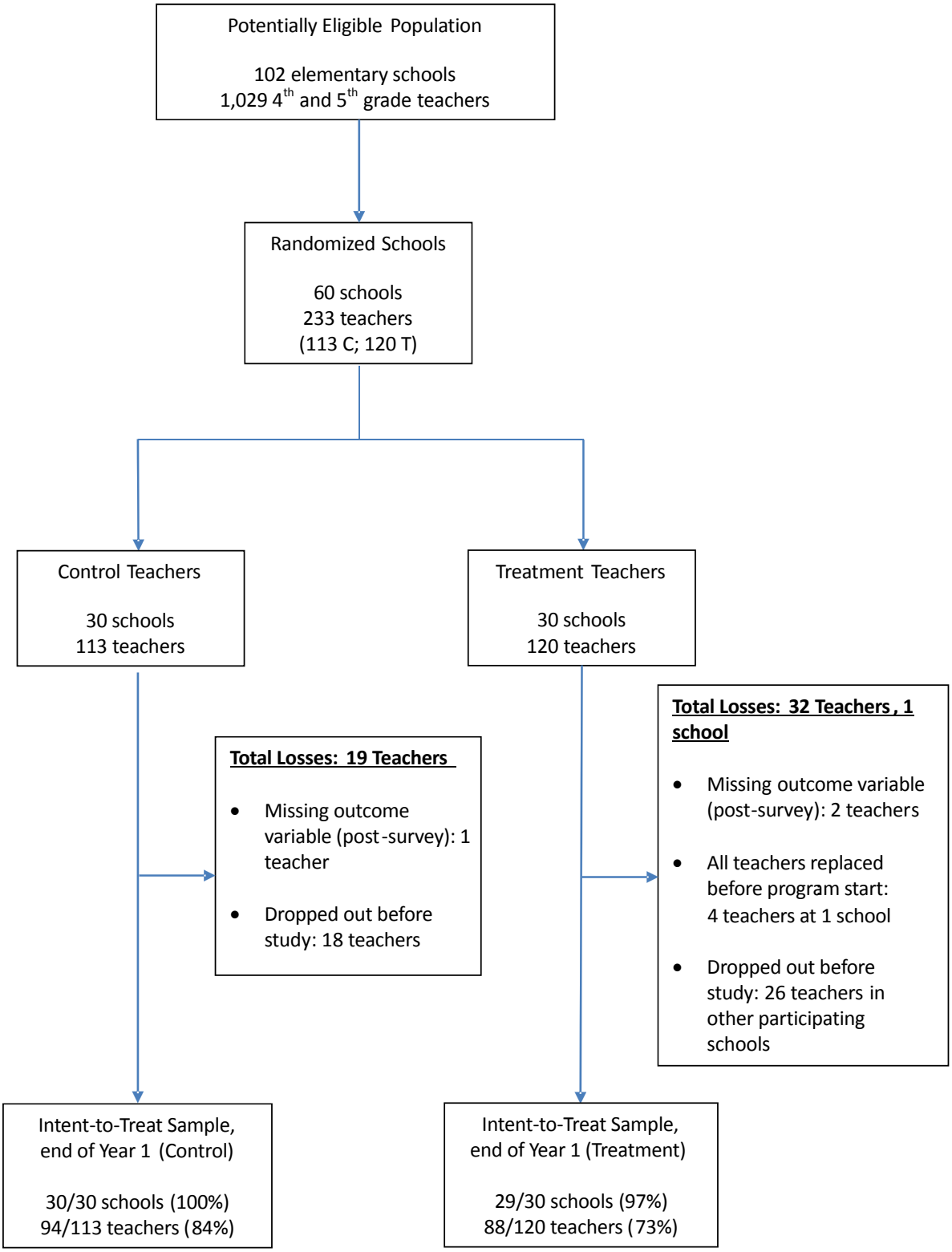
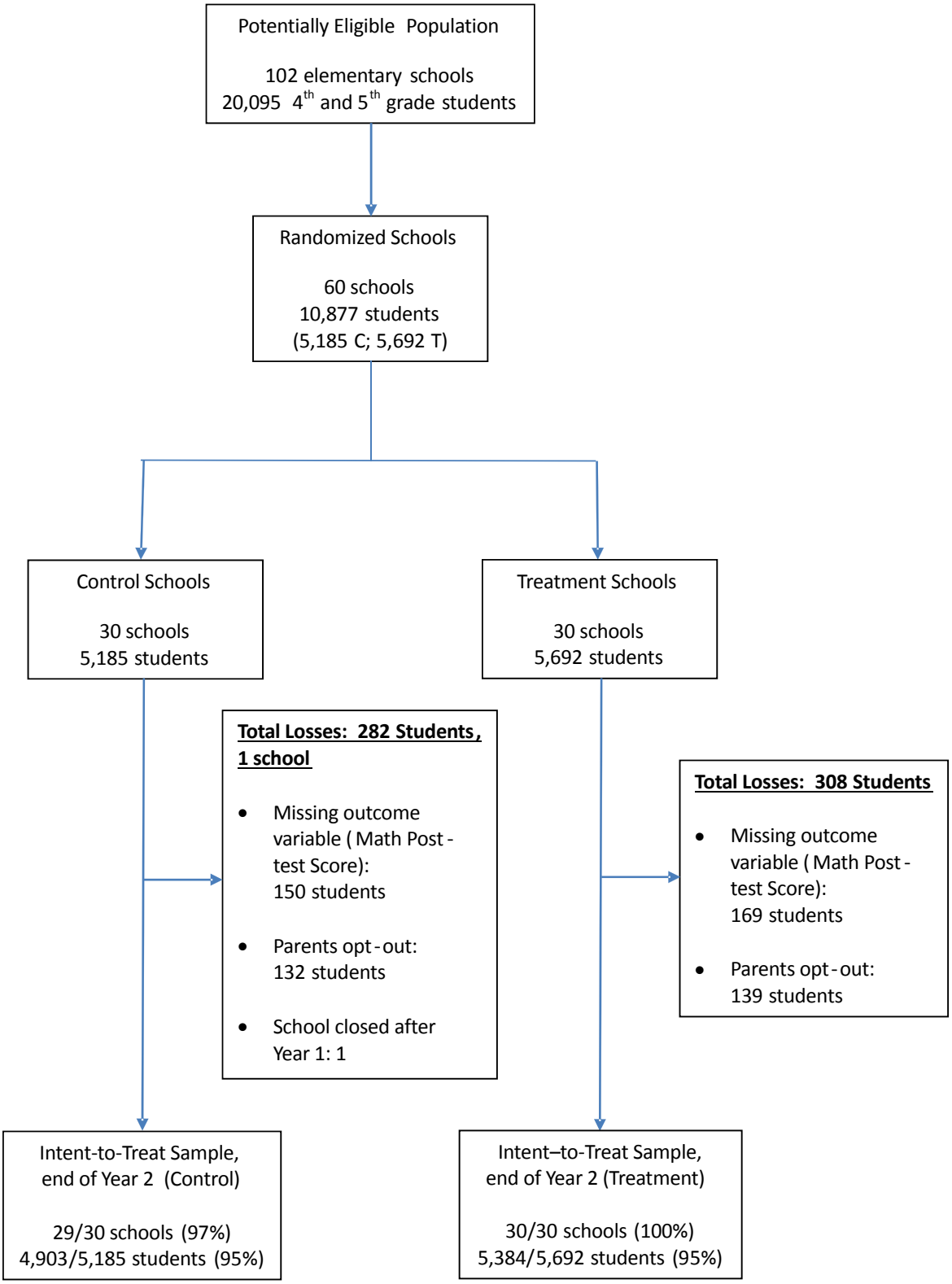


Figure 2.2. CONSORT Table, Student Sample, Year 2



2.4.3. Student Attrition

Our expectation is that impacts on students will not be observed until year 2, because year 1 is a learning year for teachers. Therefore, our primary confirmatory analysis for program effects on student achievement is based on the sample of grade 4 and 5 students in year 2. The second CONSORT table (figure 2.2, above) shows school and student attrition levels for the sample that includes both grade 4 and 5 students at study schools.

At the student level, 271 students had parents who signed opt-out forms, and another 319 students did not have an end-of-year state assessment math score,[6] so they dropped out of the sample. The resulting overall attrition rate was 5 percent, equally distributed between treatment and control schools.

These levels of attrition are acceptable by WWC standards, even under conservative assumptions.

2.5. Descriptive Statistics for the Analytic Samples

2.5.1. Descriptive Statistics and Baseline Equivalence of Teacher Sample

Table 2.4 provides descriptive statistics for the year 1 ITT teacher sample. Because teacher-level attrition in year 1 was high (17 percent for the control group, and 27 percent for the treatment group), we have to ensure baseline equivalence of model covariates for the treatment and control groups in order for our results to meet WWC standards (with reservations).

Using Hedges' *g* as the standardized mean difference measure, the equivalence analysis shows that the differences between treatment and control-group scores on all three of the pretests, the experience covariate, and all three of the missing pretest flags are all above .05 in absolute value, but below .25.

[6] In cases in which no math posttest was available for a given student, this could be because the student had left the district to attend a school elsewhere, or because the student was absent on testing days.

Table 2.4. Descriptive Statistics: Teacher Characteristics, Teacher Intent-to-Treat Sample, Year 1

Variable	Study condition	N	Mean	Std. dev.	Range		Std. mean diff.
					Min.	Max.	
Data Use pretest score (SY 10–11)	Treatment	71	-0.01	0.97	-3.19	2.95	0.05
	Control	78	-0.06	1.05	-2.83	2.72	
Knowledge and Skills pretest score (SY 10–11)	Treatment	88	-0.05	1.18	-2.05	3.17	0.05
	Control	92	-0.10	0.87	-2.05	3.11	
Attitudes and Beliefs pretest score (SY 10–11)	Treatment	88	-0.19	0.82	-3.21	2.53	-0.09
	Control	92	-0.10	1.05	-2.25	3.95	
Experience	Treatment	87	10.2	7.5	0	35	-0.16
	Control	92	11.4	7.8	2	36	
Missing Data Use pretest score	Treatment	88	0.193	0.397	0	1	0.06
	Control	94	0.170	0.378	0	1	
Missing Knowledge and Skills pretest score	Treatment	88	0.000	0.000	0	0	-0.20
	Control	94	0.021	0.145	0	1	
Missing Attitudes and Beliefs pretest score	Treatment	88	0.000	0.000	0	0	-0.20
	Control	94	0.021	0.145	0	1	

Following WWC standards, we control for these covariates in the impact analysis to remove any bias caused by this difference, to reduce unexplained variation in outcomes, and to improve the precision of impact estimates.

Table 2.5. Descriptive Statistics: School Characteristics, Teacher Intent-to-Treat Sample, Year 1

Variable	Study condition	N	Mean	Std. dev.	Range		Std. mean diff.
					Min.	Max.	
% Title I	Treatment	29	27.6	45.0	0	1	0.02
	Control	30	26.7	45.5	0	1	
% African American	Treatment	29	46.3	24.3	13	92	-0.02
	Control	30	46.8	29.4	8	96	
% FRL eligible	Treatment	29	64.8	20.9	22	96	-0.05
	Control	30	65.9	25.2	10	96	
Math scale score	Treatment	29	1575.1	87.8	1418	1816	0.16
	Control	30	1558.3	116.9	1335	1783	

Table 2.5, above, provides descriptive statistics of the school-level variables for the year 1 teacher ITT sample. There are differences in characteristics, such as the schoolwide state math test score, that exceed the WWC standard of .05 standard deviations. We adjust for this by controlling for school block (which is determined by each of these four variables) in our teacher analysis regressions.

2.5.2. Descriptive Statistics of Year 2 Student Sample

The year 2 student sample consists of all grade 4 and grade 5 students in SY 12–13 from the 59 treatment and control schools for whom we have an end-of-year state math test score. As in year 1, data on student test scores and student characteristics for the student sample were provided by DCPS. Table 2.6, below, provides descriptive statistics for students in treatment and control schools during year 2.

Table 2.6. Descriptive Statistics: Student Characteristics, Student Sample, Year 2

Variable	Study condition	N	Mean	Std. dev.	Range	
					Min.	Max.
Math scale score, SY 12–13	Treatment	5,384	217.6	20.2	155	279
	Control	4,903	218.3	21.4	155	279
Baseline math scale score (SY 10–11)	Treatment	2,213	346.9	57.6	100	500
	Control	1,914	346.1	63.5	100	500
Grade 4, SY 12–13	Treatment	5,384	0.488	0.500	0	1
	Control	4,903	0.507	0.500	0	1
Female	Treatment	5,384	0.494	0.500	0	1
	Control	4,903	0.486	0.500	0	1
African American	Treatment	5,384	0.441	0.497	0	1
	Control	4,903	0.387	0.487	0	1
Hispanic	Treatment	5,384	0.089	0.284	0	1
	Control	4,903	0.091	0.288	0	1
English language learner status	Treatment	5,384	0.053	0.225	0	1
	Control	4,903	0.064	0.244	0	1
FRL eligible	Treatment	5,384	0.601	0.490	0	1
	Control	4,903	0.596	0.491	0	1
Gifted student status	Treatment	5,384	0.046	0.210	0	1
	Control	4,903	0.049	0.216	0	1
Learning disability or condition status	Treatment	5,384	0.099	0.299	0	1
	Control	4,903	0.116	0.320	0	1
Missing prior math scale score	Treatment	5,384	0.593	0.491	0	1
	Control	4,903	0.613	0.487	0	1

An examination of the descriptive statistics presented in table 2.6 shows that the treatment and control groups are, on the whole, similar. However, the treatment group has a larger share of African American students, by 5.4 percentage points; whereas the control group has a larger share of grade 4 students, English language learners, and learning disabled students, by 1.9, 1.1, and 1.7 percentage points, respectively. We point out, as well, that there is a large share of missing pretest scores in both groups (about 60

percent). This is because a pretest score is not available at baseline (that is, in grade 2) for grade 4 students in year 2 of the study. Our statistical models below control for each of the student variables listed in table 2.6.

2.5.3. Descriptive Statistics of Year 1 Student Sample

An exploratory examination of year 1 data for students provides a thorough analysis of the available data and additional insight into the conditions under which *Using Data* can provide meaningful effects. Table 2.7 provides descriptive statistics for the year 1 student sample. As in year 2, year 1 student attrition levels were low (8 percent overall, 7 percent for treatment schools and 9 percent for control schools), so the internal validity of the research design is preserved.

Table 2.7. Descriptive Statistics: Student Characteristics, Student Sample, Year 1

Variable	Study condition	N	Mean	Std. dev.	Range	
					Min.	Max.
Math scale score (SY 11–12)	Treatment	5,408	219	20.6	155	279
	Control	5,195	220	22.0	155	279
Baseline math scale score (SY 10–11)	Treatment	5,104	339	56.9	100	500
	Control	4,870	339	61.7	100	500
Grade 4 in SY 2011–12	Treatment	5,408	0.503	0.500	0	1
	Control	5,195	0.481	0.500	0	1
Female	Treatment	5,408	0.490	0.500	0	1
	Control	5,195	0.485	0.500	0	1
African American	Treatment	5,408	.0445	0.497	0	1
	Control	5,195	.0376	0.484	0	1
Hispanic	Treatment	5,408	0.081	0.273	0	1
	Control	5,195	0.086	0.280	0	1
English language learner status	Treatment	5,408	0.058	0.233	0	1
	Control	5,195	0.063	0.243	0	1
FRL eligible	Treatment	5,408	0.589	0.492	0	1
	Control	5,195	0.554	0.497	0	1
Gifted student status	Treatment	5,408	0.034	0.181	0	1
	Control	5,195	0.044	0.206	0	1
Learning disability or condition status	Treatment	5,408	0.109	0.312	0	1
	Control	5,195	0.119	0.323	0	1
Missing prior math scale score	Treatment	5,408	0.056	0.230	0	1
	Control	5,195	0.063	0.242	0	1

This page intentionally left blank

3. Data and Methods

Our *primary confirmatory analyses* of the effects of the *Using Data* intervention on teacher behavior and student achievement are as follows:

- For **teacher behavior**, the analyses of the effect of *Using Data* on teacher data use, knowledge and skills, and attitudes and beliefs, based on teacher self-reports at the end of year 1
- For **student achievement**, the analyses of the effect of *Using Data* on student performance on the end-of-year state math assessment at the end of year 2

In addition, we conduct a number of *exploratory analyses* with respect to student achievement:

- **School context**—an examination of whether estimated *Using Data* treatment effects on student achievement differ by school "block" (the school's level of socioeconomic need and schoolwide performance on the state math assessment at baseline)
- **Dosage**—an exploration of the effect on student achievement of the school having teachers with different levels of exposure to the *Using Data* program, and of students with different levels of exposure to *Using Data*–trained teachers

3.1. Quantitative Analysis

3.1.1. Teacher Data Use Behavior Analysis

The research questions we seek to answer with respect to teacher behavior are as follows:

Compared with study teachers in control-group schools, do study teachers in treatment schools at the end of intervention year 1 –

1. Report more frequent use of data? (**data use model**)

2. Have greater data-use knowledge and skills? (**knowledge and skills model**)

3. Report more positive attitudes and beliefs about the value of data to inform instruction and improve student learning? (**attitudes and beliefs model**)

This year 1 teacher analysis represents our primary confirmatory analysis of the effects of the *Using Data* program on teacher behavior with respect to data use for instructional improvement.

Estimation Method

To answer the research questions regarding the effect of *Using Data* on teacher behavior, we will estimate a two-level hierarchical linear model (HLM) with teachers nested in schools:

Level 1 model (teacher level): $Y_{ij} = \pi_{0j} + \pi_{1j} Y^{*}_{ij} + \pi_{2j} X_{ij} + \pi_{3j} (X_{ij})^2 + e_{ij}$

Level 2 model (school level): $\pi_{0j} = \beta_{00} + \beta_{01} T_j + \Sigma\ \beta_{0n} S_{nj} + r_{0j}$

Variable list:

Y_{ij} is the score on the teacher outcome for teacher i, in school j at the end of year 1

π_{0j} is the mean outcome for teachers in school j

Y^{*}_{ij} is the pretest score for teacher i in school j

X_{ij} is the number of years' experience the teacher has with the district (note the equation includes a squared term)

e_{ij} is the error term associated with each teacher

β_{00} is the school-level intercept for school j

β_{01} is the coefficient of interest, which measures the treatment effect

T_j is the school-level treatment indicator, where 1 represents treatment status and 0 represents control status

S_{nj} is a vector that represents the block the school was assigned to prior to randomization (n = 2, 3, 4 corresponding to blocks 1, 2, and 3, respectively)

r_{0j} is the school-specific random effect

Dependent variables:

The same model design will be used to address each teacher question; however, the dependent variable will differ by research question:[7]

- Data use—the scale score on the 16-question Data Use survey, taken at the end of year 1
- Knowledge and skills—the scale score on the 25-question Knowledge & Skills survey taken at the end of year 1
- Attitudes and beliefs—the scale score on the 37-question Attitudes & Beliefs survey taken at the end of year 1

Control variables:

For each of the three outcomes corresponding to the three research questions, we will control for

- the teacher's baseline pretest scale score for that outcome;
- the teacher's baseline Knowledge & Skills pretest score in both the attitudes and beliefs and the data-use models (based on our sample equivalence analysis; see below);
- the pre-randomization block to which the school was assigned, to account for the school's context in terms of racial/ethnic minority and low-income students served, and baseline schoolwide student achievement; and
- number of years' experience the teacher has with the district.

The coefficient of interest is β_{01}, the treatment effect. Because we have established baseline equivalence of the treatment and control groups (see below) we are confident that this part of the analysis will meet WWC evidence standards (with reservations).

We calculated a minimum detectable effect size (MDES) of .42 based on a power analysis for the two-level HLM model, the block design, the indicated covariates, and the sample size of 59 schools and 182 teachers (see appendix C).

[7] Item response theory (IRT)–based scale scores for the dependent variables were developed from the raw scores.

3.1.2. Student Achievement Analysis

The research questions we seek to answer with respect to student achievement are as follows:

After the second year of program implementation –

1. Do grade 4 and 5 students attending treatment schools have higher levels of mathematics achievement than grade 4 and 5 students in control schools?
2. Do treatment effects vary by school context (socioeconomic need and mathematics performance at baseline)?
3. Do treatment effects vary by level of exposure students have to UD-trained teachers?

Research question 1 and 2 are experimental. Research question 3 is quasi-experimental, because students were assigned to teachers within schools after the initial randomization of schools.

Answering research question 1 leads to an overall impact estimate that is a weighted average of the direct effect on students instructed by *Using Data*–trained teachers *and* spillover effects, if any, on students attending treatment schools but whose teachers are not directly receiving *Using Data* professional development. Addressing research question 2 examines the effect on grade 4 and 5 students of having teachers in year 2 with different levels of exposure to UD training.

From the standpoint of an efficacy trial, the treatment effect estimate on students of study teachers (research question 2) is of particular interest, because it allows us to measure the effectiveness of UD without potential loss of effect associated with diffusion to teachers who were not directly engaged in the professional development.

In addition, for the year 2 analysis we augment the overall impact model by estimating treatment effects that are specific to school blocks. (Recall that prior to randomization, schools were assigned to one of four blocks based on the socioeconomic status of their students and schoolwide performance on the state mathematics assessment at baseline.) Block 1 represents the group of schools with the lowest level of socioeconomic need and the highest academic

performance at baseline, whereas Block 4 is the group of schools with the highest level of socioeconomic need and the lowest academic performance at baseline. We refer to this model as the school context model.

Estimation Method

The schoolwide analytical model is a two-level hierarchical linear model (HLM) with students nested in schools, and a school-level random effect:

(Students) Level-1 Model: $Y_{ij} = \beta_{0j} + \Sigma \beta_{mj} X_{mij} + \varepsilon_{ij}$

(School) Level-2 Model: $\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + \Sigma \gamma_{0n} S_{nj} + \mu_{0j}$

Variable list:

Y_{ij} is the math scale score for student i in school j on the state test in grade 4 or 5 in school year 2012–13 (end of year 2)[8]

X_{mij} is a set of (m = 1, . . . , 8) student-specific baseline covariates:

- Pretest score at baseline (state math assessment, SY 10–11)[9, 10]
- Racial/ethnic group indicators, one each for
  - African American
  - Hispanic
- FRL eligibility indicator
- English language learner indicator
- Gifted indicator
- Learning disability/condition indicator
- Missing pretest score indicator

[8] Within a school-year cohort, we pool grade 4 and grade 5 students in the same model. Grade 4 and grade 5 students take grade-specific versions of the state math assessment that are vertically scaled. We include as a covariate in our regression models a variable that indicates whether the student is in grade 4 or 5, so students are being compared only with students who took the same grade-level test.

[9] Note that state math assessment scores were re-scaled after the 2010–11 school year. This explains the difference in table 2.6 in the sample means between the prior score from 2010–11 and the outcome score from 2011–12.

[10] For the year 2 (SY 12–13) student achievement analysis, pretest scores are taken at baseline, two years earlier (SY 10–11). For grade 5 students in 2012–13, the pretest is their grade 3 math assessment score from 2010–11. We do not have pretest scores for grade 4 students in 2012–13.

T_j indicates whether the student attended in year 2 a school that was assigned at randomization as a treatment school (T=1) or a control school (T=0)

S_j is a set of three dummy variables (n = 1, . . . , 3) corresponding to three of the four blocks to which schools were assigned pre-randomization

β_{0j} is the outcome mean for students in school j

γ_{00} is the school-level intercept for school j

μ_{0j} is a school-level random effect

ε_{ij} is the student-level error term

The coefficient of interest is γ_{01}, which captures the effect of the program on students after two years of program implementation in schools, compared with students in control schools in year 2. Effect sizes based on estimated treatment effects is calculated using Cohen's *d*.[11]

A power analysis conducted based on the two-level HLM, the block design, the indicated covariates, and the sample size of 10,287 students in 59 schools calculated an MDES of .17 for the year 2 impact model (see appendix C).

Primary Confirmatory Analysis

1. (Primary Confirmatory Analysis) **Schoolwide impact model—**to estimate the school-level treatment effect, we use a two-level HLM with students nested in schools. This is our primary confirmatory analysis with respect to the effects of *Using Data* on student achievement.

Exploratory Analyses

2. (Exploratory Analysis) **School context model—**to estimate the block-specific school-level treatment effects, we add to the two-

[11] Cohen's *d* takes the treatment effect estimated in the regression model (the "coefficient of interest") and converts it to standard deviation units by dividing the coefficient by the pooled (across treatment and control groups) standard deviation of the outcome variable. For the student achievement model described above, $d = \gamma_{01}/\{[(n_T - 1)s_T^2 + (n_C-1)s_C^2]/(n_T + n_C - 2)\}^{½}$ where γ_{01} is the estimated treatment effect ("coefficient of interest"); s_T and s_C represent the standard deviations of the outcome variable for the treatment and control groups, respectively; and n_T and n_C represent the sample size for the treatment and control groups.

level HLM of schoolwide impact interaction terms between the treatment effect and membership in Block 1, 2, or 3.

3. (Exploratory Analysis) **Dosage model—**to estimate the teacher-level treatment effect of instruction by a *Using Data*–trained teacher.

Year 1 Student Achievement Analysis (Exploratory Analysis)

For student achievement in year 1 of the study, we address research questions analogous to those in year 2. We estimate two models:

1. **Schoolwide model—**as described above, to estimate an overall treatment effect

2. **Dosage model—**this model includes treatment indicators at both the school level and the teacher level to estimate separate treatment effects at each level. The teacher-level treatment indicator equals 1 if the teacher is either a member of the study's ITT sample or a teacher who replaced an ITT teacher who dropped out of the study.[12] This is a three-level HLM with students nested with teachers nested in schools, with school-level and teacher-level random effects.

3.2. Data Sources for the Quantitative Analysis

3.2.1. Source Files

We constructed the teacher and student datasets from four source files.

Student Demographic Variable File (CNA2012DemoAssessCrsz(NoConsentRemoved).xlsx)

This file contains administrative data provided by DCPS, including the following:

- A unique student identification (ID) number for each student
- The student's assigned school, by name and four-digit school ID number
- Grade-level assignment
- Demographic information, including

[12] All year 1 replacement teachers entered the program prior to its start.

- Gender
- Racial/ethnic background
- English language learner status
- FRL eligibility
- Learning disability or condition status (including gifted status)
- The local (school-specific) teacher identification number of the student's mathematics teacher

Student Assessment File (CNA-2012AssessCrosswalk.xslx, tab 1)

This file, also provided by DCPS, contains students' state math assessment scale scores for SY 10–11 and SY 11–12, along with the unique student ID number. The district provided an additional file in September 2013 containing the SY 12–13 math scale scores.

Local-District Teacher ID Crosswalk File (CNA-2012AssessCrosswalk.xslx, tab 2)

This file, provided by DCPS, matches each teacher's school-specific teacher ID number to her or his district-level personnel number.

Teacher Participants Data File (UsingData_MasterTeacherSpreadsheet4.22.2013.xlsx)

This file contains information on teachers in the study, some data provided by the district but most collected by study researchers. Information includes the following:

- Researcher-collected data:
  - Teachers' school assignment at randomization and at year 1 and year 2
  - Study condition (treatment, control) at randomization, year 1 and year 2
  - Date at which the teacher entered the study
  - Date at which the teacher was dropped (if at all)
  - Timing and nature of any changes in study status/condition for teachers who switched schools

- — Responses to Attitudes & Beliefs and Knowledge & Skills surveys for the baseline administration (Fall 2011), year 1 administration (Spring 2012), and year 2 administration (Spring 2013)
- — Summary variables constructed from survey responses: Attitudes & Beliefs, Knowledge & Skills, and Data Use

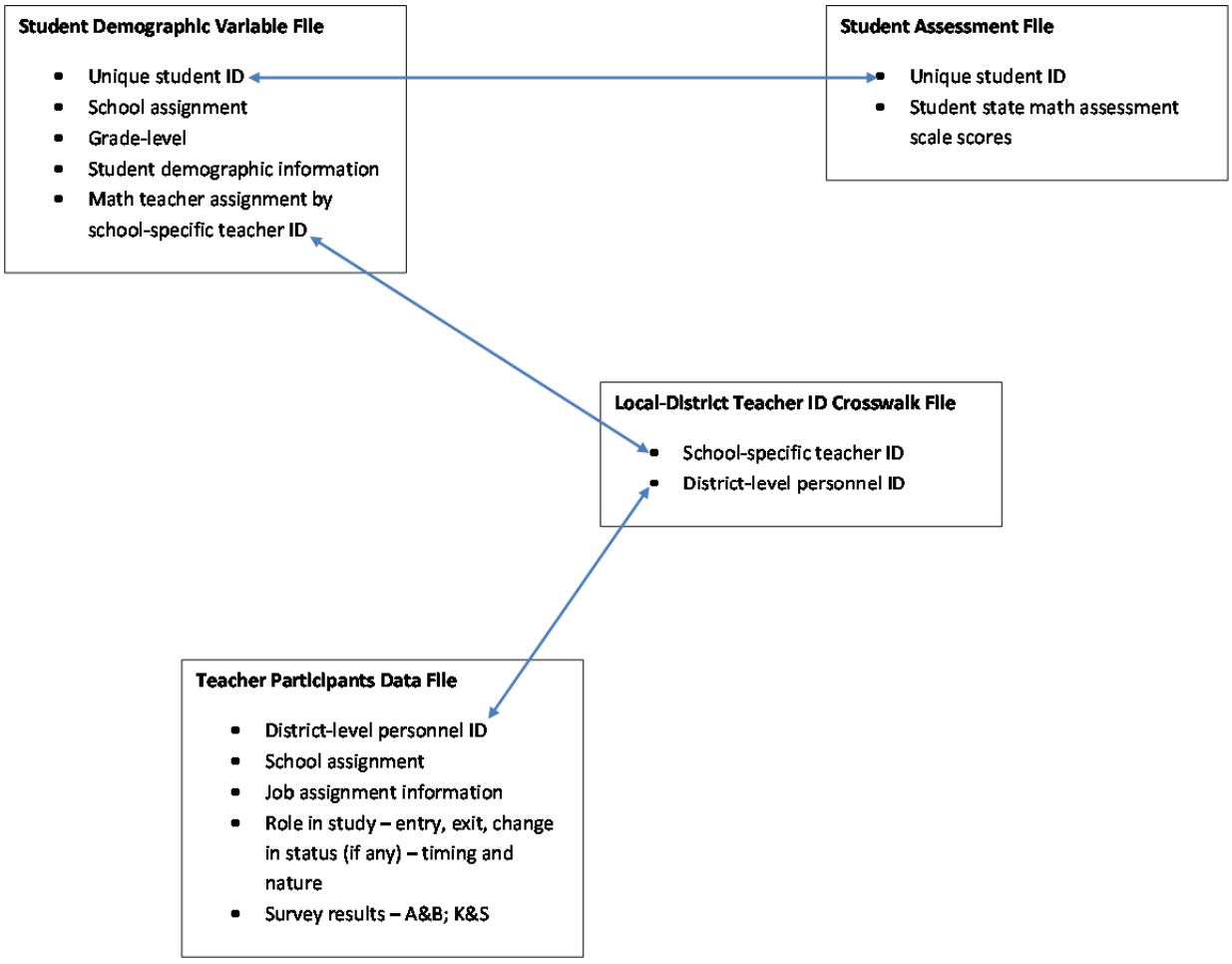
- DCPS-provided administrative data for each teacher:
  - — Grade assignment[13]
  - — Years of experience
  - — District-level personnel number

This Teacher Participants Data File serves as the dataset for the teacher analysis.

To create the student dataset, students' demographic and assessment data were merged using the unique student ID number, which is included in both student files. Student records could then be linked to teacher information by using the Local-District Teacher ID Crosswalk File, which matches the school-specific teacher identification number included in the student files to the district-level personnel number included in the Teacher Participants Data File. The process is illustrated in figure 3.1, below, with the arrows indicating the specific data fields that link the individual data files together.

[13] Students in the sample have a single grade 4 or 5 teacher for most subjects.

Figure 3.1. Student Dataset: Process for Matching Student and Teacher Information



3.2.2. Teacher Surveys

For the teacher behavior analysis, two questionnaires were developed by the research team for this study. The questionnaires are designed to measure three distinct outcomes: data use, knowledge and skills pertaining to effective data use, and attitudes and beliefs about the value of data analysis as an effective tool for instructional improvement. Sample questions from the questionnaires are provided in appendix A.

Knowledge & Skills Survey

The Knowledge & Skills questionnaire was developed by the research team to measure a teacher's knowledge of data-use concepts and

application of data-use skills when interpreting tables and graphs. Items were written to measure general data literacy, knowledge of types of data and appropriate use (aggregate data, disaggregate data, strand data, item data), and processes when interpreting data displays. A respondent's raw score on the 25-question instrument is the number of correct responses out of 25.

A reliability analysis conducted in the summer of 2013 indicated that Cronbach's alpha, a statistic calculated to indicate how consistently sets of items measure an underlying construct, was equal to .67, which exceeds the What Works Clearinghouse minimum reliability standard of .50.[14] At the same time, a validity study that compared the Knowledge & Skills scale with that of the Number Concepts and Operations (NCOP) subtest of the Learning Mathematics for Teaching assessment (LMT) was conducted.[15] The analysis showed that the Knowledge & Skills score had a moderate correlation (.42) with the LMT item response theory (IRT) scale, confirming the validity of the Knowledge & Skills scale.[16]

Attitudes & Beliefs Survey

The Attitudes & Beliefs questionnaire was developed by the research team to measure teacher agreement with core assumptions of the *Using Data* process,[17] the basis of the *Using Data* treatment intervention with teachers. The researchers wrote five to seven items for each of six dimensions, or subscales, of what the UD program calls "collaborative inquiry": *encouraging a collaborative culture among teachers, equitable treatment of diverse students, nurturing trust, collaborating with other teachers, using data to focus on learning problems of students,* and *instructional improvement.* In a pilot test conducted in Fall 2011, five of the six subscales had internal consistency reliabilities ranging from .70 to .91, well above the What Works Clearinghouse minimal

[14] See Cronbach (1951) and What Works Clearinghouse, *Procedures and Standards Handbook (Version 3.0),* http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19.

[15] See appendix B for more information about the LMT assessment.

[16] Note that the Knowledge & Skills scale and the LMT were not designed to capture the same content. Knowledge & Skills provides a measure of a teacher's data literacy, whereas LMT attempts to measure a teacher's mathematical knowledge that can be applied to teaching.

[17] See the *Using Data* project website at http://usingdata.terc.edu.

standard of .50. One subscale, Instructional Improvement, had a reliability of .31.

The Attitudes & Beliefs score was derived from the responses to 37 questions on the survey measuring teacher agreement with *Using Data* core assumptions. Responses were measured on a Likert-type scale, with 1 corresponding to strong disagreement with the statement, and 5 corresponding to strong agreement. The Attitudes & Beliefs raw score was derived from the sum of the respondent's responses on the Likert scale for each question, for a maximum possible score of 185.[18]

We developed the Data Use raw score from a separate set of responses to 16 questions on the Attitudes & Beliefs survey. Each question represents a specific data-use activity. These questions were taken from the U.S. Department of Education's National Educational Technology Trends Study (NETTS).[19]

There are four possible responses for each question/activity, corresponding to the frequency with which the activity was conducted over the course of the previous school year. We scored these responses as follows: never (0 points), a few times (0.5 points), once or twice a month (1.5 points), and once a week (4 points). The points therefore roughly correspond to the approximate frequency per month with which the respondent reports performing the activity. The Data Use raw score is the sum of these points across the 16 questions, for a maximum possible score of 64.

A reliability analysis conducted in the summer of 2013 on both the Attitudes & Beliefs and Data Use scales indicated that both of the items had acceptable reliability. Cronbach's alpha was equal to .95 for the Attitudes & Beliefs scale and .87 for the Data Use scale. Again, this exceeds the WWC minimum standard of .50.

---

[18] Of the 37 questions, 13 were worded in such a way that stronger agreement with the statement actually represented stronger disagreement with *Using Data* core principles. The scoring on these questions was reversed so that higher scores were consistent with UD principles.

[19] U.S. Department of Education, National Educational Technology Trends Study, Local-level Data Summary, 2008.

The questionnaires were administered to study teachers in Fall 2011, Spring 2012, and Spring 2013. The Fall 2011 administration was given just prior to the start of the professional development program. The posttest of the questionnaires administered in Spring 2012 included all of the available teachers who had been randomly assigned to the treatment group the previous fall, as well as replacement teachers. However, only randomized teachers are included in the teacher analysis.

See appendix A for sample questions from the questionnaires and appendix B for psychometric analysis of their reliability and validity.

3.2.3. Learning Mathematics for Teaching Assessment (LMT)

To assess validity of the Knowledge & Skills and Attitudes & Beliefs scales, we used an established measure of a teacher's content knowledge of number concepts and operations (NCOP) for elementary school. This established measure is a subscale of the Learning Mathematics for Teaching assessment (LMT), developed by the Learning Mathematics for Teaching Project.[20]

LMT is a tailored computer adaptive test, where the set of items presented is determined by a computer algorithm to best match the ability of the person taking the test. Developed as a general measure of a narrow subset of the elementary mathematics curriculum, LMT is meant to measure a teacher's ability to use mathematics knowledge to effectively teach mathematics skills to elementary school students.

The subscale was administered concurrently with the Knowledge & Skills and Attitudes & Beliefs questionnaires in Spring 2013. There was no baseline pretest given for LMT.

3.3. Missing Data

Missing data are handled according to recommendations laid out in Puma et al. (2009). There are two cases to consider: (1) missing posttest or other outcome variables, and (2) missing pretest variables or other covariates.

[20] For more about the project, see Schilling, Blunk, & Hill (2007) and the Learning Mathematics for Teaching Project website at http://sitemaker.umich.edu/lmt/home.

1. **Missing Posttest/Outcome Variables: Casewise Deletion**. When we have a missing outcome variable for an observation, we remove that observation from the analytic sample.

2. **Missing Pretest Variables/Covariates: Dummy Variable Adjustment**. When an observation is missing a pretest variable or other covariate, we retain the observation in the analytic sample and apply the dummy variable adjustment method. A missing data "flag" (a dummy variable equal to 1 if the covariate is missing for that observation and 0 otherwise) is added to the list of covariates for each type of data that is missing in one or more observations. For each observation with a missing covariate, the value of that covariate is set to an identical constant equal to the sample mean for that observation's study condition.

Table 3.1. Number of Missing Variables, By Type

	Treatment		Control	
Variable	**Sample**	**Missing**	**Sample**	**Missing**
Outcome variables				
Missing teacher outcomes	90	2	95	1
Missing math scale score	5,553	169	5,053	150
Covariates				
Teachers				
Missing Data Use pretest score	88	17	94	16
Missing Knowledge & Skills pretest score	88	0	94	2
Missing Attitudes & Beliefs pretest score	88	0	94	2
Students				
Missing math scale pretest score	5,553	3,340	5,053	3,139

3.4. Qualitative Data Collection and Analysis

In addition to our quantitative data collection and analysis, we collected extensive qualitative data throughout the *Using Data* study to explore the contextual factors that may shape the implementation of the intervention.

We used interviews and surveys. All interviews were recorded and transcribed. The interview data were either summarized in analytic memos or analyzed using a constant comparative analysis approach. Through this iterative process we established a coding framework

that enabled us to identify key patterns and themes that were consistent across treatment schools.

3.4.1. School Selection

In year 1, the qualitative work focused on eight treatment schools that demonstrated high, medium, and low overall school performance and achievement trajectories. We defined achievement trajectory as the average change in grade 4 math scores on the state's annual standardized assessment over the three-year period prior to the intervention. We hypothesized that schools' overall performance and achievement trajectories may account for variation in program implementation.

In year 2, the qualitative data collection focused on a subset of four treatment schools. These four schools were high data-use schools that had implemented *Using Data* with fidelity, as determined by interviews with teachers in year 1 and the observations and recommendation of TERC facilitators. We selected these schools from the sample of eight schools in order to examine the "best case" scenario of the implementation of *Using Data* over time.

3.4.2. Data Team Interviews – Year 1

In the two years of the *Using Data* study, we conducted 61 interviews with data team members at the eight treatment schools, including multiple interviews with data team members at the subset of four high data-use schools (table 3.2, below). The first set of interviews with the eight treatment schools occurred at the end of year 1. The purpose of the interviews was to learn how the data team was functioning at each school and the perceived effects of *Using Data* on teaching and learning.

We followed a semi-structured interview protocol that asked how data team teachers organized and carried out their work, how the intervention was affecting classroom instruction, and how the data team shared their work with other teachers. We also conducted interviews with principals at the treatment schools using a semi-structured interview protocol that asked how the principal supported the data team, how *Using Data* had influenced the teachers' work and the school overall, and what the principal perceived to be the most valuable contribution of the intervention.

3.4.3. Data Team Interviews – Year 2

Toward the beginning and end of year 2, we conducted interviews with data team members at the subset of three schools. The purpose of the year 2 interviews was to observe implementation of the intervention over time and to examine in-depth how individual teachers use data for classroom instruction. As in year 1, we developed and followed a semi-structured interview protocol for each round of interviews. These year 2 protocols followed up on year 1 themes of gathering information on data team composition and activities, and asked about the teams' goals for year 2.

In the interviews early in year 2, we asked teachers at three of the treatment schools how data use may have changed at their school, and how it was affecting instructional decision making. We also asked about the strengths and weaknesses of data-use practices and the barriers data team members had encountered in implementing the intervention. We also interviewed principals about change in data use at the school, and whether the data team had achieved what it set out to accomplish. Similar to the teacher interview, we asked principals about the strengths and weaknesses of data-use practices, and whether *Using Data* was influencing the instructional culture of the school.

Toward the end of year 2, we conducted a last round of interviews at two treatment schools. While we had intended to revisit all three schools we went to at the beginning of year 2, two of the schools were not receptive to scheduling yet another round of interviews. As a result, we revisited just one of the three schools, and chose another treatment school where we had conducted interviews in year 1 and that had demonstrated high data use and implementation fidelity.

The purpose of this last round of interviews was to examine in-depth how teachers actually used data to inform their instruction. Prior to the interview, we asked teachers to bring a couple of examples of how they used student data to make instructional decisions. During the interview, teachers showed us how they pulled the data, analyzed, and interpreted it. Then we asked teachers to describe the actions they took based on their analysis. Throughout the process, we asked how, if at all, *Using Data* had informed teachers' data-use examples.

3.4.4. Central Office Interviews – Year 2

We also conducted interviews with two central office staff members who are responsible for professional development in DCPS in order to learn how *Using Data* complements or overlaps with current district training on data use in the classroom. We followed a semi-structured interview protocol that asked them to describe complementary DCPS data-use programs and how *Using Data* contributes to or contradicts these district initiatives.

3.4.5. End-of Program Survey

We administered a survey at the end of year 2 to data teams at the 30 treatment schools. The purpose of this survey was to capture how frequently data teams met that year, the focus of the data team meetings, and the value the teams ascribed to the *Using Data* program. Twenty-seven (27) of the treatment schools completed the survey (table 3.2).

Table 3.2. Qualitative Data Collection Activities in Year 1 and Year 2

Method	Data collected –Year 1	Data collected – Year 2
Data team interviews	End of year 1: 36 interviews at eight treatment schools	Beginning of year 2: 16 interviews at three high data-use schools End of year 2: 9 interviews at two high data-use schools
Central office interviews		2 interviews with two central office staff
Data team survey		End-of-program survey administered to 30 treatment school data teams; 90% response rate (27 of 30 schools)
Data coach logs	Each time the data team met – 202 logs collected	Each time the data team met – 139 logs collected
Facilitator logs	Each time a TERC facilitator conducted a PD event or TA session	Each time a TERC facilitator conducted a PD event or TA session

4. Estimated Impacts of the Intervention

This chapter presents our study results, including the primary confirmatory analyses of the effects of *Using Data* on teacher behavior and student achievement, as well as additional exploratory analyses of the effects of the intervention on student achievement.

For teacher behavior, the primary confirmatory analyses measure the effect of the *Using Data* intervention on teacher data use, knowledge and skills, and attitudes and beliefs at the end of year 1 (section 4.1). For student achievement, the primary confirmatory analysis measures the effect of the intervention on student performance on the state mathematics assessment at the end of year 2 (the schoolwide impact model) (section 4.2).

The exploratory analyses extend our schoolwide impact analysis of student achievement by considering how the following factors may influence the *Using Data* treatment effect (section 4.3):

- School performance and socioeconomic need (the school context model)
- The amount of exposure a student's teacher has had to the *Using Data* program, and the number of years a student is exposed to a *Using Data*–trained teacher (the dosage model)
- Student exposure in year 2 to a teacher trained in UD from the beginning, by school overall and by school context
- Students in year 2 of ITT teachers in treatment schools, overall and by school context

We also conduct an exploratory analysis of the effect of *Using Data* on student achievement at the end of year 1 of the intervention (section 4.4).

4.1. Primary Confirmatory Analysis – Teacher Behavior Results, Year 1

This section presents results for the *Using Data* study, based on year 1 (school year 2011–12) teacher data. The purpose of this section is to estimate the effect of the *Using Data* program on teacher behavior in three areas: data use, knowledge and skills with respect to using data to improve mathematics instruction, and attitudes and beliefs about the value of data to inform instruction and improve student learning.

4.1.1. Estimation Results

Data Use Model

Table 4.1 reports our HLM estimation results for the data use research question when the dependent variable is the Data Use scale score, which ranges from -1.54 to 3.58. The results indicate a positive and statistically significant treatment effect of participation in the *Using Data* program on the frequency with which teachers engage in data-use activities.

Table 4.1. Two-Level HLM Regression Results for Teacher Data Use Scale Score Model, Year 1

	Data Use scale score			
Variable description	Coeff.	Std. error	z	p
Treatment effect				
Overall – teacher in treatment group	0.29	0.11	2.67	0.01
Pretest controls				
Data Use pretest score	0.37	0.06	6.65	0.00
Knowledge & Skills pretest score	-0.11	0.05	-2.28	0.02
Attitudes & Beliefs pretest score	0.16	0.05	2.83	0.01
Experience controls				
Yrs. experience in district	0.01	0.02	0.47	0.64
Yrs. experience squared	0.000	0.001	-0.63	0.53
School controls				
Block 1	-0.16	0.16	-1.03	0.30
Block 2	-0.24	0.16	-1.48	0.14
Block 3	0.17	0.16	1.09	0.28
Missing Data Use pretest score	-0.27	0.13	-2.10	0.04
Missing Knowledge & Skills pretest score	0.41	0.52	0.79	0.43

Variable description	Data Use scale score			
	Coeff.	Std. error	z	p
Missing experience	0.14	0.41	0.35	0.73
Constant term	0.30	0.19	1.54	0.12
N	182			

The coefficient on the treatment indicator is statistically significant at the 1 percent level. Given the standard deviation of the outcome variable, this coefficient corresponds to an effect size of about .37.

Looking at the control variables: as expected, the Data Use pretest score has a positive and statistically significant coefficient. Teachers who use data more frequently at baseline also tend to use data more frequently after year 1. The Knowledge & Skills pretest variable has a negative correlation with the dependent variable. There are no statistically significant differences in teacher outcome by block.

Knowledge and Skills Model

Table 4.2 reports our HLM estimation results for the knowledge and skills model. The dependent variable is the Knowledge & Skills scale score on the posttest at the end of year 1, which ranges from -2.52 to 4.45. The estimated treatment effect is positive, but not statistically significant at the 5 percent level. It is significant at the 10 percent level (p = .06).

Table 4.2. Two-Level HLM Regression Results for Teacher Knowledge & Skills Scale Score Model, Year 1

Variable description	Knowledge & Skills scale score			
	Coeff.	Std. error	z	p
Treatment effect				
Overall – teacher in treatment group	0.33	0.18	1.86	0.06
Pretest controls				
Data Use pretest score	-0.18	0.10	-1.73	0.08
Knowledge & Skills pretest score	0.57	0.09	6.56	0.00
Attitudes & Beliefs pretest score	0.09	0.10	0.87	0.38
Experience controls				

Variable description	Knowledge & Skills scale score			
	Coeff.	Std. error	z	p
Yrs. experience in district	0.03	0.04	0.78	0.44
Yrs. experience squared	-0.0003	0.0012	-0.22	0.83
School controls				
Block 1	0.49	0.26	1.93	0.05
Block 2	0.04	0.27	0.15	0.88
Block 3	0.24	0.26	0.91	0.36
Missing Data Use pretest score	-0.13	0.24	0.60	0.86
Missing Knowledge & Skills pretest score	-0.05	0.97	-0.06	0.96
Missing experience	-0.52	0.77	-0.67	0.51
Constant term	-0.37	0.34	-1.10	0.27
N		182		

The magnitude of the treatment effect coefficient, 0.33, is equivalent to an effect size of about .25 given the standard deviation of the Knowledge & Skills posttest score.

The coefficients on the blocking indicators are positive, and the coefficient on Block 1 is nearly statistically significant at the 5 percent level. This result suggests that teachers at the lowest-needs schools (Block 1) tended to score higher relative to the highest-needs schools (Block 4) in terms of knowledge of data-use concepts and application of data-use skills at the end of study year 1.

Attitudes and Beliefs Model

Table 4.3 reports results for the attitudes and beliefs model. The dependent variable is the Attitudes & Beliefs scale score, which ranges from -5.68 to 4.83. The estimated treatment effect on teachers' attitudes and beliefs is positive and statistically significant at the 5 percent level (p = .02).

Table 4.3. Two-Level HLM Regression Results for Teacher Attitudes & Beliefs Scale Score Model, Year 1

Variable description	Attitudes & Beliefs scale score			
	Coeff.	Std. error	z	p
Treatment effect				
Overall – teacher in treatment group	0.44	0.19	2.36	0.02
Pretest controls				
Data Use pretest score	-0.04	0.10	-0.44	0.66
Knowledge & Skills pretest score	-0.04	0.08	-0.52	0.60
Attitudes & Beliefs pretest score	0.58	0.10	5.87	0.00
Experience controls				
Yrs. experience in district	0.08	0.04	1.88	0.06
Yrs. experience squared	-0.002	0.001	-1.90	0.06
School controls				
Block 1	-0.45	0.27	-1.67	0.09
Block 2	-0.23	0.28	-0.82	0.41
Block 3	0.19	0.27	0.71	0.48
Missing Data Use pretest score	-0.45	0.23	-1.95	0.05
Missing Knowledge & Skills pretest score	0.27	0.93	0.30	0.77
Missing experience	0.45	0.74	0.60	0.55
Constant term	-0.22	0.33	-0.66	0.51
N		182		

The size of the coefficient, 0.44, corresponds to an effect size of approximately .34. This is the only outcome in which teacher experience seems to have any effect; more-experienced teachers tend to have attitudes and beliefs that are more closely aligned with the principles of the *Using Data* intervention (the effect is statistically significant at the 10 percent level).

Summary

In sum, the teacher behavior analysis shows that *Using Data* had a positive and statistically significant treatment effect on data use (effect size = .37, p = .01) and on attitudes and beliefs (effect size = .34, p = .02). The point estimate is also positive on the knowledge and skills outcome (effect size = .25, p = .06), but the estimate is less precise, falling somewhat short of 95 percent confidence that the effect was caused by the intervention.

4.2. Primary Confirmatory Analysis – Student Achievement Results, Year 2

4.2.1. Estimation Results, Schoolwide Impact Model

Table 4.4 provides results for the year 2 schoolwide impact model. The dependent variable (student outcome) is the scale score on the 2012–13 state math assessment. The range of scores is from 155 to 279.

Table 4.4. Two-Level HLM Regression Results of Student State Math Assessment Score, Year 2 (SY 12–13), Schoolwide Impact Model

	Schoolwide impact model			
Variable description	**Coeff.**	**Std. error**	**z**	**p**
Treatment effect				
Overall – student attends treatment school	0.16	1.06	0.15	0.883
Student controls				
Prior math scale score	0.19	0.00	40.95	0.000
Socioeconomic characteristics				
Racial/ethnic minority				
African American	-5.91	0.40	-14.67	0.000
Hispanic	-1.29	0.62	-2.08	0.037
FRL eligible	-5.94	0.38	-15.60	0.000
Educational characteristics				
English language learner	-2.17	0.77	-2.82	0.005
Gifted	14.80	0.78	18.86	0.000
Learning disability or condition	-11.22	0.52	-21.49	0.000
Grade 4 student	0.10	0.51	0.19	0.847
Missing prior math scale score	-5.34	0.53	-10.15	0.000
School controls				
Block 1	2.23	1.55	1.44	0.151
Block 2	0.97	1.52	0.64	0.520
Block 3	-0.33	1.47	-0.22	0.823
Constant term	162.5	2.05	79.13	0.000
N		10,287		

We find no overall treatment effect at the end of year 2. The effect size is small (.007) and not statistically significant.

4.3. Exploratory Analyses – Student Achievement Results, Year 2

4.3.1. School Context Model

Next, we estimate a variant of the schoolwide impact model that permits differential treatment effects by baseline school performance and socioeconomic need. This is accomplished by adding to the schoolwide impact model multiplicative interaction terms between the treatment effect and the school block indicator variables. Table 4.5 provides the results.

Table 4.5. Two-Level HLM Regression Results of Student State Math Assessment Score, Year 2 (SY 12–13), School Context Model

	School context model			
Variable description	**Coeff.**	**Std. error**	**z**	**p**
Treatment effect				
Student attends treatment school (Block 4 treatment effect, highest-needs schools)	3.13	1.96	1.60	0.110
Interaction terms (differential effects)				
Block 1(lowest-needs schools)	-9.15	2.82	-3.24	0.001
Block 2	-2.26	2.78	-0.81	0.417
Block 3	-1.10	2.72	-0.40	0.687
Student controls				
Prior FCAT math score	0.19	0.00	40.94	0.000
Socioeconomic characteristics				
Racial/ethnic minority				
African American	-5.89	0.40	-14.63	0.000
Hispanic	-1.29	0.62	-2.08	0.037
FRL eligible	-5.93	0.38	-15.58	0.000
Educational characteristics				
English language learner	-2.21	0.77	-2.87	0.004
Gifted	14.78	0.78	18.85	0.000
Learning disability or condition	-11.22	0.52	-21.49	0.000
Grade 4 student	0.10	0.51	0.20	0.844

Variable description	School context model			
	Coeff.	Std. error	z	p
Missing prior FCAT math score	-5.33	0.53	-10.14	0.000
School controls				
Block 1	7.11	2.07	3.43	0.001
Block 2	2.18	2.00	1.09	0.275
Block 3	0.27	1.95	0.14	0.891
Constant term	161.0	2.20	73.29	0.000
N	10,287			

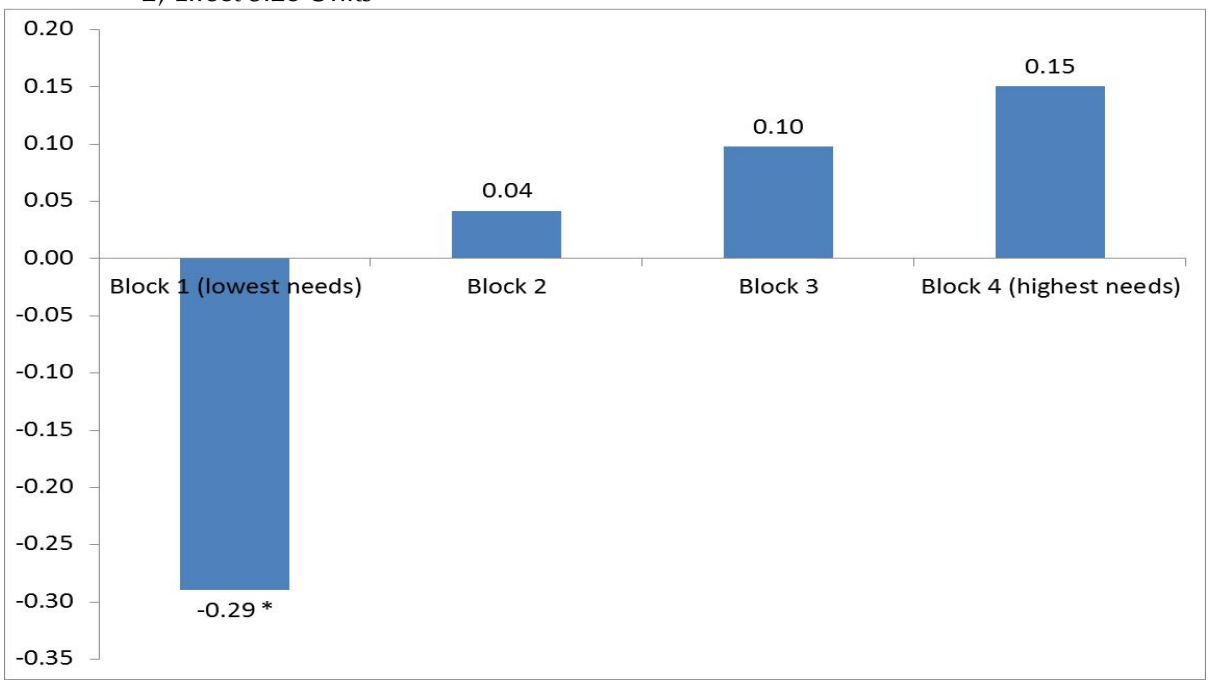
With this specification, the interpretation of the coefficient on the treatment indicator (equals 1 if the student attends a treatment school) is the treatment effect for students attending Block 4 (highest-needs) schools, because Block 4 is the omitted group in the model. The interpretation of the coefficients on the three treatment indicator–school block interaction terms is the difference between the treatment effect for that block and Block 4. As a result, the overall block-specific treatment effects can be recovered by adding the Block 4 treatment to the appropriate indicator variable. These block-specific treatment effects can be tested for statistical significance.

Table 4.6, below, provides the results for the block-specific treatment effect estimates. Figure 4.1 displays these block-specific treatment effect estimates in effect size units.

Table 4.6. Treatment Impact Estimates of *Using Data* Program, by School Block, Year 2

School block	Treatment effect	Std. error	z	p
Block 1 (lowest needs)	-6.02	2.03	-2.96	0.00
Block 2	0.87	1.97	0.44	0.66
Block 3	2.03	1.89	1.08	0.28
Block 4 (highest needs)	3.13	1.96	1.60	0.11

Figure 4.1. Student Achievement Treatment Effect Estimates of *Using Data*, by School Block, Year 2, Effect Size Units



* Statistically significant at the 5 percent level.

Table 4.6 and figure 4.1 show that the treatment effect for Block 1 schools is negative, with an effect size of -.29, and statistically significant at the 5 percent level. Block 1 schools are the group of schools with the lowest level of socioeconomic need and the highest performance on the state mathematics exam at baseline. As we move from lower-needs to higher-needs blocks, the treatment effect becomes positive and increases in size block by block, although the effect is not statistically significant for any but Block 1. Block 4, the highest-needs schools, has the highest treatment effect point estimate, with an effect size of .15.

4.3.2. Dosage Models

4.3.2.1. Detailed Dosage Model

To estimate the effects on student performance of having teachers with different levels of exposure to the *Using Data* intervention, we estimate a model that includes both school-level and teacher-level indicators in a three-level HLM. The year 2 version of this dosage model includes a school-level treatment effect and separate treatment effects for

- students having a UD-trained teacher in year 1;
- students having a replacement teacher in year 2 (who would have been exposed only to the second year of *Using Data* training); and
- students having a teacher in year 2 who was exposed to the full two years of *Using Data* training

and interaction terms between the year 1 and year 2 teacher-level treatment effects.

These interaction terms will measure the additional effect on student achievement, if any, of having *Using Data*–trained teachers in both year 1 and year 2 of the study. Note that because randomization took place at the school level, not the teacher level, estimates from this dosage model do not have the same causal implications as do those from the two-level model.

Table 4.7 presents the estimation results for this model. None of the differential treatment effects or interaction effects is statistically significant.

Table 4.7. Three-Level HLM Regression Results of Student State Math Assessment Score, Year 2 (SY 12–13), Dosage Model

	Dosage model			
Variable description	**Coeff.**	**Std. error**	**z**	**p**
Treatment effects				
Student attends treatment school	-0.59	1.24	-0.47	0.636
Student had a UD-trained teacher in year 1 (SY 11–12)	-0.04	0.92	-0.04	0.967
Student had a UD-trained teacher in year 2 (SY 12–13) who received two years of PD	0.99	1.20	0.83	0.407
Student had a UD-trained teacher in year 2 (SY 12–13) who received one year of PD (year 2 replacement)	0.13	1.97	0.07	0.947
Interaction terms				
Year 1 teacher and year 2 teacher	-0.07	1.35	-0.05	0.961
Year 1 teacher and year 2 replacement	0.46	1.95	0.24	0.813
Student controls				
Prior math scale score	0.19	0.00	40.54	0.000

Variable description	Dosage model			
	Coeff.	Std. error	z	p
Socioeconomic characteristics				
Racial/ethnic minority				
African American	-5.92	0.40	-14.76	0.000
Hispanic	-1.43	0.62	-2.32	0.021
FRL eligible	-5.35	0.38	-14.20	0.000
Educational characteristics				
English language learner	-1.35	0.80	-1.68	0.092
Gifted	14.24	0.77	18.49	0.000
Learning disability or condition	-11.44	0.54	-21.07	0.000
Grade 4 student	-0.20	0.79	-0.25	0.805
Missing prior math scale score	-5.13	0.57	-8.97	0.000
School controls				
Block 1	2.61	1.61	1.62	0.105
Block 2	1.15	1.58	0.73	0.466
Block 3	-0.51	1.60	-0.32	0.749
Constant term	160.2	2.18	73.40	0.000
N	9,863			

By summing the coefficients in table 4.7 on the appropriate variables, we can derive overall treatment effects for students who had teachers with different timing or levels of exposure to the *Using Data* treatment.

These overall effects are summarized in table 4.8, which shows that all of the different dosages of treatment effect are small and none is statistically significant. Although the estimates are imprecise, we note that the coefficients are positive only for students in year 2 with teachers trained in UD from the beginning.

Table 4.8. Treatment Effect of *Using Data* Program by Treatment Dosage, Year 2 (SY 12–13)

Dosage of treatment		Treatment effect	Std. error	z	p
No *Using Data* teacher	Student in treatment school	-0.59	1.24	-0.47	0.64
Using Data teacher during year 1 only	Student in treatment school who was taught by *Using Data* teacher during year 1 only	-0.62	1.44	-0.43	0.66
Using Data teacher during year 2 only	Student in treatment school who was taught during year 2 only by *Using Data* teacher who had been in the program for one year	-0.45	2.02	-0.23	0.82
	Student in treatment school who was taught during year 2 only by *Using Data* teacher who had been in the program for two years	0.41	1.32	0.31	0.76
Using Data teacher during both year 1 and year 2	Student in treatment school who was taught both by *Using Data* teacher during year 1 and by *Using Data* teacher who had been in the program for one year during year 2	-0.03	2.31	-0.01	0.99
	Student in treatment school who was taught both by *Using Data* teacher during year 1 and by *Using Data* teacher who had been in the program for two years during year 2	0.30	1.52	0.20	0.84

4.3.2.2. Streamlined Dosage Model

The detailed dosage model presented in section 4.3.2.1 allows us to measure differential treatment effects for every level of student and teacher exposure in the sample. But because few teachers fall into several of the exposure subgroups, the model asks a lot of the data. In this next model, we use a simpler specification and ask whether students in year 2, of teachers trained from the beginning in UD, have higher levels of achievement than other students in study schools. The students in this treatment group include students of ITT teachers *and* students of year 1 replacement teachers. This second group of teachers joined the study before the intervention began, but after the initial randomization of schools.

Table 4.9. Treatment Effect of *Using Data* Program for Students Taught by a Teacher Trained for 2 years in the Program, Year 2 (SY 12–13)

Variable Description	Streamlined Dosage Model			
	Coeff.	Std. error	z	p
Treatment effect				
Overall - student has teacher trained for 2 yrs in UD	1.10	0.54	2.04	0.042
Student controls				
Prior math state test score	0.19	0.00	40.59	0.000
Socioeconomic characteristics				
Racial/ethnic minority				
African-American	-5.94	0.41	-14.56	0.000
Hispanic	-1.25	0.62	-2.01	0.045
Free and reduced-price lunch eligible	-5.73	0.38	-14.95	0.000
Educational characteristics				
English language learner	-2.21	0.77	-2.85	0.004
Gifted	14.63	0.79	18.62	0.000
Learning disability/condition	-11.31	0.53	-21.26	0.000
Fourth grader	0.00	0.53	0.00	1.000
Missing prior math state test score	-5.06	0.54	-9.42	0.000
School controls				
Block 1	2.08	1.59	1.31	0.190
Block 2	0.74	1.55	0.48	0.635
Block 3	-0.53	1.51	-0.35	0.724
Constant term	162.4	2.01	80.67	0.000
N	10,019			

Table 4.9, above, finds a small but statistically significant effect for students in year 2 who had a teacher trained in UD from the beginning of the program (effect size = .05, p = .04). Table 4.10 displays this effect by school block.

Table 4.10. Treatment Effect of *Using Data* Program for Students Taught by a Teacher Trained for 2 years in the Program, with Block Interaction terms, Year 2 (SY 12–13)

Variable Description	Streamlined Dosage Model (School Context Version)			
	Coeff.	Std. error	z	p
Treatment effect				
Student has teacher trained for 2 yrs in UD (Block 4 treatment effect)	3.20	1.21	2.66	0.008
Interaction Terms (differential effects)				
Block 1	-3.01	1.61	-1.87	0.062
Block 2	-1.01	1.50	-0.67	0.502
Block 3	-5.17	1.72	-3.00	0.003
Student controls				
Prior FCAT math score	0.19	0.00	40.65	0.000
Socioeconomic characteristics				
Racial/ethnic minority				
African-American	-5.93	0.41	-14.53	0.000
Hispanic	-1.25	0.62	-2.01	0.044
Free and reduced-price lunch eligible	-5.73	0.38	-14.95	0.000
Educational characteristics				
English language learner	-2.34	0.78	-3.02	0.003
Gifted	14.64	0.79	18.64	0.000
Learning disability/condition	-11.34	0.53	-21.32	0.000
Fourth grader	0.09	0.53	0.17	0.869
Missing prior FCAT math score	-5.06	0.54	-9.43	0.000
School controls				
Block 1	2.69	1.61	1.67	0.095
Block 2	0.95	1.57	0.61	0.544
Block 3	0.73	1.56	0.47	0.640
Constant term	161.8	2.03	79.84	0.000
N	10,019			

Table 4.10 shows that the association between having in year 2 a teacher trained in UD from the beginning differs by the socioeconomic characteristics and baseline performance of schools, as indicated by school block. Positive effects were found for Block 2 and Block 4 schools. The largest effect is for Block 4 schools; that is, schools with the highest needs at baseline. The coefficient of 3.20 translates to an effect size of .15, and is statistically significant at the 1 percent level. Table 4.11, below, shows treatment effects by block.

Table 4.11. Treatment Effect Estimates of *Using Data* Program, Streamlined Dosage Model, by School Block, Year 2

School block	Treatment effect	Std. error	z	p
Block 1 (lowest needs)	0.19	1.07	0.18	0.86
Block 2	2.19	0.90	2.43	0.02
Block 3	-1.97	1.23	-1.59	0.11
Block 4 (highest needs)	3.20	1.21	2.66	0.01

A drawback to the analysis presented above is that it includes students of ITT teachers as well as students of teachers who joined the study after randomization. In the next section, we examine the effect on achievement for students of ITT teachers in treatment schools, compared with students of ITT teachers in control schools at the end of year 2. We then look at the estimated effect for this sample by school context (block).

4.3.2.3. Dosage Model, Treatment Students of ITT Teachers

To estimate the model, we restrict the sample to students in year 2 of ITT teachers. The sample includes 55 schools with ITT teachers and their 4,109 students. This comparison is between students in treatment and control schools of teachers who joined the study prior to randomization, increasing the internal validity of the sample, in comparison with the previous analysis (section 4.3.2.2), but reducing statistical power.

The overall treatment effect for this sample remains small and is statistically insignificant (effect size = .04, p = .59). Allowing the treatment effect to vary by school context yields similar results to those presented for the full experimental sample. Effect size estimates increase in size as school needs at baseline increase. As shown in table 4.12, below, the effect size is statistically significant and largest for Block 4 schools (effect size = .40, p = .01). These results provide additional evidence suggesting that the UD program is an effective program in low-performing schools.

Table 4.12. Treatment Effect Estimates of *Using Data* Program, Students of ITT Teachers by School Block, Year 2

School block	Treatment effect	Std. error	z	p
Block 1 (lowest needs)	-2.08	1.93	-1.08	0.28
Block 2	1.04	1.84	0.56	0.58
Block 3	2.91	3.11	2.62	0.26
Block 4 (highest needs)	8.13	3.10	2.62	0.01

4.4. Exploratory Analysis – Student Achievement Results, Year 1

4.4.1. Estimation Results

Schoolwide Model

Table 4.13 provides results for the schoolwide model. The dependent variable (student outcome) is the scale score on the 2011–12 state math assessment. The range of scores is from 155 to 279.

Table 4.13. HLM Regression Results of Student State Math Assessment Score, Year 1 (SY 11–12), Schoolwide Model

Variable description	Schoolwide model			
	Coeff.	Std. error	z	p
Treatment effect				
Overall – student attends treatment school	-0.56	0.76	-0.74	0.458
Student controls				
Prior math state test score	0.25	0.00	98.03	0.000
Socioeconomic characteristics				
Racial/ethnic minority				
African American	-3.52	0.32	-10.84	0.000
Hispanic	-1.54	0.51	-3.02	0.003
FRL eligible	-2.38	0.30	-8.02	0.000
Educational characteristics				
English language learner	-1.95	0.60	-3.24	0.001
Gifted	6.45	0.69	9.30	0.000
Learning disability or condition	-4.26	0.41	-10.43	0.000
Grade 4 student	-7.89	0.25	-30.94	0.000
Missing prior math state test score	-5.97	0.54	-11.12	0.000
School controls				
Block 1	1.10	1.11	0.99	0.322

Variable description	Schoolwide model			
	Coeff.	Std. error	z	p
Block 2	0.83	1.10	0.76	0.448
Block 3	1.71	1.07	1.60	0.110
Constant term	140.6	1.30	108.2	0.000
N	10,603			

The treatment effect is not statistically significant at the 5 percent level.

Dosage Model

Finally, we estimate a model that includes both the school-level and the teacher-level indicators. This is a three-level HLM with students nested with teachers nested in schools, with school-level and teacher-level random effects. The results of estimating this model using the whole school sample are summarized in table 4.14.

Table 4.14. Three-Level HLM Regression Results of Student State Math Assessment Score, Year 1 (SY 11–12), Dosage Model

Variable description	Dosage model			
	Coeff.	Std. error	z	p
Treatment effects				
Student attends treatment school	-1.47	0.96	-1.53	0.125
Student has a UD-trained teacher	1.76	0.91	1.93	0.054
Student controls				
Prior math state test score	0.25	0.00	97.86	0.000
Socioeconomic characteristics				
Racial/ethnic minority				
African American	-3.66	0.32	-11.43	0.000
Hispanic	-1.73	0.50	-3.48	0.001
FRL eligible	-2.01	0.29	-6.95	0.000
Educational characteristics				
English language learner	-1.37	0.62	-2.20	0.028
Gifted	6.47	0.67	9.60	0.000
Learning disability or condition	-3.63	0.42	-8.70	0.000
Grade 4 student	-7.73	0.58	-13.41	0.000
Missing prior math state test score	-5.23	0.53	-9.79	0.000
School controls				
Block 1	1.32	1.18	1.12	0.263

Variable description	Dosage model			
	Coeff.	Std. error	z	p
Block 2	1.26	1.17	1.08	0.280
Block 3	1.52	1.20	1.26	0.207
Constant term	140.1	1.39	101.1	0.000
N	10,048			

There is some difference between the school-level and teacher-level treatment effects. The school-level effect on students in treatment schools who are not taught directly by *Using Data*–trained teachers is negative (-1.47 scale score points), small (about .05 standard deviations), and not statistically significant at the 5 percent level. The coefficient on the teacher-level treatment variable, interpreted as the difference between the treatment effect for students in treatment schools taught by teachers who are not *Using Data* trained versus the treatment effect for students in treatment schools whose teachers are *Using Data* trained is, however, positive (1.76 scale score points), although small (about .08 standard deviation), and statistically significant at the 5 percent level.

To measure the full treatment effect on a student who both attends a treatment school and is taught math by a *Using Data*–trained teacher, we add the two treatment variable coefficients together. The results are summarized in table 4.15.

Table 4.15. Treatment Effects of *Using Data* Program by Treatment Dosage, Year 1 (SY 11–12)

Group	Treatment effect	Std. error	z	p
Student in treatment school	-1.47	0.96	-1.53	0.13
Student in treatment school who is taught by a *Using Data* teacher	0.29	0.89	0.33	0.75

Neither dosage of treatment has a statistically significant estimated effect on student performance. For students in treatment schools who are taught by *Using Data*–trained teachers, the two effects (school-level plus additional teacher-level effect) seem to cancel each other out.

The results for year 1 are consistent with the expectation that schools are unlikely to see the effects of the intervention during the first year. In that first year, teachers are being introduced to *Using Data* processes and practicing them for the first time, over the course of the school year.

4.4.2. Conclusion, Year 1 Results

The teacher behavior analysis shows that *Using Data* had a positive and statistically significant treatment effect on data use and attitudes and beliefs. The point estimate is also positive on teachers' knowledge and skills, but this estimate is not statistically significant at conventional levels.

The exploratory analysis of student achievement at the end of year 1 shows that there is some difference between the school-level and teacher-level treatment effects. For students in treatment schools, there is a negative effect on their state math assessment scores that is not statistically significant. For students in treatment schools who are also taught by *Using Data*–trained teachers, there is an additional positive treatment effect that is not quite significant at the 5 percent level. The two effects tend to cancel each other out, so that the overall treatment effect (school plus teacher) for this latter group of students at the end of year 1 is small and not significant.

4.5. Discussion of Results

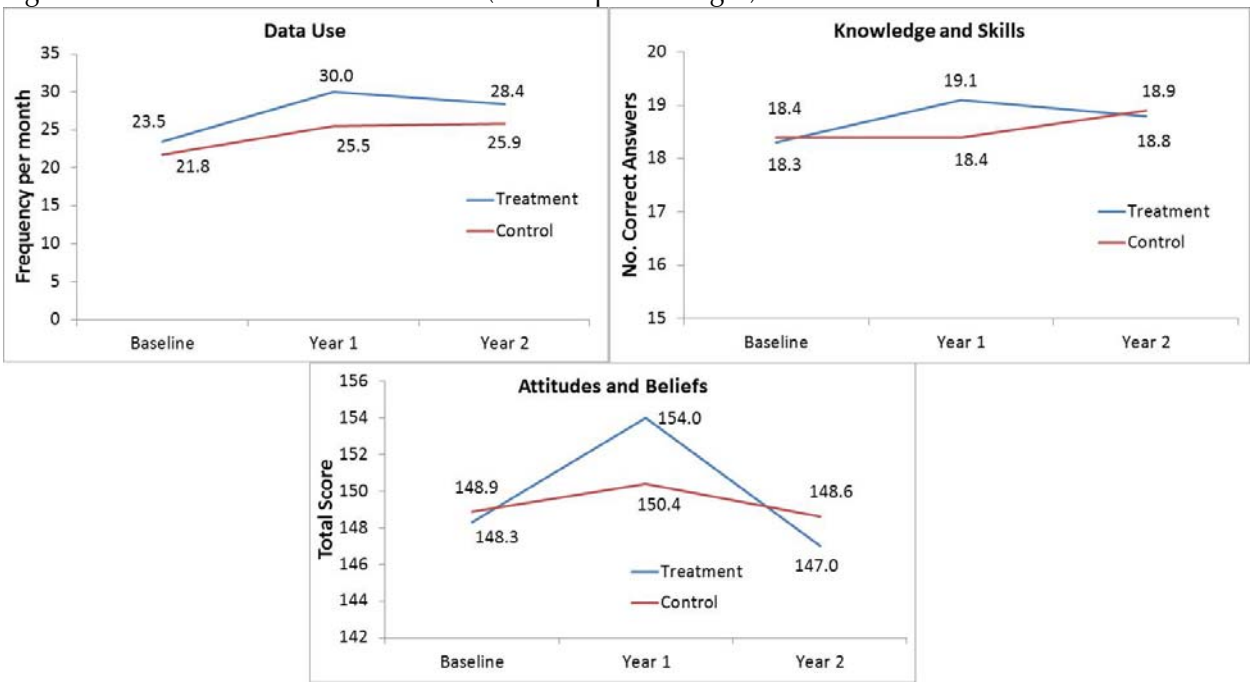
In this section we summarize the main results of the study, and explore them further. We are especially interested in better understanding why there seems to be no overall treatment effect on student achievement, and in better understanding the nature of the block-specific differences in the treatment effect on student achievement.

4.5.1. Teacher Data-Use Behavior

As stated, we find a positive treatment effect of the *Using Data* intervention on teacher behavior after year 1 (school year 2011–12). Compared with control teachers, treatment teachers used data more frequently (effect size = .37, p-value = .01), held attitudes and beliefs that were more aligned with *Using Data* principles (effect size = .34, p = .02), and displayed greater data literacy (effect size = .25, p = .06).

This effect, however, does not appear to persist through year 2. If we compare the baseline, year 1, and year 2 values for data use, knowledge and skills, and attitudes and beliefs for the 142 teachers present in the study at randomization and remaining in the study through the end of year 2, we see that most of the gains that treatment teachers made relative to control teachers in year 1 disappear by the end of the study (figure 4.2). The figure shows that differences between the treatment and control group narrow because of some loss in gains made during year 1 for treatment teachers *and* because of gains made by teachers in control schools.

Figure 4.2. Teacher Behavior Measures (ITT Sample Averages)



These data suggest that by the end of year 2, there may have been only a limited amount of difference in teacher behavior between treatment and control schools with respect to data use. Qualitative information gathered during the study (as described in section 3.4) supports this idea. Interviews with DCPS administrators and with the data teams in the eight qualitative focus schools during year 1 make clear that the district was already cultivating a data-driven culture prior to implementation of the *Using Data* intervention. As described in section 1.2.3, for the past several years, for example, DCPS had offered two professional development programs to K–12 teachers, each of which has a strong data-use component.

Further evidence that DCPS was already a relatively intensive data-use district, even before implementation of the *Using Data* intervention, comes from the baseline results (Fall 2011, before implementation began) of the Data Use portion of the Attitudes & Beliefs questionnaire. Because these 16 Data Use questions were taken from the NETTS study, we can compare the responses of teachers in DCPS with a larger national sample.

The sample from DCPS comprises the 220 teachers who were in the study at randomization and answered the 16 Data Use questions that were part of the baseline Attitudes & Beliefs questionnaire in Fall 2011. The national sample is a group of 4,935 teachers who were surveyed as part of NETTS in SY 04–05. This comparison is made in table 4.16. A limitation of the comparison is that the NETTS sample includes K–12 teachers, whereas our sample includes just upper-elementary school teachers.

Table 4.16. Comparison of Baseline Data Use in DCPS and Data Use in a National Sample of Teachers, *Using Data* SY 10–11 Versus NETTS SY 04–05

Individual Data Use: In school year 20xx–xx, how often did you use data for each of the following purposes?

	Never		A few times		Once or twice a month		Once a week or more	
In 20xx/xx, I used data…	**DCPS**	**NETTS**	**DCPS**	**NETTS**	**DCPS**	**NETTS**	**DCPS**	**NETTS**
To inform curriculum changes	7	50	25	33	37	13	30	4
To identify individual skill gaps for individual students so that you could give each student material tailored to his/her skill profile	6	45	24	35	34	14	37	6
To determine whether your class or individual students were ready to move on to the next instructional unit	14	51	16	29	30	14	40	6
To evaluate promising classroom practices	20	59	20	26	35	11	24	4
To decide to give your students test-taking practice	26	64	21	22	33	10	20	4
To estimate whether your students would make adequate yearly progress (AYP)	24	63	36	24	31	10	10	3
To track standardized test scores by grade	15	56	44	32	30	9	11	3
To track individual student test scores	5	35	29	38	36	16	30	11
To track other measures of student progress	12	40	28	34	28	15	32	12
To inform student placement in courses or special programs	31	58	34	29	20	9	15	4
To inform parents about student progress	8	30	27	34	39	22	26	14

Collaborative Data Use: In school year 20xx–xx, how often did you work with data in the following contexts to make instructional decisions?

	Never		A few times		Once or twice a month		Once a week or more	
	UD	NETTS	UD	NETTS	UD	NETTS	UD	NETTS
On your own	4	24	22	37	35	18	39	21
Working with colleagues in your grade level	8	29	35	48	43	18	15	5
Working with colleagues from other grade levels	14	43	54	46	28	9	4	2
As part of a district-level activity with staff from other schools	49	73	33	22	17	4	2	1
In another setting	55	95	28	3	15	1	3	0

Note: Table entries are percentages.

Sample sizes: *Using Data* (N = 220), NETTS (N = 4,935).

We can use the percentages from the table to calculate a rough "average" monthly frequency of data use for the two samples.[21] Such a calculation suggests that the typical teacher in the sample from DCPS engaged in data-use activities an average of 23 times per month *at baseline*, more than double the average of 9 times per month for the national sample in 2004–05.

Overall, then, we can conclude that the control schools in this study were not standing still—they were making their own gains in data use and data literacy over the course of the study, in the context of a school district that already heavily emphasized data use. It may also be the case that treatment schools immersed in the *Using Data* intervention may have missed out on other professional development that would have been beneficial for them.

4.5.2. Student Achievement on End-of-Year Math Assessments

In contrast to teacher behavior, with respect to student performance on the end-of-year statewide mathematics assessment, we find no

[21] We weight the responses as follows: "never" = 0, "a few times" = 0.5, "once or twice a month" = 1.5, and "once a week or more" = 4.0. We then sum the responses across questions to come up with the estimated number of data-use activities per month.

evidence of an overall treatment effect in either year 2 (on which the primary confirmatory assessment is based) or year 1.

One potential reason for this lack of a student achievement treatment effect is that treatment schools appear to have adopted only part of the full *Using Data* model. Our qualitative data collection suggests that while teachers visualized data and examined it collaboratively, as called for by the program, they may have struggled to take the next step—identifying student learning problems and changing instruction to address those problems.

Our data suggest that, although the data teams continued to meet throughout the duration of the intervention, they weren't as effective at identifying causes of student learning problems and changing their practice in meaningful ways to successfully address the issue. The results of our year 2 end-of-year survey of data teams, for example, suggest that data teams across blocks adhered less frequently to the parts of the *Using Data* model related to conducting research to improve instruction and following the four-phase data-driven dialogue protocol (see section 1.2.2). Table 4.17, below (and also table 1.3), displays those survey results.

Table 4.17. Data Team End-of-Year 2 Survey Response Means

					How closely did you follow the components of the *Using Data* program?								
	# of meetings	Avg. meeting length	% of time productive	Success in addressing learning problem and in achieving student learning goal	Analyze your own students' data	Collaborate with other teachers about data	Identify student learning problem	Verify cause of student learning problem	Use data to monitor student progress	Conduct research to improve your instruction	Follow the four-phase dialogue	Rate your TERC facilitators (1–5)	Rate the *Using Data* program overall (1–5)
Block 1 (n=6)	7.3	54.2	83%	3.2	2.8	2.8	2.7	2.5	2.8	1.5	2.2	4.5	4.2
Block 2 (n=7)	10.0	56.1	83%	2.7	2.7	2.6	2.7	2.3	2.7	2.0	1.9	4.3	3.7
Block 3 (n=7)	11.4	43.6	89%	2.9	3.0	2.3	2.3	2.3	2.9	2.0	2.1	4.6	4.4
Block 4 (n=7)	14.1*	49.6	89%	3.0	3.0	2.9	2.7	2.4	3.0	2.1	2.4	4.4	4.3

*After dropping a 40-meeting outlier, the average number of meetings held by Block 4 is 9.83.

Scales: Success of data team (1–4 scale, 4=very successful). Degree adhered to UD model (0–3 scale, 3=very closely). Facilitator rating (1–5 scale, 5=very helpful). *Using Data* rating (1–5 scale, 5=very useful).

Additionally, in interviews we conducted at a subset of three treatment schools at the beginning of year 2, teachers in each school described the challenge of implementing these program components. Data team members at School 1 confidently analyzed student data together, but their knowledge of what to do with the analysis and ability to make a connection to their teaching practices were evolving. Describing teachers' difficulty with using data to change their instruction, the data coach said:

> Looking at [data] and driving where the kids are—I think we're pretty good at looking at data and seeing. Whether we change what we do, I don't know that that's the case. That's the whole point of it. Looking to see where we need to change things or what we need to continue that worked with that. I think we're good looking at it, but whether we implement change, I don't know.

The data coach at School 2 expressed,

> Teachers are good at pulling and looking at data, but they're not strong in understanding the causes of student performance.

A data team member at School 2, who is also a school instructional coach, believed teachers were not taking this next step because they might find themselves to be the cause of the student learning problem:

> I think, for some people, it's still an openness of really looking at themselves and their own practices and being afraid that when we really start to break down that data, that it's more of a reflection on them—that people aren't quite ready for that next step still. I think that they're opening up more and more, but it's still hard for them to reflect on themselves as a cause for a learning problem.

Block-Specific Differences in Treatment Effect on Student Achievement

Although our evaluation did not find an overall treatment effect on student achievement, we did estimate differences across blocks in the size and sign of treatment effects. Referring back to figure 4.1, Block 1 schools (the lowest-needs, highest-performing schools at baseline) showed a negative, statistically significant treatment effect (effect size = -.29, p = .004); whereas Block 4 schools (the highest-needs, lowest-

performing schools at baseline) showed a positive effect that approached but did not reach statistical significance at the 5 percent level (effect size = .15, p = .11). Block 2 and Block 3 schools showed treatment effect estimates that were intermediate in size between those of Block 1 and Block 4 but were statistically nonsignificant.

The treatment effect, therefore, appears to be negative for lower-needs schools, and increasing with the level of school socioeconomic need and underperformance. When we restricted the analysis sample to students of ITT teachers in year 2, we found the same pattern of results: effects grew in size as we moved from higher-performing to lower-performing schools at baseline. But results were only statistically significant for Block 4, where we found an effect size of .40 (p = .01).

Our qualitative data shed some light on these differential treatment effects by block. Among the eight focus schools in which we collected qualitative data (see section 3.4), we found systematic differences in how the higher-performing and lower-performing treatment schools described their data-use histories prior to their participation in *Using Data*. The focus schools were selected based on their preintervention achievement trajectories, rather than block membership, so the eight-school sample included four Block 1 schools, two Block 2 schools, one Block 3 school, and one Block 4 school.

In our interviews with principals, data coaches, and teachers at the eight schools, we found that the Block 1 and 2 schools uniformly reported that data use was part of their school cultures before the intervention commenced. Teachers and principals at these schools generally reported that they already examined benchmark test data in tables or graphs each time the data became available and discussed strengths and weaknesses, and that they were already using norms of collaboration similar to those prescribed by *Using Data*.

Principals at two of the Block 1 schools stated specifically that they felt the UD intervention would have been more useful at lower-performing schools that lacked a strong history of data use. One of the eight schools was part of a national network of data-focused schools and already had an accountability team in place. Staff at that school reported that the addition of a separate data team for *Using Data* (which was limited to grades 4 and 5) felt particularly

redundant, and the school chose to withdraw from most of the year 2 UD professional development activities.

This is not to say that the Block 1 and 2 treatment schools in the qualitative sample reported no benefit from *Using Data*. Interviewees at these schools did describe learning new strategies from the intervention, most notably including the four-phase dialogue, UD's structured protocol for discussing data and drawing evidence-based inferences. Teachers at several Block 1 and 2 schools said the protocol had pushed them to systematically investigate reasons behind the data patterns observed. These schools also reportedly appreciated the structure *Using Data* provided, saying that the data team meetings provided a reliable "venue" for advancing data-driven decisions.

One Block 1 school described *Using Data* as "a good next step for our school, because we are very rich in data, but this really helped to narrow down a focus and to keep that focus on track." Staff at two of the schools, however, said they would have liked more information from the *Using Data* program about instructional strategies and best practices. To summarize what they believed the program had given them, staff at one school said "it wasn't anything new that we learned," but that "we never really have gone into as much depth" as they had in the *Using Data* professional development. Staff at another school said *Using Data* gave them "some little tools and tricks" that were useful in conjunction with their existing approach.

In contrast, the Block 3 and 4 treatment schools in our eight-school qualitative sample described their schools' extant data cultures differently. Staff from the Block 3 school reported that there was not a strong history of data use in the school, and that "three years ago, this school didn't use data." The principal described a concerted but slow-going effort to build a data-driven culture in more recent years. Indeed, even during the *Using Data* intervention period, the school's efforts were reportedly hampered by turnover among data team teachers, the data coach, and the principal.

The Block 4 school in the sample was under pressure from the district to raise student performance, and was already engaged in data-use practices. The principal and teachers interviewed described the school as "drowning in data" and accustomed to collaboration.

However, the coach and principal reported that the teachers needed help with data literacy and achievement data analysis, and that *Using Data* had been particularly useful in providing that support. Staff also reported that teachers had previously been asked to focus their instructional efforts on "bubble kids"—that is, students just below the proficiency threshold. However, they had learned that this approach was inconsistent with the *Using Data* philosophy of making data-driven instructional choices that benefitted all students.

In summary, while higher-performing schools reported that data literacy was an entrenched part of their culture, our qualitative evidence suggests that this culture may have been less firmly established, or at least less deeply internalized, in the lower-performing schools in DCPS. Of course, it is difficult to generalize from our small qualitative sample. Beyond just the eight-school sample, we observed that the teachers' data-use knowledge and skills in year 1 were modestly lower in Block 4 schools than in other blocks, despite a *frequency* of data use that was just as high.

This pattern is consistent with the qualitative reports from the Block 4 versus Block 1 and 2 schools, suggesting that the difference lay not only in whether the schools were using data, but in how the data were being used. Further substantiating this notion, we find that the levels of mathematical content knowledge of teachers in Block 4 schools as measured by performance on the LMT survey tended to be lower than those in higher-performing, lower-needs schools (see tables 4.18 and 4.19, below).

Table 4.18. Means of School-Level Teacher Behavior Mediators, by Treatment Condition and by Block, Year 2 (SY 12–13)

	Year 2							
	Block 1		Block 2		Block 3		Block 4	
School-level metric	Treat.	Cont.	Treat.	Cont.	Treat.	Cont.	Treat.	Cont.
Implementation metrics								
Data use	27.0	21.7	26.6	21.0	32.3	29.5	35.0	37.6
Knowledge and skills	18.4	19.8	19.1	18.9	19.2	19.3	18.9	19.1
Attitudes and beliefs	141.0	148.8	149.1	146.0	153.6	150.0	154.8	157.4
Teacher math knowledge								
LMT	-0.058	0.836	0.184	-0.178	-0.143	0.053	-0.229	-0.123
N	7	6	7	7	8	8	8	8

Perhaps as a result of a relatively low level of institutional motivation at baseline to use data, lower-performing, higher-needs schools were more likely to report that *Using Data* protocols were teaching them new ways of analyzing data and providing them with new tools for examining data collaboratively as a team of professionals. This may be primarily an issue of school culture rather than individual motivation to use data.

Table 4.19. Teachers' Level of Mathematical Knowledge and Data Literacy, by Block

	School-level average				Teacher-level average			
Block	LMT[a]	n	Baseline K&S[b]	n	LMT[a]	n	Baseline K&S[b]	n
Block 1	0.355	13	19.1	14	0.346	55	19.1	62
Block 2	0.003	14	18.3	13	0.045	59	18.3	63
Block 3	-0.045	16	18.5	16	-0.094	66	18.5	61
Block 4	-0.176	16	18.2	15	-0.125	61	18.2	56

a. Mathematical knowledge measured by LMT score (mean = 0, s.d. = 1).

b. Data literacy measured by baseline Knowledge & Skills score (s.d. = 1.6 for schools, 2.9 for teacher sample).

Additionally, during the *Using Data* intervention we observed differing levels of engagement with the program by schools in different blocks. These differences in engagement may be related to the differences in data-use cultures. Table 4.18 shows that teachers in Block 3 and 4 schools, both treatment and control, reported engaging in data-use activities somewhat more frequently (31 to 36 times per month) than did teachers in Block 1 or 2 schools (24 to 25 times per month) in 2013–13, as measured by the data-use questions from our Attitudes & Beliefs questionnaire. On our end-of-year year 2 survey, team members in Block 1 schools reported holding fewer data team meetings (7) than did respondents from other school blocks (who reported, on average, 10 or 11 meetings).

In sum, the various evidence sources suggest that *Using Data* professional development may have been more novel and useful for the lower-performing, higher-needs schools in DCPS than for higher-performing, lower-needs schools. Insofar as our eight-school subsample is representative of the broader treatment group, our evidence suggests that *Using Data* may actually have distracted higher-performing Block 1 schools that already had a strong data use culture

from other improvement efforts, while bringing a new culture of data use to lower-performing, higher-needs Block 4 schools.

5. Conclusion: Summary, Limitations, and Implications for Research and Practice

5.1. Summary of Results

We undertook this study to evaluate the efficacy of the *Using Data* intervention to change teachers' data-use practices and improve student learning outcomes in mathematics. Our evaluation found a positive effect of the intervention on teacher behavior after year 1 of the study. Specifically, teachers in the treatment group at the end of year 1 reported using data more frequently, held attitudes and beliefs that were more favorable toward data use for instructional improvement, and exhibited higher levels of data literacy compared with control teachers.

We did not, however, find an overall treatment effect of the program on student performance on annual state math assessments. This result may reflect diminishing returns from additional data use in a district that had already undertaken relatively high levels of data use. We did find some evidence that the UD intervention was more effective when implemented by schools that at baseline performed relatively poorly on the state math assessment and had high levels of socioeconomic need. It is possible that *Using Data* professional development may have been more useful for such schools due to relatively less well developed data-use cultures compared with schools with lower levels of socioeconomic need.

5.2. Limitations of the Study

A critical advantage of this study is that it employed school-level random assignment to evaluate the causal impact of the *Using Data* intervention on teacher skills and behavior and on student outcomes at the school level. Because the UD intervention is designed to effect a gradual whole-school change, this research design, with school-level randomization, is the most rigorous approach available for estimating

causal impacts. Despite this important strength, the study is subject to a number of limitations.

One limitation lies in the generalizability of the research context. As discussed above, DCPS was already using professional development to develop school cultures of data-driven instruction before the UD intervention commenced, and our interviews with district officials suggest that these efforts continued throughout it. On one hand, the district's receptiveness to a professional development program focused on data use provided fertile ground in which an intervention such as *Using Data* might take root; on the other hand, it provided a context in which the control and treatment conditions both had high exposure to core tenets of collaborative data use, thereby possibly attenuating UD's treatment effect. Other randomized trials that have focused on the use of assessment data in the modern school-accountability context have encountered similar challenges (e.g., Konstantopoulos et al. 2013).

Our finding of differential effects for lower-needs versus higher-needs schools corroborates this explanation, since our interviews in eight schools suggest that a strong culture of collaborative data use may have been more firmly established in DCPS's higher-performing schools than in its lower-performing ones. Consistent with that pattern, we estimate positive *Using Data* effects in the block of lower-performing schools, with increasingly negative effects as school performance increases. It is possible, as some administrators even suggested in our interviews, that the overall program effects are larger in districts with less-established cultures of data use already in place.

Another limitation is that the schools were tracked for only two years. Research on implementation timelines suggests that performance may initially dip with new reform implementation, as staff members must acclimate to new processes and expectations (Herold & Fedor 2008; Jellison 2006). However, the findings we report here are not entirely consistent with an implementation dip, since we find evidence of stronger impacts on teacher skills and behavior in the first year than in the second year. This suggests that the impact of the reform was actually wearing off as the intensity of the external professional development declined. Still, the summative event in Spring 2013, at which treatment schools created displays reporting on

their progress and future directions, suggested that some schools maintained momentum in continuing the work of their data teams, and it remains possible that improvements will become apparent in the long run.

Though we collected extensive data about implementation and data practices in both treatment and control schools, these data were also subject to limitations due to logistical constraints. We do have extensive and rich data on implementation and data-use histories in a subset of eight treatment schools; but we were not able to collect such data from all the schools in the sample, and thus we are constrained in our ability to generalize from the qualitative data. Nevertheless, we did observe all treatment schools during *Using Data* training workshops in years 1 and 2, and we did collect log data from all treatment schools periodically about the activities of their data teams.

By triangulating information from the teacher surveys, the eight qualitative focus schools, the workshop observations, and the school logs, we can craft a reasonably useful picture of data use and school practices during the implementation period. But we are still unable to say exactly how the implementation variables of interest changed from the preintervention period to the end of year 2 in all schools.

5.3 Implications for Policy and Practice

Our study suggests that the *Using Data* intervention may be effective in raising student performance in schools without well-established cultures of collaborative and informed data use for instructional improvement. It further suggests that teachers may be most responsive to the UD intervention in terms of their skills, beliefs, and practices when the professional development is most intensive, and that face-to-face maintenance of professional development effort beyond the first implementation year may be warranted to see a continued or expanded impact on teacher behavior. We can only speculate on whether teachers' implementation of the model would have been more consistent in year 2 had the professional development been as intensive in the second year as in the first. It does, however, appear that implementation intensity tracked with professional development intensity across the two implementation years we analyzed.

An important lesson for policymakers is that efforts to reform school culture through collaborative data teams may be hampered by staff turnover at the principal or teacher level. In a number of schools in the sample leadership changed, and with it school priorities, between the point of randomization and the conclusion of the study. Teachers experiencing such changes reportedly found it difficult to find time to make *Using Data* implementation a priority.

Similarly, most of the teams in the study experienced turnover of member teachers between the two implementation years. This turnover disrupted some data teams' group coherence and their collective level of data analysis skills by year 2. Consequently, the lighter-touch professional development effort employed by the *Using Data* facilitators in the second year may have been too light for teams who had not already cohered and mastered the core material from year 1.

At the same time, our qualitative data suggests that some of the higher-performing treatment schools found the UD professional development redundant and insufficiently challenging for the kinds of improvement efforts they wanted to pursue.

In other words, the needs of the schools were diverse, and that diversity may have actually increased during the course of the study. A model in which the professional development was more tailored to the instructional improvement foci and needs of each school might have elicited more consistent engagement and instructional reform. Again, this remains speculation in the current study.

The study should not be taken to suggest that collaborative use of data is ineffective for raising student achievement. To the contrary, a culture of collaborative data use corresponds with high student performance in the study—but that may owe as much to other features of socioeconomically advantaged schools as to the practice of collaborative data use itself. Still, what the study suggests is that, in an environment in which collaborative data use is already a strong practice, the marginal benefit of a program such as *Using Data* may be limited.

This finding still leaves open at least two avenues for future research and practice. First, an evaluation of the intervention in sites with

weaker habits and skills for collaborative data use would be informative to the field. Second, a version of the intervention that is tailored to the needs of schools with collaborative data cultures already in place may be particularly useful in many large school systems, where a focus on data use has become commonplace during the No Child Left Behind era. In particular, participants reported wanting additional support with modifying their instruction in nuanced and promising ways and culling through research to identify evidence-based best practices. These challenges fall within the scope of *Using Data,* but could perhaps merit additional emphasis in a more advanced-level sequence.

In the meantime, the study reinforces the longstanding notion that context matters for professional development implementation and impact (Berman & McLaughlin 1978). Schools and teachers, like students, have heterogeneous needs and growth trajectories. And just like students, those schools and teachers benefit from support tailored to their needs with data, skillfully deployed.

5.4 Ideas for Further Research

Based on our results and discussions with our Advisory Board, the research team proposes to analyze additional issues, which could shed further light on the findings of this *Using Data* evaluation. These issues could include the following:

- Performance of higher-needs students in lower-needs schools
- Grade-level differences in the treatment effect
- Quantity and quality of professional development received by teachers
- School leadership
- Data team attrition
- Linking data use to changes in instructional practices
- Additional fidelity analysis

Performance of Higher-Needs Students in Lower-Needs Schools

One of the main findings of our study is that the *Using Data* program appears to be more effective in schools that have higher levels of

socioeconomic need at baseline. This finding naturally leads to the question of whether there is a general beneficial effect of the program for high-needs students in any school, including low-needs schools. To investigate this question, we plan to modify our school context model (that estimates different treatment effects by school block) by including interaction terms among treatment, school block, and indicators of student need (racial/ethnic minority status, FRL eligibility) and student performance (baseline score on the state math assessment). Adding these terms to the model will allow us to estimate differential treatment effects within blocks for students with different levels of need or performance.

Grade-Level Differences in the Treatment Effect

We also found some evidence that grade 5 students exposed to teachers with two years of *Using Data* training in year 2 performed better than other students. To further examine this issue, we will look at school-level distributions of average student test score growth to see whether there are any differences across grade levels. We will also analyze information from our Attitudes & Beliefs survey on professional development received by teachers, and teacher instructional practices, looking for any differences in these measures by grade.

Quality and Quantity of Professional Development Received by Teachers

Another important question is whether students of teachers who receive more, or better-quality, professional development perform better than other students. To examine this question, we plan to look at a set of questions from our Attitudes & Beliefs survey that ask data team members about the types of professional development they received over the course of the school year. We can break down these data by treatment condition, school block, and grade, to investigate whether differences in the type of PD teachers received plays any role in the overall lack of a treatment effect, and the differential treatment effects we find by block and grade. For example, if control-group teachers were receiving PD that treatment teachers were not, that could reduce the size of any overall treatment effect.

School Leadership

Our Advisory Board suggested we consider ways to integrate the quality of school leadership into our analysis. One way we can do this is to analyze a subset of questions on school leadership on our Attitudes & Beliefs survey. We will take the responses and turn them into a set of school-level leadership scores that can be interacted with the treatment effect in our estimation models. This will allow us to estimate a treatment effect that can differ with this measure of school leadership quality.

Data Team Attrition

It is possible that changes in school-level data team personnel over the course of the study reduced the effectiveness of the *Using Data* program. We will consider the possible effects of data team attrition by calculating, by school, the percentage of data team members at randomization who were replaced, or left the team by the end of year 2. We will examine the distribution of the school-level turnover metric by treatment condition and block. We can also interact this metric with the treatment effect in our regression models, again allowing us to estimate a differential treatment effect by attrition rate.

Linking Data Use to Changes in Instructional Practices

Our Advisory Board emphasized the importance of linking improved data use to actual changes in instructional practices in order to see changes to student achievement. Our Attitudes & Beliefs survey does include several questions related to teachers' instructional practice. We will examine the distribution of responses by treatment condition, grade, block, and year.

Additional Fidelity Analysis

Over the course of the study we accumulated a variety of measures of implementation fidelity. These include, for treatment schools, information on the number of meetings, average meeting length, and attendance rates for school-level data teams. For treatment and control schools, we have survey responses that can be averaged into school-level metrics of teacher data literacy, data use, and attitudes and beliefs toward data use for instructional improvement, as well as the supplemental indicators of data team attrition and school

leadership quality mentioned above. We will examine distributions of these measures by treatment condition and block to better understand differences between treatment and control schools, and between schools with different baseline levels of student performance and socioeconomic need. We will also look at distributions of the measures within treatment schools, as a means of identifying schools that may have implemented the *Using Data* program more or less effectively. We can also include these fidelity measures as covariates, or interaction terms in our regressions estimating the treatment effect, as a means of assessing the extent to which different aspects of implementation fidelity act as mediators or moderators of program effectiveness.

Appendix A: Study Questionnaires – Sample Questions

A1. Attitudes and Beliefs Questionnaire

Attitudes and Beliefs Questions

Please indicate your level of agreement with the following statements by rating each one from Strongly Agree to Strongly Disagree [5-point Likert Scale: 1–strongly disagree, 2–disagree, 3–neither agree nor disagree, 4–agree, 5–strongly agree]:

1. All teachers in a school share responsibility to ensure that every student learns.
2. Research on best teaching practices gives me helpful suggestions about how to improve my teaching.
3. It is important that teachers discuss with each other the achievement differences of students from different races, cultures, and economic classes.
4. I find it helpful to discuss with colleagues relevant research that can inform my practice.
5. It is important to validate my intuition about students' learning needs by examining data.
6. I share responsibility for the learning of other teachers' students.
7. When I meet with other teachers to examine data it is beneficial to my students' learning.
8. Students at all levels of achievement equally need instructional attention.
9. I feel comfortable telling my colleagues when I am struggling with my teaching.
10. I find it difficult to diagnose my students' learning needs in a way that helps me adapt the curriculum to meet their needs.
11. When making instructional decisions, it is important to examine more than one source of information about students.

12. I learn more about my students' needs when I can manipulate and graph their performance data rather than just looking at a score report.
13. Teachers in this school openly discuss student data without fear of retribution.
14. I am only responsible for the learning of the students in my classroom.
15. I can improve my instruction when I work collaboratively with other teachers to analyze data and then generate ideas about ways to change my teaching.
16. I feel confident in my ability to analyze my students' learning data.
17. It is asking too much for teachers to analyze and interpret data.
18. It is important for me to know if my instruction is aligned with state standards.
19. It is important to evaluate student progress on a regular basis throughout the school year and to adjust instruction accordingly.
20. I feel comfortable talking with other teachers about differences in the achievement of students from different races, cultures, and economic classes.
21. Most teachers at this school believe that their intuitive understanding of students' learning needs provides them with sufficient information for planning instruction.
22. All students, regardless of family background, are capable of achieving at high levels.
23. I would prefer to have someone else analyze and interpret data and give me the findings rather than me analyzing and interpreting the data myself.
24. Research on instruction rarely provides me with useful information.
25. The progress of my students is the most important guide for planning instruction.
26. When my colleagues and I discuss trends in data, I can gain useful insights from my colleagues.
27. At this school, teachers are expected to help each other with instruction.
28. Analyzing and interpreting data takes more time than it's worth.
29. When planning instruction, I find it more helpful to work on my own than to work as a team with other teachers.
30. I can ask other teachers for help with instructional practices.
31. My curriculum is the most important guide for planning instruction.

32. I would prefer not to share my students' data with other teachers.

33. When it comes to interpreting graphs and tables, I feel anxious.

34. Teachers at this school focus their instruction on the students who are close to meeting curriculum standards.

35. I make time to help colleagues who are struggling with their teaching.

36. It is impossible to eliminate the achievement gaps between different ethnic/racial groups.

37. Monitoring student learning on a weekly basis takes more time than it's worth.

Data Use Questions

In school year 2012-2013, how often did you use data for each of the following purposes? (Mark one box per row.) In 2012/13, I used data... [Possible answers: Never, A few times, Once or twice a month, Once a week or more]

a. to inform curriculum changes

b. to identify individual skill gaps for individual students

c. to determine whether your class or individual students were ready to move on to the next instructional unit

d. to evaluate promising classroom practices

e. to decide to give your students test-taking practice

f. to estimate whether your students would make adequate yearly progress (AYP)

g. to track standardized test scores by grade

h. to track individual student test scores

i. to track other measures of student progress

j. to inform student placement in courses or special programs

k. to inform parents about student progress

In school year 2012-2013, how often did you work with data in the following contexts to make instructional decisions? (Mark one box per row) [Possible answers: Never, A few times, Once or twice a month, Once a week or more]

a. on your own

b. working with colleagues in your grade level

c. working with colleagues from other grade levels

d. as part of a district-level activity with staff from other schools

e. in another setting

A2. Sample Questions from Knowledge & Skills Questionnaire

Directions: Please read each question and select the best answer among the four multiple choices.

Please use Table 2 to answer questions 16-19.

Table 2: An example of item-level data from a state test for a class of students.

		Percent of students in class choosing each answer * denotes correct answer				Percent correct		
Item #	**Content Strand**	**A**	**B**	**C**	**D**	**School**	**District**	**State**
1	Probability & Statistics	2	90*	6	2	90	85	93
2	Probability & Statistics	23	7	67*	3	70	62	65
3	Measurement	2	78*	5	14	78	79	93
4	Measurement	45	10	37*	8	37	27	44

16. On what item do students in the class perform the best? [Potential answers: 1, 2, 3, or 4]

17. On what item do students in the district outperform students in the class? [Potential answers: 1, 2, 3, or 4]

18. On what item do most students give the same **incorrect** answer? [Potential answers: 1, 2, 3, or 4]

19. How does the school-level performance on item 2 compare to the state-level performance? [Potential answers: the same; school performance is better than state performance; school performance is worse than state performance; there is not enough information on the table to answer]

Appendix B: Psychometric Analysis of *Using Data* Scales[22]

B1. Introduction

Two questionnaires were developed by CNA as part of the *Using Data* study. The questionnaires measured three aspects of elementary teachers' mathematics instruction:

- **Knowledge and skills**: The Knowledge & Skills questionnaire measured a teacher's data literacy skills for determining gaps in student learning that should influence instructional decisions.

The Attitudes & Beliefs questionnaire measured two aspects of instruction:

- **Attitudes and beliefs**: The questionnaire explored a teacher's attitudes and beliefs about using data for instructional improvement in the context of teaching and collaborating with colleagues.
- **Data use**: The questionnaire also asked about a teacher's actual experience concerning use of data resources in the school.[23]

This psychometric report presents information on the reliability of these three scales—Knowledge & Skills, Attitudes & Beliefs, and Data Use. Its final section evaluates the validity of the Knowledge & Skills scale.

These questionnaires were developed to measure the effectiveness of a professional development curriculum called *Using Data,* designed to

[22] This section was written by Frank Jenkins, statistician at Westat, 1600 Research Blvd., Rockville, MD 20850.

[23] Questions pertaining to level of data use were partially based on a survey conducted as part of the National Educational Technology Trends Study (NETTS). U.S. Department of Education, National Educational Technology Trends Study, Local-level Data Summary (2008).

help teachers use student data to inform instructional decision making. CNA Education, in partnership with TERC, the program developer, conducted a randomized controlled trial to examine the impact and efficacy of the *Using Data* (UD) intervention on teachers' ability to use data and improve student math performance in grades 4 and 5 in a large urban school system. The evaluation is funded through the Institute of Education Sciences's Teacher Quality: Mathematics and Science Education grant program.[24]

The questionnaires were developed in a pilot study carried out in Spring 2011 by CNA, in which draft versions of the questionnaires were administered to 53 elementary mathematics teachers in 34 schools in Maryland and Virginia. Revisions were made to the instruments as a result of this pilot.

The final versions of the questionnaires were administered in Fall 2011 and Spring 2012 to grade 4 and 5 mathematics teachers in DCPS. The Fall 2011 administration was given just prior to the random assignment of teachers to the *Using Data* treatment group and to a control group where the treatment was not administered. Treatment was started just after randomization. A posttest of the questionnaires was administered in Spring 2012 to all of the available teachers who had been randomly assigned to the treatment and control groups the previous fall.

Comparisons of the posttest and the pretest responses were used to evaluate the effectiveness of the *Using Data* intervention for increasing the frequency and quality of data use in mathematics teaching. The study also will evaluate whether the intervention improves students' mathematics performance in grades 4 and 5.

A final administration of the *Using Data* questionnaires was given in Spring 2013 to study participants, in part for the purposes of evaluating the psychometric properties of the scales and establishing the validity of two of them, Knowledge & Skills and Attitudes & Beliefs. In order to assess validity of these two scales, an established measure of a teacher's content knowledge of number concepts and operations for elementary school was administered concurrently. This

[24] Now the Effective Teachers and Effective Teaching topic area of the IES Education Research Grant program.

established measure is a subscale of the Learning Mathematics for Teaching Project assessment (LMT).[25,26]

Further details of the two scales that are the subject of the validity study are given below.

Knowledge & Skills Questionnaire

The Knowledge & Skills assessment was developed to measure a teacher's knowledge of data-use concepts and application of data-use skills when interpreting tables and graphs. Items were written to measure general data literacy, knowledge of types of data and appropriate use (aggregate data, disaggregate data, strand data, item data), and processes when interpreting data displays.

The Spring 2011 pilot administration consisted of 39 items for the Knowledge & Skills questionnaire. After the pilot, 14 questions were dropped, leaving a total of 25 questions on the revised questionnaire.

Attitudes & Beliefs Questionnaire

The Attitudes & Beliefs assessment was developed to measure teacher agreement with core assumptions of the *Using Data* process, the basis of the *Using Data* treatment intervention with teachers. The researchers wrote five to seven items for each of six dimensions, or subscales, of what the UD program calls "collaborative inquiry":

- Encouraging a collaborative culture among teachers
- Equitable treatment of diverse students
- Nurturing trust
- Collaborating with other teachers
- Using data to focus on learning problems of students
- Instructional improvement

[25] See Schilling, Blunk, & Hill (2007).

[26] The Learning Mathematics for Teaching project website can be found at http://sitemaker.umich.edu/lmt/home.

In the pilot test, five of the six subscales had internal consistency reliabilities ranging from .70 to .91, well above the What Works Clearinghouse minimal standard of .50.[27] One subscale, Instructional Improvement, had a reliability of .31, below the standard of .50.

B2. Psychometric Analysis of Spring 2013 Administration of the *Using Data* Questionnaires

Item Analysis

Frequencies

Frequencies were run on the 25 items of the Knowledge & Skills questionnaire, after the data was scored as 1=right or 0=wrong. There were 259 cases. Of the 259, there were three individuals missing responses on every item.

Overall, the items were rather easy, with an average percent correct over all items of 76 percent. Ten items had percent correct of 85 percent or higher, and only two items had a percent correct of less than 50 percent. Table B.4 gives the percent correct for each item.

For the Attitudes & Beliefs questionnaire, items were scored from 1=strongly agree to 5=strongly disagree, with five categories of agreement. Some items were worded in a positive way and some in a negative. More than half of the items were recoded to make higher scores positive (5=strongly agree). The three same individuals were missing for each case. Table B.5 lists the frequencies for the 37 items, reflecting recoding to make all items positive.

Overall, the items tended to be answered positively, with 16 of the 37 items having 85 percent or higher in the most positive two categories.

The 16 Data Use questions in the Attitudes & Beliefs questionnaire were scored on a four-category Likert scale. The general question for the set of items was "In the past school year how often did you [do the following]...." In order to better reflect the spacing of categories in the underlying time dimension, the categories were scored as

[27] What Works Clearinghouse, *Procedures and Standards Handbook (Version 3.0),* http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19.

follows for purposes of creating total scores and for reliability analysis: 0.0=never, 0.5=a few times, 1.5=once a month, and 4.0=once a week.

Only one case was missing, the same person for each item. The majority of responses were in the Once a Month or Once a Week category. Table B.6 lists the frequencies for the 16 items.

Reliability Analysis

A reliability analysis was done on all three of the *Using Data* scales. In summary, all of the items had acceptable reliability. An internal consistency reliability, or Cronbach's alpha,[28] was calculated to indicate how consistently sets of items measured an underlying construct. Note, however, that the final reliability results occurred after several changes were made in the scales, some of them already mentioned:

- Knowledge & Skills: Items 15, 20, and 24 were deleted because they had a near zero correlation with the total scale score. That left 22 items in the final scale.
- Attitudes & Beliefs: Item 21 was deleted because it had a near zero correlation with the total scale score. That left 36 items in the final scale. As mentioned before, more than half the items (23) were reverse coded to make them positive with respect to the measured construct.

The final reliabilities are presented in table B.1. Note that the standardized alpha reliability scores are the reliabilities that result when all of the items are standardized. The standardized scores are nearly identical to the raw scores.

Table B.1. Reliabilities of the Three *Data Use* Scales

Scale	Knowledge & Skills	Attitudes & Beliefs	Data Use*
Raw	0.67	0.95	0.87
Standardized	0.69	0.95	0.87

* No items were deleted or recoded.

[28] See Cronbach (1951).

All scales had reliabilities well above the What Works Clearinghouse minimum standard of .50. All were above or close to the .70 rule of thumb commonly used as a more stringent standard for cognitive measures.

Confirmation of Simple Structure of Scales

Table B.7 shows the item to total scale correlations for the three scales in the *Using Data* study. The yellow highlighted regions of each column indicate the items that were hypothesized to measure the scale named in the heading of the column. For example, all of the $NKSQ_n$ items were assumed to measure the Knowledge & Skills scale. The item-scale correlation of these items with the Knowledge & Skills scale is highlighted for ease of comparison.

Indeed, we see that in column 1, the highlighted correlations are much higher than the unhighlighted correlations, indicating that every item in the Knowledge & Skills scale has a much higher correlation with the scale it was intended to measure than with other scales. If we look at the highlighted correlations in columns 2 and 3, we see the same pattern—that is, the $AttBlfP_n$ items and $DataUse_n$ items also strongly measure the scales they were hypothesized to measure.

This confirms the "simple structure" assumption of the study, that every item measures only the scale it was hypothesized to measure and no other scales.[29]

B3. IRT Scaling of the Three *Using Data* Scales

There was a sufficient number of respondents (259) to estimate scales using item response theory (IRT).[30, 31] An advantage of IRT scaling versus observed total scores is that IRT scales are approximately normally distributed, and each item is weighted by

[29] Muliak, S. A. (2005). Looking back on the indeterminacy controversies in factor analysis. In *Contemporary psychometrics*. Mahwah, NJ: Erlbaum.

[30] Baker, F. B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques* (2d ed., rev.). New York: Marcel Dekker.

[31] du Toit, Mathilda, ED. (2006). IRT from SSI, Scientific Software International, Inc.

how strongly it correlates with the scale being measured, resulting in more precise scale estimates than with item averages.[32]

In addition, with an IRT analysis, more information is available about how each item measures an underlying construct because the analysis produces an item response function for each item, which is a graph of the probability of a response to an item versus the scale score (theta). Figures B.1 through B.3 present matrices of the item response functions for each of the three CNA scales. The graphs in the figures are numbered by row; that is, item 1 is at the top left, item 2 is adjacent to it on the right, etc. At the end of a row, the next item is the leftmost item in the row below.

The graphs in figure B.1 show the modeled probability of a correct response (blue line) and the parallel probability of a wrong response (black line) for the 22 Knowledge & Skills scale items. Except for items 16 and 19, which show a flat curve indicative of weak measurement, these graphs show acceptable fit to the IRT model. A flat curve means that the probability of a correct response is very much the same, regardless of the person's overall scale score. This implies that the item has a small correlation with the scale and gives little information about the person's scale score. In other words, the item is a weak measure of the scale.

Figure B.2 displays the item response functions for the 36 items in the Attitudes & Beliefs scale. These show the probability of being in one of the five categories of the items. Instead of right or wrong, these items indicate degrees of propensity for a person to manifest qualities of the underlying scale. For example, a high probability of being in category 1 (1=strongly disagree) indicates that the person has a low propensity for collaborative inquiry. Conversely, a high probability of being in category 5 (5=strongly agree) indicates the person has a high propensity for collaborative inquiry. Note, however, that some items were collapsed because of small frequencies in a category. This was necessary because about 20 responses in each category are necessary for the scaling program to converge estimate probabilities of being in that category.

---

[32] Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

All but two of the items demonstrated item responses where the probability of being in one category or the other is well defined. In items 10 and 30, the item responses are relatively flat, indicating that they are weak measures of the underlying construct. Again, with the exception of a couple of items, the IRT model seems to fit well.

In figure B.3 we find the item response functions for the Data Use scale. All 16 of these items have well-defined probabilities of being in a category of response as a function of the underlying construct. This indicates that all items are strong measures of the scale.

A practical advantage of IRT is that it accommodates missing data, in that a scale score is estimated for a respondent if any item is responded to. However, since there was little missing data, this was not an issue.

B4. Validity of the Knowledge & Skills and the Attitudes & Beliefs Scales

During the 2013 administration of the *Using Data* scales, the Number Concepts and Operations (NCOP) subtest of the Learning Mathematics for Teaching (LMT) assessment was concurrently administered. This was a tailored computer adaptive test, where the set of items presented to the respondent is determined by a computer algorithm to best match the ability of the person taking the test. A computer-generated scale score is available on the data file.

The main focus of the validity study is the comparison of the Knowledge & Skills scale with the LMT. Both of these assessments are meant to measure a teacher's ability to use mathematics knowledge and skills to effectively teach mathematics skills to elementary school students. It should be noted, however, that the Knowledge & Skills scale and the LMT were not designed to capture the same content. The Knowledge & Skills scale is meant to measure elements of the *Using Data* curriculum being assessed in the CNA study. The LMT was developed as a general measure of a narrow subset of the elementary mathematics curriculum.

The second *Using Data* scale, Attitudes & Beliefs, will also be included in the validity comparison with the LMT, with the caveat that the LMT was not designed to measure the attitudes and beliefs about

teaching and teaching practice of elementary school teachers. As a result, this validity comparison is of limited use.

Table B.2 shows the correlations between the two *Using Data* scales and the LMT. The numbers in parentheses are the probability of the significance test that the correlation is zero. Probabilities less than .05 mean that the correlation is significantly different from zero.

Table B.2. Observed Scale Correlations with LMT Scale

	K&S score	A&B score	LMT score
K&S score	1.00	0.15	0.42
		(0.016)	(<.000)
A&B score	0.15	1.00	0.08
	(0.016)		(.181)
LMT scale	0.42	0.08	1.00
	(<.000)	(.181)	

KEY: K&S score=observed total score for the Knowledge & Skills assessment. A&B score=observed total score for the Attitudes & Beliefs assessment. LMT scale=IRT scale score for the Learning Mathematics for Teaching assessment.

The Knowledge & Skills score has a small correlation with the Attitudes & Beliefs score (.15). It also has a moderate correlation (.42) with the LMT IRT scale. While this provides some evidence for the validity of the Knowledge & Skills scale, it underlines the fact that these two assessments were focused on different content.

It is our opinion that the validity of the Knowledge & Skills scale is moderately confirmed. The degree to which the two assessments are designed to capture different aspects of content and teaching practices needs to be further explored.

As mentioned, the Attitudes & Beliefs total score has a small correlation with the Knowledge & Skills total score. However, it has a correlation with the LMT that is not significantly different from zero.

Table B.3 shows the correlation among the same measures, but this time the Knowledge & Skills and Attitudes & Beliefs measures are IRT scales. The only difference between this and the previous table is that the correlation between the Knowledge & Skills scale and the Attitudes & Beliefs scale is smaller (.11 vs. .15) and is not significantly different from zero.

Table B.3. IRT Scale Correlations with LMT Scale

	K&S scale	A&B scale	LMT scale
KS scale	1	0.11	0.42
		(.075)	(<.000)
AB scale	0.11	1	0.08
	(.075)		(.188)
LMT scale	0.42	0.08	1
	(<.000)	(.188)	

KEY: K&S scale=IRT score for the Knowledge & Skills assessment. A&B scale=IRT score for the Attitudes & Beliefs assessment. LMT scale=IRT scale score for the Learning Mathematics for Teaching assessment.

B5. Conclusions

The *Using Data* scales are reliable. The internal consistency reliabilities are at or above the stringent standard for educational cognitive scales of .70.

The *Using Data* scales can be successfully scaled with the item response theory model. However, the IRT scales result in similar scale correlations with each other and with the LMT scale as the observed total scores.

The validity analysis yielded mixed results. While the Knowledge & Skills assessment did not have such a high correlation with LMT as to be redundant—that is, providing identical information as the LMT and therefore not a useful addition to the set of current teacher assessments—it had only a moderate correlation with the LMT. As a result, the degree that the already established validity of the LMT can be extended to the Knowledge & Skills assessment is limited.

As explained above, the validity correlation of .42 is moderate confirmation of the validity of the Knowledge & Skills assessment. It would be beneficial to explicate to what extent the CNA and LMT assessments are intentionally focused on different content and teacher practices.

As a next step, it would be useful to explore the validity of the Knowledge & Skills scale for predicting the mathematics achievement of students of the teachers who took the assessment.

B.6. Item Analysis Tables and Graphs

Table B.4. Percent Correct for the Knowledge & Skills Items

Note: Shaded questions were not retained in the final scale.

Question	% Correct
KSQ1	95
KSQ2	53
KSQ3	92
KSQ4	56
KSQ5	95
KSQ6	87
KSQ7	75
KSQ8	93
KSQ9	86
KSQ10	63
KSQ11	71
KSQ12	64
KSQ13	63

Question	% Correct
KSQ14	82
KSQ15	28
KSQ16	97
KSQ17	95
KSQ18	84
KSQ19	96
KSQ20	27
KSQ21	77
KSQ22	82
KSQ21	77
KSQ22	82
KSQ23	85
KSQ24	79
KSQ25	75

Table B.5. Frequencies on the Attitudes & Beliefs Scale

Note: Original responses were: 1=strongly agree, 2=agree, 3=neither, 4=disagree, 5=strongly disagree. Items where disagreement was a negative indicator of the underlying construct were reverse coded. The frequencies reflect this reverse coding.

AttBlf1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7	2.71	7	2.71
2	18	6.98	25	9.69
3	13	5.04	38	14.73
4	88	34.11	126	48.84
5	132	51.16	258	100.00

AttBlf2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	2.33	6	2.33
2	7	2.71	13	5.04
3	8	3.10	21	8.14
4	91	35.27	112	43.41
5	146	56.59	258	100.00

AttBlf3	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	2.33	6	2.33
2	9	3.49	15	5.81
3	14	5.43	29	11.24
4	82	31.78	111	43.02
5	147	56.98	258	100.00

AttBlf4	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	3	1.16	3	1.16
2	8	3.10	11	4.26
3	10	3.88	21	8.14
4	91	35.27	112	43.41
5	146	56.59	258	100.00

AttBlf5	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7	2.71	7	2.71
2	6	2.33	13	5.04
3	4	1.55	17	6.59
4	62	24.03	79	30.62
5	179	69.38	258	100.00

AttBlf6	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7	2.71	7	2.71
2	7	2.71	14	5.43
3	29	11.24	43	16.67
4	96	37.21	139	53.88
5	119	46.12	258	100.00

AttBlf7	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	9	3.49	9	3.49
2	3	1.16	12	4.65
3	11	4.26	23	8.91
4	84	32.56	107	41.47
5	151	58.53	258	100.00

AttBlf8	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	9	3.49	9	3.49
2	10	3.88	19	7.36
3	14	5.43	33	12.79
4	55	21.32	88	34.11
5	170	65.89	258	100.00

AttBlf9	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	5	1.94	5	1.94
2	19	7.36	24	9.30
3	26	10.08	50	19.38
4	108	41.86	158	61.24
5	100	38.76	258	100.00

AttBlf10	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	24	9.30	24	9.30
2	42	16.28	66	25.58
3	35	13.57	101	39.15
4	117	45.35	218	84.50
5	40	15.50	258	100.00

AttBlf11	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	8	3.10	8	3.10
2	7	2.71	15	5.81
3	5	1.94	20	7.75
4	51	19.77	71	27.52
5	187	72.48	258	100.00

AttBlf12	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	5	1.94	5	1.94
2	17	6.59	22	8.53
3	32	12.40	54	20.93
4	96	37.21	150	58.14
5	108	41.86	258	100.00

AttBlf13	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7	2.71	7	2.71
2	31	12.02	38	14.73
3	50	19.38	88	34.11
4	105	40.70	193	74.81
5	65	25.19	258	100.00

AttBlf14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	22	8.53	22	8.53
2	30	11.63	52	20.16
3	30	11.63	82	31.78
4	92	35.66	174	67.44
5	84	32.56	258	100.00

AttBlf15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	9	3.49	9	3.49
2	7	2.71	16	6.20
3	4	1.55	20	7.75
4	84	32.56	104	40.31
5	154	59.69	258	100.00

AttBlf16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	8	3.10	8	3.10
2	11	4.26	19	7.36
3	11	4.26	30	11.63
4	103	39.92	133	51.55
5	125	48.45	258	100.00

AttBlf17	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	10	3.88	10	3.88
2	15	5.81	25	9.69
3	34	13.18	59	22.87
4	86	33.33	145	56.20
5	113	43.80	258	100.00

AttBlf18	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	8	3.10	8	3.10
2	6	2.33	14	5.43
3	2	0.78	16	6.20
4	63	24.42	79	30.62
5	179	69.38	258	100.00

AttBlf19	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7	2.71	7	2.71
2	7	2.71	14	5.43
3	2	0.78	16	6.20
4	38	14.73	54	20.93
5	204	79.07	258	100.00

AttBlf20	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7	2.71	7	2.71
2	14	5.43	21	8.14
3	24	9.30	45	17.44
4	98	37.98	143	55.43
5	115	44.57	258	100.00

AttBlf21	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	16	6.20	16	6.20
2	52	20.16	68	26.36
3	79	30.62	147	56.98
4	83	32.17	230	89.15
5	28	10.85	258	100.00

AttBlf22	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	2.33	6	2.33
2	24	9.30	30	11.63
3	20	7.75	50	19.38
4	97	37.60	147	56.98
5	111	43.02	258	100.00

AttBlf23	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	12	4.65	12	4.65
2	19	7.36	31	12.02
3	36	13.95	67	25.97
4	88	34.11	155	60.08
5	103	39.92	258	100.00

AttBlf24	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	18	6.98	18	6.98
2	18	6.98	36	13.95
3	31	12.02	67	25.97
4	118	45.74	185	71.71
5	73	28.29	258	100.00

AttBlf25	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	5	1.94	5	1.94
2	15	5.81	20	7.75
3	21	8.14	41	15.89
4	111	43.02	152	58.91
5	106	41.09	258	100.00

AttBlf26	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	5	1.94	5	1.94
2	9	3.49	14	5.43
3	18	6.98	32	12.40
4	123	47.67	155	60.08
5	103	39.92	258	100.00

AttBlf27	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	2.33	6	2.33
2	34	13.18	40	15.50
3	53	20.54	93	36.05
4	105	40.70	198	76.74
5	60	23.26	258	100.00

AttBlf28	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	14	5.43	14	5.43
2	14	5.43	28	10.85
3	32	12.40	60	23.26
4	82	31.78	142	55.04
5	116	44.96	258	100.00

AttBlf29	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	14	5.43	14	5.43
2	53	20.54	67	25.97
3	45	17.44	112	43.41
4	97	37.60	209	81.01
5	49	18.99	258	100.00

AttBlf30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	5	1.94	5	1.94
2	12	4.65	17	6.59
3	13	5.04	30	11.63
4	132	51.16	162	62.79
5	96	37.21	258	100.00

AttBlf31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	9	3.49	9	3.49
2	58	22.48	67	25.97
3	67	25.97	134	51.94
4	102	39.53	236	91.47
5	22	8.53	258	100.00

AttBlf32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7	2.71	7	2.71
2	11	4.26	18	6.98
3	24	9.30	42	16.28
4	133	51.55	175	67.83
5	83	32.17	258	100.00

AttBlf33	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	2.33	6	2.33
2	23	8.91	29	11.24
3	34	13.18	63	24.42
4	100	38.76	163	63.18
5	95	36.82	258	100.00

AttBlf34	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	8	3.10	8	3.10
2	53	20.54	61	23.64
3	80	31.01	141	54.65
4	84	32.56	225	87.21
5	33	12.79	258	100.00

AttBlf35	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	2	0.78	2	0.78
2	13	5.04	15	5.81
3	34	13.18	49	18.99
4	133	51.55	182	70.54
5	76	29.46	258	100.00

AttBlf36	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	14	5.43	14	5.43
2	24	9.30	38	14.73
3	32	12.40	70	27.13
4	96	37.21	166	64.34
5	92	35.66	258	100.00

AttBlf37	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7	2.71	7	2.71
2	16	6.20	23	8.91
3	34	13.18	57	22.09
4	88	34.11	145	56.20
5	113	43.80	258	100.00

Table B.6. Frequencies on the Data Use Scale

DataUse2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	15	5.81	15	5.81
0.5	59	22.87	74	28.68
1.5	90	34.88	164	63.57
4	94	36.43	258	100.00

DataUse3	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	4	1.55	4	1.55
0.5	29	11.24	33	12.79
1.5	101	39.15	134	51.94
4	124	48.06	258	100.00

DataUse4	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	7	2.71	7	2.71
0.5	20	7.75	27	10.47
1.5	99	38.37	126	48.84
4	132	51.16	258	100.00

DataUse5	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	18	6.98	18	6.98
0.5	58	22.48	76	29.46
1.5	100	38.76	176	68.22
4	82	31.78	258	100.00

DataUse6	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	20	7.75	20	7.75
0.5	80	31.01	100	38.76
1.5	85	32.95	185	71.71
4	73	28.29	258	100.00

DataUse7	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	24	9.30	24	9.30
0.5	91	35.27	115	44.57
1.5	97	37.60	212	82.17
4	46	17.83	258	100.00

DataUse8	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	14	5.43	14	5.43
0.5	115	44.57	129	50.00
1.5	98	37.98	227	87.98
4	31	12.02	258	100.00

DataUse9	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	4	1.55	4	1.55
0.5	39	15.12	43	16.67
1.5	108	41.86	151	58.53
4	107	41.47	258	100.00

DataUse10	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	10	3.88	10	3.88
0.5	46	17.83	56	21.71
1.5	92	35.66	148	57.36
4	110	42.64	258	100.00

DataUse11	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	70	27.13	70	27.13
0.5	94	36.43	164	63.57
1.5	71	27.52	235	91.09
4	23	8.91	258	100.00

DataUse12	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	8	3.10	8	3.10
0.5	49	18.99	57	22.09
1.5	139	53.88	196	75.97
4	62	24.03	258	100.00

DataUse13	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1	0.39	1	0.39
0.5	17	6.59	18	6.98
1.5	61	23.64	79	30.62
4	179	69.38	258	100.00

DataUse14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5	1.94	5	1.94
0.5	48	18.60	53	20.54
1.5	128	49.61	181	70.16
4	77	29.84	258	100.00

DataUse15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	52	20.16	52	20.16
0.5	109	42.25	161	62.40
1.5	64	24.81	225	87.21
4	33	12.79	258	100.00

DataUse16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	81	31.40	81	31.40
0.5	128	49.61	209	81.01
1.5	35	13.57	244	94.57
4	14	5.43	258	100.00

DataUse17	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	92	35.66	92	35.66
0.5	122	47.29	214	82.95
1.5	29	11.24	243	94.19
4	15	5.81	258	100.00

Table B.7. Item to Total Scale Correlations

Note: Table B.7 shows the item to total scale correlations for the three scales in the *Using Data* study. The yellow highlighted regions of each column indicate the items that were hypothesized to measure the scale named in the heading of the column. For example, all of the NKSQn items were assumed to measure the Knowledge & Skills scale. The item-scale correlation of these items with the Knowledge & Skills scale is highlighted for ease of comparison. Indeed we see that in column 1, the highlighted correlations are much higher than the unhighlighted correlations, indicating that every item in the Knowledge & Skills scale has a much higher correlation with the scale it was intended to measure than with other scales. If we look at the highlighted correlations in columns 2 and 3, we see the same pattern—that is, the AttBlfPn items and DataUsen items also strongly measure the scales they were hypothesized to measure. This confirms the "simple structure" assumption of the study, that every item measures only the scale it was hypothesized to measure and no other scales.

	Knowledge & Skills	Attitudes & Beliefs	Data Use
NKSQ1	0.22	-0.04	-0.02
NKSQ2	0.39	0.11	0.03
NKSQ3	0.22	0.05	-0.10
NKSQ4	0.41	0.05	-0.03
NKSQ5	0.29	0.08	0.02
NKSQ6	0.43	0.00	-0.07
NKSQ7	0.36	0.02	-0.12
NKSQ8	0.36	0.07	-0.03
NKSQ9	0.54	-0.02	-0.13
NKSQ10	0.43	0.14	0.04
NKSQ11	0.30	0.00	0.01
NKSQ12	0.37	0.10	0.13
NKSQ13	0.37	0.01	0.01
NKSQ14	0.47	0.03	-0.05
NKSQ16	0.25	-0.02	-0.04
NKSQ17	0.19	0.05	-0.10
NKSQ18	0.39	0.04	-0.14
NKSQ19	0.37	0.12	0.02

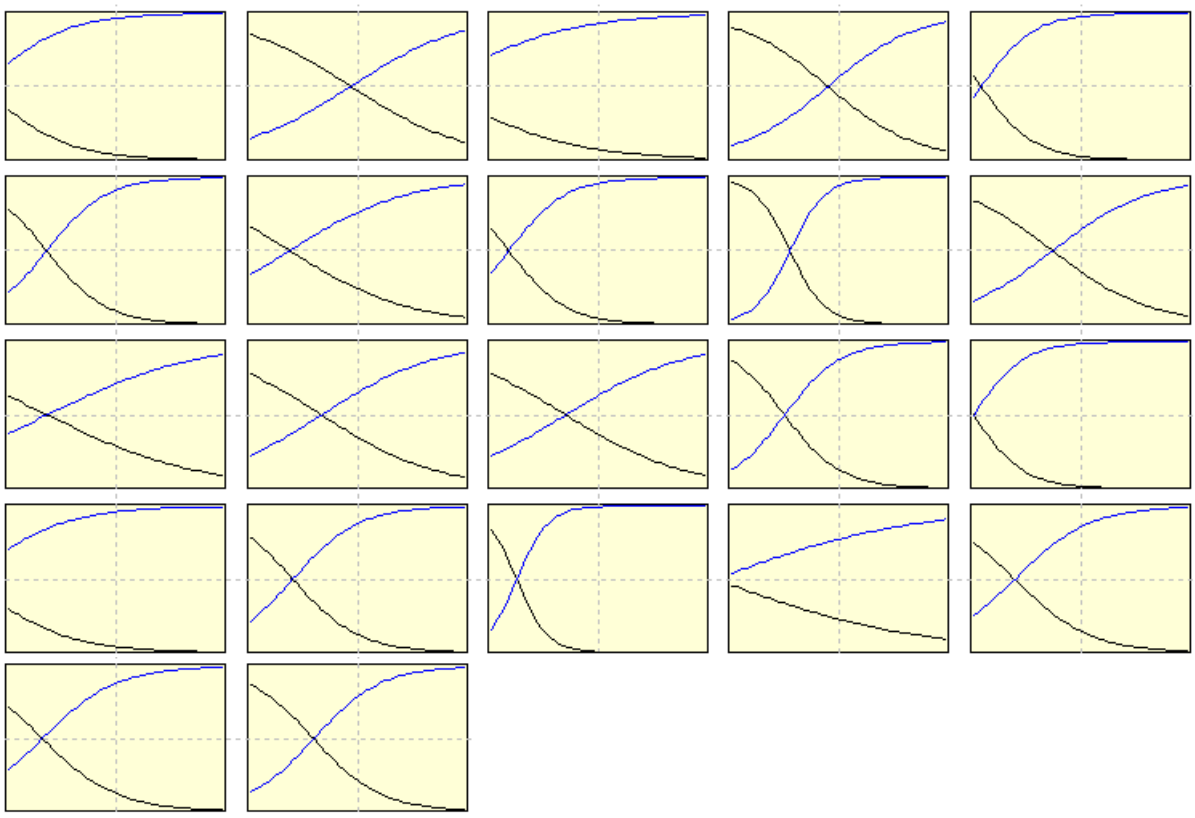
	Knowledge & Skills	Attitudes & Beliefs	Data Use
NKSQ21	0.28	0.12	0.22
NKSQ22	0.38	0.10	-0.04
NKSQ23	0.42	0.04	-0.01
NKSQ25	0.47	0.10	-0.12
AttBlfP1	0.06	0.69	0.17
AttBlfP2	0.04	0.76	0.18
AttBlfP3	0.04	0.67	0.08
AttBlfP4	0.07	0.77	0.18
AttBlfP5	0.14	0.80	0.16
AttBlfP6	0.16	0.67	0.20
AttBlfP7	0.14	0.77	0.21
AttBlfP8	0.07	0.66	0.13
AttBlfP9	0.00	0.62	0.10
AttBlfP10	0.11	0.31	0.09
AttBlfP11	0.04	0.77	0.16
AttBlfP12	0.03	0.67	0.22
AttBlfP13	-0.11	0.46	0.14
AttBlfP14	0.08	0.47	0.03
AttBlfP15	0.05	0.80	0.17
AttBlfP16	0.06	0.64	0.27
AttBlfP17	0.19	0.63	0.19
AttBlfP18	0.12	0.79	0.14
AttBlfP19	0.03	0.84	0.15
AttBlfP20	0.05	0.69	0.17
AttBlfP22	-0.01	0.58	0.18
AttBlfP23	0.20	0.52	0.26
AttBlfP24	0.15	0.60	0.11
AttBlfP25	-0.01	0.62	0.17
AttBlfP26	-0.02	0.77	0.26
AttBlfP27	-0.03	0.48	0.17

	Knowledge & Skills	Attitudes & Beliefs	Data Use
AttBlfP28	0.18	0.67	0.19
AttBlfP29	0.14	0.45	0.07
AttBlfP30	0.02	0.62	0.15
AttBlfP31	0.12	0.27	0.02
AttBlfP32	0.17	0.61	0.17
AttBlfP33	0.27	0.49	0.06
AttBlfP34	0.13	0.37	0.22
AttBlfP35	0.10	0.60	0.24
AttBlfP36	0.24	0.49	0.06
AttBlfP37	0.20	0.65	0.29
DataUse2	0.10	0.14	0.65
DataUse3	0.02	0.29	0.68
DataUse4	-0.01	0.20	0.67
DataUse5	-0.09	0.18	0.72
DataUse6	-0.08	0.11	0.67
DataUse7	-0.11	0.18	0.62
DataUse8	-0.09	0.07	0.59
DataUse9	-0.04	0.16	0.62
DataUse10	0.08	0.16	0.69
DataUse11	-0.15	0.01	0.56
DataUse12	-0.13	0.14	0.55
DataUse13	0.15	0.10	0.49
DataUse14	0.01	0.21	0.58
DataUse15	-0.08	0.19	0.58
DataUse16	-0.07	0.17	0.49
DataUse17	-0.07	0.19	0.52

Figures B1–B3. Item Response Curves

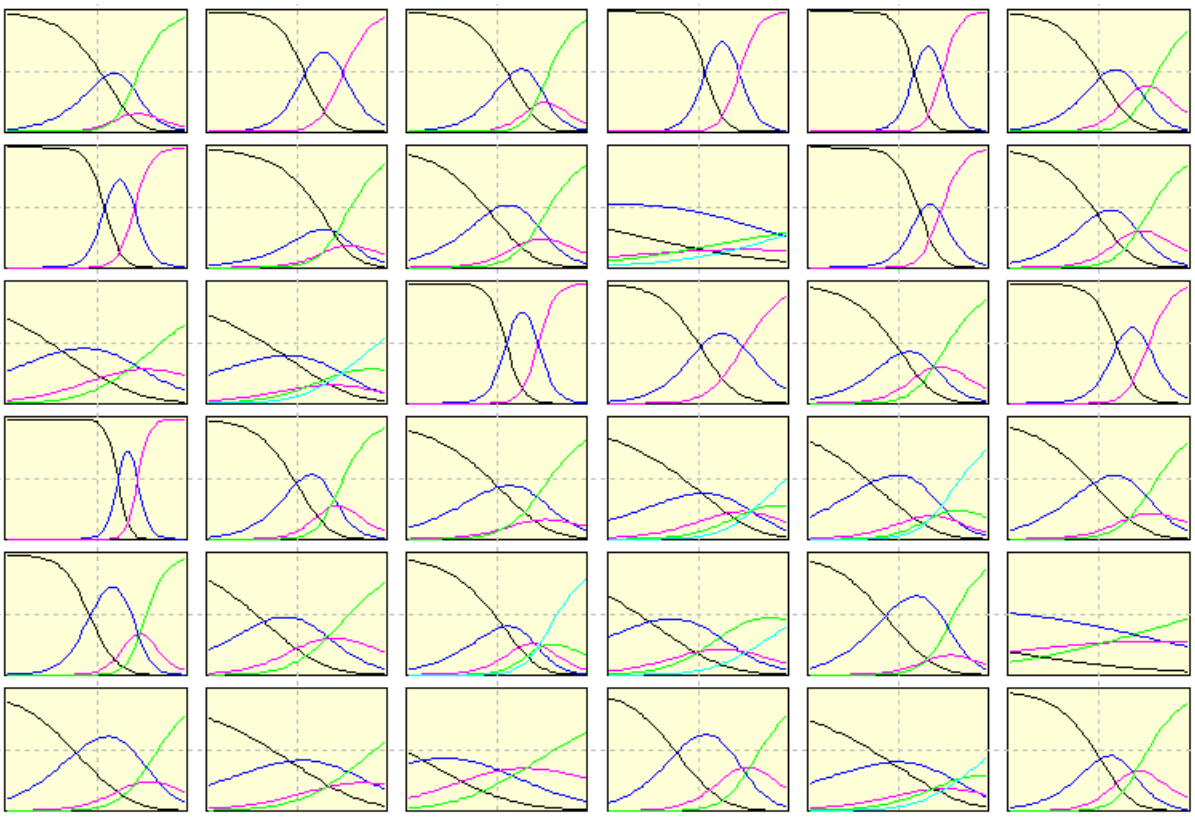
The three figures that follow present matrices of the item response functions for each of the three CNA scales. The graphs in the figures are numbered by row; that is, item 1 is at the top left, item 2 is adjacent to it on the right, etc. At the end of a row, the next item is the leftmost item in the row below. For example, in figure B1, below, the first row starts with item 1 and finishes with item 5, moving left to right; the second row starts with item 6 and finishes with item 10; and so on.

Figure B.1. Item Response Curves for the Knowledge & Skills Items



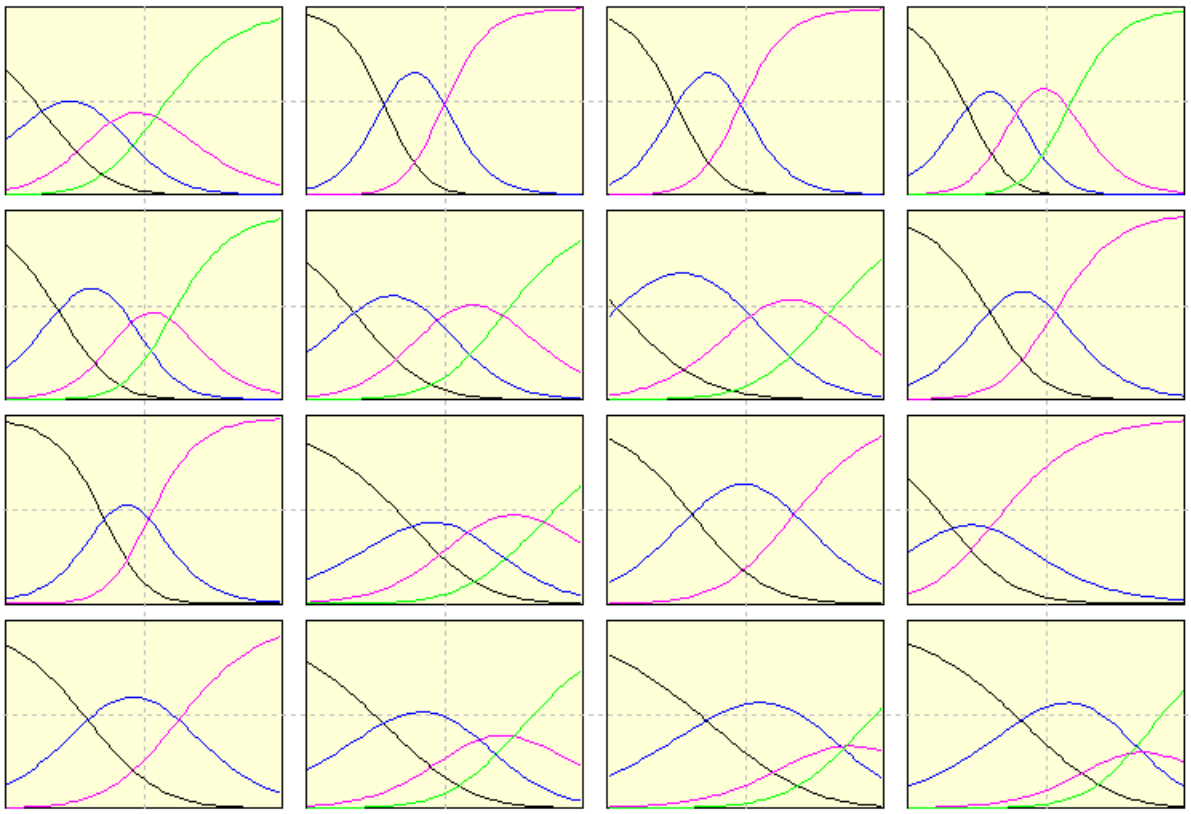
Note: The graphs in figure B.1 show the modeled probability of a correct response (blue line) and the parallel probability of a wrong response (black line). Except for items 16 and 19, which show a flat curve indicative of weak measurement, these graphs show acceptable fit to the IRT model. A flat curve means that the probability of a correct response is very much the same, regardless of the person's overall scale score. This implies that the item has a small correlation with the scale and gives little information about the person's scale score is. In other words, the item is a weak measure of the scale.

Figure B.2. Item Response Curves for the Attitudes & Beliefs Items



Note: Figure B.2 displays the item response functions for the 36 items in the Attitudes & Beliefs scale. These show the probability of being in one of five categories of the items. Instead of right or wrong, these items indicate degrees of propensity for a person to manifest qualities of the underlying scale. For example, a high probability of being in category 1 (highly disagree) indicates that the person has a low propensity for collaborative inquiry. Conversely, a high probability of being in category 5 (highly agree) indicates the person has a high propensity for collaborative inquiry. Note, however, that some items were collapsed because of small frequencies in a category. This was necessary because about 20 responses in each category are necessary for the scaling program to converge estimate probabilities of being in that category. All but two of the items demonstrate item responses where the probability of being in one category or the other is well defined. In items 10 and 30, however, the item responses are relatively flat, indicating that they are weak measures of the underlying construct.

Figure B.3. Item Response Curves for the Data Use Items



Note: Figure B.3 displays the item response functions for the third scale, Data Use. All 16 of these items have well-defined probabilities of being in a category of response as a function of the underlying construct. This indicates that all items are strong measures of the scale.

Appendix C: Power Analysis

We conducted a power analysis at the outset of the study in order to determine the appropriate sample size for this study. Our updated analysis is based on the randomized samples.

C1. Teachers

Table C.1 displays the power analysis for our model of teacher impacts. It is a two-level block-randomization model with teachers nested in schools. Assumptions about intraclass correlations and proportions of variances accounted for by groups of covariates are based on analysis of our randomized sample.

Assumptions:

1. Teachers at Level 1: The intent-to-treat (ITT) sample contains 182 teachers in 59 schools. On average, then, each school has approximately three *Using Data* participants.
2. Intraclass Correlation (ICC): Analysis of the sample suggests that about 10 percent of variation in teacher outcomes will be accounted for by clustering at the school level.
3. Cluster-Level Covariate: We plan to use four school-level covariates in the model—a treatment indicator and indicator variables for the block the school was assigned to pre-randomization. Analysis of the sample suggests that these covariates account for about 15 percent of school-level variation in teacher outcomes.
4. Level of Significance, One- vs. Two-Tailed Tests: We plan to use 0.05 as the level of significance, and all tests are two-tailed.
5. Number of Schools: We randomized 60 schools in total, 30 schools into the treatment condition (*Using Data*) and the other 30 schools into the control condition (business-as-usual). One treatment school dropped out because all of the teachers at randomization dropped out of the study prior to year 1. Schools were assigned to four blocks pre-randomization.

To calculate the minimum detectable effect size (MDES) associated with a power of .8, we used Dong and Maynard's *PowerUp!* tool.[33] Based on these assumptions, the MDES for our teacher impact analyses is estimated to be .42.

Table C.1. Power Analysis for Analysis of Teacher Behavior Impacts in Year 1

Model

Level		Covariates
2	School	Treatment indicator
		3 dummy variables for blocking
1	Classroom/teacher	Assessment pre-score (ATTITUDES, K&S, DATAUSE)
		Knowledge and Skills pre-score

Power Analysis

Assumptions		Comments
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
Rho_2 (ICC_2)	0.10	Proportion of variance among Level 2 units (V2/(V1 + V2))
P	0.48	Proportion of Level 2 units randomized to treatment: $J_T / (J_T + J_C)$
R_1^2	0.20	Proportion of variance in Level 1 outcome explained by Level 1 covariates
R_2^2	0.15	Proportion of variance in Level 2 outcome explained by Block and Level 2
g_2*	4	Number of Level 2 covariates
n (Average Sample Size for Level 1)	3	Mean number of Level 1 units per Level 2 unit (harmonic mean recommended)
J (Average Sample Size for Level 2)	15	Mean number of Level 2 units per Level 3 unit (harmonic mean recommended)
K (Sample Size [# of Level 3 units])	4	Number of Level 3 units
M (Multiplier)	2.86	Computed from T_1 and T_2
T_1 (Precision)	2.01	Determined from alpha level, given two-tailed or one-tailed test
T_2 (Power)	0.85	Determined from given power level
MDES	**0.421**	Minimum Detectable Effect Size

C2. Students

Table C.2 presents a power analysis for our year 2 student sample.

[33] Dong, N., and Maynard, R. A. (2013). *PowerUp!*: A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs. *Journal of Research on Educational Effectiveness*, *6*(1), 24–67.

Assumptions:

1. Students at Level 1: The year 2 student sample contains 10,287 grade 4 and 5 students in 59 schools. On average, then, each school has approximately 174 students.
2. Intraclass Correlation (ICC): Analysis of the sample suggests that less than 10 percent of variation in student outcomes will be accounted for by clustering at the school level.
3. Cluster-Level Covariate: We plan to use four school-level covariates in the model—a treatment indicator and indicator variables for the block the school was assigned to pre-randomization. Analysis of the sample suggests that these covariates account for about 50 percent of school-level variation in student outcomes.
4. Level of Significance, One- vs. Two-Tailed Tests: We plan to use 0.05 as the level of significance, and all tests are two-tailed.
5. Number of Schools: We randomized 59 schools in total, 30 schools into the treatment condition (*Using Data*) and the other 29 schools into the control condition (business-as-usual; one control school closed prior to year 2). Schools were assigned to four blocks pre-randomization.

The baseline estimation model is a two-level block-randomization model with students nested in schools. Based on a sample of 10,287 students, we estimate an MDES of about .17.

Table C.2. Power Analysis for Analysis of Student Performance Impacts (Two-Level Model)

Model

Level		Covariates
2	School	Treatment indicator
		3 dummy variables for blocking
1	Student	Race is African-American - dummy variable
		Race is Hispanic - dummy variable
		FRL status dummy variable
		LEP status dummy variable
		Gifted status dummy variable
		LD status dummy variable

Power Analysis

Assumptions		Comments
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power ($1-\beta$)	0.80	Statistical power (1-probability of a Type II error)
Rho_2 (ICC_2)	0.10	Proportion of variance among Level 2 units (V2/(V1 + V2))
P	0.51	Proportion of Level 2 units randomized to treatment: $J_T / (J_T + J_C)$
R_1^2	0.20	Proportion of variance in Level 1 outcome explained by Level 1 covariates
R_2^2	0.50	Proportion of variance in Level 2 outcome explained by Block and Level 2 covariates
g_2*	4	Number of Level 2 covariates
n (Average Sample Size for Level 1)	174	Mean number of Level 1 units per Level 2 unit (harmonic mean recommended)
J (Average Sample Size for Level 2)	15	Mean number of Level 2 units per Level 3 unit (harmonic mean recommended)
K (Sample Size [# of Level 3 units])	4	Number of Level 3 units
M (Multiplier)	2.86	Computed from T_1 and T_2
T_1 (Precision)	2.01	Determined from alpha level, given two-tailed or one-tailed test
T_2 (Power)	0.85	Determined from given power level
MDES	**0.173**	Minimum Detectable Effect Size

References

Allensworth, E., & Easton, J. Q. (2007). *What matters for staying on-track and graduating in Chicago Public Schools: A close look at course grades, failures, and attendance in the freshman year.* Chicago: Consortium on Chicago School Research.

Berman, P., & McLaughlin, M. W. (1978). *Federal programs supporting educational change. Vol. VIII: Implementing and sustaining innovations.* Santa Monica, CA: RAND Corporation. http://www.rand.org/content/dam/rand/pubs/reports/2006/R1589.8.pdf

Black, P. J., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan,* 86(1), 8–21.

Black, P. J., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice,* 5(1), 7–75.

Black, P. J., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan,* 80(2), 139–148.

Brookhart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education,* 8(2), 153–169.

Brunner, C., Fasca, C., Heinze, J., Honey, M., Light, D., Mandinach, E., & Wexler, D. (2005). Linking data and learning: The Grow Network study. *Journal of Education for Students Placed at Risk,* 10(3), 241–267.

Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Education Evaluation and Policy Analysis,* 33(3), 378–398.

Christman, J. B., Neild, R. C., Bulkley, K., Blanc, S., Liu, R., Mitchell, C., & Travers, E. (2009). *Making the most of interim assessment data: Lessons from Philadelphia.* Philadelphia: Research for Action. Retrieved January 22, 2012, from http://eric.ed.gov/PDFS/ED505863

Confrey, J., & Makar, K. (2002). Developing secondary teachers' statistical inquiry through immersion in high-stakes accountability data. In D. Mewborn, P. Sztajn, & D. White (Eds.), *Proceedings of the Twenty-fourth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Vol. 3* (pp. 1267–1279). Athens, GA: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Confrey, J., & Makar, K. (2005). Critiquing and improving data use from high stakes tests: Understanding variation and distribution in relation to equity using dynamic statistics software. In C. Dede, J. P. Honan, & L. C. Peters (Eds.), *Scaling up success: Lessons learned from technology-based educational improvement* (pp. 198–226). San Francisco: Jossey-Bass.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.

Duncan, A. (2009). *Secretary Arne Duncan addresses the Fourth Annual IES Research Conference*. Speech made at the Fourth Annual IES Research Conference, Washington, DC. Retrieved June 12, 2009, from http://www.ed.gov/news/speeches/2009/06/06-82009.html

Faria, A., et al. (2012, Summer). *Charting Success: Data Use and Student Achievement in Urban Schools*. Washington, DC: Council of the Great City Schools and the American Institutes for Research.

Feldman & J., & Tung, R. (2001). Using data-based inquiry and decision making to improve instruction. *ERS Spectrum*, 19(Summer), 10–19.

Fullan, M. (2000). The three stories of education reform. *Phi Delta Kappan*, 81(8), 581–84.

Halverson, R., & Thomas, C. N. (Eds.). (2007). *The roles and practices of student services staff as data-driven instructional leaders*. New York: Teachers College Press.

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making* (NCEE 2009–4067). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved January 22, 2012, from http://ies.ed.gov/ncee/wwc/pdf/practice_guides/dddm_pg_092909.pdf

Hammerman, J. K., & Rubin, A. (2002) Visualizing a statistical world. *Hands On!,* 25(2).

Hammerman, J. K., & Rubin, A. (2003). *Reasoning in the presence of variability.* Paper presented at the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-3), Lincoln, NE.

Haycock, K. (2001). Closing the achievement gap. *Educational Leadership, 58*(6), 6–11. Retrieved December 2013 from http://www.ascd.org/publications/educational-leadership/mar01/vol58/num06/Closing-the-Achievement-Gap.aspx

Hayward, L., Priestley, M., & Young, M. (2004). Ruffling the calm of the ocean floor: Merging practice, policy and research in assessment in Scotland. *Oxford Review of Education,* 30(3), 397–415.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan,* 89(2), 140–45.

Herman, J., & Gribbons, B. (2001). *Lessons learned in using data to support school inquiry and continuous improvement: Final report to the Stuart Foundation* (CSE Technical Report 535). Los Angeles: UCLA Center for the Study of Evaluation.

Herold, D. M., & Fedor, D. B. (2008). *Change the way you lead change: Leadership strategies that really work.* Stanford, CA: Stanford University Press.

Honey, M., Brunner, C., Light, D., Kim, C., McDermott, M., Heinze, C., Breiter, A., & Mandinach, E. (2002). *Linking data and learning: The Grow Network study.* New York: EDC/Center for Children and Technology.

Jellison, J. (2006). *Managing the dynamics of change.* New York: McGraw-Hill.

Johnson, J. H. (1996). *Data-driven school improvement* (OSSC Bulletin Series, vol. 39, no. 5). Eugene, OR: Oregon School Study Council.

Keigher, A. (2010). *Teacher attrition and mobility: Results from the 2008–09 Teacher Follow-Up Survey* (NCES 2010–353). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Konstantopoulos, S., Miller, S. R., & van der Ploeg, A. (2013). The impact of Indiana's system of interim assessments on mathematics and

reading achievement. *Educational Evaluation and Policy Analysis*, 35(4), 481–499.

Leukens, M. T., Lyter, D. M., Fox, E. E., & Chandler, K. (2004). *Teacher attrition and mobility: Results from the Teacher Follow-up Survey, 2000–01.* Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Light, D., Wexler, D., & Henize, J. (2004). *How practitioners interpret and link data to instruction: Research findings on New York City Schools' implementation of the Grow Network.* Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

Loeb, S., Beteille, T., & Kalogrides, D. (2012). Effective schools: Teacher hiring, assignment, development and retention. *Education Finance and Policy*, 7(3), 269–304.

Love, N. (2002). *Using Data/getting results: A practical guide for school improvement in mathematics and science.* Norwood, MA: Christopher-Gordon.

Love, N. (2004). Taking data to new depths. *Journal of Staff Development*, 25(4).

Love, N. (2009a). Building a high-performing data culture. In N. Love (Ed.), *Using data to improve learning for all: A collaborative inquiry approach* (pp. 2–24). Thousand Oaks, CA: Corwin Press.

Love, N. (2009b). *Using data to improve learning for all: A collaborative inquiry approach.* Thousand Oaks, CA: Corwin Press.

Love, N. B., Stiles, K. E., Mundry, S. E., & DiRanna, K. (2008). *The data coach's guide to improving learning for all students: Unleashing the power of collaborative inquiry.* Thousand Oaks, CA: Corwin Press.

Mandinach, E. B., & Honey, M. (Eds.). (2008). *Data-driven school improvement: Linking data and learning.* New York: Teachers College Press.

Mandinach, E., Honey, M., Light, D., Heinze, J., & Rivas, L. (2005). Technology-based tools that facilitate data-driven decision-making. In C-K. Looi, D. Jonassen, & M. Ikeda (Eds.), *Toward sustainable and scalable educational innovations informed by the learning sciences,* pp. 267–274. Amsterdam: IOS Press.

Mandinach, E. B., Rivas, L., Light, D., & Heinze, J. (2006). *The impact of data-driven decision making tools on educational practice: A systems analysis of six school districts.* Paper presented at the meeting of the American Educational Research Association, San Francisco.

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education: Evidence from recent RAND research.* Santa Monica, CA: RAND Corporation.

Mason, S. (2002). *Turning data into knowledge: Lessons from six Milwaukee public schools.* Madison, WI: Wisconsin Center for Education Research.

National Research Council. (1996). *National science education standards.* Washington, DC: National Committee on Science Education Standards and Assessment.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 2009–0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved June 16, 2013, from http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE20090049

Schilling, S. G., Blunk, M., & Hill, H. C. (2007). Test validation and the MKT measures: Generalizations and conclusions. *Measurement: Interdisciplinary Research and Perspectives,* 5(2–3), 118–127.

Schmoker, M. J. (1999). *Results.* Alexandria, VA: Association for Supervision and Curriculum Development.

Shepard, L. A. (2005). Linking formative assessment to scaffolding. *Educational Leadership,* 63(3), 66–71.

Thessin, R. A. (2007). Newton North High School gets smart about data. In K. P. Boudett & J. L. Steele (Eds.), *Data wise in action: Stories of schools using data to improve teaching and learning* (pp. 29–50). Cambridge, MA: Harvard Education Press.

Wayman, J. C., & Stringfield, S. (2006). Technology-supported involvement of entire faculties in examination of student data for instructional improvement. *American Journal of Education,* 112(4), 549–571.

Zalles, D. (2005). *Designs for assessing foundational data literacy.* Retrieved September 2005 from the On the Cutting Edge: Professional

Development for Geoscience Faculty Web site:
http://serc.carleton.edu/NAGTWorkshops/assess/essays.html

List of Figures

This page intentionally left blank

List of Tables

This page intentionally left blank