The Effect of a Data-Based Instructional Program on Teacher Practices: The Roles of Instructional Leadership, School Culture, and Teacher Characteristics

Author: Beth A. Morton

Boston College
Lynch School of Education

Department of
Educational Research, Measurement, and Evaluation

THE EFFECT OF A DATA-BASED INSTRUCTIONAL PROGRAM ON TEACHER PRACTICES:

THE ROLES OF INSTRUCTIONAL LEADERSHIP, SCHOOL CULTURE, AND TEACHER CHARACTERISTICS

Dissertation
by

BETH A. MORTON

submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

August 2016

ABSTRACT

THE EFFECT OF A DATA-BASED INSTRUCTIONAL PROGRAM ON TEACHER PRACTICES:

THE ROLES OF INSTRUCTIONAL LEADERSHIP, SCHOOL CULTURE, AND TEACHER CHARACTERISTICS

Beth A. Morton

Henry I. Braun, Chair

Data-based instructional programs, including interim assessments, are a common tool for improving teaching and learning. However, few studies have rigorously examined whether they achieve those ends and contributors to their effectiveness. This study conducts a secondary analysis of data from a matched-pair school-randomized evaluation of the Achievement Network (ANet). Year-two teacher surveys (n=616) and interviews from a subset of ANet school leaders and teachers (n=40) are used to examine the impact of ANet on teachers' data-based instructional practices and the mediating roles of instructional leadership, professional and achievement cultures, and teacher attitudes and confidence.

Survey results showed an impact of ANet on the frequency with which teachers' reviewed and used data, but not their instructional planning or differentiation. Consistent with the program model, ANet had a modest impact on school-mean teacher ratings of their leaders' instructional leadership abilities and school culture, but no impact on individual teachers' attitudes toward assessment or confidence with data-based

instructional practices. Therefore, it was not surprising that these school and teacher characteristics only partially accounted for ANet's impact on teachers' data practices.

Interview findings were consistent. Teachers described numerous opportunities to review students' ANet assessment results and examples of how they used these data (e.g., to pinpoint skills on which their students struggled). However, there were fewer examples of strategies such as differentiated instruction. Interview findings also suggested some ways leadership, culture, and teacher characteristics influenced ANet teachers' practices. Leaders' roles seemed as much about holding teachers accountable for implementation as offering instructional support and, while teachers had opportunities to collaborate, a few schools' implementation efforts were likely hampered by poor collegial trust. Teacher confidence and attitudes varied, but improved over the two years; the latter following from a perceived connection between ANet practices and better student performance. However, some teachers were concerned with the assessments being too difficult for their students or poorly aligned with the curriculum, resulting in data that were not always instructionally useful.

ACKNOWLEDGEMENTS

There are many people to whom I owe thanks for supporting me during the dissertation process. First, I would like to thank my committee members who were always willing offer advice on whatever issue had me stuck at the moment. This dissertation is better thanks to the valuable methodological and substantive expertise of Dr. Lauren Saenz and Dr. Vincent Cho. I also owe tremendous thanks to Dr. Henry Braun, my committee Chair. His thoughtful feedback ensured that my dissertation would not only have a high level of technical rigor, but make a real contribution to the field.

I would also like to extend my gratitude to Dr. Marty West and Corinne Herlihy at the Center for Education Policy Research at Harvard University. This dissertation would not have been possible, nor would it have had such insight, without the opportunity to be part of the evaluation team. Marty, I sincerely appreciate all of your advice and for being an unofficial and invaluable advisor. I also owe thanks to Marty, Corinne, Dr. Barb Gilbert, and Hilary Bresnahan for their feedback and guidance on my qualitative analysis.

Most importantly of all, this dissertation would not have been possible without the support of my family and friends. I would especially like to thank my parents who, as teachers, influenced my choice to pursue a career in education research in the first place. To my mother, thank you for being my resident expert and sounding board for many education related questions. Special thanks are owed to my extended family and friends for their frequent words of encouragement that often came via social media when I was at my peak of procrastination. Finally, thank you to my best writing partner, Sophie.

Every bit of support helped. I did it!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EXHIBITS

CHAPTER ONE: INTRODUCTION

Efforts to promote the use of student assessment data to inform teachers' instructional decisions are widespread. These efforts are not unique to the classroom; they are part of a larger movement toward using data to improve educational decision making at all levels of the system. The trend stems from a focus on accountability-driven strategies to improve student achievement in American schools (Dembosky, Pane, Barney, & Christina, 2005; Marsh, Pane, & Hamilton, 2006; Christman, et al., 2009; Bulkley, Oláh, & Blanc, 2010; Faria, et al., 2012; Hargreaves & Braun, 2013); an imperative both in terms of improving education quality and closing achievement gaps. However, the strategies that emphasize teachers' use of student data to achieve these ends have attracted some criticism. Specifically, the implementation of these data-driven strategies is occurring despite limited empirical knowledge of whether and how they contribute to changing teacher practices and improving student outcomes (Carlson, Borman, & Robinson, 2011; Cordray, Pion, Brandt, Molefe, & Toby, 2012; Konstantopoulos, Miller, & van der Ploeg, 2013).

One strategy to improve teaching and learning is the administration of interim assessments – assessments given at regular intervals during instruction to gauge students' progress toward mastering content standards and to inform educators' instructional decisions at the classroom, school, or district level (Perie, Marion, & Gong, 2009). Despite a common purpose, interim assessment programs may differ in the types of products and supports that are provided to educators. Some program providers supply only the assessments, where others offer complementary services such as training or

professional development, coaching in data use and instructional strategies, and supplemental materials such as sample lesson plans or curriculum alignment guides.

The Achievement Network (ANet), a Boston-based organization, provides all of these services as part of their "data-based instructional program." In 2010, ANet was awarded a U. S. Department of Education Investing in Innovations (i3) development grant to subsidize the expansion of its program, as well as to inform program developers and the wider education community on effective data-based instructional practices. To achieve the latter, ANet partnered with the Center for Education Policy Research (CEPR) at Harvard University which conducted an independent evaluation of the program.

This dissertation draws on data from the i3 evaluation of ANet's data-based instructional program to examine how instructional leadership, school culture, and teacher characteristics are related to the use of interim assessment data and instructional practices. Leadership refers specifically to the subset of practices performed by instructional leaders that support the improvement of teaching and learning. The aspects of school culture that are the focus of this dissertation include teacher professional culture and the presence of a culture of achievement. Also of interest are the roles played by individual teacher attitudes toward, and confidence in, using interim assessment data and various instructional planning strategies. As chapter two illustrates, each of these is frequently cited as potential mediators of instructional data use in observational research.

THE PROBLEM

The use of data to improve outcomes is a growing trend in many industries, including business, health care, and government. Advances in technology have improved the collection and analysis of data, as well accountability and transparency within these sectors (Howard, 2012). With the current strong focus on accountability, it is not surprising that this trend has extended to the U.S. education system. Fueled by mandates such as No Child Left Behind and programs like Race to the Top, data are an integral component of monitoring progress in our current education accountability systems (Wayman, 2005; Halverson, Grigg, Prichett, & Thomas, 2007; Mandinach & Honey, 2008; Mandinach, 2012). Data influence decision-making at all levels of education: from system-level decisions about funding and resources, to classroom decisions about instruction and student placement.

In classrooms, teachers have long used the results of class assignments, homework, and informal assessments to judge their students' understanding of topics and make adjustments to their teaching. However, the increased focus on data-driven decisions and accountability for student achievement has formalized the use of external, standardized assessment programs. Many schools are adopting interim assessment programs that include a series of periodic assessments – often in mathematics and English-language arts – that are aligned to content standards and paired with training for school leaders and teachers. The intention is to build educators' capacity to use data to improve instruction.

Despite their prevalence, the evidence of whether and how interim assessment programs and related data-driven practices impact teacher practices and student outcomes is relatively sparse. Few studies have utilized designs that are strong enough to make causal linkages between interim assessment programs and changes in teacher instruction and student achievement. Two quasi-experimental studies found no evidence of a relationship between interim assessment practices and student outcomes (Henderson, Petrosino, Guckenburg, & Hamilton, 2007; 2008; Quint, Sepanik, & Smith, 2008).

Findings from studies of interim assessment programs that employed stronger empirical designs have been mixed. Some have found an effect on certain teacher outcomes, but no overall effect on student outcomes (Cordray, Pion, Brandt, Molefe, & Toby, 2012; Randel et al., 2011; Cavalluzzo, et al., 2014). Others found positive and/or negative effects on achievement in certain grades and subjects (Konstantopoulos, Miller, & van der Ploeg, 2013; Carlson, Borman, & Robinson, 2011; Konstantopoulos, Miller, van der Ploeg, & Li, 2014, 2016). However, these experimental studies failed to collect (or report) data on implementation fidelity or contextual conditions that would improve our understanding of the mediating mechanisms by which interim assessment programs and related data-driven practices impacted, or failed to impact, teacher practice and student achievement.

The presence of an interim assessment program is unlikely to impact student outcomes on its own; conditions that facilitate analyzing, interpreting, and using data are likely necessary. Many observational studies have hypothesized that the factors contributing to effective data use include particular program components such as

professional support (i.e., professional development and coaching) as well as characteristics of schools, leaders, and teachers. While prior research on effective data use supports makes note of "best practices," there is little evidence of a causal connection between various types of support and effective instructional data use practices or higher student achievement. Observational evidence suggests that the frequency of coaching around data use is associated with teachers' self-reported changes in instructional practice, their likelihood of attributing instructional changes to coaching, and small increases in student reading and math achievement (Marsh, McCombs, & Martorell, 2010). Numerous other descriptive studies indicate a relationship between characteristics of school culture, leadership, and teachers and the use of student assessment data for instructional purposes (e.g., Dembosky, et al., 2005; Marsh, Pane, & Hamilton, 2006; Christman, et al., 2009; Goertz, Oláh, & Riggan, 2009a; Faria, et al., 2012; Datnow & Park, 2014).

In summary, the field lacks evidence linking interim assessment programs, teachers' data use and instructional practices, and student achievement. Specifically, it remains unclear what school conditions and teacher characteristics mediate the impact of interim assessment programs on teaching and learning. The adoption of data-based instructional strategies is outpacing evidence that shows whether and how they impact teacher and student outcomes.

THE ACHIEVEMENT NETWORK

The Achievement Network (ANet) is a Boston-based non-profit that provides its data-based instructional program to schools that serve high-need students in grades 3-8. Its mission is to help teachers use interim assessment data to identify and close achievement gaps. ANet was founded in 2005 when it began working with a small group of charter schools in Boston. By the start of the 2015-16 school year, the program was serving over 500 schools in ten geographic networks across the United States (Achievement Network website, 2015).

The i3 Evaluation

Combining interim assessments aligned to content standards with data tools and analysis protocols, coaching, and networking opportunities, ANet's data-based instructional model is an example of a comprehensive, data-driven instructional initiative in use in schools nationwide. In 2010, CEPR began a five-year, i3-funded evaluation which utilized a matched-pair school-randomized design to examine the effect of ANet on student achievement in mathematics and English language arts in grades three through eight. The evaluators utilized a mix of survey-based quantitative data and school site visit-based qualitative data to examine the intermediate effects on leader and teacher practices and school outcomes. Surveys were administered to all eligible school leaders

and teachers in both treatment and control schools; whereas, qualitative site visits were conducted in a subset of treatment schools in each of four geographic school networks.[1]

Schools were recruited from five mid- to large-size urban districts: Boston (MA), Chelsea (MA), Springfield (MA), Jefferson Parish (LA), and Chicago (IL). Individual elementary and middle schools in these districts were invited to apply to receive ANet services at a rate that was subsidized by the i3 grant. All schools were screened on criteria thought to facilitate successful implementation of the program: e.g., leadership capacity and support, and school priorities for a standards-based curriculum and time for teacher collaboration. Although the racial and ethnic composition of the students enrolled in these schools varied across districts, all of the schools served high proportions of students eligible for a subsidized lunch and who were not performing at proficient levels on state math and reading assessments.

The Intervention

As part of the evaluation design process, the i3 program officers asked each grantee to develop a logic model for its program. The concept of a logic model originates from program theory evaluation; a type of evaluation that is focused on testing "an explicit theory or model of how the program causes the intended or observed outcomes" (Rogers, Petrosino, Huebner, & Hacsi, 2000, p. 5). In this way, logic models are useful

[1] The schools in the sample represent five school districts, but are part of four geographic "networks": Boston and Chelsea are part of the Eastern Massachusetts network, Springfield participates in the Western Massachusetts network, Jefferson Parish is a part of the Louisiana network, and Chicago schools are part of the Illinois network. Networks consist of i3 and non-i3 schools; not all of the ANet-partnered schools in each network participated in the study.

tools for evaluation because they make manifest program theory (Bickman, 2000). A logic model represents a hypothesis for program processes; i.e., how the program inputs are expected to influence the intended outcomes (McLaughlin & Jordan, 1999).

The ANet data-based instructional program is a whole-school – and increasingly whole-district – reform model designed to embed data-driven decision-making in school leaders' and teachers' everyday practice. Their logic model specifies the hypothesized pathways by which program inputs support more effective use of interim assessment data and increase student achievement (Achievement Network, 2012). It is important to note that the logic model for ANet is not fixed, but is based on a program theory of action that is continually evolving from the lessons learned by the organization. The ANet program model presented here reflects the organization's thinking at the start of year two of implementation; the year in which analyses in this study are focused.

The ANet logic model includes: the *inputs* or resources provided as part of the program, the *activities* that are necessary to achieve the intended outcomes, and the expected *outcomes* of the program (categories adapted from McLaughlin & Jordan, 1999). Specifically, the logic model first documents the *inputs* of this program: 1) quarterly interim assessments, 2) logistical support, 3) professional development, and 4) school networking opportunities (exhibit 1.1). It also summarizes the *activities* (actions and structures) by which these four inputs are hypothesized to lead to *outcomes* of improved student achievement.

Prior to the school year, ANet works with districts and schools to develop interim assessments that are aligned with their curriculum and curricular scope and sequence.

ANet also sets a schedule for assessment administration and regular coach visits at key points in data cycle such as planning and data review. Coaches work to build school leaders' capacity to support teachers' data-based instructional practices. After the interim assessments are administered and student results are returned, schools hold data meetings. In these meetings, teachers analyze their students' data and plan instruction to meet identified learning gaps. The ANet coach is present to support this work, but the intention is that it is led by school leaders and data leadership team members. Initially, meetings are scheduled for a three-hour block – typically at the end of the school day. Over time, however, data meetings may change format; for example, they may take place during grade-level common planning time. Elements of the program logic model are detailed below as they are intended to be implemented. Appendix D provides a glossary of key ANet terms.

Intervention Inputs. The core inputs are quarterly standards-aligned assessments in English Language Arts and mathematics in grades 3 through 8 that are administered at regular intervals during the school year (Aligned Assessments).[2] Within 48 hours of sending completed assessments to ANet, school leaders and teachers receive students' results via an online platform called MyANet. Reports provide aggregate results at the network, school, grade, and classroom level, and teachers receive student-level results that detail strengths and misunderstandings at the item and standard level. The MyANet online platform also provides teachers with resources on planning and instructional tools

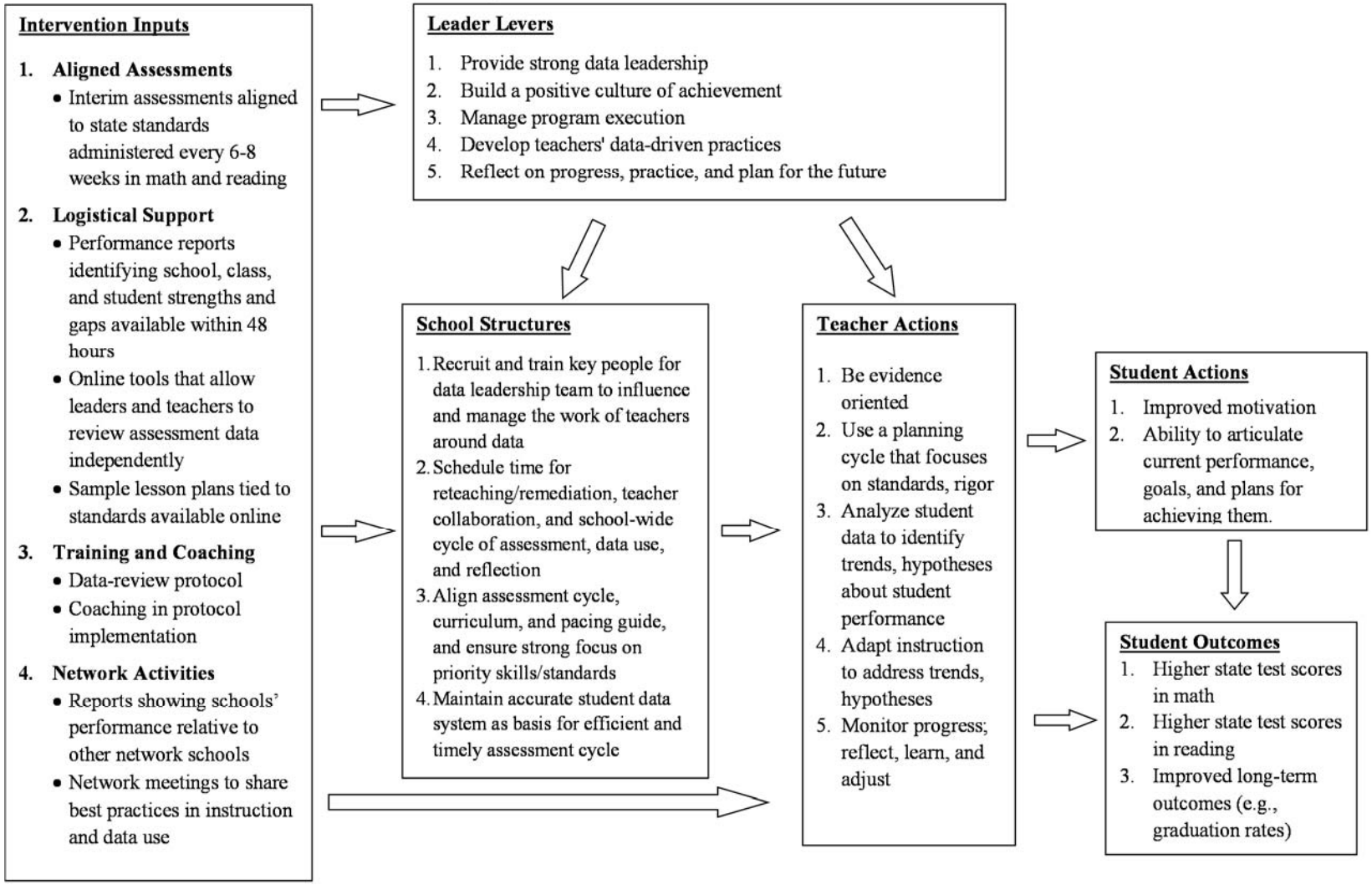
[2] Over the course of the evaluation, the program's interim assessments shifted from aligning with state content standards to Common Core State Standards, as appropriate within each state. However, the evaluation did not validate this through an alignment study.

such as schedules of standards to be covered on future interim assessments, misconception guides, templates for planning their reteaching, and a quiz building tool (Logistical Support).

Considered to be the program's chief "value-add" over other programs, ANet data coaches visit schools an average of 19 times per school year at key points in the data cycle (e.g., pre-assessment planning, when results are released, and after reteaching has occurred). The total number of visits may be higher or lower depending on individual school needs. Most school visits involve coaches and school leaders, and focus on supporting leaders' data-based work with teachers. During group data meetings, coaches meet with teachers, supported by a school-based data leadership team, to coach them on how to analyze the data for patterns in responses, use the data to draw conclusions about their students' performance on particular standards, skills, or subskills, and develop reteaching plans to address identified learning gaps (Training and Coaching).

Another of the more unique aspects of ANet is the geographically connected networks of participating schools. ANet reports show schools' performance on the interim assessments relative to other schools in the network. The school data teams are invited to participate in two annual meetings of schools in their network. These network events are an opportunity for professional development, but they also provide the opportunity to learn and share successful practices (Network Activities).

Exhibit 1.1. The Achievement Network Logic Model



The Achievement Network program logic model for the i3 evaluation (2012).

Leader Actions. With these inputs in place, the ANet model is focused on building leadership capacity in five key areas. First, school leaders are supported by ANet coaches in making data use a priority and engaging directly in data work in their schools. Leaders are expected to build a positive school-wide culture of achievement, manage the implementation of the data cycle, develop teachers' data-driven skills and practices, and reflect on student and school progress, taking action where needed. ANet's approach to coaching aims to build leader capacity; the intent is that desired actions are modeled by the ANet coach, eventually becoming the practice of the school leader and data leadership team.

School Structures. ANet also works with school leaders to put structures in place that facilitate program implementation. School leaders establish a data leadership team, typically including themselves, an assistant principal, and grade-level or content-area instructional leaders. With help from their ANet coach, the data leadership team is expected to lead and support teachers' data-based practices. School schedules are arranged to include time for activities such as teacher collaboration around data, and planning for and carrying out reteaching. Leaders also ensure that the assessment cycle is aligned to curricular resources and planning. Finally, they must maintain an accurate student data system that supports the adoption of the assessment cycle.

Teacher Actions. Leaders and ANet coaches support teachers in using data to gauge student progress and adjust teaching. Teachers are expected to use backward planning to align their instruction to state content and performance standards. The goal is to develop a planning cycle that focuses on instructional alignment and rigor. Teachers

are also expected to analyze assessment results to determine students' progress toward standards and to use this information to develop and implement reteaching plans that address students' gaps in knowledge. Once they have implemented these reteaching plans, teachers are encouraged to reassess and reflect on their impact on student learning.

Student Actions and Outcomes. Through the sharing of interim assessment results, it is expected that students will exhibit greater motivation to learn, as well as the ability to articulate their own performance goals and plans to achieve them. The primary outcome of interest for the i3 evaluation was higher state summative assessment scores in math and reading. However, it is expected that any short-term impacts on test scores will translate into improvements in longer-term student outcomes, such as high school graduation and post-secondary success.

PURPOSE OF THE STUDY

This study draws on an existing dataset from the larger i3 evaluation of the Achievement Network's data-based instructional program to examine whether and how the program impacts teacher practices. The purpose of this study is to explore the process of instructional data use, conditions that may mediate this process, and the relationship between these potential mediators and teacher practices. The process of instructional data use by teachers is said to be under-conceptualized, taking place within a black box (Black & Wiliam, 1998; Little, 2012; Spillane, 2012). Opening the black box is key to understanding the impact of data-based instructional programs on teachers and students, and improving the effectiveness of these programs.

Research Questions

The conceptual framework for this study builds on the program processes described by the logic model. The study's purpose to explore this conceptual framework; specifically, the potentially mediating roles of school culture, instructional leadership, and teacher characteristics in the relationship between the implementation of a data-based instructional program and teachers' data use and instructional practices. The research questions are:

1. Are teachers' data use and instructional **practices** different in ANet (treatment) schools from those in control schools?
2. Are levels of school culture, instructional leadership, and teachers' attitudes towards and confidence with data-based practices (**hypothesized mediators**) different in ANet (treatment) schools from those in control schools?
3. Do the hypothesized mediators account for differences in ANet and control-school teachers' data use and instructional practices?
4. Does the effect of ANet on teachers' data use and instructional **practices** vary by schools' baseline implementation "readiness" ratings?

The first two research questions exploit the randomized design to examine the effect of ANet on school culture, instructional leadership, teacher characteristics, and teacher practices. These questions focus on whether ANet has an effect on the proposed mediators and teacher practice outcomes highlighted after adjusting for various observed school and teacher characteristics. It is important to note that the counterfactual is not the

absence of data-based instructional practices. Since all control schools were known to have interim assessments in some grades and subjects, as well as varying types of support, any differences that are found would represent the unique effects of ANet over other data-based practices.

The third and fourth research questions move beyond the more evaluative question of whether the program has the intended effect on teachers' data-based instructional practices. They offer insight into the "black box" by providing evidence of whether and how school leadership and culture, and teacher characteristics facilitate (or possibly inhibit) the relationship between a data-based instructional program and teacher practices. Specifically, research question three examines whether certain school-level conditions – considered important both by ANet and the larger field – and teacher characteristics predict or explain teachers' data-based instructional practices.

Research question four examines whether baseline school characteristics moderate the effect of ANet on teachers' data use and instructional practices. During recruitment, schools' "readiness" to implement ANet was assessed on nine categories (see chapter 3, "School Recruitment") using a survey collected from all schools. The subscores across the most relevant subset of these categories were used to classify schools into "higher" and "lower" categories. Models taking into account the school's readiness score at baseline provide evidence of whether the ANet program is more or less effective in schools with varying levels of readiness.

Conceptualizing the Measures

The focal measures in this dissertation are not new to the research on interim assessment. However, their definition and operationalization vary widely across studies. Defined briefly in this section, each of these measures is fully explored in chapter two, including a discussion of relevant prior research.

Instructional leadership is primarily concerned with the role of the principal or other school leaders in defining and managing the school's mission and goals; managing instruction through supervision, coordinating the curriculum, and monitoring student progress; promoting a positive learning climate by protecting instructional time and professional development; maintaining visibility; enforcing academic standards; and providing incentives to students and teachers (Hallinger & Murphy, 1985). In this study, instructional leaders are hypothesized to play a key role in teachers' adoption of data-based instructional practices.

School culture is an important factor in the adoption of new programs. It has been characterized by "a set of beliefs, values, and assumptions that participants share." (Page, 1987, p. 82) This dissertation focuses on two aspects of school culture: teacher professional culture and the presence of a culture of achievement. First, the beliefs, values, and habits of communities of teachers can be said to constitute their ***professional culture*** or a school's culture of teaching (Hargreaves, 1994). A school's culture of teaching facilitates the transmission of norms through shared discussions of teaching practices, occasions to observe one another's work, and collaborations around planning, selecting, or designing teaching materials (Hargreaves & Dawe, 1990). Distinct from

teacher professional culture, a school's ***achievement culture*** relates to its focus on clear goals, high academic standards and expectations for student performance, as well as frequent monitoring of teacher efficacy and student progress toward meeting these goals (Purkey & Smith, 1983; Zigarelli, 1996).

This study also examines the roles played by teachers' ***attitudes or beliefs***, and ***confidence*** around instructional data use and the adoption of data-based instructional practices. Prior research has examined educators' perceptions of data as a valid, reliable and useful tool for improving instructional practice. Data quality issues and perceived barriers, such as the time commitment required for analyzing and using assessment results to inform instruction, have also been explored in the research (Luo, 2008; Wayman, Cho, Jimerson, & Spikes, 2012). Teachers' skills and facility with assessment and assessment data are often labeled pedagogical data literacy (Mandinach, Gummer, & Muller, 2011; Mandinach & Gummer, 2013). Pedagogical data literacy is a skill set that combines teachers' knowledge and use of assessment data with their content area expertise to inform and improve their teaching practice. Although this study lacks a direct measure of teachers' data literacy, respondents' ***confidence*** in data use and instructional planning are hypothesized to be related to data literacy.

Data-based instructional practice, or instructional data use, is regarded as the process of reflecting on student data as a way to improve teaching and learning through specific goals and actions (Halverson, Grigg, Prichett, & Thomas, 2007). Importantly, this definition highlights two separate practices. First, it encompasses analytic activities such as reviewing and analyzing interim assessment results to identify gaps in students'

knowledge (e.g., ***data review*** and ***data use***). From that follows an instructional response: identifying and implementing appropriate instructional interventions to address students' learning gaps (e.g., ***instructional planning*** and ***instructional differentiation***). Each of these practices were shown to be positively related to students' math and reading achievement in the larger i3 evaluation (West, Morton, & Herlihy, 2016).

OVERVIEW OF THE METHODS

This study utilizes quantitative and qualitative data from the larger i3 evaluation for a secondary mixed methods analysis of the ANet program on teachers' data use and instructional practices. Given the nature of the research questions, this study relies predominantly on year-two survey data. The year-two data are used because of the expectation that an intensive program, such as ANet, would take at least two years to be fully implemented. Due to high levels of school leader survey nonresponse (discussed in chapter three), the primary data source for the quantitative analyses in this study is the year-two teacher surveys (n = 616). The quantitative results for each of the research questions are supported by an analysis of qualitative site visit interview data. This mixed methods approach to data analysis provides a depth of understanding that could not be achieved by quantitative analysis alone (Sammons, 2010).

Scale Validation

This study was conducted in parallel with the larger evaluation on which I was the lead analyst (under the direction of the Principal Investigator). Because of my interest in

exploring the relationships between measures of school culture, instructional leadership, and teacher characteristics, and teachers' data-based instructional practices, I had the opportunity to use data from baseline and year-one surveys to improve upon the measurement of these focal constructs in the year-two survey. Although the dissertation does not include a complete discussion of the survey revision work, chapter four reviews the characteristics of the revised year-two scales that measure instructional leadership, school culture, teachers' attitudes towards and confidence with various data-based practices, and the frequency of teachers' data use and instructional practices. Details are provided on the items within each scale, the overall scale reliabilities, and their validation.

Quantitative Analysis

Given the nested design of the study, the analysis of survey data to estimate the effects of ANet uses multilevel regression modeling (MLM). Failure to model the nesting of teachers within schools can lead to violations of the assumptions of homoscedasticity and independence appropriate to the use of ordinary least squares (OLS) regression, increasing the likelihood of type I errors. Multilevel modeling addresses the issue of correlated errors by modeling the relationship at the various levels of the data (e.g., school and teacher) instead of constraining the model to a single level (as in OLS). The estimation procedures used in multilevel modeling generate standard errors that are not inflated due to nesting (Bickel, 2007).

Qualitative Analysis

After completing the quantitative analyses, all leader and teacher interview transcripts were fully coded and analyzed. A first round of coding identified portions of leader and teacher interviews that address the focal measures and research questions in this study (Leech & Onwuegbuzie, 2008; Saldaña, 2009; Hesse-Biber, 2010). The second round of coding entailed both finer-grained coding and analysis. Coding was informed by a conceptual framework developed from the ANet logic model and prior research. Operationally, the second round of coding utilized a constant comparative approach to extract themes and provide explanations for the quantitative results (Leech & Onwuegbuzie, 2008).

The mixing of analytic strategies is meant to take advantage of the strengths of both methods (Teddlie & Tashakkori, 2003; Johnson & Onwuegbuzie, 2004) and maximize the likelihood of collecting evidence of the relationship between school culture, teacher characteristics, and teachers' instructional practices. The "mixing of methods" takes place at the interpretation stage. Given the causal nature of research questions, the results from the quantitative analyses take precedence. The qualitative results serve to explain the quantitative findings and provide explanatory context. In particular, they: 1) provide context for the impacts, or the lack thereof, on the teacher practices and key mediators in this study, 2) explore the validity of the conceptual framework and causal linkages (Yin, 2009), and 3) offer evidence of why ANet may be more effective at changing teachers practices in some contexts than others.

SIGNIFICANCE OF THE STUDY

A very considerable amount of time and resources are spent each year on data-based instructional strategies, including interim assessments (Lazarin, 2014; Hart, et al., 2015). In fact, evidence suggests that district-mandated tests – such as interim assessments – make up a larger proportion of testing time than state tests, especially in urban districts (Lazarin, 2014). This is despite the fact that empirically sound research on the impacts of these practices is sparse and results are varied. Given the recent call by the Obama administration to reduce time spent on testing in American schools and a provision to allow states to set limits on time spent on testing as part of the *Every Student Succeeds Act* (ESSA), evidence of the quality of interim assessments and their utility in improving teaching and learning may become more important than ever (U.S. Dept. of Education, Fact Sheet, 2015; ESSA, 2015).

This study addresses two main problems in the current research on interim assessment and data-driven instruction. First, it fills an empirical need for research on interim assessment programs and data-driven instructional practices that combines empirically sound research designs with rich process and outcome data. This design allows for the study to explore the data-based instructional process and address major gaps in our current understanding of whether and how data-based initiatives have an impact on teachers' practices. In particular, the study explores the oft-cited, but not well-understood roles played by certain school conditions and teacher characteristics. The combination of quantitative and qualitative evidence, collected as part of a randomized

evaluation, provides a unique opportunity to address the empirical and substantive gaps in prior research on teachers' data-based instructional practices.

To make findings more useable, some have suggested that researchers align their work with the current challenges that administrators are facing (Honig & Coburn, 2007). In terms of its practical importance, the hope is that the results of this study will provide district- and school-level practitioners and policymakers looking to implement data-based instructional strategies with useable insights on where and how to focus their energies in order to foster change without unintended, negative consequences for teachers and students. In the longer term, the results have the potential to inform the development of interim assessment programs; specifically, implementation and training targeted to the conditions in schools and characteristics of educators that support the adoption of effective data-based instructional practices.

CHAPTER TWO: LITERATURE REVIEW

The use of student data has become widespread despite a limited body of evidence linking specific programs, conditions, or practices to improvements in teaching and learning. This dissertation explores the impact of the Achievement Network's (ANet) data-based instructional program on teachers' data use and instructional practices. On a broader level, it also explores the process of instructional data use and how the adoption of data-based instructional practices is related to instructional leadership, school culture, and teacher characteristics.

This chapter provides a review of relevant prior research and a summary of the contribution of this study to the field. It begins by defining the concept of instructional data use, the key outcome of interest in this study. Next, the context for instructional data use is set within a discussion of current accountability systems in education. Components of data-based instructional programs are briefly discussed, though the formal review of prior research begins with a reflection on recent quasi-experimental and experimental studies of the data-based instructional programs most like the ANet. Although the outcomes of these studies are most often student achievement, the variation in results is an argument for exploring intermediate impacts such as teachers' instructional data use and the conditions that may affect these practices. Consequently, research on the potential mediators of instructional data use that are central to this study, namely instructional leadership, school culture, and teacher attitudes toward and confidence using data, are explored in the final section.

DEFINING INSTRUCTIONAL DATA USE

Some suggest that programs such as ANet can contribute to student learning by supporting a system of organizational learning or continuous improvement (Bulkley, Oláh, & Blanc, 2010; Halverson, 2010). Models of continuous improvement of instruction typically include the steps of: 1) planning or goal setting based on standards, 2) providing instruction, 3) assessment of learning, 4) analysis and use of assessment results, 5) planning and reteaching, and 6) reassessment (Deming, 1993; Datnow, Park, & Wohlstetter, 2007; Flumerfelt & Green, 2013).[1] Embedded in this cycle is *instructional data use* which comprises two distinct, but related practices which are the focus of this dissertation: (1) data analysis and use and (2) instructional planning and remediation. It involves using student assessment results to identify areas of student need and improve teaching and learning by implementing appropriate instructional actions or responses (Faria, et al., 2012). In the larger i3 evaluation, the frequency with which teachers reviewed and used student data, used various instructional planning strategies, and used instructional differentiation were each positively related to students' math and reading achievement scores (West, Morton, & Herlihy, 2016).

Reflecting on student data consists of a variety of tasks such as reviewing and making sense of data, alone or collaboratively, with the purpose of informing instructional actions (Faria, et al., 2012). Data review and reflection requires skills such as knowing and being able to navigate data reports, accessing and synthesizing available

[1] Many private organizations provide models for continuous improvement in education and other sectors. For example: Plan, Do, Study, Act (PDSA); Six Sigma (DMAIC); Lean; Results-Oriented-Cycle of Inquiry (ROCI); and Data Wise.

forms of student data, understanding which data are appropriate to answer particular instructional questions or guide instructional decisions, reviewing and analyzing student data to identify the instructional needs of specific students or groups of students (e.g., gaps in learning, misconceptions), and communicating accurate results and inferences to team teachers, school leaders, or other stakeholders (Faria, et al., 2012; Mandinach, 2012).

Instructional actions refer to the way teachers respond to the "knowledge and information generated by their review of student data" (Faria, et al., 2012, p. 14) including judgments about (1) the use of instructional time; (2) allocating additional instruction for individuals or groups students who are struggling with particular topics; (3) addressing students' weaknesses with instructional interventions; (4) gauging overall instructional effectiveness of classroom lessons, and (5) refining instructional methods by selecting instructional approaches that address the situation identified through the data (Hamilton et al., 2009). Examples include establishing or adjusting student groupings, changing the curricular scope, sequence, or pacing, altering instructional strategies or materials, adjusting or reteaching particular lessons to address students' skills gaps, and providing supplemental resources to targeted students (Heritage, Kim, Vendlinski, & Herman, 2009; Coburn & Turner, 2011; Faria, et al., 2012).

EDUCATIONAL DATA USE IN CONTEXT

While the focus of this dissertation is on classroom-level use of interim assessment data for improving teaching and learning, there is a broader movement toward

data-driven decision making (DDDM) in public education. The phrase is used to describe decision-making processes at all levels of the education system that are informed by various sources of data. DDDM involves the systematic and ongoing collection, analysis, interpretation, and use of educational data for various ends such as improving instruction, better allocating resources (i.e., material and human capital), and informing policies (Mandinach, 2012). Since DDDM can be used at every level and in every role, it incorporates a variety of educational data from student assessments and demographics, to administrative, financial, personnel, and multiple other data sources (Mandinach, 2012).

Despite its growing prevalence, data use in education is not a new practice and has its roots in the growth of measurement and accountability for student achievement (Dembosky, et al., 2005; Marsh, Pane, & Hamilton, 2006; Christman, et al., 2009; Bulkley, Oláh, & Blanc, 2010; Faria, et al., 2012). However, recent trends in accountability policies have provided the impetus for a more formal process of data use, including a more systematic use of external, standardized assessments as a key source of data on student learning. Data use and accountability have become "inextricably" linked (Mandinach & Honey, 2008, p. 2).

There has been criticism over the phrase "data-driven" and some of the practices falling under this umbrella, criticism that highlights the range of these practice. Specifically, critics of the term contend that to be data-driven both oversimplifies the process and implies one in which data drive the focus of education reform at the macro level and the focus of instruction at the micro level (Shirley & Hargreaves, 2006). Instead, some experts in the field propose that the process should be "evidence

informed": the collection of *evidence* that *informs* educational decisions (Shirley & Hargreaves, 2006; Hargreaves & Braun, 2013). Furthermore, others contend that although student data are a useful tool, the process should be combined with and guided by values and professional judgment (Hertiage & Yeagley 2005; Shirley & Hargreaves, 2006; Knapp, Copland, & Swinnerton, 2007; Wayman, Snodgrass Rangel, Jimerson, & Cho, 2010; Wayman, Jimerson, & Cho, 2012; Hargreaves & Braun, 2013; Hargreaves, Morton, Braun, & Gurn, 2014). The fact that this argument is part of the conversation on interim assessment programs illustrates the range of philosophies on which these programs are based: from programmed and prescribed, to adaptable and open to professional judgment.

It is not the purpose of the dissertation or literature review to evaluate where ANet or other data-based instructional programs fall on this range of data-driven or evidence-informed. Throughout this chapter, the terms data-driven or data-based are used to encompass the range of practices and programs related to instructional data use; from the provision of periodic interim or benchmark assessments, to more comprehensive systems that include tools (e.g., protocols and data systems), professional development and support, and new technology (e.g., data dashboards). Whenever possible, characteristics of the programs examined in prior research are described.

Origins of & Influences on Data Use in Education

Utilizing data has become a key practice in almost any industry that values productivity and continuous improvement: public sectors like health care and

government, or private sector business and finance. In setting the context for instructional data use in education, one could argue its evolution has not only been influenced by the advent of high-stakes educational accountability systems, but also by the increasing role of the business sector in educational management (Marsh, Pane, & Hamilton, 2006; Young & Kim, 2010). The private sector has long promoted management systems that monitor productivity, improve performance, and evaluate systems at all levels (Stecher, Kirby, Barney, Pearson, & Chow, 2004). Data have become an important component of these systems; see, for example, the booming industry around "big data." Successful businesses are said to empower their employees and one way this can be achieved is by providing real-time, relevant data that allow them to take ownership over decision making (Stecher, et al., 2008; Hargreaves, Morton, Braun, & Gurn, 2014). Given evidence of the success of these practices in other industries (Manyika, et al., 2011), policymakers and reformers have advocated for education to adopt similar processes (Tyack, 1995; Stecher, et al., 2008).

Though the influence of the business sector has had an impact, the proliferation of data use in education has had as much to do with test-based accountability policies that are meant to increase student achievement relative to specific content standards. These accountability systems rely heavily on student assessments which have provided a constant stream of achievement data. From the 1970s through the 1990s assessments were used to monitor whether Title I funds were improving the educational outcomes of disadvantaged students, maintain minimum competency for graduation or grade

promotion, and ensure schools were achieving high levels of performance to improve global competitiveness (Linn, 2000).

The modern era of test-based accountability was ushered in when the penultimate reauthorization of the Elementary and Secondary Education Act, the *No Child Left Behind Act* (2002) (NCLB), was passed in 2001. NCLB maintained a focus on standards and test-based accountability by setting annual achievement goals aiming, ultimately, at proficiency for all students. Whether schools met these goals was determined by annual testing in English language arts (ELA), math, and science. Schools and districts that failed to meet annual targets (i.e., Annual Yearly Progress) were initially subject to sanctions ranging from providing supplemental services to students, to school restructuring. NCLB was scheduled for reauthorization in 2007. While the U.S. House and Senate debated proposals for reauthorization, states were granted waivers by the federal Department of Education from some portions of the bill's original requirements in an effort to avoid further sanctions (U.S. Dept. of Education, ESEA flexibility website, 2014).[2]

In its signature educational reform effort, the Obama administration earmarked grant funding through Race to the Top (RttT) to encourage education reform through improvements in four key interrelated areas: standards and assessment; data systems, collection, and use; teacher effectiveness; and turning around low-performing schools (U.S. Dept. of Education, RttT Executive Summary, 2009). RttT also included a $350

[2] President Obama signed the newest iteration of the bill, called the *Every Student Succeeds Act*, on December 10, 2015. The new law upholds the testing requirements of NCLB, but allows states more flexibility to set annual accountability targets which are reviewed by the U.S. Department of Education.

million assessment program competition that funded two multi-state consortia to design assessment systems that include a combination of features such as diagnostic, interim, and summative assessments that are, in some cases, administered in a computer-adaptive format (SMARTER Balanced Consortium website, 2014; PARCC Consortium website, 2014). According to the Department of Education, the intention was to

> develop assessments that are valid, support and inform instruction, provide accurate information about what students know and can do, and measure student achievement against standards designed to ensure that all students gain the knowledge and skills needed to succeed in college and the workplace. These assessments are intended to play a critical role in educational systems; provide administrators, educators, parents, and students with the data and information needed to continuously improve teaching and learning…. (U.S. Dept. of Education, RttT Assessment Program website, 2014)

The effectiveness of current accountability systems rests on a theory of action that posits that student achievement will be positively impacted by a system that holds teachers and school leaders accountable to raising student achievement, as measured by student assessments, and through a series of sanctions and incentives (Hamilton, Stecher, & Klein, 2002). The problem is that an accountability system based on summative assessments that measure achievement against a proficiency benchmark cannot provide school leaders and teachers with timely data at the level of detail necessary to draw inferences about student learning, make "actionable" decisions, and adjust instruction as necessary (Mandinach & Jackson, 2012, p. 16). In fact, some argue that such a system actually has limited educational value (Bennett & Gitomer, 2008) and that improvements in teaching and learning will only be realized by "aligning curriculum, instruction, and professional development and by supplementing mere access to data with opportunities

for educators to analyze data with colleagues in the light of curricular objectives." (Bulkley, Oláh, & Blanc, 2010, p. 115).

In its 2001 report "Knowing What Students Know," the National Research Council (NRC) made several recommendations on student assessment. To be instructionally useful to classroom teachers, they contend that assessment systems should include classroom assessments that are integrated with instruction and make students' cognitive processes evident; e.g., teachers should be able to infer from students' assessment results both the strategies students used, as well as their thought processes. Tasks should be developed with consideration to the skills students need to answer an item correctly, the context in which it is presented, and whether it requires transfer of knowledge from other contexts. To increase the likelihood of student learning, results should be timely and teachers should be adequately trained in both theory and practice to use this information (NRC, 2001).

The two consortia's assessment systems were designed to address the limitations of the current accountability system by shifting from assessment *of* learning to a system that attempt to include assessment *for* learning (Pellegrino, 2006; Bennett & Gitomer, 2008; Mandinach & Jackson, 2012) which allows teachers to use the results in some of the ways recommended by the NRC. In a system of assessment for learning, assessments do not merely check on learning summatively, they provide on-going evidence of what students have and have not mastered (see also, Stiggins, 2005). While districts and schools awaited the roll-out of these new assessment systems, many turned to interim assessment programs to improve teaching and learning (Herman & Baker, 2005).

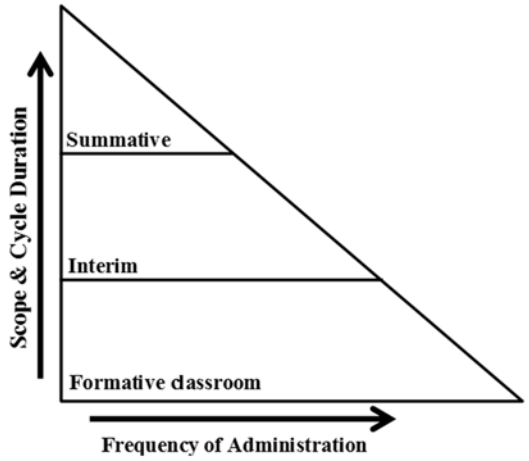
COMPONENTS OF DATA-BASED INSTRUCTIONAL PROGRAMS

This section provides a typology for three types of student assessments. Their purposes, characteristics, and summaries of the research on their utility for informing teaching and learning are discussed. The research reviewed in this chapter focuses mainly on efforts to implement interim assessments and other data-based programs. However, to promote the use of student assessment data for instructional decision making, several programs also include related supports such as professional development, coaching, and guides and resources. The research on these supports is reviewed in brief.

Interim Assessments

Perie, Marion, and Gong (2009) provide a useful framework for distinguishing among assessment types, and for defining and evaluating interim assessment programs, specifically. They organize assessments into three main categories: summative, interim, and formative. Two criteria are used to distinguish among assessment types: (1) the scope (e.g., coverage, purpose), and (2) the frequency of administration (figure 2.1).

Figure 2.1. Tiers of Assessment



Perie, Marion, & Gong, 2009.

Summative assessments are the broadest in curricular scope, but the least frequent in administration. As part of an accountability system, the results are used to inform policy, and to determine rewards and sanctions at the classroom (i.e., teacher), school, and district levels (Perie, Marion, & Gong, 2009). However, summative assessments have limited instructional use. Results are received too late, too infrequently, and are typically not granular enough to provide teachers with the type of data needed to inform their instruction (Dembosky, et al., 2005; Stecher & Hamilton, 2006; Marsh, Pane, & Hamilton, 2006; Supovitz, 2009).

At the other end of the spectrum are formative assessments, the narrowest type in terms of scope. Formative assessments can vary widely, but they are characterized by a short assessment cycle (i.e., frequent assessments) and are often embedded within the current lesson or unit of instruction (Perie, Marion, & Gong, 2009). Their purpose is to inform teachers of students' mastery of skills related to only one or several content standards; diagnosing student learning, gaps in understanding, and often misconceptions (Perie, Marion, & Gong, 2009). However, they are typically not standardized for comparison across classrooms, grades, or schools.

One of the key inputs of the ANet program is interim assessments which fall between formative and summative assessments on the continuum. These assessments are also referred to as benchmark, predictive, diagnostic, or, in some cases, even formative assessments. Interim assessments that serve an instructional purpose tend to be most similar to formative assessments, but with a longer assessment cycle and greater coverage of content standards. They are "administered during instruction to evaluate students'

knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level" (Perie, Marion, & Gong, 2009, p. 6). These assessments are often standardized for comparison across schools and built around a bank of items aligned to standards and curriculum. Results are reported quickly and often disaggregated by student and standard. When part of an assessment system, the assessments are often paired with support for interpreting the results and making decisions about instructional interventions.

Other Program Components

Prior research contends that the effectiveness of educational reforms in general, and interim assessment programs in particular, is dependent on leaders and teachers having the necessary skills and knowledge to properly implement such programs (Borko, Mayfield, Marion, Flexer, & Cumbo, 1997; Christman, et al., 2009; Blanc, et al., 2010). Furthermore,

> "[w]hile Benchmarks may be helpful, they are not in themselves sufficient to bring about increases in achievement without a community of school leaders and faculty who are willing and able to be both teachers and learners." (Christman et al., 2009, p. 44)

Data-based instructional programs vary widely in respect to the types of support and resources offered to teachers. In her review of data-based interventions, Marsh (2012) found evidence that multiple, linked supports may be necessary to support effective data use: e.g., data systems or tools that are supported through professional development. Unfortunately, the existing research on these supports tends to be observational and

focused on data-use strategies specific to one program, reform effort, or district, making it difficult to connect specific supports to teacher outcomes.

Observational research has shown that the content and reported usefulness of data-based professional development (PD) varies substantially (Marsh, Pane, & Hamilton, 2006; Means, Padilla, DeBarger, & Bakia, 2009; Stecher & Hamilton, 2006). Even when satisfied with the PD they have been offered (Means et al., 2009), some studies concluded that teachers were not offered enough content around the effective use of data or data systems (Mason, 2002; Dembosky, et al., 2005; Clune & White, 2008; Means et al., 2009; Means, Padilla, & Gallagher, 2010). Support at various points in the instructional data use cycle also appears to be lacking. In some cases, support was insufficient during implementation (Jacobs, Gregory, Hoppey, & Yendol-Hoppey, 2009). In others, initial support on data systems access and operations was provided, but failed to assist teachers' analysis, interpretation, and use of data (Means, Padilla, & Gallagher, 2010; Jimerson & Wayman, 2011). Particularly lacking from PD is content to help teachers bridge the gap between interpreting student assessment data and making appropriate instructional decisions (Clune & White, 2008; Goertz, Oláh, & Riggan, 2009a).

Despite extensive literature suggesting best practices for professional development processes and content,[3] there have been few studies that link specific PD models with teachers' use of data, instructional strategies, or student achievement. One

[3] See Borko, et al., 1997; Cohen & Hill, 2000; Garet, Porter, Desimone, Birman, & Yoon, 2001; Desimone, Porter, Garet, Yoon, & Birman, 2002; Lee & Wiliam, 2005; Wayman, 2005; Young, 2006; Goertz, Oláh, & Riggan, 2009b; Young & Kim, 2010; Jimerson & Wayman, 2011.

such study utilized a school-randomized design to explore the effects of the Classroom Assessment for Student Learning (CASL) program, a "self-executing" professional development program aimed at improving teachers' knowledge and practices around classroom and formative assessments through learning teams (Randel, et al., 2011, p. 11). After two years, no detectable difference was found in the quality of teachers' classroom assessment practices or in students' math achievement. However, they did find significant differences in teachers' knowledge of classroom assessment, with intervention teachers answering about 2.78 more items correctly (0.42 standard deviations, $p = 0.01$) (Randel, et al., 2011).

Several observational studies have examined the relationship between PD and teachers' data-based beliefs and practices. Chen, Heritage, and Lee (2005) found that training on the usage of a particular data system was related to improvements in educators' perceived value of student data, as well as increases in their collection, analysis, and use of data for understanding student learning. Case study research in three urban, high-need districts found a similar positive relationship between support provided and teachers' instructional data use (Kerr, Marsh, Ikemoto, Darilek, & Barney, 2006). Means, Padilla, and Gallagher (2010) found a moderate, positive correlation between teachers' perceptions of support for data use and the frequency they used data in various ways ($r = 0.40$).

Prior research cites the importance of coaching as a specific form of PD (Lachat & Smith, 2005; Denton, Swanson, & Mathes, 2007; Marsh, McCombs, & Martorell, 2010). Coaching models can vary widely, but typically include coaches who are experts

in one content area, instructional model, or skill (e.g., literacy, math, data use), and provide in-person support one-on-one or to small groups of teachers. Means, Padilla, and Gallagher (2010) found that 32 percent of the districts they surveyed reported providing some type of data coaching to all schools and 26 percent of districts reported that instructional coaches were required to include elements of data use in the support they provided to teachers in all schools.

As with other forms of professional support, there is a recognized need for experimental studies that examine the impacts of coaching on various teacher and student outcomes, and identify the particularly effective models.[4] Observational evidence has shown that the frequency of coaching around data use is associated with both teachers' self-reported changes in instructional practice and with teachers' likelihood of attributing instructional changes to coaching (Marsh, McCombs, & Martorell, 2010). The frequency with which coaches review data with teachers has also been associated with a small, but positive and significant increase in student reading and math achievement (Marsh, McCombs, & Martorell, 2010). Effective coaches focused on teachers' specific needs, modeled data use, observed teachers during the data-use cycle, provided feedback and expertise, supported dialogue and questions around data and instruction, and helped bridge the gap between data and instruction (Huguet, Marsh, & Farrell, 2014).

[4] The author found no experimental studies of data/instructional coaching on teacher or student outcomes. However, two RCTs offer more general support of coaching on teacher and student outcomes. Blank, Smithson, Porter, Nunnaley, and Osthoff (2006) found evidence of a positive impact of an instructional improvement professional development model on middle-school math and science teachers' alignment of instruction with standards. Campbell and Malkus (2011) found that, over three years, math coaching had a positive impact on student achievement in grades 3 through 5.

The research on supplemental tools and resources to support teachers' data use and instructional practices is extremely limited. These tools often include instructional materials, model lesson plans, curriculum frameworks and guides linked to the interim assessment, and protocols for organizing and analyzing student data, and developing instructional plans. Means, et al. (2009) report that only three of their ten case study districts provided tools as part of their data systems. Goertz, Oláh, & Riggan (2009a) found that the districts in their study used a data analysis protocol to set and reinforce expectations for the analysis and use of interim assessment data. Both teachers and leaders were required to complete the protocol for their respective roles. Leaders also reviewed teachers' protocols during grade-level team meetings, often inserting a level of accountability by asking for evidence that the reteaching plan captured in the protocol actually took place (Datnow, Park, & Wohlstetter, 2007; Goertz, Oláh, & Riggan, 2009a). When part of teachers' instructional communities, tools – such as score reports, curriculum guides, and lesson plans – can provide a starting point for conversations about student performance, as well as structure and routine around data-based practices such as instructional planning and practices (Brunner, et al., 2005; Blanc, et al., 2010; Turner & Coburn, 2012). Datnow, Park, and Wohlstetter (2007) found that protocols helped teachers and principals interpret student data correctly, make appropriate instructional plans based on the data, and ensure follow through on reteaching.

RESEARCH ON INTERIM ASSESSMENTS AND OTHER DATA-BASED INSTRUCTIONAL PROGRAMS

To examine the landscape around teachers' use of data for instructional decision making, experimental, quasi-experimental, and observational research on programs like ANet was consulted. Where the quasi-experimental and experimental research can provide evidence of the causal impacts of these programs, observational studies are more likely to provide descriptive evidence of the processes involved in instructional data use. This section sets the stage for the exploration of the impact of ANet on teacher practices, as well as provides context for the focus on school culture, instructional leadership, and teacher characteristics as potential mediators.

Quasi-Experimental and Experimental Evaluations of Data Use

Quasi-experimental and experimental research on data-based instructional programs often focuses on student achievement outcomes (see exhibit 2.1). As a result, they tend to offer more limited conclusions about the role of mediators (e.g., instructional leadership and school culture) and intermediate outcomes (e.g., teachers' data use and instructional practices). However, the presence of an interim assessment or other data-based instructional program is unlikely to impact teacher practices and student achievement on its own. Therefore, these studies provide evidence that impacts may be the result of some mediating mechanism or mechanisms; e.g., leadership, culture, collaboration, or teacher characteristics.

Several quasi-experimental studies have attempted to link the use of interim assessments and related supports with improved student achievement. While these

designs provide evidence of the relationship between interim assessments and teaching and learning, they can only control for observable characteristics and cannot rule out the influence of other factors. In their study of the Formative Assessments of Student Thinking in Reading (FAST-R) program, Quint, Sepanik, and Smith (2008) used a comparative interrupted time series design to test the program's impact on reading achievement in third and fourth grade in 21 schools in Boston. FAST-R consisted of periodic, short reading assessments aligned to content standards and the summative state test. They were paired with data coaching and professional development aimed at helping teachers interpret and use the assessment results. The study matched the FAST-R schools with other schools in the same district on a number of factors and used five years of baseline achievement scores to predict students' scores in both groups of schools for two years post intervention.

The study garnered mixed findings that were largely not statistically significant. FAST-R teachers found the program's coaches to be helpful, improving their understanding of data and ability to work with students' strengths. However, teachers in matched control schools reported as much professional development, found it as useful, and analyzed data as much or more often than their FAST-R counterparts. Similarly, although the gains in achievement for students in FAST-R schools were larger than those of students in non-FAST-R schools, the difference was not statistically significant. The evaluators hypothesized that the program's training and coaching were not intensive enough, nor were they sufficiently different from professional development in the

comparison schools to have an impact on teaching and learning (Quint, Sepanik, & Smith, 2008).

Henderson, Petrosino, Guckenburg, and Hamilton (2007, 2008) used a similar methodology to examine the impact of quarterly benchmark math assessments on middle school math achievement. In a study that matched 22 treatment with 44 comparison schools, a comparison of post-intervention results found positive, but not statistically significant differences in school mean achievement in intervention and non-intervention schools after one and two years (Henderson, Petrosino, Guckenburg, & Hamilton, 2007, 2008). They hypothesized that the lack of statistically significant results may be attributable to both an underpowered design and too little time for the assessments to impact student achievement. They also cite an unknown "counterfactual": no data were collected on practices in control schools.

There is a recognized need for longitudinal, randomized-controlled experimental designs that study the impact of data-driven strategies on teacher and student outcomes (Chen, Heritage, & Lee, 2005; Wayman & Stringfield, 2006; Marsh, Pane, & Hamilton, 2006; Hamilton et al., 2009) which the research community is beginning to address. Unlike quasi-experimental designs, experimental designs offer stronger internal validity. These designs can be difficult to implement in education as randomizing schools, classes, or students to treatment conditions is not always possible. However, they provide the highest level of evidence of the effectiveness of data-driven instructional programs.

The impact of the instructional improvement model "Data on Enacted Curriculum" (DEC) on instructional leadership and teacher practices was tested in an

randomized controlled trial (RCT) that included 50 middle schools in 5 districts around the United States (Blank, Smithson, Porter, Nunnaley, & Osthoff, 2006). Schools in the treatment group were required to form a leadership team that received training on how to lead data use in their schools. Specifically, the program trained them to assist teachers in using data to reflect on their teaching practices and their students' achievement, using the data to address weaknesses in their practice and gaps in student knowledge, and identify areas for professional development. After two years, math and science teachers' alignment to standards increased in both groups; however, only math teachers in treatment schools showed greater alignment of instruction to standards as compared to the control-school counterparts. This was most notable among the math teachers who were also part of the leadership team (Blank, et al., 2006).

The Using Data program, developed by TERC, is described as a professional development and technical assistance program that helps teachers use data in collaboration with peers to address students' learning needs (Cavalluzzo, et al., 2014). Cavalluzzo and colleagues (2014) used a block-randomized design including 30 treatment and 30 control schools to estimate impact of Using Data on teacher and student outcomes. Using a two-level multilevel model, they found a positive impact after one year on the frequency with which teachers used data (ES = 0.37, $p = 0.01$), as well as their attitudes about the value of data for improving instruction (ES = 0.34, $p = 0.02$). They also found a marginally significant, positive impact on teachers' data literacy (ES = 0.25, $p = 0.06$). Despite this, no detectable difference in students' overall math achievement was found between treatment- and control-school students after two years.

However, students in the lowest performing block of schools (i.e., "highest need" at baseline) did score higher than their control-school counterparts after two years (ES = 0.40, $p = 0.01$) (Cavalluzzo, et al., 2014).

Carlson, Borman, and Robinson (2011) examined the impact of whole-district data-driven reform initiative developed by the Center for Data-Driven Reform in Education (CDDRE). The intervention focused on the targeted use of quarterly predictive benchmark assessments in reading, writing, and math with support from consultants in data analysis and interpretation. Consultants also provided district and school leaders with assistance in reviewing and interpreting the results of the interim assessments and other available data, as well as assistance with selecting and adopting appropriate evidence-based reforms. The study included 549 schools in 59 districts and 7 states. Districts were randomly assigned to treatment and control groups within each state. Using multilevel modeling, the authors found small, but statistically significant positive effects on mathematics achievement after the first year of implementation; school mean achievement was 0.06 standard deviations higher in treatment schools. However, the positive impact on school achievement in reading of 0.03 standard deviations was not statistically significant. The authors contend that since district-level achievement is less

variable than student-level achievement, the results have the potential to be substantively meaningful despite the small effect sizes (Carlson, Borman, & Robinson, 2011).[5, 6]

Indiana was the first state to institute a statewide program of voluntary technology-supported benchmark assessments consisting of two programs: mCLASS and Acuity (Konstantopoulos, Miller, & van der Ploeg, 2013). With mCLASS, teachers in grades kindergarten through two administered periodic assessments in the form of face-to-face language tasks or short (one-minute) probes. Acuity provided online assessments in reading, math, science, and social studies for grades 3 through 8. The multiple-choice tests were aligned to state standards and intended to predict performance on the state summative test. The system also provided teachers with item banks to construct on-demand assessments and access to instructional tools.

A cluster-randomized trial was conducted with 31 treatment schools and 18 control schools that had volunteered to participate in the benchmark assessments (RCT-1). Treatment schools received mCLASS and/or Acuity. Control schools were subject to business-as-usual which included the use of some type of assessment data to monitor student learning in 88 percent of schools (Konstantopoulos, Miller, & van der Ploeg, 2013, p. 486). Using multilevel modeling, they found a significant treatment effect on

[5] Since the outcome was standardized for comparison across states, the impact estimates are already in effect size units. However, the authors cite two methods for calculating effect sizes in cluster-randomized trials: using the estimated within- or between-group variation in the outcome. These methods result in effects sizes ranging from 0.20 to 0.21 in math and 0.12 to 0.14 in reading, respectively. These adjusted effect sizes are the basis for their claim that the results are substantively meaningful.

[6] A follow-up to the Carlson, Borman, and Robinson evaluation (2011) show some potential long-term, positive impacts of CDDRE; however, after year one, control schools were provided the treatment as well. Differences in math and reading achievement after four years were relatively large though, due to smaller sample sizes, not always statistically significant (Slavin, et al., 2013). The authors suggest that data-driven initiatives take time to implement and require that data use go beyond informing instruction to include the adoption of proven methods of addressing the gaps in knowledge revealed by the assessments (Slavin, et al., 2013).

school mean mathematics and reading achievement in grades 3 through 8 (ES = 0.26, $p \leq 0.05$), but not grades K-2. The impact was most notable in 5th- and 6th-grade math where results showed impacts greater than one-quarter of a standard deviation and 3rd- and 4th-grade reading where results showed impacts of about one-seventh of a standard deviation (all $p \leq 0.05$) (Konstantopoulos, Miller, & van der Ploeg, 2013).

Konstantopoulos, Miller, van der Ploeg, and Li (2014, 2016) conducted a RCT of a separate sample of schools as a replication study (RCT-2). No overall impact on achievement was found in grades 3-8 for either math or reading. However, a negative impact was found on student achievement in math and reading in grades K-2. The authors conclude that, based on combined estimates from RCT-1 and RCT-2, the lack of overall impacts in grades K-8 may be the result of offsetting negative results in grade K-2 (mCLASS) and positive impacts in grade 3-8 (Acuity), especially in math.

Cordray, Pion, Brandt, and Molefe (2012) conducted an evaluation of the Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) benchmark assessment program in Illinois. This program consists of a series of computer-adaptive interim assessments intended to measure student growth and mastery of specific skills and standards, as well as predict performance on state summative assessments. In addition to the assessments, educators were provided with online instructional resources, and on-site and on-demand training throughout the year. The evaluation employed a cluster-randomized trial with grade-level assignment. Recruitment resulted in 32 schools that were randomly assigned to receive the MAP benchmark program either in grade 4 or grade 5, with the non-MAP grade assigned to the control group.

Multilevel analyses found no evidence of an impact on student achievement in reading, as measured by the state test and the MAP composite score.[7] Observations and teachers logs showed no evidence that MAP teachers were more likely to differentiate instruction than their counterparts in the control group; however, teacher self-reports of differentiation in grade 5 did show a relatively large positive impact (ARSI[8] = 0.89, $p < 0.001$). The authors acknowledge that there was high variation in teacher dosage; i.e., the number of MAP components (training and resources) teachers completed or used. They also found that control-group teachers had access to other professional development programs and assessment resources (Cordray, et al., 2012).

Finally, the larger i3 evaluation of ANet found no impact of the program on students' math or reading achievement after two years (West, Morton, & Herlihy, 2016). Program effects were found to vary by geographic school network and by schools' baseline ratings of their readiness to partner with ANet. In particular, treatment schools that worked with ANet for two years, but were rated the least "ready" to partner with the program had significantly lower achievement in math and reading than their matched-pair control-school counterparts (both $p < 0.01$). Given the overall null impacts, it is not surprising, therefore, that schools that were rated as most "ready" had significantly higher achievement in both subjects when compared with their matched pairs in the control group (both $p < 0.05$).

[7] Control-group teachers were asked to administer the final MAP assessment to their students, but no score or diagnostic report was provided.
[8] The ARSI or Achieved Relative Strength Index is based on Hedges' g with a correction for clustering at the classroom level.

Exhibit 2.1. Summary of Quasi-Experimental and Experimental Studies of Data-Based Instructional Interventions

Authors	Program	Duration	Location	Sample	Design	Impact Measure	Key Findings
Quasi-experimental Designs							
Quint, Sepanik, & Smith (2008)	**Formative Assessments of Student Thinking in Reading (FAST-R):** Interim assessments (reading) paired with data coaching	Two years	Boston, MA	21 "treatment" schools, 36 "control" schools	Interrupted time series w/ school matching	Reading achievement and percent proficient in 3rd & 4th grade	No impact on student achievement.
Henderson, Petrosino, Guckenburg, & Hamilton (2008)	Quarterly benchmark assessments (math)	Two years	Massachusetts	22 "treatment" schools, 44 "control" schools	Interrupted time series w/ school matching	Math achievement in middle school (8th grade)	No impact on student achievement.
Randomized Designs							
Blank et al. (2006)	**Data Enacted Curriculum (DEC):** a PD program using practice data and student data to improve instruction	Two years	Charlotte-Mecklenburg, Chicago, Miami-Dade, Philadelphia, Winston-Salem	50 middle schools	School-randomized	Degree of alignment of instruction to state standards and state or district assessments	**Positive impact on alignment of math teachers' instruction to standards.** No difference in science teachers' instructional alignment.
Cavalluzzo, et al. (2014)	**Using Data:** a PD and technical assistance program that helps teachers use data to address students' learning needs	Two years	Duval County, Florida	60 schools with grades 4 and 5	Block-randomized assignment of schools	Teacher skills, beliefs, and practices (Y1); student math achievement (Y2)	**Positive impact on teachers' data use and attitudes,** marginally positive impact on teachers' data literacy; no impact on overall achievement in math, positive impact on math achievement of "highest need" block.
Carlson, Borman, & Robinson (2011)	**Center for Data-Driven Reform in Education (CDDRE) data-driven reform program:** Interim assessments, and educator training and support	One year	Alabama, Arizona, Indiana, Mississippi, Ohio, Pennsylvania, and Tennessee	Final estimation sample: Reading: 524 schools in 59 districts Math: 514 schools in 57 districts	District-randomized	Reading and math achievement (all tested grades)	**Positive impact on math.** No impact on reading.

Exhibit 2.1. Summary of quasi-experimental and experimental studies on data-based instructional interventions, cont.

Authors	Program	Duration	Location	Sample	Design	Impact Measure	Key Findings
Randomized Designs, Continued							
Konstantopoulos, Miller, van der Ploeg, & Li (2011) (RCT-1)	**Indiana benchmark assessment program:** mCLASS and Acuity with training on both systems	One year	Indiana	31 treatment schools, 18 control schools	School-randomized	Reading and math achievement in grades K-8	**Positive impact in gr 3-8 math and reading (notably 3rd & 4th grade reading, 5th & 6th grade math);** No impacts in K-2.
Konstantopoulos, Miller, van der Ploeg, & Li (2014, 2016) (RCT-2)	**Indiana benchmark assessment program:** mCLASS and Acuity with training on both systems	One year	Indiana	Including RCT-1 sample: total of over 100 schools.	School-randomized	Reading and math achievement in grades K-8	**Negative impact in math & reading in gr K-2.** The overall treatment effect from gr K-8 (RCT-1 & RCT-2) is not significant in either subject.
Cordray, Pion, Brandt, & Molefe (2013)	**Measures of Academic Progress (MAP):** computer-adaptive benchmark assessments with training & resources for teachers	Two years	Illinois	Total of 32 schools	Grade-randomized	Reading achievement in 4th & 5th grade	No impact on student achievement overall. **Positive treatment effect on 5th-grade teachers' survey self-reported instructional differentiation.**
West, Morton, & Herlihy (2016)	**Achievement Network (ANet):** data-based instructional program offering interim assessments and other supports	Two years	Illinois, Louisiana, Massachusetts	Total of 89 schools	Within-district, matched-pair school randomized assignment	Math and reading achievement in grades 3-8	No impact on student achievement overall. **Positive treatment effect in schools rated most "ready" to partner with ANet at baseline.**

These quasi-experimental and experimental studies show mixed impacts of interim assessment programs specifically, and data-driven instructional practices more generally, on student achievement. Among them, several studies found no impact on student achievement, a handful of studies found small positive impacts on achievement, particularly in math, and one found a negative impact on K-2 reading and math (exhibit 2.1). One shortcoming of these studies is that they often fail to collect process data on the thoughts, behaviors, or actions of educators' that may mediate the pathway between data use and student achievement (i.e., the "black box"). As a result, explanations for lack of teacher and student impacts are often conjectural.

Observational Studies of Impact of Data Use on Teacher Practice

To address the limitations of quasi- experimental and experimental studies, this review turns to observational studies of instructional data use. The strength of observational research is its focus on relating data use and instructional practices to the contextual conditions of schools and characteristics of teachers. Although observational studies of data-driven initiatives are limited by an inability to draw causal conclusions, they can offer more detailed and descriptive evidence of the mediators of instructional data use. This section reviews frequently cited observational studies that yield evidence of a relationship between data-based initiatives and the key outcomes of the research questions in this dissertation: teachers' (1) analysis of assessment data and use in planning instruction, and (2) instructional practices.

Evidence shows that data use is a nearly universal practice in schools; however, there is considerable variation in the frequency with which data are used to inform instruction. Some studies suggest that there is wide variation in the adoption of practices between districts (Marsh, Pane, & Hamilton, 2006; Christman, et al., 2009; Goertz, Oláh, & Riggan, 2009a). However, the greatest variation in the frequency of data use appears to be among teachers within the same school (Marsh, Pane, & Hamilton, 2006).

Descriptive findings also suggest that data use is more prevalent in the elementary grades than the middle grades and, not surprisingly, in math and reading (or ELA) compared to non-tested subjects (Dembosky et al., 2005; Faria, et al., 2012). However, some studies found high data use across all grades (Dembosky et al., 2005; Christman, et al., 2009). Where data use is higher among elementary-level teachers, it may be attributable to their smaller self-contained classes (Faria, et al., 2012), the availability of assessment data in math and reading, and the greater prevalence of whole-group instruction in high-school classes (Dembosky, et al., 2005).

Examples of Data Analysis & Use in Instructional Planning

In a study of math interim assessments and data use, Goertz, Oláh, & Riggan (2009a) found that all teachers reported reviewing their students' interim assessment results and nearly all teachers examined the data both at the student level and by content area (Riggan & Oláh, 2011). Across multiple studies, the majority of teachers were satisfied with the data that their assessments provided, and cited the usefulness of the data

not only for identifying and monitoring what students were learning, but also for shaping and modifying instruction, and preparing for and predicting performance on the state test (Dembosky et al., 2005; Stecher & Hamilton, 2006; Christman, et al., 2009; Clune & White, 2008; Goertz, Oláh, & Riggan, 2009a).

Several studies discussed the use of interim assessment content for instructional planning, including the use of content on future interim assessments for backward planning and to identify gaps in, or misalignment with, the curricular scope and sequence (Clune & White, 2008; Christman, et al., 2009). Backward planning a lesson or unit begins by making the desired learning goals concrete and specific; e.g., what students will understand, know, and be able to do (Wiggins & McTighe, 2005). Although a review of prior research showed some evidence that teachers were using these practices (Clune & White, 2008; Christman, et al., 2009), they do not appear to be particularly widespread.

Using data to inform *reteaching* appears to be far more common practice. Prior research suggests that the vast majority of teachers analyze interim assessment results by looking for weak areas in performance or gaps in student leaning – topics on which students failed to meet expectations or achieve mastery – in order to target their reteaching (Supovitz & Klein, 2003; Goertz, Oláh, & Riggan, 2009a; Stecher & Hamilton, 2006[1]; Stecher et al., 2008; Riggan & Oláh, 2011). In one study, teachers reported that benchmark assessment data not only allowed them to identify gaps in

[1] In this study, teachers responded both in reference to summative state tests and "progress" tests.

student knowledge, but gaps they wouldn't have otherwise known about (Christman, et al., 2009).

However, teachers' thresholds for determining which learning gaps were large enough to require an instructional response (e.g., reteaching) were often individually set and varied by student, class, timing of the assessment, and overall range of student responses. Some teachers flagged students whose scores were below a minimum performance level. Others weighed the results of assessments and professional judgments of other forms of "data" such as their students' backgrounds, prior performance, and the placement of the assessed content within the district curricular scope and sequence (Goertz, Oláh, & Riggan, 2009a; Oláh, Lawrence, & Riggan, 2010; Hargreaves, Morton, Braun, & Gurn, 2014). The authors found that setting thresholds was a critical strategy; the process allowed teachers to group students and prioritize what they could reteach within the available instructional time (Goertz, Oláh, & Riggan, 2009a; Oláh, Lawrence, & Riggan, 2010).

Beyond identifying gaps in learning, interim assessment data often helped identify students' misconceptions; a feature that a majority of teachers found useful (Christman, et al., 2009). Examining the underlying cause of students' incorrect responses provides teachers with critical information for guiding the focus of reteaching. In their aforementioned study, Goertz, Oláh, and Riggan conducted multiple interviews and classroom observations of 45 elementary school teachers in 9 schools in Philadelphia and Cumberland (PA). As part of the interview process, the researchers asked teachers to

diagnose students' misconceptions based on incorrect responses to assessment items (Oláh, Lawrence, & Riggan, 2010). Teachers in Philadelphia more commonly attributed incorrect responses to students' procedural errors than conceptual errors (Oláh, Lawrence, & Riggan, 2010). Teachers in Cumberland were more likely to cite explanations of student errors that fell on a symptom-etiology continuum (Goertz, Oláh, & Riggan, 2009a). The differences they found between teachers in Philadelphia and Cumberland may be representative of the different ways teachers interpret misconceptions in various districts and schools.

Misconceptions were not the only explanations given for students' incorrect responses to assessment items, however. Teachers in both districts cited other cognitive "weaknesses" such as a student's attention deficit problems or limited English proficiency, as well as contextual or external factors such as their students' lives at home. Cumberland teachers tended to interpret student errors as a reflection of the curriculum design such as overly-complicated lessons that mixed various math algorithms. They were also less likely than their Philadelphia counterparts to cite "non-explanations" such as "fractions are hard for them." (Goertz, Oláh, & Riggan, 2009a, p. 126)

Although the vast majority of teachers use interim assessment data to identify and target students' learning needs for reteaching, there appears to be substantial variation in the strategies teachers used to analyze and interpret data. It may be that some of the strategies are more effective at identifying the students who would benefit most from reteaching and their skill gaps, misconceptions, and procedural errors. For example, it is

possible that thresholds for reteaching should be set lower or that attributing students' poor performance to "non-explanations" can be counterproductive and fails to take full advantage of the information contained in the student data.

Examples of Instructional Practice

The instructional actions that teachers take based on the results of interim assessments encompass a range of decisions such as what should be taught or re-taught, to whom, and when (Hamilton, et al., 2009). At least one study reported that a majority of teachers felt that benchmark assessments had improved their instruction for students who had not mastered particular skills (Christman, et al., 2009). However, other research suggests that data-based initiatives may result in changes to *what* teachers taught, but not necessarily *how* they taught it (Marsh, 2012). Data were used to determine what needed to be re-taught and to whom, but often the same instructional strategies were used (Goertz, Oláh, & Riggan, 2009a).

In practice, most teachers used student assessment data for tailoring instruction to meet the needs of the whole class, small groups, or individual students, or to provide "developmentally appropriate lessons" as indicated by their students' results (Supovitz & Klein, 2003; Dembosky et al., 2005[2]; Marsh, Pane, & Hamilton, 2006). Teachers often responded to widespread gaps in learning using whole-class instruction and employed small-group instruction – often outside of regular class time – for less pervasive issues

[2] In this study, teachers were using prior year summative data and mid-year assessment results.

(Oláh, Lawrence, & Riggan, 2010), Although half of the teachers in one study reported weekly use of classroom assessment data to individualize instruction (Stecher & Hamilton, 2006), other studies either found greater variation in the frequency that teachers reported targeting instruction to students' particular gaps in knowledge – including individualizing instruction – or less frequent use of differentiation in general (Supovitz & Klein, 2003; Dembosky et al., 2005; Marsh, Pane, & Hamilton, 2006; Goertz, Oláh, & Riggan, 2009a; Oláh, Lawrence, & Riggan, 2010). Individualized instruction may occur more frequently in the presence of greater instructional resources and support staff (e.g., instructional coaches, curriculum specialists, and other school-based aides) who provide support both in planning and carrying out instruction (Goertz, Oláh, & Riggan, 2009a).

Observational research on interim assessment and other data-driven instructional programs has shown some consistency in the ways in which teachers review, analyze, and use interim assessment to inform their instruction, as well as the instructional strategies that teachers employ. There is also evidence that providing teachers with interim assessment data can help them to identify and target students' learning needs. However, data-based instructional practices vary in frequency and evidence that they lead to instruction that impacts student learning is sparse. Ultimately, many reteaching strategies were considered to be superficial; failing to diagnose and address student misconceptions, alter instructional strategies based on the data, or select adequate instructional interventions (Goertz, Oláh, & Riggan, 2009a; Blanc, et al., 2010; Shepard,

2010). In some cases, changes in practice attributed to data use may actually be unproductive or lead to negative impacts on student achievements; e.g., a focus on "bubble" students, a narrowing of the curriculum, or superficial changes in instruction such as test preparation (Brunner, et al., 2005; Stecher & Hamilton, 2006; Diamond & Cooper, 2007; Blanc, et al., 2010; Christman, et al., 2009; Booher-Jennings, 2005; Coburn & Turner, 2011). These practices can be more common in low-performing schools that seek quicker ways to meet accountability targets rather than longer-term efforts to improve instruction (Diamond & Cooper, 2007).

Mediators of Effective Instructional Data Use

The existing research on the relationship between assessment, data use, instructional practice, and student learning is said to be a poorly understood process or "black box" (Little, 2012; Spillane, 2012; Bulkley, Oláh, & Blanc, 2010). Efforts to understand what mediates or moderates effective instructional data use is important to providing context for experimental studies and potential "insight into when and under what conditions data use acts as a productive pathway to educational improvement and when it does not." (Coburn & Turner, 2012, p. 100) This study examines roles of school culture, instructional leadership, and teacher characteristics in teachers' data use and instructional practices. The remainder of this chapter explores prior research on these mediators of instructional data use.

School Culture & Instructional Leadership

In educational research, the term *culture* is often used to describe qualities or characteristics of the schools in which reform efforts unfold. Various researchers have identified school culture as consisting of collective or shared beliefs, vision, knowledge, norms, values, customs, expectations, and language (Senge, 1990; Hargreaves, 1995; Faria, et al., 2012). Similarly, DuFour and Eaker (1998) contend that school culture "…is founded upon the assumptions, beliefs, values, and habits that constitute the norms for that organization—norms that shape how its people think, feel, and act." (p. 131) Although these beliefs are often implicit, they "…provide a powerful foundation for members' understanding of the way they and the organization operate." (Page, 1987, p. 82)

Describing its complex role in school reform, Hargreaves (1995) noted that "[s]chool culture may be a *cause*, an *object* or an *effect* of school improvement." (Hargreaves, 1995, p. 41 [emphasis in original]) School culture can influence the degree to which educational reform efforts, including improvements in teaching and learning, are realized, with the culture of organizations potentially acting as both an enabler or inhibitor (Johnson, Berg & Donaldson, 2005; Datnow, Park, & Wohlstetter, 2007; Coburn & Turner, 2011). In fact, some contend that past school reform movements have failed due to a lack of attention to the power dynamics and culture in schools, and how they affect educational change (Sarason, 1996; DuFour & Eaker, 1998; Fullan, 2007). DuFour and Eaker argue that to sustain any educational reform, the "change must be

embedded within the culture of the school." (1998, p. 133) Accountability-based reforms assume that simply giving educators information about their performance will motivate them to improve.

> The assumption is that teachers will try harder and become more effective in meeting goals for student performance when the goals are clear, when information on the degree of success is available, and when there are real incentives to meet the goals. (Newmann, King, & Rigdon, 1997, p. 43)

Sarason (1996) and others would argue that these assumptions are severely flawed and reforms are destined to fail unless the complexities of school culture are taken into account. These complexities include the power relationships within schools; educational reforms must be introduced in a way that respects all roles and gives all stakeholders a sense of ownership over the process.

School culture has a number of facets such as the teaching culture, leadership culture, pupil culture, and parent culture (Stoll, 1998). Each facet can be shaped by its own goals, norms, expectations, processes, and attitudes (Faria, et al., 2012). Likewise, each facet can impact the success or failure of school reform (Stoll, 1998). Since these facets of culture exist in schools simultaneously, they can also interact in complex ways.

This study explores the role of two facets of school culture that are hypothesized to encourage more effective use of student assessment data for instructional purposes: achievement culture and teacher professional culture (e.g., collaboration) (Dembosky et al., 2005; Borman, et al., 2005; Christman, et al., 2009; Goertz, Oláh, & Riggan, 2009a; Purkey & Smith, 1983; Zigarelli, 1996; Little, 1999; Datnow & Park, 2014). This study

also explores the role of instructional leadership. Though leadership is not a facet of culture, school leaders may facilitate or inhibit positive school cultures (see below: Stoll, 1998; Copland, 2003; Supovitz & Klein, 2003; Dembosky et al., 2005; Halverson, et al., 2007; Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010).

The motivation for exploring culture and leadership in this study stems from their frequent inclusion in discussions of data culture (Mason, 2002; Love, 2008; Faria, et al., 2012; Mandinach, 2012). Data culture has been defined as one that specifically

> espouses the importance of using data to inform practice. The environment contains attitudes and values around data use, recognized behavioral norms and expectations to use data, and objectives for why data are to be used, informed by a district-level or school-level vision for data use. (Mandinach & Jackson, 2012, p. 141)

A school's data culture is one of the main ways in which the vision, responsibilities, and expectations around instructional data use are conveyed to teachers. School leaders are often responsible for disseminating this vision (Mandinach & Jackson, 2012). Evidence suggests that the effectiveness of interim assessments in improving student achievement may be strongest when there is "concomitant attention to developing strong school leaders who promote data-driven decision making within a school culture focused on strengthening instruction, professional learning, and collective responsibility for student success." (Blanc, et al., 2010, p. 206) These conditions can be difficult to disentangle and impact one another in important ways.

Achievement Culture. The collective achievement culture within a school is touched upon in the literature, but is not particularly well researched. Achievement

culture is often conceptualized as an orientation toward student achievement based on high expectations for student performance, a shared vision that all students can learn, a focus on high academic standards and quality instruction, and frequent monitoring of teacher efficacy and student progress toward meeting these goals (Purkey & Smith, 1983; Zigarelli, 1996; Little, 1999).

The extent to which schools emphasize an achievement orientation has been shown to relate to higher student achievement (Zigarelli, 1996). It may also relate to teachers' individual and collective responsibility for student achievement, something which Little (1999) advocates emphasizing through professional development focused on inquiry around student learning. This recommendation was grounded in research that found a positive correlation between student achievement and teachers' levels of collective responsibility for learning (Lee & Smith, 1996). Prior research has also shown that high achieving charter schools were more likely to have high academic expectations for students; defined as "a relentless focus on academic goals and having students meet them." (Dobbie & Fryer, 2011, p. 9)

As it relates to instructional data use, the hypothesis is that teachers in schools with a stronger achievement culture also use data more effectively for improving student learning and meeting learning standards. Datnow and Park (2014) acknowledge the difficulty of addressing teachers' low expectations for student achievement, especially when working with underprivileged students. In case studies of six elementary and

secondary schools in three districts in different states[3], the researchers found that as teachers began to examine disaggregated student data, their conversations shifted from placing blame on the students to discussions of instructional gaps and strategies to address them. To support this shift in focus, districts made a conscious effort to create a culture of high expectations and ownership over the success of all students, and to develop programs to address students' needs.

This strategy was also central to education reform efforts in Ontario. Intended to improve the educational outcomes of special education students, the reforms led many districts to rely on data-based instructional strategies. These strategies were embedded within data cultures that espoused collective responsibility for student learning (Hargreaves & Braun, 2012, 2013).

Professional Culture & Collaboration. As another facet of school culture, professional cultures or the cultures of teaching "comprise beliefs, values, habits and assumed ways of doing things among communities of teachers." (Hargreaves, 1994, p. 165) Cultures of teaching also give "… meaning, support, and identity to teachers and their work." (Hargreaves, 1994, p. 165) Like other forms of culture, cultures of teaching can take various forms and each has direct implications for teachers' work and the success of educational programs and reforms (Hargreaves, 1994). For example, school data culture can both foster and be fostered by collaboration among teachers (Chen,

[3] Interviews were conducted with district leaders (n = 9), school leaders (n = 10), and teachers (n = 76).

Heritage, & Lee, 2005; Lachat & Smith, 2005; Wayman, 2005; Wayman & Stringfield, 2006; as cited in Wayman & Cho, 2009).

Hargreaves (1994) identifies four cultures of teaching: individualism, collaboration, contrived collegiality, and balkanization. Historically, the teaching profession was one of isolation, with teachers operating alone or in silos. However, recent research and school improvement efforts show that most teachers not only access data alone, but also with members of their teacher team (Means, et al., 2009). This has put a focus on the potential benefits of pairing a collaborative professional culture with data use to improve teaching and learning (Datnow, Park, & Wohlstetter, 2007; Hargreaves & Braun, 2012).

Hargreaves' (1994) defines collaborative teacher cultures as spontaneous, voluntary, and development-oriented. More than camaraderie or congeniality, collaborative teacher cultures encourage teachers to work together to analyze and improve their practices by engaging in an ongoing cycle of continuous improvement; a cycle that promotes higher student achievement through team learning and collective inquiry (DuFour, Eaker, & DuFour, 2005, p. 36). In contrast, contrived collegiality is more likely to be regularly scheduled, compulsory, implementation-oriented, fixed in time and space, and predictable. In settings where collaboration is contrived, teachers' joint work is less likely to result in meaningful change (Hargreaves, 1994).

Datnow (2011) addresses the fact that collaboration among teachers around instructional data use often resembles contrived collegiality. In her research, though

teacher collaboration around data was often scheduled, prescribed, and organized around predetermined questions and goals, few negative consequences were observed. Datnow attributes this to the high capacity of schools and the role of the school leader in creating a positive school culture for data use and continuous improvement, and teacher capacity and trust (Datnow, 2011). Furthermore, contrived teacher cultures may still effect change when they are based on a collective sense of responsibility for a "commitment to shared goals and targets for improvement" that includes student success (Hargreaves, Morton, Braun, & Gurn, 2014, p. 8; Little, 1999). Ultimately, what begins as contrived collegiality can become a collaborative environment much like Hargreaves (1994) advocated (Datnow, 2011; see also Blanc, et al., 2010).

Key to building a culture of data use, providing opportunities for collaborative inquiry gives teachers a forum for analyzing data and using results to inform instruction (Gerzon, 2015). The amount of collaboration around instructional responses to student benchmark data have been shown to be positively related to growth in student achievement (Christman et al., 2009). Further, the use of benchmark assessment data was more likely to lead to instructional improvement when "school leaders focused on developing robust instructional communities that supported teachers in interpreting benchmark data in the light of the learning goals." (Blanc, et al., 2010, p. 206)

However, not all collaboration is strong, purposeful, or even regular enough to affect change (Halverson, et al., 2007; Dembosky, et al., 2005). Although teacher teams that include a school leader may collaborate more frequently and purposefully

(Halverson, et al., 2007), collaboration between teachers and leaders, or teachers across different departments or roles, are more likely to encounter conflicting ideas, interpretations of data, and responses to results than when teachers interact with other teachers in their grade or subject (Coburn & Turner, 2011). Datnow and Park (2014) found that trust and common goals were key to ensuring that disagreements during collaboration were healthy debates and not impediments to data use. They also found that trust was engendered when the data culture was non-threatening and non-punitive, but existed expressly for the purpose of instructional improvement (Datnow & Park, 2014).

Instructional Leadership. Recent educational reform efforts that focus on accountability for student achievement have put increased focus on school leaders in their role as the schools' instructional leaders (Leithwood & Montgomery, 1982; Grubb & Flessa, 2006). In this role, school leaders play an important part in shaping school cultures and championing school improvement efforts that target teaching and learning (Stoll, 1998; Copland, 2003; Supovitz & Klein, 2003; Dembosky et al., 2005; Halverson, et al., 2007; Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010). They can influence the success of school reforms, including data-based instructional initiatives (Copland, 2003; Wayman & Stringfield, 2006; Young, 2006).

As leaders of data-based initiatives, school leaders play a role in the effectiveness with which teachers interpret and use data, and alter their practices to improve student achievement (Johnson, Berg & Donaldson, 2005; Wayman & Stringfield, 2006; Datnow, Park, & Wohlstetter, 2007; Diamond & Cooper, 2007; Clune & White, 2008). They do

this by shaping school culture and vision around data use, as well as how it is operationalized through practices and structures (Copland, 2003; Wayman, 2005; Goertz, Oláh, & Riggan, 2009a; Blanc, et al., 2010; Faria, et al., 2012; Datnow & Park, 2014).

Various researchers describe the role of the principal as an instructional leader as both a model of instructional data use and a participant (Young, 2006; Blanc, et al., 2010; Coburn & Turner, 2011). Evidence suggests that instructional leaders who promote and model a school-wide commitment to data use are more likely to foster data use among their teachers (Mason, 2002; Lachat & Smith, 2005; Murnane, Sharkey, & Boudett, 2005; Marsh, Pane, & Hamilton, 2006; Kerr et al., 2006; Knapp, Copland, & Swinnerton, 2007; Blanc, et al., 2010). Leaders influence their school's culture by establishing norms, expectations, and purpose around data-based instructional practices (Heritage & Yeagley, 2005; Marsh, Pane, & Hamilton, 2006; Datnow, Park, & Wohlstetter, 2007; Goertz, Oláh, & Riggan, 2009a; Blanc, et al., 2010; Coburn & Turner, 2011; Datnow & Park, 2014; Gerzon, 2015), monitoring teachers' data-based practices (Goertz, Oláh, & Riggan, 2009a), and creating "accountable learning systems" in their schools (Halverson, Grigg, Pritchet, & Thomas, 2005, p. 3).

Operationally, they create structures, routines, and time for data analysis (Heritage & Yeagley, 2005; Datnow & Park, 2009; Coburn & Turner, 2011). For example, they promote distributed leadership and data teams (e.g., developing content-area or data leaders who share in leading the work around data) (Copland, 2003; Lachat & Smith, 2005; Wayman & Stringfield, 2006; Knapp, Copland, & Swinnerton, 2007;

Gerzon, 2015), offer feedback on the instructional plans teachers develop from student data (Dembosky, et al., 2005), and provide supports for data use such as collaboration time and professional development (Wayman & Stringfield, 2006; Young, 2006; Halverson, et al., 2007; Daly, 2012).

School leaders wear a variety of hats; they develop and manage human capital (i.e., their staff and teachers), are the disciplinarian, and are responsible for internal processes (e.g., schedules, school budgets, school management, and external relations (e.g., community and parent). One strategy is to prepare school leaders to fulfill each of these demands. However, finding candidates who are capable of filling each of these leadership roles well can be difficult (Grubb & Flessa, 2006). Consistent with promoting distributed leadership, the ANet model advocates for the development of a data leadership team that supports teachers' instructional use of data. The role of instructional leader in this study is not only filled by the school principal, but also may be filled by other school staff members such as content or grade level lead teachers or specialists. For this reason, it is important to note that instructional leadership in this study may refer to leadership provided by the school principal or other school leaders.

Teacher Characteristics: Confidence & Attitudes

Despite most data-based interventions being implemented at the school level, findings from prior research (Marsh, Pane, & Hamilton, 2006) and the larger i3 evaluation suggest that teachers' data-based practices vary widely within schools (West,

Morton, & Herlihy, 2016). In fact, data-use and instructional practices vary among teachers within a school more so than across schools. This suggests that there are important teacher-level characteristics that influence data-based practices. This study explores the roles of confidence, as well as beliefs and attitudes, in predicting teachers' data use and instructional practices.

Confidence. Pedagogical data literacy is a term with specific meaning (Mandinach, 2012). It is defined as the ability to translate evidence of student learning into "actionable instructional knowledge and practices by collecting, analyzing, and interpreting all types of data." (Mandinach, Friedman, & Gummer, 2015) The process requires understanding of standards, expertise in the academic discipline, pedagogical knowledge, as well as knowledge of how students learn (Mandinach & Gummer, 2013). With the emergence of data systems to support instructional data use, it also includes knowledge of and ability to use technological data tools (Supovitz & Klein, 2003; Wayman & Cho, 2009).

In discussions of data use, it is often acknowledged that some level of data literacy is required to support effective practices; e.g., interpreting assessment results accurately, using student data to draw accurate inferences about student performance, and making decisions about appropriate instructional interventions (Webb, 2002). However, much of the research has called attention to insufficient pre-service and in-service teacher training around interpreting and using data for instructional improvement as a barrier for effective data use (Schafer & Lissitz, 1987; Daniel & King, 1998; Massell, 2001;

Stiggins, 2002; Mandinach & Honey, 2008; Means, Padilla, & Gallagher, 2010; Mandinach & Gummer, 2013; Mandinach, Friedman, & Gummer, 2015).

A recent survey of schools of education found that the large majority reported offering at least one course in data use or integrating data-driven decision making skills into other course offerings. However, a review of syllabi found that these courses were cursory in coverage and tended to focus on assessment literacy – how to select or design classroom assessments – rather than data analysis and use (Mandinach, Friedman, & Gummer, 2015). Teachers also report gaps in their own knowledge with respect to the provision of external supports, for skills such as asking the right questions of student data, interacting appropriately with data systems, data literacy, incorporating data use into practice, and sharing and codifying knowledge (Jimerson & Wayman, 2015).

A lack of capacity is a frequently cited barrier to effective data use (Mason, 2002; Supovitz & Klein, 2003; Means, et al., 2009; Coburn & Turner, 2011). Teachers' preparedness to interpret and use data has been shown to enable data use and predict the frequency with which they use data in decision-making and to adapt their teaching (Kerr, et al., 2006; Marsh, Pane & Hamilton, 2006). Without sufficient training, practitioners are not easily able to make judgments about the alignment between assessment techniques and curriculum standards, or the validity and reliability of particular instruments (Heritage & Yeagley, 2005).

In contrast, teachers who are able to align and map student performance with learning standards are more likely to create "developmentally appropriate lessons."

(Supovitz & Klein, 2003, p. 16) Goertz, Oláh, & Riggan (2009a) found that teachers who exhibited a greater understanding of using assessment data to uncover students' conceptual misunderstandings were more likely to make meaningful and appropriate modifications to their instruction. Along with culture and leadership, they argue that effective data use is enabled through professional development that focuses on interpreting data and using it to make appropriate instructional decisions. Further, Gerzon (2015) suggests that professional learning opportunities should focus on building teachers' knowledge of varying instructional strategies to ensure they are able to remediate gaps in student learning that are uncovered by the data.

Although prior research has established some evidence of the positive relationship between teachers' data literacy and effective data-based instructional practice, this study lacks a direct measure of data literacy. Instead, teachers were asked to self-report their confidence using data and various instructional strategies. These composite measures of data confidence and instructional confidence are hypothesized to be correlates of data literacy. While there is little prior research that can establish this relationship empirically, studies have shown a link between teaching self-concept (i.e., confidence in teaching abilities) and personal efficacy (Guskey, 1988).

Attitudes. Prior research has shown that teachers' attitudes towards assessments and assessment data may play a role in their adoption of educational innovations, including data-driven instructional practices (Luo, 2008; Guskey, 1988; Kerr, et al., 2006; Marsh, Pane, & Hamilton, 2006). Attitudes towards assessments have been defined

widely, encompassing perceptions of clarity, congruence, cost, validity, reliability, utility, and alignment of the assessments to content standards and the curriculum. While leaders' attitudes toward data use are often overwhelmingly positive, teachers' attitudes toward data and assessment, though generally positive, often vary more widely; for example, teachers are more likely to display skepticism (Dembosky, et al., 2005; Wayman, Cho, Jimerson, & Spikes, 2012). Still, Wayman, Cho, Jimerson, and Spikes (2012) found that teachers and leaders held generally positive attitudes toward data and data use even when faced with barriers.

Unless challenged, our thoughts and actions are often influenced by automatic responses, i.e., they are made intuitively. Unless time is taken to consciously process information, intuitive conclusions and resulting decisions are typically based on prior beliefs and evidence, and biased toward confirming them (Kahneman, 2011). In relation to data use, Coburn and Turner (2011) found that teachers' prior beliefs, assumptions, and experiences may influence which sources of data they notice, as well as how they interpret and take action based on data. Teachers tend to notice data that are congruent with their beliefs, interpret them through that lens, and ignore data that contradict or challenge their beliefs (Coburn & Turner, 2011, p. 177). Buy-in to data-driven practices was greater when teachers thought interim assessments data were useful and valid measures of their students' knowledge and ability (Kerr, et al., 2006; Marsh, Pane, & Hamilton, 2006).

When faced with interpreting multiple forms and sources of data, beliefs about their "educational significance" – e.g., the consistency with other knowledge of students – may impact the weight they assign to each (Young & Kim, 2010, p. 13: citing Young, 2008). Teachers often ignored state assessment data when they perceived classroom assessments and student work to be more valid measures of student knowledge (Kerr, et al., 2006). When teachers feel overwhelmed by data, they may narrow the sources based on what is consistent with their preconceptions (Coburn & Turner, 2011). In some cases, teachers "base their decisions on experience, intuition and anecdotal information (professional judgment)" instead of systematically collected information (Ingram, Louis, & Schroeder, 2004, p. 1281).

The compatibility of a reform effort with teachers' beliefs, educational philosophies, or "personal metric" can play a part in the successful adoption of practices (quoting Ingram, Louis, & Schroeder, 2004, p. 1281; Borko, et al., 1997; Hochberg & Desimone, 2010). Consistent with the aforementioned findings in Datnow and Park (2014), Cho and Wayman (2013) found that individuals' attitudes in one district – particularly regarding the idea that all students could learn at high levels – were difficult to change and required the intervention of district leaders who enforced expectations around student achievement and promoted the use of data for that end.

CONCLUSION

To date, experimental studies of the impacts of data-based instructional programs on student outcomes have yielded mixed results; when statistically significant impacts have been found, the results vary by subject, grade level, and direction. Observational studies of data-based instructional programs show variation in teachers' data use and instructional practices across grades, subjects, and context. Although all teachers appear to be using data in some way, the research suggests that this typically has not resulted in meaningful changes in instruction that leads to improved student achievement. The process by which teachers link the results of their data analysis with decisions to alter their instructional practice is not well understood and may be a one of the reasons that the theory of action behind data-based instructional practice breaks down.

It is also clear that school leadership, culture, and collaboration can each take on a range of characteristics; some that appear to support teachers' instructional data use and others that may inhibit it. Likewise, teachers' own attitudes and aptitudes likely play a role in their ability to improve their instruction. Current strategies may fail in the absence of strong leadership, positive achievement and professional cultures, and teachers who see the value in and are able to do the difficult work of instructional data use.

Turner and Coburn (2012) point out that prior "studies have tended to examine either the outcomes or the processes of data use interventions, but not both." (p. 3). Furthermore, they contend that research on the linkages between pathways and outcomes is crucial. This study adds to the existing research through a secondary analysis of data

from the larger i3 evaluation of ANet. The inclusion of quantitative and qualitative data allows a thorough characterization of each of the key mediators and their relationship with teacher practices, enhancing our understanding of whether and how these factors relate to teachers' data use and instructional practices. The design of the larger evaluation and the proposed analyses, discussed in the following chapter, allows for potentially stronger conclusions about these relationships.

CHAPTER THREE: METHODOLOGY

Data-based instructional programs are unlikely to have an impact on student achievement without in some way changing teacher practices. Because these changes are poorly understood, this study focuses on understanding the intermediate effects of ANet on teachers' data use and instructional practices. As chapter two documents, there is a need for rigorous study of the effectiveness of instructional data use and interim assessments programs in improving teacher practices, as well as the mechanisms by which these improvements take place. This study addresses the gap in the research through a secondary analysis of data from an evaluation of the Achievement Network's (ANet) data-based instructional program.

Though the evaluation takes a mixed methods approach, it was principally designed as a matched-pair, school-randomized controlled trial (RCT). This design served the larger evaluation's primary goal of measuring the effect of ANet on student achievement. Additional quantitative and qualitative data were collected to test ANet's logic model and to better understand how the program may impact student achievement by way of intermediate effects on school structures, and leader and teacher actions.

The mix of data collection modes resulted in an extraordinarily rich dataset. In this study, quantitative data from year-two teacher surveys are the primary data source for estimating the effect of ANet on teacher practices, and for exploring potential school- and teacher-level mediators. These findings are supplemented by qualitative data from interviews with leaders and teachers in a subset of year-two treatment schools. The remainder of this chapter discusses the design of the larger i3 evaluation ("the

evaluation" or "the larger evaluation"), as well as the secondary analysis of the evaluation data for the purposes of this dissertation ("the dissertation" or "this study").

MIXED METHODS FRAMEWORK

Mixed methods research is operationally defined as combining quantitative and qualitative approaches to design and data collection in a single study. The purpose is three-fold; to capitalize on complementary strengths and non-overlapping weaknesses, to ensure adequate representation or comprehensiveness of information, and to legitimate the validity of inferences from the data (Onwuegbuzie & Teddlie, 2003; Johnson & Onwuegbuzie, 2004; Bamberger, Rugh, & Mabry, 2006). However, mixed methods designs can also be valuable in the context of an RCT (Spillane et al., 2010).

It is not uncommon for RCTs to produce weak or null findings on the effects of educational interventions (Coalition for Evidence-Based Policy, 2013). Therefore, there is a renewed effort to ensure that researchers, program developers, educators, and policymakers learn from null findings by exploring possible reasons for these results, particularly design or methodological issues, flaws in the program logic model or causal chain, failures of implementation, or contextual factors that act as barriers to implementation (White, 2013; Jacob, Jones, Hill, & Kim, 2015). Spillane, et al. suggest that "mixed method designs increase the probability that such studies will generate other valuable empirical knowledge in addition to evidence of the absence of a treatment effect." (2010, p. 23) Among RCTs that find significant program effects, a common criticism is that they tell us whether an intervention works, but not how or why (White,

2013). Whether or not ANet has an impact on teachers' data-based instructional practices, the mixed methods approach of the evaluation provides an opportunity to learn about the context and processes involved.

Although the larger evaluation did not set out with a particular mixed methods design, it most closely resembles an embedded experimental model. In an embedded design, data are collected in a single phase, but one mode plays a supplemental role in analysis and interpretation. In both the larger evaluation and this study, the qualitative data alone would not allow for valid causal conclusions about the effect of ANet on educator practices and student achievement and, therefore, are a supplement to the quantitative data (Creswell & Plano Clark, 2007).

EVALUATION DESIGN

As noted above, the larger evaluation was primarily designed as a matched-pair school-randomized controlled trial. Randomized-controlled trials are often considered the gold standard in education and other fields, and have distinct, but interrelated advantages over observational research designs. With randomization, units such as schools are assigned to treatment and control conditions based only on chance and have a known probability of assignment to the treatment group. Successful randomization provides a counterfactual which enables the measurement of what would have happened to the treatment group in the absence of treatment (Puma, Olsen, Bell, & Price, 2009). More importantly, randomization addresses the issue of internal validity by ruling out plausible threats to the validity of observed treatment effects such as selection bias (Shadish, Cook,

& Campbell, 2002). In sufficiently large samples, treatment and control groups will be equivalent on all observable and unobservable characteristics, except for the receipt of treatment, up to statistical sampling error. As a result, randomization facilitates the unbiased estimation of the average treatment effect.

School Recruitment

ANet's program is a school-wide initiative. Therefore, schools – not individual principals or teachers – were recruited into the i3 evaluation. In the fall of 2010, CEPR and ANet undertook the recruitment of schools that were willing to accept the study conditions in exchange for receiving subsidized ANet services during the 2011-12 and 2012-13 school years (i.e., treatment schools) or at the conclusion of the two-year implementation period (i.e., control schools).[1] Schools from Boston, Chelsea, and Springfield (MA), Jefferson Parish (LA), and Chicago (IL) were invited to apply. ANet chose these locations because of pre-existing relationships with these districts, as well as the goal of expanding their network within these geographic areas.

In order to assess their readiness to implement the program, each school completed an ANet-developed "screener" survey that assessed its readiness to partner with the program. The screener included questions about conditions that ANet felt were related to a school's ability to engage in data-based instructional practices: whether data use was prioritized by district and school leaders, the presence of or willingness to create a dedicated data leadership team, a standards-based curriculum and aligned curricular

[1] A second wave of schools, discussed below, began the study in the fall of 2012, with treatment schools receiving services in 2012-13 and 2013-14.

scope and sequence, and dedicated time for data meetings. Schools' responses were scored by ANet staff members according to a rubric (appendix C). All schools that expressed interest in participating in the study were determined to be ready to implement ANet and none were screened out.

The research team at CEPR set a goal of recruiting 120 schools based on calculations of statistical power for the purpose of detecting a small effect on student achievement. In total, 101 schools were recruited to participate in the expansion of ANet's data-based instructional program. Because the initial recruitment efforts fell short of the goal, ANet recruited a second wave of schools during the spring of 2012 with the primary purpose of improving statistical power (n = 18). Schools in this second wave began receiving services one year later than the schools in the initial, wave-one sample. All aspects of the design that were applied to the first wave of schools were largely applied to the second wave, including recruitment, screening, randomization, data collection, and data analysis. There were two exceptions: no baseline survey was administered to the wave-two schools and no site visits were made.

These recruitment and screening procedures impact the generalizability of the dissertation findings and the types of inferences that can be made to schools not in the sample. First, the sample generally consists of low-performing, urban, elementary and middle schools. These schools are likely to differ systematically from schools in other settings (e.g., suburban or rural, high performing) in terms of resources and, potentially, leader and teacher quality. Second, because schools were invited to apply and subjected to a screening process, results may only be generalizable to other schools that (1) would

apply for ANet services when given the opportunity to do so at a subsidized rate; (2) are assessed by ANet to have sufficient readiness to implement the intervention; and (3) are willing and able to pay the unsubsidized portion of the ANet fee.[2] Given the prevalence of interim assessment and other instructional data-use practices, and the fact that no schools that applied were screened out of the study, the first two conditions may not meaningfully reduce external validity. However, the current school funding climate makes the third condition a possible limitation, as some schools would not be able allocate even the heavily subsidized costs of partnering with ANet.

School Sample

Because ANet is a school-level intervention, the evaluation employed a cluster randomized trial; assigning schools to treatment or control conditions. It was not appropriate to randomly assign individual teachers due to the dangers of contamination (Raudenbush, 1997). Prior to randomization, matched-pairs of schools were created within each of the five districts. Formally, matching entails the grouping of units with similar values on one or more matching variables so that treatment and control groups are balanced on these characteristics (Shadish, Cook, & Campbell, 2002). Matching variables are selected because of their known correlation with the outcome of interest.

Although matching is often discussed in the context of quasi-experimental designs as a way of reducing selection bias (Rubin, 1973; Rubin, 1974), it has several important applications for randomized designs (Rubin, 2007). In a randomized experiment,

[2] During the study period, the annual cost of ANet services was $30,000. Ninety percent of that was subsidized through the i3 funds, leaving schools to pay about $3,000 per year.

matching can increase the likelihood that pretest means and variances on matching variables, and any others with which they are highly correlated, are similar. This leads to an improvement in the power to detect a treatment effect (Shadish, Cook & Campbell, 2002; Imai, King, & Nall, 2009). Additionally, if there is contamination or attrition within a school, the school and its matched pair can both be excluded from analyses when appropriate. When differential attrition is a threat, this can improve internal validity by helping to maintain equivalence in the treatment and control groups.

The evaluation team at CEPR created matched pairs of accepted schools on: prior achievement (by subject, overall and by grade level), grade span, enrollment (in grades 3-8), and demographics such as percent enrollment by race/ethnicity, and the percentage of students eligible for free or reduced-priced lunch, who are ELL, or who have been identified for special education services (Jencks & Phillips, 1998; Bloom, 2003; Sirin, 2005; Hannaway, 2005; Henderson, Petrosino, Guckenburg, & Hamilton, 2007; 2008). Matches were created using simple blocking procedures. After matching, schools within each pair were randomly assigned to one of two conditions: schools that received ANet services (treatment group) and those that did not (the control group). Both the matching and randomization was done using the blockTools package for the R software (Moore & Schnakenberg, 2013).

Data-based instructional programs like ANet are intended to positively impact student achievement and are in widespread use. To avoid possible disadvantages to teachers and students in the control group, no restrictions were placed on control schools' ability to use or solicit non-ANet assessment services during the two-year evaluation

period. However, control schools were barred from participating in any part of the ANet program until after the two-year evaluation period. Based on data that were collected from educator surveys and interviews with district administrators, all control schools took part in some type of periodic assessment program and educators had access to some type of data-based support. Therefore, the impact of ANet in the treatment schools is compared to alternative practices in the control schools. Because of this, it is possible that any treatment effects are smaller than they would be relative to the absence of the use of data-based instructional practices, including interim assessments.

Prior to randomization of wave-one schools, one school was dropped from the study by the evaluation team because it served an alternative student population. The school still received ANet services in year-one, but is not included in the impact sample and closed prior to year-two. After randomization, but prior to any implementation of the ANet program, a total of 11 schools withdrew from the study due to leadership turnover (e.g., new leader's disinterest or prioritization of other initiatives) or district reorganization (resulting in the loss of discretionary funding to cover the school's share of the cost of services). These schools are excluded from both the student impact sample (in the larger evaluation) and the survey impact sample (in this study). One additional school refused to participate in survey data collection. This school is excluded from the survey impact sample only. With the matched-pair design, these schools (n = 13) and their pairs (n = 13) were dropped from the survey impact sample, resulting in a year-one survey impact sample of 75 schools: 38 treatment and 37 control. The uneven number is

due to one Chelsea "pair" consisting of 3 schools, with two randomized into the treatment group.[3]

Prior to the start of the second year of implementation for wave-one schools (2012-13), nine treatment schools indicated that they would not renew their partnership with ANet. Of these schools, attrition was most often due to a change in leadership with the new leader lacking interest in continuing working with ANet. One school lost discretionary funding to use toward ANet services. Another was put into turnaround status with the entire staff removed, effectively severing ANet's relationship with the school; ANet was not asked to partner with the school in year two. Although these schools were included as partial compliers in the larger evaluation's analysis of year-two student achievement impacts because extant student data exist, no survey data were collected from these schools in year two and, therefore, they are not part of the year-two survey impact sample. Additionally, one control school closed and no survey data are available. The loss of these 10 schools and their matched pairs results in a wave-one, year-two survey impact sample of 55 schools: 28 treatment and 27 control.

Of the 18 wave-two schools recruited from Springfield and Jefferson Parish, one closed, one opted-out of survey administrations, and one dropped out after randomization (but prior to implementation). Consistent with decisions made in wave one, these schools and their pairs are dropped from the all samples (n = 6). No wave-two schools attritted from the sample between the first and second years of treatment. With the addition of

[3] A dummy variable is included to account for this three school "pair" and the differential probability of assignment to the treatment group in this triad.

wave-two schools, the full year-two survey impact sample for this study includes 67 schools (34 treatment and 33 control).

In a RCT, the potential exists for contamination when units – e.g., schools – fail to adhere to their random assignments. Treatment-group schools may fail to implement the ANet program or refuse to participate (i.e., noncompliance) or control-group schools may find ways to take part in ANet (i.e., crossover). ANet opted to provide treatment to two schools that were randomized into the control group due to a prior relationship. This dissertation is focused on the intention-to-treat (ITT) sample. The ITT analyses treat these crossover schools as control schools; i.e., the treatment indicator is a measure of school *assignment*. When there is contamination, the ITT estimates offer an unbiased estimate of the impact of *offering* or *being assigned to* the treatment (Shadish, Cook, & Campbell, 2002; Bloom, 2006).

Analytically, the 67 schools in the survey impact sample are not the true intention-to-treat (ITT) sample, but represent the "modified" ITT sample since year-two teacher survey data is not available for all schools that were initially randomized at baseline. The loss of schools between year one and year two, and their subsequent lack of survey data, has direct implications for internal and external validity. Because this sample does not include schools that declined to work with ANet for a second year, all results are generalizable only to schools that would remain in the program for two years when given the opportunity to do so. Due to the matched-pair method of the initial randomization, however, the analysis provides internally valid estimates of program impacts for such

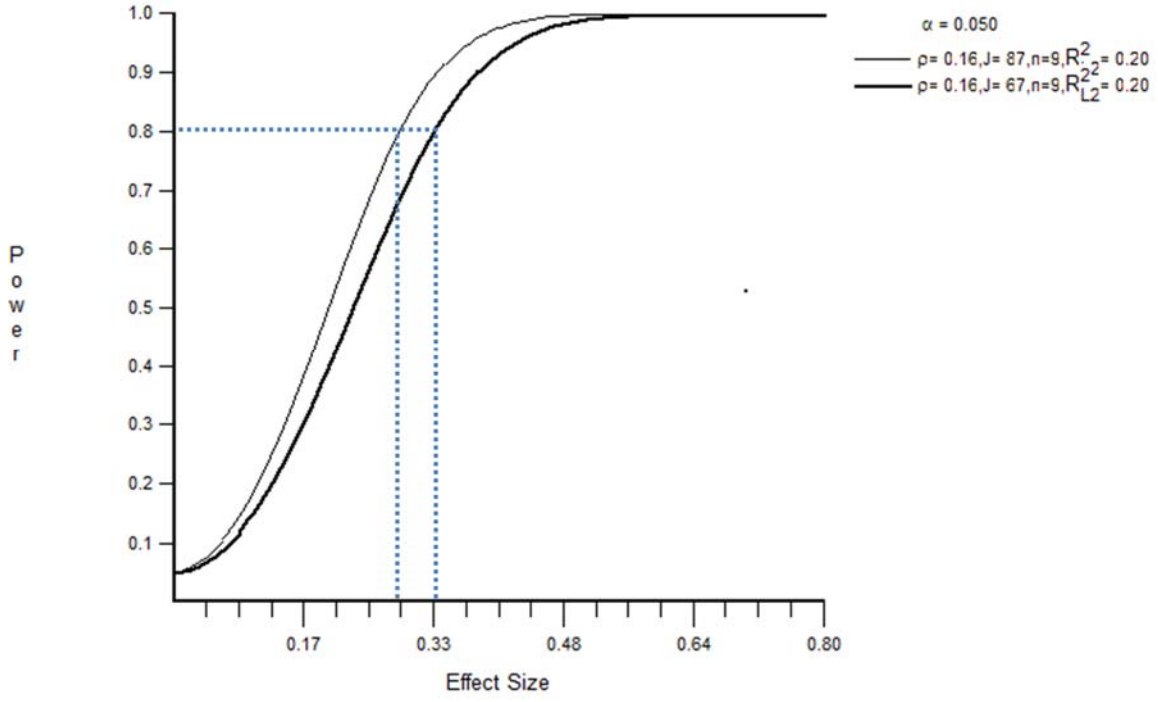
schools under the assumption that the decision to remain in the program is uncorrelated with the outcomes of interest within school pairs.

However, it is possible that a new leader or an existing leader facing budgetary constraints might look to cut what they perceive to be their least effective programs and partnerships, thereby leaving only those treatment schools that are the most motivated or that perceived ANet was having a positive impact. Supplemental analyses tested for bias introduced through attrition: systematic differences between schools in the year-two survey impact sample (n = 67) and 1) schools that were dropped from the study after randomization, but prior to implementation or that refused to take part in survey data collection (n = 32) using information from the Common Core of Data and state performance data and 2) schools that closed or attritted from the sample between year-one and year-two (n = 20) using year-one survey data. Results suggest that differences in these samples are small and could be due to chance. There were no concerns regarding the effects of school attrition on the analyses in this study (see appendix A).

Given the less than desirable recruitment and attrition characteristics of the sample, an a posteriori power analysis was conducted to determine the minimally detectable effect size for the outcomes of interest in this study. This power analysis includes the 67 year-two survey impact sample schools ($J = 67$) and specifies an average of nine responding teachers per school (n = 9). Though there is scant evidence, the existing research on the impact of data-driven interventions on teachers' assessment knowledge, data use, and instructional practices indicates that the intraclass correlation ranges from about 7 percent to 27 percent (Faria, et al., 2012; Randel, et al., 2011). This a

posteriori power analysis uses an average of $\rho = 0.16$, the average ICC across teacher practice outcomes in the teacher sample. A final, and likely very conservative, assumption was made that teachers' baseline practices would explain about 20 percent of the variance in teachers' year-two practices ($R^2 = 0.20$).

Figure 3.1. A Posteriori Power Analysis for Teacher Outcomes



With these parameters, and at a level of power of .80 – or an 80 percent chance of observing a treatment effect when it occurs – the minimally detectable effect size for teacher practices in year two (compared to year one) increases from 0.29 to 0.33 due to loss of schools (figure 3.1).[4] An effect of this size is within the range of effects found in

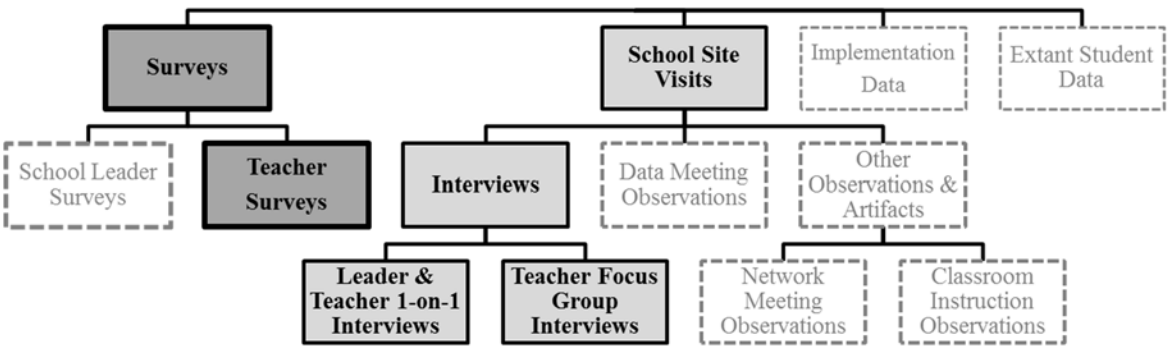
[4] Due to baseline survey nonresponse, the primary quantitative models will not control for teachers' pretest scores, however. The removal of this parameter ($R^2 = 0.20$) only increases the MDES from 0.31 to 0.35 in the year-two survey impact sample of 67 schools.

the existing research and is likely to represent an intermediate impact that is large enough to affect student outcomes (Hill, Beisiegel, & Jacobs, 2013).

DATA COLLECTION & SAMPLES

The main source of data for this study come from the teacher surveys that were administered in treatment and control schools. These survey data are supplemented by qualitative, one-on-one and group interviews with educators in a subset of treatment schools. Figure 3.2 shows the data collected for the larger i3 evaluation. The darker shaded boxes represent the quantitative data and lighter shaded boxes represent the qualitative data that are used in this study. Data sources that were collected as part of the larger evaluation, but are not utilized in this study, are unshaded.

Figure 3.2. Dissertation Data Sources



Quantitative Data

As part of the larger project, surveys were developed to inform the evaluation purposes and research questions. The use of surveys in the larger evaluation allowed for the collection of a wealth of data on attitudes, beliefs, and practices from leaders and teachers in the i3 sample of treatment and control schools. For this study, teacher surveys

provide the main source of data on individual-level mediators (attitudes, confidence) and outcomes (data and instructional practices). Due to school leader survey nonresponse and because there was at least one responding teacher in all 67 year-two survey impact sample schools, teacher-level data are also aggregated within schools to generate school-level indicators of culture and perceived instructional leadership.

School leaders and teachers in treatment and control schools were sent surveys at baseline and again at the end of year one and year two.[5] The surveys were developed by CEPR using several strategies. First, items were included to answer the i3 evaluation questions related to changes in school leaders' and teachers' data-based practices. For purposes of measuring implementation fidelity, items were included to capture variation in model implementation across treatment schools. Finally, a review of the literature and surveys of similar topics was undertaken to ensure that the surveys included items of relevance to the field. CEPR administered the surveys through an online survey platform.

The year-one surveys were kept largely the same as those administered at baseline, save for the addition of questions pertaining to issues or ideas that emerged during year-one site visits. In preparation for this dissertation research, efforts were made to improve the year-two measures of school culture, instructional leadership, teacher attitudes and confidence, and teachers' data-based instructional practices. In some cases, items were added or deleted after a thorough review of the relevant literature on each of

[5] Baseline surveys were not administered to teachers in seven wave-one schools or any wave-two schools. See the section "Quantitative Sample," below, for an explanation.

the constructs, as well as an examination of item and scale reliabilities and component structures.[6]

Revisions to the year-two surveys also included adding items based on early findings from year-one surveys and site visits. Across all surveys, major themes remained largely the same. School leaders were asked about the culture of their school and their attitudes towards data use; the presence of data-based instructional programs and their implementation; and background information on school leadership. Teacher surveys focused on attitudes towards, and use of, data in the classroom; awareness, understanding, and satisfaction with data-based instructional programs, and their implementation; their school culture; and teacher background. A set of ANet-specific items was directed to the treatment group respondents to measure fidelity of implementation, as well as their satisfaction with various ANet program components.

Quantitative Sample

Surveys are typically undertaken with a sample of respondents whose answers are used to generalize to the larger population. However, because of the nature of the larger evaluation, the goal was to survey the universe of eligible educators in all treatment and control schools. Seven wave-one control schools were not fully surveyed in the baseline year due to confusion over their participation in the evaluation (the two crossover schools) and failure to secure necessary contact information to administer the baseline teacher survey. No baseline survey was administered to leaders and teachers in wave-two

[6] These revisions are beyond the scope of the dissertation. However, scale items and characteristics are reviewed in chapter four.

schools. The lack of baseline teacher survey data for these schools presented some issues for analysis. The implications are discussed later in this chapter in the section titled "Quantitative Analysis."

School Leaders. The target population of school leaders included the principal of any i3 school included in the survey impact sample. If a school principal was unavailable to complete a survey due to a leave of absence or retirement, the interim or assistant principal was surveyed as his or her representative. A total of 62 school-leader survey responses were received from the 75 schools in the baseline analytic sample; 27 control schools and 35 treatment schools. This represents a combined response rate of 83 percent (73 percent control-school leader and 92 percent treatment-school leader response rate) (table 3.1).

Table 3.1. School Leader Survey Response Rates (Percentages), by Survey Year and Treatment Assignment

	Baseline[1] (W1)			Year Two (Both Waves)		
	Total	Treatment	Control	Total	Treatment	Control
Overall Response Rate	83	92	73	90	97	82
Network Response Rate						
Eastern MA	90	93	86	92	100	83
Western MA	88	100	75	92	83	100
Chicago	65	80	50	67	100	33
Jefferson Parish	89	100	78	92	100	83

[1] No baseline survey was administered to the wave-two schools.
Note: Leader response rates were calculated based on the requirements that the respondent be: from a survey impact school (BL n = 75; Y2 n = 67), was the principal or a stand-in such as assistant principal or lead teacher, gave consent, and responded to at least some portion of the survey beyond the consent items. Treatment and control groups are determined by group assignment.

In year-two, responses were received from 60 school leaders of the remaining 67 survey impact sample schools (a 90 percent response rate). Response rates were poorest among control-school leaders in Chicago.

Teachers. The target population of teachers included individuals in the i3 schools with an assignment of mathematics, English language arts (including reading and writing), or general elementary in one or more of grades 3 through 8. The base-year teacher universe was generated from lists of school staff provided by treatment and control schools, including their teaching assignments. When schools failed to respond, rosters were found on school or district websites and other public sources. In these schools especially, teacher assignments were often unknown or outdated. In year one and year two, the process was repeated; however, treatment-school rosters were supplemented by rosters of participating leaders and teachers provided by ANet.

Several issues arose as a consequence of constructing the target teacher frame in this manner. First, the rosters and, therefore, survey responses included some teachers who were not in the teacher target population (i.e., out-of-scope), but had been included in the population frame erroneously. Second, the evaluation team at CEPR recognized that supplementing the school rosters with ANet rosters for the treatment group might have introduced coverage bias (discussed below) by better identifying the target population of teachers in treatment schools. Finally, detailed roster information – including the grade level and subject area – was not available for all teachers. These issues have implications for the calculation of teacher response rates; therefore, teacher response rates have been calculated in a specific way.

For each survey year, teacher respondents were categorized as "in-scope" or "out-of-scope" based on their responses to three survey questions: (1) the grades they taught, (2) the subjects they taught, and (3) how many hours of ELA or mathematics instruction they reported. To be in-scope, a teacher must have reported teaching at least one of grades 3 through 8 *and* either ELA, math, or general elementary, or reported some amount of instruction in ELA or math. All other teachers were considered out-of-scope. This definition did include some special education teachers, gifted and talented teachers, English language development teachers, coaches, or other teachers or teacher leaders in grades 3 through 8 as long as they reported some amount of math or English language arts instructional time. On examination, most of the out-of-scope teachers in year-two (n = 166) only taught science or social studies, only taught in grades K-2, held a primary role as a coach, specialist, or support staff that reported no direct instructional time, or were an administrator who was erroneously sent the teacher survey.

Having defined the in-scope respondents, the challenge became identifying the in-scope nonrespondents when their teaching assignments were not always known. As a result, response rates for teachers were estimated two ways. Both of the estimated teacher response rates were calculated as follows:

$$response\ rate = \frac{\#\ \text{of in-scope respondents}}{(\#\ \text{of in-scope respondents}) + (\boldsymbol{est.\#\ of\ in\text{-}scope\ nonrespondents})} \qquad (3.1)$$

In both calculations, the numerator for the year-two response rates includes 616 in-scope teachers. The difference in the methods is the way in which the number of in-scope nonrespondents is estimated.

First, a conservative estimate is provided that assumes all nonrespondents were in-scope (table 3.2). Second, an adjusted response rate is provided that estimates an in-scope nonrespondent count based on the proportion of in-scope respondents (table 3.3). This proportion was generated by examining whether there were differences in the proportions of in-scope and out-of-scope responding teachers by (1) treatment assignment or (2) network. Although the proportions of respondents excluded as out-of-scope did not differ by treatment assignment, there were statistically significant differences by network. Therefore, the estimated number of in-scope nonrespondents was calculated based on the proportion of in-scope respondents in each network. Since some respondents were known to be out-of-scope, it is also plausible to assume that some nonrespondents were out-of-scope. Thus, it is likely that the second approach – the adjusted response rates – provides a better estimate of the actual response rates of the target teacher population.

Based on the conservative estimate, there was a combined baseline teacher response rate of 63 percent; 64 percent in control schools and 62 percent in treatment schools (table 3.2). When broken out by network, the combined response rates ranged from a high of 79 percent in Jefferson Parish to a low of 39 percent in Chicago. With the exception of western Massachusetts, responses rates were lower in control schools. For year-two, the overall response rate was 74 percent; 65 percent in control schools and 82 percent in treatment schools and. When broken out by network, the combined response rates ranged from a high of 86 percent in western Massachusetts to a low of 55 percent in Chicago. With the exception of eastern Massachusetts, responses rates were lower in control schools.

Table 3.2. Unadjusted, In-Scope Teacher Survey Response Rates (Percentages), by Survey Year and Treatment Assignment

	Baseline[1] (W1)			Year Two (Both Waves)		
	Total	Treatment	Control	Total	Treatment	Control
Overall Response Rate	63	62	64	74	82	65
Network Response Rate						
Eastern MA	65	66	63	77	75	81
Western MA	62	52	73	86	95	78
Chicago	39	42	30	55	75	37
Jefferson Parish	79	81	77	68	88	47

[1] No baseline survey was administered to the wave-two schools.

Note: Teachers were considered a respondent if they were: from an impact school (BL n = 75; Y2 n = 67), gave consent, responded to at least some portion of the survey beyond the consent items, and were "in-scope." An in-scope teacher taught some amount of either ELA or math instruction in grades 3-8. The denominator includes all in-scope respondents plus all nonrespondents who were sent a survey. Baseline response rates do not take into account teachers in schools that were not surveyed. Treatment and control groups are determined by group assignment.

When looking at the adjusted estimates, the combined baseline teacher response rate was 67 percent; 68 percent in control schools and 66 percent in treatment schools and (table 3.3). Again, these rates are higher because an estimated number of out-of-scope nonrespondents are removed from the denominator. When broken out by network, the combined response rates ranged from a high of 81 percent in Jefferson Parish to a low of 47 percent in Chicago. With the exception of western Massachusetts, responses rates were lower in control schools. For year-two, the overall response rate was 78 percent; 70 percent in control schools and 85 percent in treatment schools. When broken out by network, the combined response rates ranged from a high of 88 percent in western Massachusetts to a low of 65 percent in Chicago. With the exception of eastern Massachusetts, responses rates were lower in control schools.

Table 3.3. Adjusted, In-Scope Teacher Survey Response Rates (Percentages), by Survey Year and Treatment Assignment

	Baseline[1] (W1)			Year Two (Both Waves)		
	Total	Treatment	Control	Total	Treatment	Control
Overall Response Rate	67	66	68	78	85	70
Network Response Rate						
Eastern MA	67	68	64	81	79	84
Western MA	66	57	76	88	96	81
Chicago	47	50	38	65	82	47
Jefferson Parish	81	83	79	73	90	52

[1] No baseline survey was administered to the wave-two schools.
Note: Teachers were considered a respondent if they were: from an impact school (BL n = 75; Y2 n = 67), gave consent, responded to at least some portion of the survey beyond the consent items, and were "in-scope." An in-scope teacher taught some amount of either ELA or math instruction in grades 3-8. The denominator is adjusted to account for the likelihood that some nonrespondents were out-of-scope. Baseline response rates do not take into account teachers in schools that were not surveyed. Treatment and control group response rates are based on their group assignment.

The year-two sample of 616 teachers includes all in-scope teachers in the 67 survey impact sample schools during the second year of the study. This includes teachers who began teaching at an impact sample school or moved to an in-scope assignment in year two. The year-two sample includes at least one teacher from each of the 67 year-two survey sample schools. There is only one in-scope teacher in 3 of the 67 schools; all other schools had 3 or more responding, in-scope teachers. Overall, the range is from 1 to 23 teachers, with an average of 9.2 responding, in-scope teachers per school. As expected, due to survey non-administration or nonresponse, or teacher turnover, not all of these teachers took the baseline survey. In fact, only about 44 percent (n = 273) match to a baseline survey.

Sources of Error. Although the surveys administered as part of the ANet evaluation were targeted to all leaders and in-scope teachers in the i3 schools (i.e., a

census), survey data collection is still subject to some of the same concerns that would be present with sampling; issues of coverage, nonresponse, and measurement error (Fowler, 2009). Coverage error occurs when there is a mismatch between the target population and the frame population (Couper, 2000). In defining the target population for the evaluation, coverage error would occur if the ANet-school rosters and control-school staff lists systematically differed from the target population overall or between treatment and control schools. For example, the ANet rosters often included out-of-scope 2nd-grade teachers who were participating in an ANet pilot program, while the control-school rosters did not. To minimize these issues, the research team at CEPR compared the rosters or lists that were received from ANet or control schools to publically available information on recently updated school or district websites. If anything, over-coverage (i.e., surveying some out-of-scope teachers) of the target population may be more of an issue than under-coverage. However, while the proportion of out-of-scope teachers varied by network, it did not differ in treatment and control groups.

Nonresponse error is related to both the rate of responses and their representativeness of the population. When nonresponse is random, the implications are mainly an issue of statistical power. However, a perennial concern in survey research is nonresponse *bias* or systematic differences between respondents and nonrespondents that result in biased estimates of the outcome (Fowler, 2009). For example, nonresponding teachers may also be those who engaged in data-based instructional practices less frequently.

In addition, teacher-level data were used to generate school-mean measures of school leadership and culture due to the high rate of school-leader nonresponse. Limitations in the tracking of the teacher sample and their survey responses made it difficult to assess the teacher survey response rate by school. It is assumed that in at least some schools, teacher response rates were less than 100 percent. In these cases, generating a school-mean composite score from individual teacher data assumes that the sample of responding teachers is a random sample of the target population of in-scope teachers. This is unlikely and, therefore, an unknown amount of bias may have been introduced. Nonresponse bias can be difficult to detect without information on the nonrespondents. Efforts to improve response rates in year two were undertaken to minimize nonresponse; however, low response rates in control schools in some networks may introduce bias.

Finally, survey researchers must consider measurement error. In this study, school culture, instructional leadership, and teacher characteristics and practices are characterized by survey scales. In classical test theory, a respondent's observed score on a scale is a function of his or her true score and measurement error. Measurement error in the survey scales can attenuate bivariate correlations with the outcome, introduce bias in multiple regression estimates, and reduce statistical power (Shadish, Cook, & Campbell, 2002; DeVellis, 2003; Ree & Carretta, 2006). Therefore, year-two survey items and scales were designed to improve validity and minimize measurement error by designing reliable estimates of the focal measures (Fowler, 2009). Descriptive statistics for each scale are presented in chapter four.

Qualitative Data

Where survey research offers breadth, data collected through interviews and observations provides depth. Alternatively, "qualitative research is concerned with identifying the presence or absence of something and with determining its nature or distinguishing features (in contrast to quantitative research, which is concerned with measurement)." (Watson-Gegeo, 1988, p. 576) In an effort to collect rich, descriptive information, researchers involved in the larger evaluation conducted site visits each year with three treatment schools in each of the four networks (wave one only). In each of these 12 schools, interviews were scheduled with school leaders, data or instructional leaders, and teachers.[7] In addition, the research team observed data meetings in these schools; meetings where teachers and school leaders came together – usually with their ANet coach – to discuss the students' most recent interim assessment results. A teacher group interview was conducted in each network and open to teachers from all i3 schools.[8]

In year one, researchers visited schools in late January and early February of 2012 during the third assessment cycle and data meeting. Year-two site visits took place during the same period in late winter of 2013. The development of site visit data collection protocols was guided by some initial evaluation goals: to gather detailed data on program implementation and fidelity, adaptations to the model, barriers to and facilitators of implementation, and program effects on educators, students, and school culture. Specifically, the protocols were developed with the purpose of gathering data to test the

[7] An attempt was also made to observe teachers' classroom instruction during their scheduled reteaching. Too few classroom observations were conducted to use this as a data source.
[8] Two group interviews were conducted with teachers in the Eastern Massachusetts network, one each in Boston and Chelsea. No other site visit data collection took place in Chelsea.

logic model and focused on conditions that are hypothesized to be related to the successful use of data to inform instruction: accountability structures, leadership practices, provision of support and resources, the impact of the program on school culture and teacher practices. Like the surveys, revisions to the year-two protocols were made to address early findings from the year-one results. In particular, questions shifted focus from program implementation to perceived program effects, as well as differences between implementation in years one and two.

One-on-one interview protocols were designed to be relatively standardized and scripted, and to be completed in about 40 minutes. In the event that a school leader or teacher had less time, priority questions were identified. This tight script allowed for comparable information to be collected across sites. The group interview, or focus group, was more open ended and had fewer predetermined questions. Prompts were included in the event that teachers were not forthcoming with information. Though this sacrificed some comparability, the open-ended style allowed the team to collect information that might otherwise have been missed.

Qualitative Sample

In year one, site visit schools were selected from the sample of schools participating in the i3 study and assigned to the treatment conditions. In this way, the qualitative sample and data collection are embedded within the quantitative design. Selection was based on the recommendations of ANet coaches who were asked to nominate one school from each geographic network that fit each of these categories: 1)

high implementation fidelity with high coach support, 2) high implementation fidelity with low coach support, and 3) low implementation fidelity with high coach support. This corresponds to a maximum variation approach to site sampling (Bamberger, Rugh, & Mabry, 2006). This was done to ensure that a range of variation in implementation was documented and is intended to enhance inference validity.

For year-two, the goal was to revisit as many of the year-one schools as possible. However, two of the schools had left the study between year one and year two, and another had merged with a non-i3 school. In their place, ANet provided substitute schools chosen for their similar levels of implementation fidelity and coach support. Once schools were selected, ANet coaches were asked to provide the CEPR team with the date of the third assessment cycle data meeting. With these dates in hand, the CEPR team contacted school leaders to confirm the data meeting dates and schedule a time for a school leader interview.

When scheduling the year-two interview with school leaders, the CEPR team requested the names of two to three teachers – preferably from different grade levels or content areas – who could be contacted for an interview. Where possible, the team requested to speak with teachers who had been in the school for both the 2011-12 and 2012-13 school years. In schools that had teacher leader roles (e.g. instructional coach, grade-level or subject-area leads, master teachers), the CEPR team requested interviews with both teacher leaders and classroom teachers. In total, interviews were conducted with 10 out of 12 school leaders (table 3.4). At least one other school leader – e.g., the

assistant principal, data leader, or teacher leader – was interviewed in 8 schools (9 total interviews). At least one teacher was interviewed in 10 schools (16 total interviews).

Table 3.4. Type and Number of Year-Two Qualitative Data Points, by District

District	School	School Leader Interview	Other Leader Interview[1]	Teacher Interview(s)	Teacher Focus Group
Total	12	10	9	16	5
Boston	A	1	n/a	2	Yes
	B	n/a	n/a	1	
	C	1	1	1	
Chelsea		n/a	n/a	n/a	Yes
Chicago	A	1	1	1	
	B	n/a	1	2	Yes
	C	1	1	2	
Jefferson Parish	A	1	1	2	
	B	1	1	n/a	Yes
	C	1	1	2	
Springfield	A	1	n/a	2	
	B	1	2	n/a	Yes
	C	1	n/a	1	

[1] Other leaders were typically the assistant or vice principal, the designated data leader, or department heads/subject area leads who held responsibilities other than a classroom teacher.
Note: School names are masked for confidentiality. Because of the small number of schools in Chelsea, individual schools were not visited. Only a teacher focus group was conducted.

Focus groups were open to teachers from all i3 schools in the district and hosted at a participating school in a central location. In year-two, the Chicago focus group was the least well attended with only two teachers participating from the host school. In contrast, Chelsea was the best represented with at least one teacher from each of the treatment schools. In the other districts, about half of the treatment schools were represented by participants.

MIXED-METHODS ANALYSES

Since this study is a secondary analysis of data collected from the i3 evaluation, the design of the larger evaluation is directly relevant. However, this section details the analyses for this study: the quantitative analyses of the year-two teacher surveys and the incorporation of qualitative year-two site visit data that provide the empirical basis for a mixed methods analysis. The rationales behind the use of mixed methods research are relatively well developed and, despite their varying labels, frameworks for categorizing the types of mixed methods models tend to converge on a core set of designs. However, mixed methods analysis techniques have not yet cohered into generally and widely accepted frameworks (Greene, 2008).

What is consistent across mixed methods analytic frameworks is that, like the overall research design, analyses can be sequential, concurrent, or iterative. Because of its primary focus on quantitative data in an experimental framework, this dissertation employs a sequential approach to analysis within an explanatory framework (Creswell & Plano Clark, 2007, pp. 142-143). This approach uses the qualitative data to provide "follow-up explanations" of the quantitative impact results (Creswell & Plano Clark, 2007, p. 106) and insight into causal processes (Yin, 2009).

To be clear, the development of research questions and the analyses of quantitative and qualitative data in this study were sequential. Qualitative analysis of the site visit data took place once the quantitative analyses were complete and sought to answer new questions raised by the quantitative results. However, both data sources were collected for the larger evaluation before this study began. Given this study's purpose of

exploring the causal impacts of ANet on teachers' data-based instructional practices and potential explanatory pathways, the conclusions rely heavily on the quantitative results. Qualitative results are incorporated during the interpretation phase (chapter six).

Scale Validation

Before addressing the research questions, the reliability and validity of the relevant year-two composite measures (i.e., scales and indices) was established (see chapter four). Validation has been described as the process during which evidence is gathered in order to support the inferences drawn from measurement scales or test scores (Cronbach, 1971). The chief concerns in this study are content validity, statistical conclusion validity, and construct validity. Note that, at times, the term construct is used to characterize measures that are better defined as composites or indices.

Evaluation of content validity was addressed during revisions of the relevant scales and indices during year-two survey revisions. Though the details of the year-two survey revision process are not described in detail in this dissertation, it was one of my primary tasks on the larger evaluation. In brief, I first performed a thorough review of the literature on each of the mediator and outcome measures that are central to this study. I then developed a working definition for each measure that provided a basis for judging whether each corresponding scale or index included items that were relevant and representative of the measure (Messick, 1990). These definitions were used by myself and the evaluation team to assess the degree of construct representation of the existing baseline and year-one survey scales and indices. Where those scales were found to

contain construct-irrelevant items, existing items were removed or revised for the year-two scale or index. When gaps in construct representation were revealed, new items were added to the year-two scale or index. New items were often borrowed (with necessary permissions) from previously validated scales or indices.

Statistical conclusion validity can be affected by characteristics of the design, the measures, or the analyses. As it relates to the scale validation, the concern is measurement error (Shadish, Cook, & Campbell, 2002). Measurement error in independent variables can bias regression estimates and lead to over- or underestimation of the relationship depending on the number of variables in the model (Shadish, Cook, & Campbell, 2002). Measurement error in the outcome does not introduce bias, but can increase the standard error of the regression estimate, and reduce precision, power, and the explanatory power of predictors (Pedhazur, 1997; Raudenbush & Bryk, 2002).

In an effort to reduce measurement error, part of the year-two survey revision process included identifying items within a scale or index that demonstrated poor fit. Fit was judged by determining whether removal of the item(s) improved the reliability of the composite measure and/or the item(s) loaded more strongly on a second principal component. Items demonstrating poor fit were often the same ones judged by the research team to be construct-irrelevant during the process of establishing content validity. These items were revised in an attempt to improve fit or flagged for potential exclusion in future scale construction and analysis.

Construct validity is concerned with what a test or scale purports to measure and its relationship with other constructs (Crocker & Algina, 2008). Construct validity

assumes that an individual's score on a test or scale is an indicator of the construct of interest (Messick, 1990). Since these constructs are unobservable, one must be able to say that the scale is measuring what is intended. This is done by assessing the internal consistency of the scale, as well as the extent to which the relationships between the scale scores and other measures behave as expected (Crocker & Algina, 2008). What is "expected" is determined based on a priori theory of the relationship between the two constructs and their measured correlation. For example, convergent validity would suggest positive correlations between the various school culture measures (Christman, et al., 2009). Discriminant validity, the correlation between measures of dissimilar constructs, is expected to be low.

Scale reliability and construct validity is reported in chapter four. The internal consistency was estimated using Cronbach's alpha: a measure of reliability that estimates the proportion of variance among items within scales that can be attributed to the true score of the underlying construct (Fowler, 2009). For the few scales that do measure a respondent's level of an underlying construct (e.g., teacher confidence), the internal consistency is expected to be high. However, many of the focal survey measures are better defined as indices or composites; essentially, they are ratings of school-level conditions or the frequencies of data-based instructional practices. From the perspective of validity, it is not as critical that the internal consistency of these indices be high. However, a low estimated Cronbach's alpha has implications for the utility of a given index in a regression analysis, specifically as a signal of considerable measurement error. Correlations among scales were calculated for evidence of convergent construct validity.

Quantitative Analysis

The quantitative models rely on year-two survey responses from 616 in-scope teachers. Teacher surveys provide measures of dependent variables (teachers' data use and instructional practices), as well as all hypothesized school- and teacher-level mediators. Each of these measures is constructed as either the individual teacher-mean score (level one) or aggregate teacher-mean score within schools (level two) on a set of Likert-scaled survey items (exhibit 3.1). There are several issues worth reviewing as they relate to the creation of composite measures.

First, many of the item sets are based on Likert-scale response options (e.g., "*agree*" versus "*strongly agree*"). Others item sets are measured on an underlying frequency response scale (e.g., "*rarely*" versus "*occasionally*"). As a result, all measures are ordinally scaled. Although the analysis of ordinal data using interval-based statistical tests (i.e., parametric tests) has long been debated, there is consensus that parametric tests are appropriate when aggregating individual ordinally-scaled items into a composite scale (DeVellis, 2003).

Second, there are some assumptions in the interpretation of results based on school-mean measures constructed from teacher-level responses. Substantively, aggregation of individual teacher responses to a school-level measure assumes that the aggregate score is measuring the same construct as individual responses. Statistically, it assumes that there is invariance in the measurement model at the teacher and school levels (Schweig, 2014). Lack of invariance implies that different factor solutions could be obtained by analyzing the data at the different levels. The threat of violations of the

statistical assumption is moot since this study used school-mean scores generated across all items in an item set; factors were not extracted.

More relevant to this study is the substantive issue. In unreported analyses, the relationships between leaders' responses and school-mean teacher responses (in the same school) were explored where similar scales were available for both respondent types. The results showed very small, positive correlations which suggests that the leader and teacher scales either measure different constructs or that leaders and teachers have very different patterns in practices and perceptions. Ultimately, only two of these school-mean scales, generated from aggregated teacher-level responses, truly replaced a measure that otherwise could have come from the school leader survey: instructional leadership and achievement culture.

Exhibit 3.1 also includes school- and teacher-level covariates used in the models. School-level covariates account for differences associated with study waves and geographic networks, as well as the ex-ante probability of some schools being assigned to treatment (i.e., the Chelsea "pair" of three schools). To address chance differences at baseline and some limitations of the year-two teacher survey sample (e.g., attrition), statistical controls for teaching experience and education are included in all models. With increasing attention to data use in education, new teachers may have more training in, or be more open to, instructional data use. Veteran teachers might be resistant to changing their practices (see Hochberg & Desimone, 2010).

Exhibit 3.1. Summary of Variables Used in the Quantitative Models

Measures	Description
TREATMENT INDICATOR (level two)	0 = school assigned to control, 1 = school assigned to treatment
MEDIATORS	
Level Two	
Instructional leaders' abilities	Teacher-reported ratings of their instructional leader(s) abilities on various tasks: e.g., setting high standards for learning, participating in instructional planning with teachers.
Professional culture	
CPT discussions	Frequency various topics are discussed during common planning time: e.g., student test results, instructional methods/pedagogy, developing lesson plans.
General collegiality	Agreement with statements about colleagues: e.g., teachers feel responsible for helping each other do their best, teachers respect other teachers who take the lead in school improvement.
Achievement culture	Proportion of teachers who say or do various things: have high expectations for students' academic work, reteach to students who weren't successful the first time.
Level One	
Assessment/data attitudes	Agreement with statements about interim assessments: e.g., make teachers feel accountable to other teachers, are a useful instructional tool.
Confidence	
Data use	Confidence in abilities to use data in a range of ways: e.g., measure student progress toward learning goals, adjust teaching plans.
Instructional planning	Confidence in abilities to use a range of strategies for instructional planning: e.g., create differentiated learning plans, use curricular scope and sequence to design lessons.

Exhibit 3.1. Summary of Variables Used in the Quantitative Models, Continued

Measures	Description
OUTCOMES (level one)	
Data Practices	
Data review	Frequency teachers report reviewing data alone or with various others at their school (administrators, teacher teams, or all teachers).
Data use	Frequency teachers report using data in various ways: e.g., measure student progress toward learning goals, adjust teaching plans.
Instructional Practices	
Instructional planning	Frequency teachers report using various planning strategies: e.g., create differentiated learning plans, use curricular scope and sequence to design lessons.
Instructional differentiation	Frequency teachers report teaching to small groups or individual students (a subset of instructional practice items)
Covariates	
Level Two	
District	Set of dummy variables representing Chelsea (MA), Chicago (IL), Jefferson Parish (LA), and Springfield (MA) (reference group = Boston (MA))
Data collection wave	Data collection wave: 0 = wave one, 1 = wave two
Unbalanced pair dummy	0 = not in "uneven" Chelsea triad, 1 = part of Chelsea "uneven" triad
Level One	
Years of teaching experience (total)	Total years of teaching experience (year two)
Highest degree	0 = bachelor's degree, 1 = master's or higher degree (year two)
Baseline measure of corresponding outcome	*Only in appendix models*

Finally, although randomization ensures an unbiased estimate of the treatment effects of ANet, baseline measures of the outcome of interest are often included to improve the precision of estimates and, therefore, the power to detect a treatment effect. Models in appendix B are restricted to the subset of teachers for whom a baseline measure of the outcome is available.

With proper multilevel analytic methods, cluster-randomized trials can provide unbiased estimates of the average treatment effect on individuals, just as do designs that randomize individuals (Donner & Klar, 2004; Bloom, 2006). In view of the nested design of the study, the analysis of survey data uses multilevel regression modeling (MLM). MLM accounts for the clustering of teachers within schools that, if ignored, can lead to violations of the assumptions of homoscedasticity and independence appropriate in an ordinary least squares (OLS) analysis. The estimation procedures used in multilevel modeling also generate standard errors that are not inflated due to nesting and allow for more accurate determinations of significance (Bickel, 2007).

Modeling the relationship between the predictors and outcome as a separate regression for each level-1 unit mitigates aggregation bias. The coefficients of the level-1 equation are jointly modeled at level 2 with the possibility of including level-2 covariates. This method takes account of the correlational structure and appropriately partitions the variance in the outcome at each level. The solution is an overall regression coefficient that provides a more accurate representation of the relationship between the independent variable(s) and the outcome of interest by using the weighted average of the relationships within each group.

Unconditional Models

Before addressing the main research questions, unconditional multilevel models were run for each outcome in order to obtain the unconditional intraclass correlation (ICC). The ICC is a measure of the dependence of observations within groups or the proportion of total variance that is between groups. This model also provided an estimate of the grand mean. The statistical model for a one-way ANOVA with random effects is:

Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$ (3.2)

Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$ (3.3)

In this model, Y_{ij}, represents a composite measure of teachers' data use or instructional practices in year two for teacher *i* in school *j*, and β_{0j} represents the mean school-level practice for school *j*. The random deviation associated with teacher *i* in school *j* is represented by r_{ij}. The grand mean for teacher practices is γ_{00} and u_{oj} is the random deviation from the grand mean associated with school *j*.

This model includes two other parameters: τ_{00}, which represents the between-school variance, and σ^2, which represents the common within-school variance. Using equation 3.4, the ICCs were estimated overall and for the treatment group.

$$\hat{\rho} = \hat{\tau}_{00} / (\hat{\sigma}^2 + \hat{\tau}_{00}) \quad (3.4)$$

The model and the ICC are considered unconditional because no level-1 or level-2 predictors have been included. However, the unconditional models were re-run with only a treatment indicator (γ_{01}) at level-2 to obtain the conditional ICCs before including the hypothesized mediators and any covariates (equation 3.6).

Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$ (3.5)

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(treat)_j + u_{0j}$ (3.6)

Like models 3.2 and 3.3, Y_{ij}, represents a composite measure of teachers' data use or instructional practices in year two for teacher *i* in school *j*, and β_{0j} represents the mean school-level practice for school *j*. The random deviation associated with teacher *i* in school *j* is represented by r_{ij}. The grand mean for teacher practices is γ_{00}, $(treat)_j$ is a dichotomous indicator of whether the school was assigned to treatment (partnering with ANet) or control conditions, therefore, γ_{01}, is the estimate of the treatment effect. Finally, u_{oj} is the random deviation from the grand mean associated with school *j*.

Teacher Practice Impact Models

Research Question One. Are teachers' data use and instructional practices different in ANet (treatment) schools from those in control schools?

Teacher practices are measured by four scales and indices reporting the frequency with which teachers: 1) review data, 2) use data in various ways, 3) use various instructional planning strategies, and 4) use various instructional practices. In equation 3.7, the outcome (Y_{ij}), represents one of the four scales measuring teachers' data use or instructional practices in year two for teacher *i* in school *j* (the model is repeated for each of the four teacher practice outcomes of interest). β_{0j} represents the (adjusted) intercept for mean school-level teacher practice for school *j*. The level-one model also includes a block of year-two teacher demographics $\boldsymbol{X}_{ij}^{(1)}$ (i.e., years of teaching experience, highest degree). Because of the lack of baseline data for over half the sample, the main models do not include any baseline teacher covariates. However, these models are re-run on the

sample of teachers for whom baseline measures of data use and instructional practice ($\boldsymbol{X}_{ij}^{(2)}$) are available in order to account for possible differences in the teacher sample at baseline (appendix B). Finally, r_{ij} represents the random deviation associated with outcome for teacher *i* in school *j.* The full level-one model with covariates is shown in equation 3.7.

Level 1: $$Y_{ij} = \beta_{0j} + \boldsymbol{\beta}'_{1j}\left(\boldsymbol{X}_{ij}^{(1)}\right) + \boldsymbol{\beta}'_{2j}\left(\boldsymbol{X}_{ij}^{(2)}\right) + r_{ij} \quad (3.7)$$

At level-2, the adjusted intercepts from level-1 (β_{0j}) are modeled as a sum of a grand mean (γ_{00}), a school-level treatment assignment indicator ($treat$)$_j$, a block of network dummy variables, $\boldsymbol{W}_j$, representing each school district (with Boston as the reference group), Z_j a dummy variable accounting for the unequal probably of assignment to treatment in one Chelsea triad of schools, T_j a dummy variable for the wave of the study, and u_{0j} which is the school-level random deviation. The coefficient (γ_{01}) on the treatment indicator represents the treatment effect: the difference in the mean outcome of the treatment (($treat$)$_j$ = 1) and control groups (($treat$)$_j$ = 0) after statistically controlling for both level-1 and level-2 covariates. The level-1 slopes are fixed across level-2 schools.

Level 2: $$\beta_{0j} = \gamma_{00} + \gamma_{01}(treat)_j + \boldsymbol{\eta}'_{\mathbf{1}}\boldsymbol{W}_j + \eta'_2 Z_j + \eta'_3 T_j + u_{0j}$$
$$\boldsymbol{\beta}'_{1j} = \boldsymbol{\gamma_{10}}$$
$$\boldsymbol{\beta}'_{2j} = \boldsymbol{\gamma_{20}} \quad (3.8)$$

School and Teacher Mediator Impact Models

Research Question Two. Are levels of school culture, instructional leadership, and teachers' attitudes towards and confidence with data-based practices (hypothesized mediators) different in ANet (treatment) schools from those in control schools?

The second research question asks whether ANet had an impact on the proposed mediators: instructional leadership and school culture, as well as teacher attitudes and confidence related to assessment and data use. Recall that these variables were derived from teacher surveys, but because they are measured both as school-mean teacher responses at the school level (instructional leadership and culture) and teachers' individual responses (attitudes and confidence), the models are run two ways.

School-Level Mediators. School-mean scale scores for measures of instructional leadership and culture were calculated based on individual teacher scale scores within each school. These values are then used in estimating the impact of ANet on school-level mediators using single-level ordinary least squares models:

$$Y_j = \alpha + \beta_1(treat_1) + \beta_2(\boldsymbol{W_J}) + \beta_3(Z_J) + \beta_4(T_J) + \epsilon_i \quad (3.9)$$

where Y_j is school-mean teacher-reported: 1) instructional leader abilities, 2) each of the two measures of professional culture, and 3) achievement culture. The intercept – or grand mean – is represented by α, the treatment effect (β_1) is the difference in the mean outcome of the treatment and control groups after statistically controlling for district ($\boldsymbol{W_j}$), the unequal probability of assignment to treatment in schools in Chelsea (Z_j), and the wave of data collection (T_j). Finally, ϵ_J represents the random deviation.

Teacher-Level Mediators. Models estimating the impact of ANet on the teacher-level mediators are identical to the multilevel models shown in research question one

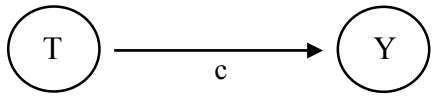
(equations 3.7 and 3.8). The only difference is the use of each of the following teacher-reported measures as the outcome Y_{ij}: 1) attitudes toward data and assessment; and 2) confidence in various a) data-use and b) instructional practices.

Teacher Practice Mediation Models

Research Question Three. Do the hypothesized mediators account for differences in ANet and control-school teachers' data use and instructional practices?

The traditional approach to simple mediation analysis (e.g., with one mediator) was influenced by the work of Baron and Kenny (1986) who proposed a four-step process of establishing mediation effects. In this example, *T* represents the treatment (i.e., assignment to ANet), *Y* represents the outcome, (i.e., teachers' data-based instructional practices), and *M* represents a single, potential mediator. A non-zero correlation between the outcome of interest and the treatment indicator provides an estimate of the total effect of ANet on teacher practices (path *c*, figure 3.3). Though a regression of the outcome on the treatment indicator is often proposed as the first step, a non-zero direct effect is not required for mediation. For example, off-setting directional estimates of the relationships represented by pathways *a* and *b* (figure 3.4) can result in a null direct effect.

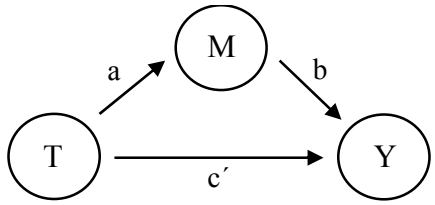
Figure 3.3. Direct Effect



Steps two and three, however, are required. Step two establishes a non-zero correlation between the treatment and the hypothesized mediator (path *a*, figure 3.4). For example, the mediation of the impact of ANet on teacher practices cannot occur without there first being an effect of ANet on the proposed mediator. Step three must also

establish a non-zero correlation between the hypothesized mediator and outcome (path *b*, figure 3.4) (Kenny, 2014).[1] Pathways *a* and *b* represent the indirect effect or the estimate of the mediation effect. If the direct effect of *T* on *Y* is reduced to (near) zero after controlling for *M* (path *c′*, figure 3.4), that is evidence that the treatment effect is completely mediated by *M*. If path *c′* is smaller than *c* after controlling for *M*, but still non-zero, that is evidence of partial mediation of *T* on *Y* by *M*.

Figure 3.4. Mediation Effect



Research question three explores the extent to which hypothesized school- or teacher-level mediators explain the relationship between ANet implementation and teachers' data-based instructional practices. Put another way, research question three asks whether ANet's impact on teachers' data and instructional practices occurred by way of the program's impact on instructional leadership, school culture, or teachers' attitudes toward or confidence with data. Blocks of school- and teacher-level mediators are first tested on their own, then simultaneously, in order to estimate their "effect" controlling for other mediators.

School-Level Models. This model provides an estimate of the upper-level mediation of perceived instructional leadership and school culture on the treatment effect

[1] Structural equation modeling (SEM) approaches can also be used to estimate the various pathways simultaneously rather than independently (Zhao, Lynch Jr., & Chen, 2010).

of ANet on teacher practices. The level-1 equation remains the same as in research question one.

$$\textbf{Level 1}: \quad Y_{ij} = \beta_{0j} + \boldsymbol{\beta}'_{1j}\left(\boldsymbol{X}_{ij}^{(1)}\right) + \boldsymbol{\beta}'_{2j}\left(X_{ij}^{(2)}\right) + r_{ij} \tag{3.10}$$

However, the level-2 model adds a block of school leadership and culture variables, $\boldsymbol{S}_j$, – representing school-mean teacher-reported instructional leadership, professional culture, and achievement culture – in order to explain the variation in level-1 intercepts. The level-1 regression coefficients (slopes) on teacher covariates are fixed across schools.

$$\begin{aligned} \textbf{Level 2}: \quad & \beta_{0j} = \gamma_{00} + \gamma_{01}(treat)_j + \boldsymbol{\eta}'_{\mathbf{1}}\boldsymbol{W}_j + \eta'_2 Z_j + \eta'_3 T_j + \boldsymbol{\eta}'_{\mathbf{4}}\boldsymbol{S}_j + u_{0j} \\ & \boldsymbol{\beta}'_{1j} = \boldsymbol{\gamma_{10}} \\ & \boldsymbol{\beta}'_{2j} = \boldsymbol{\gamma_{20}} \end{aligned} \tag{3.11}$$

Again, γ_{00} represents the grand mean. The coefficient on the treatment indicator (γ_{01}) represents the treatment effect after statistically controlling for the level-two covariates which now include measures of instructional leadership and school culture.

Teacher-Level Models This model provides an estimate of the lower-level mediation of teacher characteristics on the treatment effect of ANet on teacher practices. The level-1 model is the same as specified in research question one, but adds a block of teacher characteristics ($\boldsymbol{X}_{ij}^{(3)}$) representing measures of teacher attitudes and confidence to explain variation in instructional practices among individuals.

$$\textbf{Level 1}: \quad Y_{ij} = \beta_{0j} + \boldsymbol{\beta}'_{1j}\left(\boldsymbol{X}_{ij}^{(1)}\right) + \boldsymbol{\beta}'_{2j}\left(X_{ij}^{(2)}\right) + \boldsymbol{\beta}'_{3j}\left(\boldsymbol{X}_{ij}^{(3)}\right) + r_{ij} \tag{3.12}$$

The level-2 model remains the same as in research questions one and two. The level-1 regression coefficients (slopes) on teacher covariates are constrained across schools:

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(treat)_j + \boldsymbol{\eta_1'W}_j + \eta_2'Z_j + \eta_3'T_j + u_{0j}$

$\boldsymbol{\beta_{1j}' = \gamma_{10}}$

$\boldsymbol{\beta_{2j}' = \gamma_{20}}$

$\boldsymbol{\beta_{3j}' = \gamma_{30}}$ (3.13)

Estimates of the mediating effect of school and teacher characteristics are simply reported as the proportion by which the direct effect (*c*, estimated in RQ1) is reduced after controlling for the mediators of interest (*c´*, estimated in RQ3). In summary, models addressing research question three are repeated with each teacher outcome on which ANet had an impact and with blocks of hypothesized school- and teacher-level mediators first included individually, then simultaneously.

Mediation analysis requires that certain assumptions be met: specifically, that there is 1) no unmeasured confounding of the relationship between the treatment and outcome, 2) no unmeasured confounding of the relationship between the mediator and outcome, 3) no treatment-mediator interaction (treatment-mediator confounding), and 4) no mediator-outcome confounder that is affected by the treatment. There is also an assumption that the treatment, mediator, and outcome are temporally ordered (Valeri & VanderWeele, 2013; Kenny, 2014).These assumptions, in the context of this study, are discussed in chapter 6. Alternative methods of estimating mediation effects and alternative designs for strengthening causal claims are also discussed in chapter 6.

Exploratory Pre-Conditions Model

Research Question Four. Does the effect of ANet on teachers' data use and instructional practices vary by schools' baseline implementation "readiness" ratings?

Based on data from the school readiness survey used by ANet during school recruitment, the 67 year-two schools were ranked on the sum of their subscores across the five most relevant categories to this study – school opt-in, program priority and organization, dedication of leadership, standards and alignment, and scheduling – and sorted into two groups (i.e., high and low).[2] Models from research question one are then re-run, for each readiness group, in order to explore whether the estimates of the impact of ANet on teacher practices vary by group. Although these analyses may be underpowered, patterns in effect sizes provide useful evidence of the moderating effect of initial school readiness on the impacts of ANet on teachers' data-based instructional practices.

Statistical Conclusion Validity

The quantitative analyses have additional implications for statistical conclusion validity (Shadish, Cook, & Campbell, 2002), some of which were addressed in prior sections. In terms of power, school matching and the inclusion of covariates in the analyses should improve the precision of the estimated treatment effects (Bloom, 2006) and ensure that any covariation between the treatment and outcome is easier to detect. Also, recall that the results of the a posteriori power analysis determined that the number of schools remaining in the study in year two should provide a sufficient sample to detect

[2] Details on the construction of the two readiness groups are provided in chapter four.

an effect on teacher practices that is consistent with what prior literature deems necessary to see a subsequent impact on student outcomes (see this chapter, "School Sample"; Hill, Beisiegel, & Jacobs, 2013). Also discussed above, year-two scale revisions were done with goal of maximizing reliability to prevent the attenuation of results. Finally, the use of multilevel regression modeling addresses the need to properly calculate standard errors by accounting for the school-level clustering that is present in the data.

Two other important points must be noted. First, tests of significance are not corrected for simultaneous inference. This is acceptable since analyses explore different outcome domains and are considered exploratory in nature (What Works Clearinghouse, 2014). Second, Hedges' *g* effect sizes are not calculated since all scale variables were standardized prior to analysis. In each analysis sample, pooled (treatment- and control-group) standard deviations – used in calculating Hedges' *g* – are close to one; therefore, effect sizes are generally the same or only slightly different from standardized results (i.e., on a magnitude of a few hundredths of a standard deviation).

Qualitative Analysis

For this study, year-two site visit school leader and teacher interviews (including teacher focus groups) comprised the data source for the qualitative analysis, which was conducted after the quantitative analyses had been completed. The secondary analysis of non-naturalistic qualitative data – e.g., transcripts of interviews and focus groups – implies the analysis of qualitative data for the purpose of answering new or additional research questions (Heaton, 2008). This use of qualitative data is a relatively new

methodology (Leech & Onwuegbuzie, 2008). Though there is little guidance on the topic, secondary analysis of interview data is thought to be feasible when the context of the original study is known (e.g., the setting or sample), and corresponds well or provides some value to the new study (Notz, 2005; van den Berg, 2005).

Given my involvement in the i3 evaluation and knowledge of the design, the secondary use of the larger evaluation's qualitative data for this study is feasible in terms of its context. This study's research questions are well-matched to the larger evaluation's design, setting, and sample. Likewise, there is sufficient overlap in the purpose and measures of interest in the larger evaluation and this study to warrant the secondary use of the interview data. That said, interview protocols did not perfectly correspond to the questions I would have asked had I designed the protocols specifically for this study. These differences, therefore, limit the amount of evidence for certain focal measures in this study. In some cases, focal measures in this study were not explicitly explored during the site visits and are less well represented in this study's findings.

To ensure that the transcripts were coded consistently and took into account my operationalization of the measures of interest in this study, all 40 interview transcripts were re-coded rather than relying on the larger evaluation's already-coded transcripts. Coding for this study took place in two rounds. The purpose of the first round of coding was to ensure that the extant site visit interviews contain sufficient explanatory evidence to support the quantitative results. Therefore, first round coding focused on identifying those portions of the leader and teacher interviews that addressed the focal measures and research questions in this study (exhibit 3.2). For example, identification of data relevant

to research question one – the impact of ANet on teacher practices – focused on coding portions of the interview in which teachers talked about their data use and instructional practices. This step is referred to as open coding where the codes that are applied are relatively literal, identifying portions of transcripts that address particular focal measures or research questions (Leech & Onwuegbuzie, 2008; Hesse-Biber, 2010).

Second round coding entailed both finer-grained coding and analysis. It was informed by a conceptual framework developed from both the ANet logic model and prior research. The goal was to test where the conceptual framework held up or broke down. Operationally, second round coding utilized a constant comparative approach to extract themes, theories, or explanations for the quantitative results (Leech & Onwuegbuzie, 2008). The specific coding schema utilized this framework and was determined after the quantitative analyses were completed, so as to provide descriptive evidence for emerging research questions. For example, coding and analysis in the second round sought confirming or disconfirming evidence of how ANet may work to affect teacher practices through instructional leadership or other hypothesized mediators.

Second round coding and analysis focused on several goals. First, it explored how leaders and teachers described the focal measures and the context around them. Second, similarities and differences in how respondents defined or experienced focal measures were explored: e.g., difference among teachers or between teachers and leaders in aggregate across schools. Finally, a judgment of magnitude or merit was applied to focal measures: e.g., identifying effective or ineffective data use or instructional practices (Saldaña, 2009). First and second round codes are shown in exhibit 3.2.

Exhibit 3.2. Qualitative Codes

First Round Codes	Second Round Codes
General Codes	**General Codes**
Exemplar Quote	Exemplar Quote
School Culture	School Culture (+)
	School Culture (-)
Data Culture	Data Culture (+)
	Data Culture (-)
Hindrance (other)	
Facilitator (other)	
School Characteristics	**School Characteristics**
Professional Culture	Professional Culture (+)
	Professional Culture (-)
Achievement Culture	Achievement Culture (+)
	Achievement Culture (-)
Instructional Leadership	Instructional Leadership (+)
	Instructional Leadership (-)
Teacher Characteristics	**Teacher Characteristics**
Data/Assessment Confidence (Ability)	Data/Assessment Confidence (+)
	Data/Assessment Confidence (-)
Instructional Confidence (Ability)	Instructional Confidence (+)
	Instructional Confidence (-)
Data/Assessment Attitudes & Beliefs	Data/Assessment Attitudes (+)
	Data/Assessment Attitudes (-)
Teacher Outcomes	**Teacher Outcomes**
Data Use (& Analysis)	Data Analysis (+)
	Data Analysis (-)
	Data Use (+)
	Data Use (-)
Instructional Practice (& Planning)	Instructional Planning (+)
	Instructional Planning (-)
	Instructional Practice (+)
	Instructional Practice (–)

Note: A (+) sign indicates a quality such as high, positive, effective; a (-) sign indicates a quality such as low, negative, ineffective.

Validation. As with the quantitative components, the qualitative components of this dissertation required validation. Validity of qualitative findings is especially important given the critique of researcher bias that is often directed at qualitative methods

due to its comparatively more open-ended structure (Johnson, 1997). Johnson (1997) provides a validation framework for qualitative research that includes five aspects: descriptive, interpretive, theoretical, as well as the familiar internal and external validity. Determinations of the degree to which the findings from the qualitative analysis are valid, depends greatly on the research design and analytic process (Saenz, personal communication, February 25, 2015; Gilbert, personal communication, March 20, 2015).

Descriptive validity is concerned with accuracy in the reporting of facts and descriptive information, whereas interpretive validity is concerned with the accuracy of inferences made from the data. To address these validity concerns, member checking – the review of analyses, interpretations, and conclusions by key stakeholders – was used to validate the accuracy of the interpretations from qualitative site visit interviews (Lincoln & Guba, 1985; Bamberger, Rugh, & Mabry, 2006). Specifically, three members of the CEPR evaluation team who engaged in the site visits also reviewed the findings of this study to ensure consistency with the larger evaluation findings: the Primary Investigator, another senior research team member, and the project manager (who had analyzed all year-two site visit data for a summary report to ANet). Committee readers, selected purposefully for their combined expertise in data-based instruction and the use of mixed methods approaches, provided valuable feedback as external debriefers.

Theoretical validity is concerned with the fit between the overarching conceptual framework of the study and its congruence with the results emerging from the research findings (Johnson, 1997). Evidence of theoretical validity often includes predicting patterns in findings, or pattern matching, based on theory and seeing if they hold true. It

also involves acknowledging cases or examples that do not fit the theoretical explanation (Johnson, 1997). A constant comparative approach to analyzing the data viewed the qualitative data in light of cases that confirm, or disconfirm, findings from prior research and the conceptual model. This prevents the study from being overly inclusive of results that fit the model, while ignoring those that do not (Johnson, 1997).

Internal validity in qualitative research is similar to its quantitative counterpart. It is "the degree to which a researcher is justified in concluding that an observed relationship is causal." (Johnson, 1997, p. 287) However, the consensus in the qualitative research community is that causation is an "inappropriate" concept in qualitative research (Maxwell, 2012, p. 655). In this study, the quantitative data provided substantial evidence of the causal *relationships*. However, as discussed above, qualitative data provide evidence of causal *hypotheses* and *theories* (e.g., confirming evidence), while exploring potential rival explanations (i.e., disconfirming evidence). Attention is paid to the degree of agreement between qualitative and quantitative findings as a means for cross-validation of results (Johnson, 1997). These uses are consistent with what Maxwell (2004, 2012) describes as a generative theory of causation in qualitative research, one that explores process and context.

Finally, external validity in qualitative research parallels its quantitative counterpart. It is the degree to which theories or inferences drawn from the evaluation sample of schools can be generalized to other settings. In other words, it asks whether the theories or explanations extracted from a particular study are useful in making sense of other, similar situations (Maxwell, 1992, p. 293). The purposive sampling of site visit

schools ensured that a range of implementation conditions were observed. However, with only about one-third of treatment schools visited, external generalizability may be limited. External validity can also be judged by the extent to which the qualitative findings are consistent across treatment schools and congruent with findings from prior research in other settings.

Meta-Analysis & Meta-Inference Validation

For each of the research questions, survey findings from the quantitative models provides the primary source of evidence, but is supplemented by and elaborated with findings from the site visit interviews with school leaders and teachers. The integration of quantitative and qualitative results at the interpretation phase serves several uses. For example, it: 1) provides explanations of impacts, or the lack thereof, on the teacher practices and key mediators in this study, 2) "tests" or explores the conceptual model and causal linkages (Yin, 2009), and 3) offers evidence of why ANet is more effective at changing teachers practices in some contexts than others.

Validation. The discussion of validation of the findings has focused separately on the quantitative and qualitative research components. The evaluation of the validity of the quantitative components of dissertation research has been outlined in discussions of internal, external, construct, and statistical conclusion validity. The discussion of the validity of the qualitative component has included descriptive, interpretive, theoretical, internal, and external validity. Validation within the mixed methods framework is equally important. Some advocate for a single framework that addresses the validity of the

separate components in addition to the validity of meta-inferences from the combined methods, also known as legitimation or inference quality (Onwuegbuzie & Johnson, 2004; Teddlie & Tashakkori, 2003). However, this study addresses the quantitative and qualitative components' validity individually; an acceptable approach because it relies on each fields' established practices (Creswell & Plano Clark, 2007; O'Cathain, 2010).

CHAPTER FOUR: QUANTITATIVE ANALYSES & RESULTS

This chapter provides a discussion of unit- and item-level missing data, as well as descriptive statistics for the teacher sample and each of the scales and indices measuring school- and teacher-level mediators and teacher practice outcomes. The validity of the scales and indices as indicators of the focal measures are also explored. Finally, the quantitative results for the four main research questions are presented.

MISSING DATA ANALYSIS

Missing data can present several difficulties, reducing the analytical sample size and, consequently, power, as well as potentially introducing bias in the results when teachers with and without missing data differ systematically on their responses to focal measures of interest. The latter can be especially difficult, if not impossible, to detect. However, a thorough analysis of missing data precedes the main analyses in this study in an attempt to address some of these concerns.

Missing Data at Level Two (Schools)

The first step in exploring missing data was to examine unit-level missingness of school leader survey data. The school-leader survey would have been a potential source for measuring hypothesized school-level mediators of interest in this study. However, year-two surveys were completed by school leaders in 60 of the 67 schools in the sample, and each of the 7 nonrespondents were from different matched pairs. If school-leader data were used to measure hypothesized school-level mediators, all analyses should exclude

the responding matched-pair school leader for each of the nonresponding school leaders in an attempt to maintain internal validity through treatment- and control-school balance. This would further reduce the available year-two school leader sample to 53, making the use of the school leader data for level-two (school-level) measures of leadership and culture problematic. Not only would it reduce the year-two school leader sample, but any multilevel analysis would also eliminate teachers in these 14 schools. This would reduce the sample of year-two teachers by 20 percent (from 616 to 492) and have implications for validity and statistical power.

This reduction in sample size was the justification for using aggregate year-two teacher survey data to generate the school-mean school-level mediators of interest in this study. This strategy ensures that there are valid values for each of the hypothesized school-level mediators for all 67 schools (i.e., no missing level-two data). The issues associated with this method were discussed in chapter three and judged to be outweighed by the benefits. Values for all other cluster-level (level-two) covariates – the treatment indicator and indicators for member district, data collection wave, and schools in the Chelsea three-school "pair" – are known for all schools.

Missing Data at Level One (Teachers)

All year-two measures of mediator and outcome scales or indices were calculated for teachers who responded to 80 percent or more of the items in the set. This rule was set after examining the distribution of missing data. Across the items sets, respondents were generally missing either one item in a set or all items in a set. Therefore, choosing an 80

percent rule versus a 50 percent rule made very little difference.[3] As a result, the choice was made to use the higher standard – at least 80 percent nonmissing responses – in the calculation of scale scores. An exception was made for any scale or index comprised of four or fewer items; in these cases, teachers must have responded to all items to generate a nonmissing composite score. In cases where these conditions were not met, the mean scale or index score is missing. Using this 80 percent rule, the number of teachers with missing scale or index values ranges from 0 to 44 (at most, about 7 percent of the teacher sample).

Imputation of missing items or scale/index scores was not performed for two reasons. First, most cases of missing scale or index scores are missing because teachers skipped every individual item in the set used to construct the scale or index. This would make imputation more difficult and would require modeling a mean scale or index score using other survey information. Second, both the proportions of teachers who are missing on any given teacher-level scale or index, or from any regression model, is less than 10 percent. When the proportion of missing data for a particular variable or scale is less than 10 percent and the number of cases with no missing data is large enough to support the selected analysis technique, missing data can generally be ignored as it is unlikely to introduce bias into the study (Hair, Black, Babin, & Anderson, 2010; Dong & Peng, 2013).

[3] For 6 of 11 scales, there is no difference in the number of cases for which a valid composite scale measure is generated using the 50 versus 80 percent rules. The remaining 5 scales would gain between 1 and 4 additional teachers (out of the possible 616) using a 50 percent rule.

A significant proportion of teachers are missing baseline data and, therefore, the main models in the study do not control for baseline measures of the outcomes of interest. Models in appendix B show results for the subset of teachers with nonmissing baseline data. Because other level-one covariates are measured in year two, the proportion of missing data is much smaller.[4] Only 2 percent of teachers are missing data on their highest degree and years of teaching experience; these data are also not imputed.

In sum, the use of teacher-level responses to generate aggregate school-level measures and the universal availability of school- and district-level covariates ensured that there was no missing data at level two. While there is missing data at level one, no imputation was performed since the rate of missingness was unlikely to introduce bias in the results. In a given analysis, individual teachers with missing values for teacher-level variables specified in the model are dropped.

YEAR-TWO TEACHER SAMPLE CHARACTERISTICS

In the year-two sample, teachers reported an average of 12.6 total years of experience with 10.9 years in the current district, 7.2 in the current school, and 6.8 in the current grade and subject (table 4.1). Treatment and control differences in mean years of experience appear to be largest in teachers' current grade and subject (7.0 and 6.4 years, respectively). However, the difference is not statistically significant.

[4] The use of post-treatment covariates is acceptable as long as they are expected to be unrelated to the treatment.

Table 4.1. Year-Two Teacher Descriptive Statistics, Overall and by Treatment Assignment

	Total			Treatment			Control		
	n	Mean	sd	n	Mean	sd	n	Mean	sd
Teaching experience	609	12.6	9.08	367	12.7	9.28	242	12.4	8.79
In district	608	10.9	8.68	367	10.8	8.72	241	11.0	8.63
In school	608	7.2	6.82	369	7.2	7.16	239	7.1	6.27
In grade & subject	611	6.8	6.73	369	7.0	7.10	242	6.4	6.11
	Total			Treatment			Control		
	n	%		n	%		n	%	
Subject area[1]									
ELA	269	43.7		161	43.2		108	44.4	
Mathematics	215	34.9		132	35.4		83	34.2	
General elem.	175	28.4		109	29.2		66	27.2	
Special ed.	117	19.0		61	16.4		56	23.0	
ESL/ELD	63	10.2		36	9.7		27	11.1	
Other	53	8.6		31	8.3		22	9.1	
Grade level									
Grades 3-5 only	475	77.1		288	77.2		187	77.0	
Grades 6-8 only	126	20.5		76	20.4		50	20.6	
Both levels	15	2.4		9	2.4		6	2.5	
Gender									
Female	535	88.4		331	89.7		204	86.4	
Male	70	11.6		38	10.3		32	13.6	
Race/ethnicity									
African American	93	15.4		57	15.7		36	14.9	
Hispanic	33	5.5		17	4.7		16	6.6	
White	442	73.2		270	74.4		172	71.4	
Other	36	6.0		19	5.2		17	7.1	
Highest degree									
Bachelors	187	30.7		121	32.8		66	27.4	
Masters	416	68.2		245	66.4		171	71.0	
Doctorate	7	1.2		3	0.8		4	1.7	
Alt. certification									
Yes	124	20.3		69	18.6		55	22.8	
No	488	79.7		302	81.4		186	77.2	

Notes: ELA=English-language arts. ESL/ELD=English as a second language/English language development.
[1] Subject area categories are not mutually exclusive; totals sum to greater than 616 teachers and 100 percent. All other estimates are calculated from valid responses in the sample of 616 in-scope teachers.

Since an in-scope teacher is defined as one reporting some amount of instruction in math or reading in grades 3 through 8, it is not surprising to find the vast majority of teachers taught one or both of these subjects or reported a general elementary assignment (43.7 percent ELA, 34.9 percent math, and 28.4 percent general elementary).[5] Most teachers instruct students in grades 3 through 5 only (77.1 percent), but 20.5 percent teach only middle grades (grades 6 through 8) and 2.4 percent teach at least one grade at each level. As with experience, the distribution of teachers by subject and grade-level assignment does not differ much between treatment and control schools.

In terms of demographics, the majority of teachers are female (88.4 percent) and white (73.2 percent). Although these proportions are slightly higher in treatment schools, they are not statistically significantly different from control schools. Finally, in the combined sample, most teachers hold a master's as their highest degree earned (68.2 percent) and entered teaching through a traditional certification route (79.7 percent). Compared to their treatment-school counterparts, the percentage of teachers in control schools with a master's or doctorate degree is slightly higher (72.7 versus 67.2 percent, respectively) and they were slightly more likely to have entered teaching through an alternative certification (22.8 versus 18.6 percent, respectively). However, as with the other measures, no statistically significant treatment-control differences in the distribution of education and certification were found.

[5] Teachers could report multiple subject areas and grade levels. The descriptive statistics for subject areas are not reported in mutually-exclusive categories (i.e., they sum to greater than 100 percent).

YEAR-TWO SURVEY SCALE CHARACTERISTICS

Organized by focal measure, this section provides a list of survey items that make up each scale or index, as well a summary of the descriptive statistics of the scale or index (e.g., means, standard deviations, and number of responses for each scale).[6] The descriptive results include Cronbach's alpha, a measure of reliability that is typically used to estimate the proportion of variance that can be attributed to a common source (Fowler, 2009). Reliability is typically applied to scales where higher values indicate better measures of the underlying trait (i.e., greater internal consistency, less measurement error). As it relates to internal consistency, Cronbach's alpha is not necessarily expected to be as high for items sets that are better described as indices: inventories of school conditions or teacher practices. Instead, reliability provides a measure of the utility of the scales and indices in the analysis models. Specifically, it serves an indicator of the amount of measurement error that is present. A full summary of the scale descriptive statistics are reported in table 4.2 at the end of this section.

Instructional Leadership (School Level)

The school-mean teacher-reported rating of school leaders' instructional abilities on a range of practices is the primary measure of instructional leadership in this dissertation. The scale consists of nine items that cover a range of activities for which an instructional leader might be responsible (exhibit 4.1). Teachers provide a rating of their

[6] Descriptive statistics for school-level mediators were generated from the aggregate school-level dataset; i.e., using school-mean teacher responses for relevant items (n = 67 schools). Descriptive statistics for teacher-level mediators were generated at the individual level.

school leader's ability based on a 5-point scale from '*very poor*' to '*excellent*.' The average school-mean rating of instructional leaders' abilities is 3.9 (*SD* = 0.55) (table 4.2). The item set has very high reliability both overall (α = 0.97) and by group (α = 0.99 treatment; α = 0.96 control).

Exhibit 4.1. Year-Two Teacher-Reported Instructional Leaders' Abilities Items

Items	Scale
Thinking about your school's instructional leader(s), how would you rate their ability to do each of the following activities?	**Very Poor - Excellent** (5 pt)
a. Communicate a clear vision for teaching and learning for this school.	
b. Set grade or classroom level instructional goals.	
c. Track students' academic progress toward school goals.	
d. Monitor the quality of teaching at this school.	
e. Set high standards for student learning.	
f. Support teachers in implementing what they have learned in professional development.	
g. Participate in instructional planning with teachers.	
h. Institute concrete practices and procedures that encourage the use of student test data by teachers to improve student learning.	
i. Provide actionable feedback on classroom instructional plans.	

Professional Culture (School Level)

Two sets of teacher survey items are used to measure teacher professional culture. The first set consists of 12 items that ask about the frequency with which various types of collegial conversations occur during common planning time and is measured on a five-point scale ('*never'* to '*almost always'*) (exhibit 4.2). Two items on the frequency of discussion of non-academic issues (behavior outside the classroom and logistical issues) are removed due to poor fit.[7] The remaining 10 items ($\bar{x}$ = 3.7, *SD* = 0.40) show high reliability (α = 0.92 overall; 0.94 treatment; 0.91 control) (table 4.2).

[7] These two items load more highly on a second component.

The second item set asks about agreement with a set of conditions related to collegiality among the teaching staff (exhibit 4.3). Items are measured on a five-point scale ranging from '*strongly disagree*' to '*strongly agree*.' The average of the school-means is 3.6 (*SD* = 0.48) and shows high reliability overall and by group assignment (α = 0.95 overall; 0.96 treatment; 0.95 control) (table 4.2).

Exhibit 4.2. Year-Two Teacher-Reported Common Planning Time Discussion Items

Items	Scale
During common planning time this year, how often have teachers discussed:	**Never - Almost Always** (5 pt)
a. The school's goals or vision for improving student achievement?	
b. Preparation for the state test?	
c. Student test results?	
d. Other student work?	
e. Developing grading rubrics?	
f. Developing class tests?	
g. Developing lesson plans?	
h. Instructional methods/pedagogy?	
i. Students who are not meeting grade level expectations?	
j. Student behavior?[1]	
k. Observations of teachers' classrooms?	
l. Logistical or other non-academic issues?[1]	

[1] Removed from scale for analyses.

Exhibit 4.3. Year-Two Teacher-Reported General Collegiality Items

Items	Scale
Please indicate the extent to which you agree or disagree with the following statements about teachers in your school:	**Strongly Disagree - Strongly Agree** (5 pt)
a. Teachers in this school respect colleagues who are expert in their craft.	
b. Teachers in this school trust each other.	
c. Teachers in this school really care about each other.	
d. Teachers respect other teachers who take the lead in school improvement efforts.	
e. Many teachers openly express their professional views at faculty meetings.	
f. Teachers in this school are willing to question one another's views on issues of teaching and learning.	
g. We do a good job of talking through views, opinions, and values.	
h. Teachers in this school feel responsible for helping each other do their best.	

Achievement Culture (School Level)

A 12-item set measures the school's achievement culture. The five-point response scale asks for teachers' estimates of the proportion of their school peers who hold various beliefs or use various practices ('*very few*' to '*nearly all*'); for example, the proportion of teachers who believe all students can learn (exhibit 4.4). The average of the school means is 4.0 (*SD* = 0.45). The item set shows high reliability at 0.97 (0.98 treatment; 0.97 control) (table 4.2).

Exhibit 4.4. Year-Two Teacher-Reported Achievement Culture Items

Items	Scale
Thinking about what the teachers in your school say and do, how many teachers would you say:	**Very Few - Nearly All** (5 pt)
a. Are invested in improving their teaching?	
b. Have a good grasp of the subject matter they teach?	
c. Feel responsible when students in this school fail?	
d. Believe that all students can learn?	
e. Have high expectations for students' academic work?	
f. Reteach content to students who aren't successful the first time?	
g. Use another instructional approach when students aren't successful the first time?	
h. Use student assessment data to identify students in need of instructional support?	
i. Use student assessment data to identify which standards students have not mastered?	
j. Provide instruction to meet individual student learning needs?	
k. Motivate students to learn?	
l. Encourage students to set and meet academic goals?	

Attitudes Toward Data and Assessment (Teacher Level)

The 6-item set measuring teachers' attitudes toward interim assessments and data use is measured on a 4-point scale. Questions ask teachers to state their level of agreement on a scale of '*not at all*' to '*very much*' (exhibit 4.5). With a mean of 3.0 (*SD* = 0.44), the set shows an overall reliability of 0.71 (0.72 treatment; 0.68 control) (table 4.2). Despite adjusting the wording on the item asking whether interim assessments take

away needed instructional time for year two, it still exhibits poor fit with the other items. Although its removal improves the reliability of the scale slightly (from about 0.71 to 0.78), the item is included in the summary scale.

Exhibit 4.5. Year-Two Teacher-Reported Assessment/Data Attitudes Items

Items	Scale
We are interested in your opinions about interim assessments, their administration, and student results. Please respond to the following:	**Not At All - Very Much** (4 pt)
a. How accountable do you feel to other teachers in your school for your students' progress on interim assessments?	
b. How accountable do you feel to your school leaders for your students' progress on interim assessments?	
c. How much does the administration of interim assessments take needed time away from classroom instruction? (r)	
d. How useful are interim assessments as an instructional tool?	
e. How consistent are interim assessment results with your own observations of student learning?	
f. How predictive are interim assessment results of students' performance on end-of-year state tests?	

Note: Item (c) is reverse coded.

Confidence in Data Use and Instructional Practices (Teacher Level)

Two sets of items measure teachers' confidence in using data ($\bar{x}$ = 4.0, *SD* = 0.63; 5-point scale) and instructional planning and practice ($\bar{x}$ = 3.1, *SD* = 0.51; 4-point scale) (exhibits 4.6 and 4.7). Scales range from '*I don't know how*' or '*not at all confident*' to '*highly confident*.' Overall, the reliability is high for the items measuring confidence in data use (α = 0.96 overall; 0.95 treatment; 0.96 control) and confidence in instructional planning and practice (α = 0.90 overall; 0.91 treatment; 0.89 control).

Exhibit 4.6. Year-Two Teacher-Reported Data Use Confidence Items

Items	Scale
When working with interim assessment data, how confident are you in your own ability to:	**I Don't Know How - Highly Confident** (5 pt)
a. Use data to set learning goals for individual students?	
b. Identify the skills students need to answer an assessment item correctly?	
c. Determine which students have not mastered specific standards or skills?	
d. Use data to measure student progress toward learning goals?	
e. Adjust your teaching plans to better meet students' learning needs based on the data?	
f. Understand if a skill should be taught or retaught to the whole class, in small groups, or with individual students?	
g. Use data to identify gaps in the school's core curriculum?	
h. Use the data to identify and target instruction to students who are scoring just below a performance cut point?	
i. Use the data to reflect on the success of past instruction?	
j. Identify new materials to address gaps in the school's core curriculum?	

Exhibit 4.7. Year-Two Teacher-Reported Instructional Planning Confidence Items

Items	Scale
How confident are you in your own ability to:	**Not At All Confident - Highly Confident** (4 pt)
a. Plan and modify instruction to meet students' learning needs?	
b. Create differentiated instruction plans to meet students' learning needs?	
c. Motivate students who show little interest in school work?	
d. Provide appropriate challenges for very capable students?	
e. Gauge individual students' mastery of specific standards?	
f. Reteach content that students did not master the first time?	
g. Use a curriculum scope and sequence to design lesson or unit plans?	
h. Use the content of upcoming interim assessments to design lesson or unit plans?	
i. Fit reteaching time into the existing curricular scope and sequence?	

Data Practices (Teacher Level)

Two sets of items comprise the outcome indices related to teachers' self-reported data practices. One 4-item set covers the frequency with which teachers review data independently and with others ($\bar{x}$ = 3.0, *SD* = 0.67) (exhibit 4.8). A second 8-item set covers the frequency with which teachers use interim assessment data in various ways ($\bar{x}$

= 3.4, *SD* = 0.76) (exhibit 4.9). The five-point scales range from '*never'* to '*more than once a week*.' The data review item set has a relatively high reliability of 0.83 (0.83 treatment; 0.83 control). The second index measuring the frequency of teachers' data use is even stronger with a reliability of 0.95 (0.94 treatment; 0.96 control) (table 4.2).

Exhibit 4.8. Year-Two Teacher-Reported Data Review Items

Items	Scale
Over this past school year, how often have you reviewed interim assessment data:	**Never - More Than Once A Week**
a. Independently?	(5 pt)
b. With other teachers in your grade or subject area?	
c. With all teachers in your school?	
d. With your principal, coach, or other instructional leader?	

Exhibit 4.9. Year-Two Teacher-Reported Data Use Items

Items	Scale
Over this school year, how often have you used interim assessment data to:	**Never - More Than Once A Week**
a. Set learning goals for individual students?	(5 pt)
b. Determine which students have not mastered specific standards or skills?	
c. Measure student progress toward learning goals?	
d. Adjust your teaching plans to better meet students' learning needs based on the data?	
e. Understand if a skill should be taught or re-taught to the whole class, in small groups, or with individual students?	
f. Identify and target instruction to students who are scoring just below a performance cut point?	
g. Reflect on the success of past instruction?	
h. Identify gaps in the school's core curriculum?	

Instructional Practices (Teacher Level)

Finally, two sets of items comprise the indices that measure the frequency of teachers' instructional planning and practices; indices are measured on a five-point scale from '*never'* to '*almost always.*' These sets cover the frequency with which teachers use various planning strategies ($\bar{x}$ = 4.0, *SD* = 0.59) (exhibit 4.10) and differentiate instruction to students ($\bar{x}$ = 4.0, *SD* = 0.69) (exhibit 4.11). Among the seven-item set of

instructional planning activities, the reliability is moderately high at 0.78 (0.78 treatment; 0.78 control). For the two items that are combined to generate the measure of frequency with which teachers differentiate instruction to students, reliability is moderately high at 0.74 (0.75 treatment; 0.72 control).

Exhibit 4.10. Year-Two Teacher-Reported Instructional Planning Items

Items	Scale
When planning instruction, how often do you:	**Never - Almost Always** (5 pt)
a. Begin by identifying the skill or goal you hope students will master?	
b. Begin by identifying the state standard you hope students will master?	
c. Create differentiated instruction plans to meet student's individualized learning needs?	
d. Use a curriculum scope and sequence to design lesson or unit plans?	
e. Schedule reteaching time into your lesson or unit plans?	
f. Schedule reteaching time outside of regular class time?	
g. Use the content of upcoming interim assessments to design lesson or unit plans?	

Exhibit 4.11. Year-Two Teacher-Reported Instructional Differentiation Items

Items	Scale
Teachers use a variety of strategies to address students' different learning needs. In your own practice, how often do you do each of the following?	**Never - Almost Always** (5 pt)
a. Teach or reteach content to small groups of students.	
b. Teach or reteach content to individual students.	

In sum, the reliability of the scales and indices measuring hypothesized school- and teacher-level mediators is very high, the exception being the scale measuring teachers' attitudes toward data and interim assessment (table 4.2). Measurement error in this predictor may result in either the upward or downward bias of multiple regression estimates in models in which this predictor is included. The reliability of the outcome measures of data review, instructional planning, and instructional differentiation are not as high. While this will not introduce additional bias in the regression estimates, measurement error in the outcome can reduce statistical power and the ability to detect a

statistically significant relationship. Again, high reliability estimates can indicate that a composite variable has relatively little measurement error. However, the high reliability estimates for some of these scales and indices are likely due in part to moderate proportions of teachers' responding to all items within a scale with the same response, a phenomenon known as straight-lining.

Table 4.2. Descriptive Statistics and Reliability for Each Scale or Index

		Descriptives			Reliability		
Measure	Items	Mean	SD	n	Overall	T_x	C_x
SCHOOL LEVEL[1]							
Instructional leaders' abilities	9	3.9	0.55	67	0.97	0.99	0.97
Professional culture							
CPT discussions	10	3.7	0.40	67	0.92	0.94	0.91
General collegiality	8	3.6	0.48	67	0.95	0.96	0.95
Achievement culture	12	4.0	0.45	67	0.97	0.98	0.97
TEACHER LEVEL							
Assessment/data attitudes	6	3.0	0.44	585	0.71	0.72	0.68
Confidence							
Data use	10	4.0	0.63	588	0.96	0.95	0.96
Instructional planning	9	3.1	0.51	615	0.90	0.91	0.89
Data Practices							
Data review	4	3.0	0.67	572	0.83	0.83	0.83
Data use	8	3.4	0.76	581	0.95	0.94	0.96
Instructional Practices							
Instructional planning	7	4.0	0.59	616	0.78	0.78	0.78
Differentiated instruction	2	4.0	0.69	608	0.74	0.75	0.72

[1] The descriptive statistics for the school-level measures are calculated after generating the school-mean teacher scale score. Likewise, reliability is calculated on the school-mean teacher response to each item within a scale.

Note: Estimates are for the valid, nonmissing responses from a possible sample of 67 impact sample schools and 616 in-scope teachers. T_x = treatment group, C_x = control group. Items comprising all scales and indices are measured on a 5-point response scale except for teacher attitudes and teacher confidence in instructional planning (4-point response scale).

MEASURE VALIDATION

Chapter three reviewed the process followed during year-two survey revisions to maximize the content validity of focal measures. Prior research has suggested that hypothesized school- and teacher-level mediators are expected to covary and should be positively correlated with mediators within the same level. This section explores the relationships among hypothesized mediators. Specifically, correlations among the scales and indices measuring hypothesized school- and teacher-level mediators are reviewed as a means for testing their convergent validity.

These calculations are based on Pearson's correlation coefficient which assumes the data are from an interval scale and normally distributed. Since measurement error can attenuate the correlation between scales, a disattenuation formula has been applied to correct for measurement error: the unadjusted correlation of the two scales is divided by the square root of the product of the two scales' reliabilities (Pedhazur, 1997). Also of note, the statistical significance of Pearson's correlations among teacher-level mediators are not adjusted for clustering of teachers within schools. In separate analyses (not shown), the results were found to be robust to both calculations of point estimates using Spearman's rank – which is appropriate for ordinal data and does not assume normality – and, for teacher-level mediators, tests of significance using a two-level regression model accounting for nesting. Pairwise deletion of cases with missing data was used, rather than listwise, to maximize the available data for each correlation.

The results suggests strong, positive correlations among school-level mediators. Teachers' mean perceptions of their school leaders' instructional leadership abilities are

all highly correlated with mean perceptions of school culture (all $p < 0.01$) (table 4.3). Among the various measures of school culture, the lowest correlations are seen between collegial discussions (frequency) and other measures, particularly achievement culture ($r = 0.47$, $p < 0.01$). Still, these are moderately strong, positive relationships. These results confirm that measures of instructional leadership, professional culture, and achievement culture do covary in the expected ways, providing evidence of convergent validity among school-level mediators.

Table 4.3. Corrected Pairwise Correlations Among School-Level Scales

	Instructional leaders' abilities	Professional culture: CPT discussions	Professional culture: general collegiality
Professional culture: CPT discussions	0.65	--	--
Professional culture: general collegiality	0.80	0.61	--
Achievement culture	0.71	0.47	0.80

Notes: All n = 67; all $p < 0.01$ (accounting for Bonferroni adjustment). Disattenuation correction has been applied.

Turning to teacher-level mediators, all of the correlations among teacher-reported scales are positive and significant at the $p < 0.01$ level (table 4.4). The correlation between measures of confidence – in data use and instructional planning – is highest ($r = 0.71$). Correlations among the two measures of confidence and teachers' attitudes are also positive, but lower ($r = 0.38$-0.39).

Table 4.4. Corrected Pairwise Correlations Among Teacher-Level Scales

	Assessment/data attitudes	Confidence: data use
Confidence: data use	0.39	--
N	*584*	--
Confidence: instructional planning	0.38	0.71
N	*584*	*587*

Notes: All $p < 0.01$ (accounting for Bonferroni adjustment). Disattenuation correction has been applied.

The results show strong, positive correlations among hypothesized mediators at each level. These results are consistent with prior research, and provide reasonable evidence of convergent validity of the school- and teacher-level mediators. In addition, the correlations among mediators are strong, but not so strong as to warrant concern about multicollinearity when simultaneously included in multiple regression models.

RESULTS: MAIN RESEARCH QUESTIONS

Recall from chapter three, the first two research questions exploit the randomized design of the larger Achievement Network (ANet) i3 evaluation to explore the effect of ANet on school and teacher characteristics, and teacher practice outcomes. The third and fourth research questions are more exploratory; they seek evidence of the role played by school and teacher characteristics on ANet's impact on teacher practice. Specifically, research question three explores these characteristics as meditators, whereas research question four examines their role as baseline moderators.

Because of the nested structure of the data – teachers within schools – these analyses generally employ two-level models to account for the possibility that teachers within the same school may be more likely to share characteristics with each other than with teachers in other schools (i.e., leading to a violation of the usual assumptions of OLS regression and greater likelihood of type I errors). A first step in analyzing nested data is to examine the intra-class correlation (ICC). Table 4.5 presents the ICCs for each of the four teacher practice outcomes. The ICC estimates in the top row are based on an unconditional model that includes no independent variables. The ICC estimates in the

bottom row are conditional on treatment assignment, a level-two indicator for whether the school was assigned to the ANet (treatment) or control group.

Table 4.5. Intraclass Correlations for Each Teacher Practice Outcome: Unconditional and Conditional on Treatment Assignment

ICC	Data Review	Data Use	Instructional Planning	Instructional Differentiation
Unconditional	0.252	0.164	0.153	0.064
Treatment assignment	0.221	0.155	0.148	0.061

In the unconditional models, the proportion of variation at the school level is highest for the frequency of teachers' data review practices – about 25 percent – which makes intuitive sense; teachers are generally either reviewing their students' interim assessment results with other teachers in their school or on their own; therefore, responses within a school should be more similar. About 15 to 16 percent of the variance in the frequency of teachers' data use and instructional planning activities is at the school level. In contrast, the frequency with which teachers differentiate instruction appears to be an individual decision. Only about 6 percent of the variation is between schools; the great majority of variation is seen within schools, at the teacher-level. Still, the proportion of between-school variance in most teacher practice outcomes is justification for pursuing multilevel analyses.

The ICCs are calculated from estimates of the variance components at levels one and two of the multilevel models. Table 4.6 reports the variance components for each teacher practice outcome, by level, for the unconditional model and the conditional model. Since treatment assignment is a school-level (L2) variable, it's inclusion in the conditional models should not reduce the level-one variance, which it does not. However,

it does reduce the level-two variance in the data review model by about 16%, calculated as:

$$\frac{\hat{\tau}_{00\,(uncond)} - \hat{\tau}_{00\,(cond)}}{\hat{\tau}_{00\,(uncond)}} \quad (4.1)$$

For the other outcomes, treatment assignment explains about 7 percent of the between-school variance in data usage, 4 percent of the variance in instructional planning, and 5 percent of the variance in instructional differentiation. Results suggest that treatment and control schools do not differ much in between-school variance in the four outcomes of interest.

Table 4.6. Variance Components for Each Teacher Practice Outcome, by Model and Level

Model	Data Review	Data Use	Instructional Planning	Instructional Differentiation
Unconditional				
Level 1 ($\hat{\sigma}^2$)	0.769	0.827	0.836	0.930
Level 2 ($\hat{\tau}_{00}$)	0.260	0.162	0.151	0.063
Conditional on treatment assignment				
Level 1 ($\hat{\sigma}^2$)	0.768	0.825	0.834	0.931
Level 2 ($\hat{\tau}_{00}$)	0.218	0.152	0.145	0.060

Note: total variance sums to approximately one because outcomes were standardized prior to the analysis.

The presentation of results now turns to each of the four main research questions. The multilevel models used in these analyses are random-intercept models with fixed slopes. All models include indicators at level two for: treatment-group assignment, district, data collection wave, and whether they belong to the Chelsea "pair" of three schools. At level one, teachers' total years of teaching experience and highest degree, measured at year two, are included in all models. The sample for a given model includes

any of the 616 in-scope teachers with nonmissing values on the predictors and outcome. Additions to the model specifications or exceptions to the sample are noted for relevant analyses. All composite measures have been standardized prior the analysis; results are reported in standard deviation units.

Research Question One: Teacher Practice Impact Models

Are teachers' data use and instructional practices different in ANet (treatment) schools from those in control schools?

In research question one, the impact of ANet on each of the four teacher practice outcomes is explored *after two years*. These outcomes are the frequency with which teachers (1) review data, (2) use data, (3) use various instructional planning techniques, and (4) differentiate instruction. Results show that ANet has a positive impact on the frequency with which teachers review and use data (both $p < 0.01$). The mean frequency with which teachers in ANet schools reviewed data was nearly a half a standard deviation higher than in control schools (0.45 *sd*). The mean frequency with which ANet teachers used data was one-quarter of a standard deviation higher than in control schools (0.25 *sd*). ANet had a small, but only marginally significant impact on the frequency of teachers' instructional planning practices (0.16 *sd,* $p < 0.10$) and no impact on the frequency they reported differentiating instruction (table 4.7). In analyses, not shown, ANet was found to have had a positive impact on teachers' use of whole-class instruction (0.22 *sd*; $p = 0.01$); a finding that will be discussed in later chapters.

Table 4.7. Teacher Practice Impact Results

Variable	Model 1 Data Review	Model 2 Data Use	Model 3 Instructional Planning	Model 4 Instructional Differentiation
Fixed effect				
Assigned to treatment: school	0.45 ‡	0.25 ‡	0.16 *	-0.10
	0.112	*0.089*	*0.096*	*0.088*
District				
Chelsea	-0.19	-0.14	-0.41	0.01
	0.300	*0.230*	*0.253*	*0.227*
Chicago	0.70 ‡	0.27	0.52 **	0.19
	0.245	*0.217*	*0.226*	*0.218*
Jefferson Parish	-0.14	-0.11	0.26 *	0.11
	0.160	*0.132*	*0.141*	*0.134*
Springfield	0.30 *	0.29 **	0.29 *	0.38 ‡
	0.169	*0.132*	*0.146*	*0.134*
Data collection wave two: school	0.40 **	0.38 ‡	0.17	0.08
	0.161	*0.126*	*0.137*	*0.126*
Unbalanced pair dummy: school	-0.35	-0.61 **	-0.12	-0.42 *
	0.353	*0.260*	*0.294*	*0.257*
Years of teaching experience	0.01 ‡	0.01 ‡	0.01 ‡	0.00
(total): teacher	*0.004*	*0.004*	*0.004*	*0.005*
Highest degree: teacher				
Master's	0.21 **	0.06	0.06	0.11
	0.099	*0.101*	*0.099*	*0.105*
Doctorate	-0.18	-0.43	0.12	0.18
	0.350	*0.363*	*0.365*	*0.386*
Random effect				
School (intercept)	-1.05 ‡	-0.77 ‡	-0.59 ‡	-0.16
	0.220	*0.185*	*0.195*	*0.186*
Variance Components				
L1	0.748	0.824	0.831	0.938
L2	0.094	0.020	0.045	0.014
Additional Variance Explained (%)				
L1	3%	0%	0%	-1%
L2	57%	87%	69%	77%
Model statistics				
n	559	569	603	596
Number of groups	67	67	67	67
Wald χ^2	70.09 ‡	81.76 ‡	53.89 ‡	27.30 ‡

Notes: Outcome scales were standardized within the teacher sample; results are reported in standard deviation units. Estimates are reported on the top row for each predictor. Standard errors are reported below, in italics. Omitted district = Boston; omitted degree = bachelor's; and data collection wave one = 1, wave two = 2. Additional variance explained is based on comparisons with conditional models in table 4.6.
‡ $p < 0.01$; ** $p < 0.05$, * $p < 0.10$.

An interesting pattern also emerges in each district. All else being equal, teachers in Chicago and Springfield report, on average, more frequent review and use of data than their counterparts in Boston (the district reference group). The difference between the frequency teachers review data in Chicago and Boston is statistically significant ($p < 0.01$) To put this is context, the estimate of the frequency with which Chicago teachers review data (compared to Boston teachers) is about 1.6 times that of the ANet (treatment) impact estimate (0.70 *sd* vs. 0.45 *sd*, respectively). In terms of instructional practices, the frequency with which teachers in Chicago reported various instructional planning activities was about a half of standard deviation higher than teachers in Boston (0.52 *sd, p* < 0.05). The frequency with which teachers in Springfield reported each of the four data-based instructional outcomes was about one-third of standard deviation higher than teachers in Boston (all $p < 0.10$).

On average, teachers in Chelsea and Jefferson Parish review and use data less frequently that their peers in Boston, though these differences were not statistically significant. In Chelsea, the pattern holds for instructional planning: on average, teachers in Chelsea report less frequent use of various instructional planning activities than their Boston peers. Though the difference is large (-0.41 *sd*), it is not statistically significant. However, teachers in Jefferson Parish reported more frequent use of various instructional planning (0.26 *sd*, $p < 0.10$) than their Boston peers. No difference was found between the frequency of instructional differentiation by teachers in these districts and Boston.

Also of note, year-two teachers in the second wave of data collection reported higher frequencies of data review and use than their counterparts in wave one ($p < 0.05$ or

better), but no difference in instructional practices was found between the two waves. Across outcomes, the magnitude of the differences in frequency of practices between teachers in waves one and two was similar to that of the difference between treatment- and control-school teachers. Since schools in wave two were recruited only in Springfield and Jefferson Parish where ANet was negotiating district-level partnerships, teachers in wave two may have benefitted from district-wide structures or cultures that facilitated their data practices.

The inclusion of teacher-level (L1) covariates measuring experience and highest degree reduces the level-one variance very little when compared to the conditional model in table 4.6. However, the additional district- and school-level covariates at level two explain a substantial proportion of between-school variance in teacher practice outcomes: ranging from 57 percent of the level-two variance in the frequency teachers review data to 87 percent of the frequency the use data. Finally, the Wald χ^2 statistic provides a test of the null hypothesis that, across the set of regression coefficients, at least one is not equal to zero (table 4.7). The fact that it is statistically significant across all models means the null hypothesis can be rejected; taken together, the coefficients in each model are statistically significant.

Because the pattern of findings pointed to interesting treatment and district effects on teacher practice outcomes, the models shown in table 4.7 were re-run to include an interaction between school treatment assignment and district. The omnibus F-test for the interaction indicates that it is statistically significant at $p < 0.05$ or better in three of the four models, those predicting teachers' frequency of data review, data use, and

instructional planning (table 4.8). With the interaction terms included, the coefficient of treatment assignment estimates the treatment effect in Boston (the district comparison group). Coefficients of treatment-by-district interaction terms estimate the difference between the treatment effects in Boston and each district. For example, in model one (table 4.8), the treatment effect of ANet on teachers' frequency of reviewing data in Boston is 0.59 ($p < 0.01$). Treatment effects of ANet on teachers' frequency of reviewing data in Chelsea and Chicago are lower than Boston (by -1.10 and -1.19 *sd*, respectively) and these differences are significant (both $p < 0.01$). More precisely, the results indicate that the treatment effect in Chelsea is -0.51 (0.59 – 1.10) and in Chicago is -0.60 (0.59 – 1.19) (*ns*). In contrast, treatment effects in Jefferson Parish and Springfield are positive (0.66 and 0.58 *sd*, respectively), though not statistically different from Boston. Only the treatment effects of ANet on the frequency teachers review data in Boston and Jefferson Parish are significant ($p < 0.05$).

Exploring other teacher practice outcomes, the treatment effect of ANet on the frequency teachers use various instructional planning strategies is lower in Chelsea and Chicago than in Boston ($p < 0.05$) and the impact of ANet on the frequency teachers reported using data in various ways is lower in Chicago than Boston ($p < 0.10$). The impact of ANet on the frequency of teachers' data use in Jefferson Parish is significant (0.52 *sd*, $p < 0.05$) and the impact of ANet on teachers' instructional planning in Springfield is significant (0.57 *sd*, $p < 0.05$).

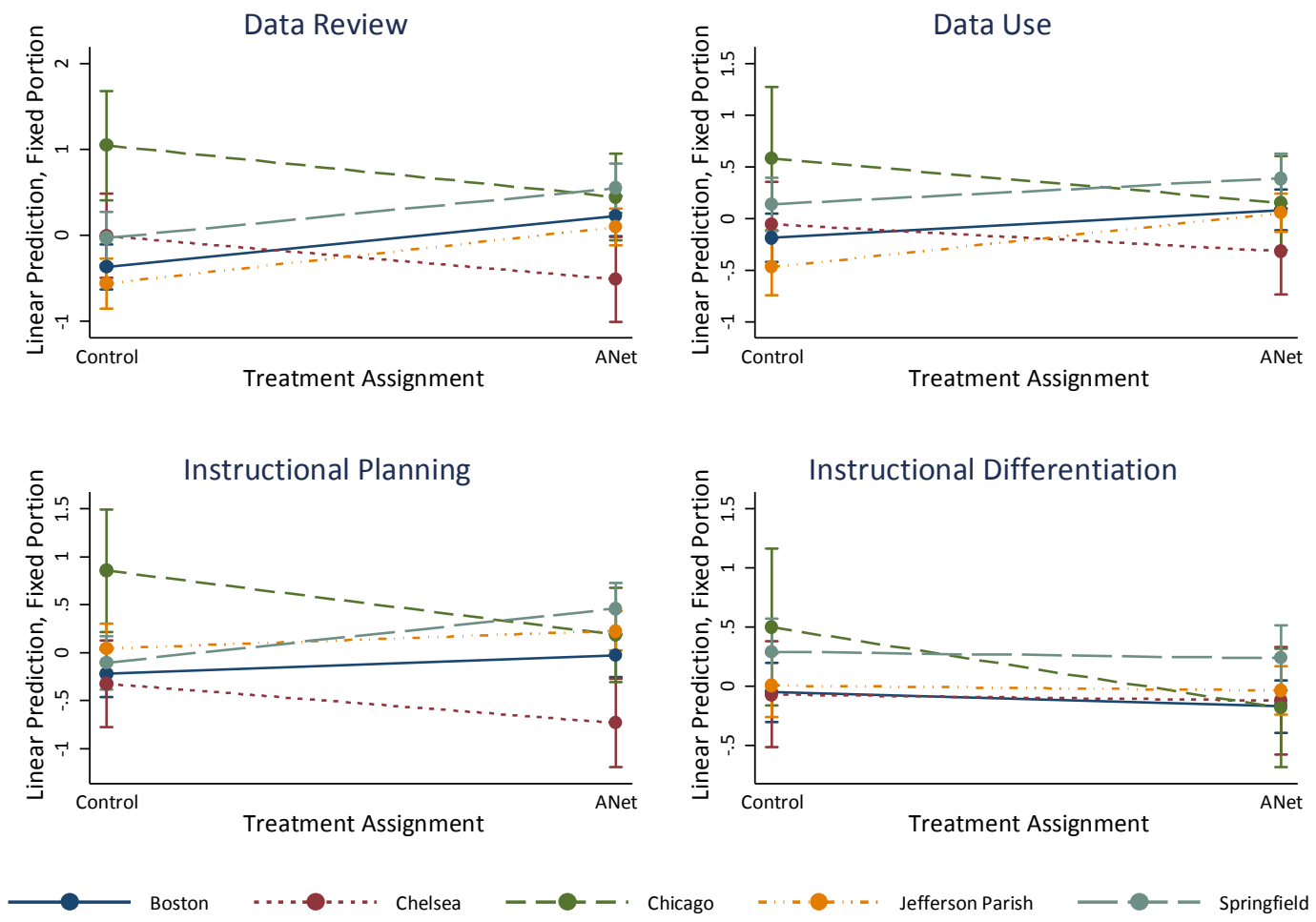
Table 4.8. Teacher Practice Impact Results with Treatment by District Interaction

Variable	Model 1 Data Review	Model 2 Data Use	Model 3 Instructional Planning	Model 4 Instructional Differentiation
Fixed effect				
Assigned to treatment: school	0.59 ‡	0.27 *	0.18	-0.12
	0.168	*0.147*	*0.158*	*0.159*
District				
Chelsea	0.36	0.13	-0.11	-0.02
	0.295	*0.249*	*0.273*	*0.270*
Chicago	1.42 ‡	0.77 **	1.07 ‡	0.55
	0.346	*0.369*	*0.346*	*0.358*
Jefferson Parish	-0.19	-0.29	0.26	0.06
	0.203	*0.186*	*0.187*	*0.191*
Springfield	0.34 *	0.32 *	0.11	0.34 *
	0.204	*0.178*	*0.190*	*0.191*
Treatment x District Interaction				
Treatment*Chelsea	-1.10 ‡	-0.53 **	-0.59 **	0.07
	0.322	*0.264*	*0.298*	*0.292*
Treatment*Chicago	-1.19 ‡	-0.70	-0.85 *	-0.56
	0.444	*0.446*	*0.439*	*0.450*
Treatment*Jefferson Parish	0.07	0.26	0.00	0.08
	0.239	*0.211*	*0.221*	*0.222*
Treatment*Springfield	-0.01	-0.02	0.38	0.06
	0.267	*0.229*	*0.249*	*0.251*
Data collection wave two: school	0.42 ‡	0.39 ‡	0.19	0.09
	0.136	*0.117*	*0.125*	*0.126*
Unbalanced pair dummy: school	-0.21	-0.53 **	-0.03	-0.44 *
	0.286	*0.236*	*0.264*	*0.258*
Years of teaching experience	0.01 ‡	0.01 ‡	0.01 ‡	0.00
(total): teacher	*0.004*	*0.004*	*0.004*	*0.005*
Highest degree: teacher				
Master's	0.21 **	0.06	0.05	0.10
	0.099	*0.100*	*0.098*	*0.105*
Doctorate	-0.25	-0.48	0.07	0.13
	0.351	*0.364*	*0.365*	*0.388*
Random effect				
School (intercept)	-1.13 ‡	-0.79 ‡	-0.63 ‡	-0.14
	0.206	*0.187*	*0.194*	*0.198*
Variance Components				
L1	0.754	0.822	0.828	0.935
L2	0.039	0.004	0.024	0.013
Additional Variance Explained (%)				
L1	-1%	0%	0%	0%
L2	58%	81%	48%	6%
Model statistics				
n	559	569	603	596
Number of groups	67	67	67	67
Wald χ^2	112.56 ‡	112.25 ‡	79.82 ‡	29.84 ‡
F (interaction)	20.92 ‡	11.22 **	13.64 ‡	2.16

Notes: Outcome scales were standardized within the teacher sample; results are reported in standard deviation units. Estimates are reported on the top row for each predictor. Standard errors are reported below, in italics. Omitted district = Boston; omitted degree = bachelor's; and data collection wave one = 1, wave two = 2. Additional variance explained is compared to models in table 4.7.
‡ $p < 0.01$; ** $p < 0.05$, * $p < 0.10$.

As a visual example, see the model for data review (model 1, table 4.8) and the plot of the interaction between treatment assignment and district for that model (top left, figure 4.1). Comparing the line for Chicago (in green, short-dashed line) with the line for Boston (in blue, solid line) in figure 4.1 (top left), the main effect of ANet on teachers' frequency of data review is higher than in Boston (i.e., comparing the height of the line, or "averaging" the treatment and control estimates within each district, Chicago is higher). This corresponds to the positive main effect for Chicago teachers' data review of 1.42 ($p < 0.01$) in table 4.8 (model 1). However, the slope of line for Chicago is negative (ANet teachers in Chicago review data less frequently than control-school teachers in Chicago) in figure 4.1 (top left) indicating a negative treatment effect of -0.60 ($p < 0.05$). A similar pattern in seen in Chicago and Chelsea for data use and instructional planning.

Figure 4.1. Interaction Between District and Treatment Assignment, by Teacher Practice Outcome



Research Question Two: School and Teacher Mediator Impact Models

Are levels of school culture, instructional leadership, and teachers' attitudes towards and confidence with data-based practices (hypothesized mediators) different in ANet (treatment) schools from those in control schools?

This analysis estimates the impact of ANet on hypothesized school- and teacher-level mediators of teacher practices. The school-level measures consist of: (1) teachers' perceptions of their instructional leaders' abilities, (2) the frequency of teachers' collegial discussions during common planning time, (3) their perceptions of the school's professional culture, and (4) achievement culture. Teacher-level measures include: (1) teachers' attitudes toward interim assessment and data use, and their confidence using various (2) assessment/data use practices and (3) instructional practices. Recall, all models include the same set of control variables: indicators for each district (with Boston as reference group), study wave, the Chelsea triad of schools, and teachers' years of experience (continuous) and highest degree (with bachelor's degree as reference group).

School-Level Mediators. In models estimating the impact of ANet on hypothesized school-level mediators, there are two points of note. First, because these analyses are modeled at the school-level they utilize OLS regression models instead of multilevel models. All covariates from research question one are included, but teacher-level covariates are aggregated to their school mean. Second, these models rely on aggregate teacher data: a school-mean composite score was generated from individual teacher data. Given teacher response rates within schools is expected to be less than 100 percent in some cases, this method assumes that the sample of responding teachers is a random sample of the target population of in-scope teachers. This is unlikely and,

therefore, an unknown amount of bias may have been introduced. As an alternative, models that apply weights, calculated as the inverse of the variance in the respective school-mean scale estimate, are included in appendix B.[1]

Although these models are underpowered, results after two years show that ANet had a small, positive impact on both teacher mean perceptions of their school leader's instructional leadership abilities (0.41 *sd*), as well as their school's achievement culture (0.45 *sd*) (both $p < 0.10$) (table 4.9, models 1 and 4). In other words, mean teacher perceptions of school leaders' instructional abilities and achievement culture were higher in ANet schools than control schools by just over four-tenths of a standard deviation. The impact of ANet on the frequency of teachers' common planning time discussions and their perceptions of the school's general collegiality were also positive and non-negligible (about three-tenths of a standard deviation), though not statistically significant.

Across the four measures of hypothesized school-level mediators, treatment-control differences fall short of conventional standards for statistical significant likely due to the small number of schools. Nonetheless, the magnitude of these effects are potentially meaningful and suggest that, in sufficiently powered analyses, ANet may have an impact on these measures. Therefore, it is reasonable to conclude that these school-level conditions could at least partially mediate ANet's impact on teachers' data-based instructional practices.

[1]In the weighted models, school means that are more precisely estimated contribute more to the overall model estimates than school means that are less precisely estimated. Inverse variance weighting makes an assumption that larger variance is due to fewer responding teachers and, therefore, higher nonresponse. However, larger variance could also indicate greater heterogeneity in teacher responses within school, regardless of response rate. If this is the case, weighted models can also introduce bias. For schools with one teacher, the weight is equal to 1/mean(variance).

Interesting patterns in the impacts of ANet on these hypothesized school-level mediators also emerge across districts. With very few exceptions, school-mean reported leader abilities, professional culture measures, and achievement culture are lowest among schools in Boston (the reference district). This is evident from the largely positive coefficients across district dummy variables for each hypothesized school-level mediator. The most notable pattern was in the frequency of collegial discussions during common planning time (model 2). All else being equal, the frequency with which collegial discussions took place during common planning time was about two-thirds of a standard deviation higher in Springfield ($p < 0.10$), over eight-tenths of standard deviation higher in Chicago ($p < 0.05$), and nearly one and a half standard deviations higher in Jefferson Parish ($p < 0.01$) compared to schools in Boston.

School-mean years of teaching experience and the proportion of teachers holding a master's degree were positively associated with the frequency of common planning time discussions ($p < 0.01$ and $p < 0.05$, respectively). These findings may be due to a growing focus on collegial collaboration in schools such that more experienced or educated staffs may be more likely to collaborate over data due to pre-service training or in-service exposure to the practice. Finally, in school-level mediator models including a treatment by district interaction, omnibus F-tests showed the interaction terms were not statistically significant for any outcome, therefore, these results are not reported.

Table 4.9. School Mediator Impact Results

Variable	Model 1	Model 2	Model 3	Model 4
		Professional Culture		
	Instructional Leaders' Abilities	CPT discussions	General Collegiality	Achievement Culture
Fixed effect				
Assigned to treatment: school	0.41 *	0.29	0.35	0.45 *
	0.232	*0.209*	*0.239*	*0.230*
District				
Chelsea	0.39	0.46	-0.40	0.25
	0.781	*0.703*	*0.804*	*0.771*
Chicago	0.39	0.84 **	0.50	0.09
	0.456	*0.410*	*0.470*	*0.451*
Jefferson Parish	0.60	1.40 ‡	0.60	0.00
	0.520	*0.468*	*0.535*	*0.514*
Springfield	0.63 *	0.65 *	0.63	0.50
	0.376	*0.338*	*0.387*	*0.371*
Data collection wave two: school	-0.19	0.00	-0.10	-0.49
	0.345	*0.310*	*0.355*	*0.341*
Unbalanced pair dummy: school	-1.49	-0.64	-0.53	-1.58 *
	0.893	*0.803*	*0.919*	*0.882*
School mean years of teaching	0.05	0.09 ‡	0.03	0.05
experience (total)	*0.035*	*0.032*	*0.036*	*0.035*
School mean highest degree				
Master's	0.39	1.22 **	1.08	0.70
	0.667	*0.600*	*0.686*	*0.659*
Doctorate	-4.40	-5.23	-1.97	-2.55
	4.478	*4.029*	*4.609*	*4.424*
School (intercept)	-1.23	-2.78 ‡	-1.53 *	-0.85
	0.797	*0.717*	*0.820*	*0.787*
Model statistics				
n	67	67	67	67
Adjusted R^2	0.13	0.29	0.08	0.15

Notes: Outcome scales were standardized within the school sample; results are reported in standard deviation units. Estimates are reported on the top row for each predictor. Standard errors are reported below, in italics. Omitted district = Boston; omitted degree = bachelor's; and data collection wave one = 1, wave two = 2.
‡ $p < 0.01$; ** $p < 0.05$, $p < 0.10$.

Teacher-Level Mediators. To test the impact of ANet on hypothesized teacher-level mediators, analyses again rely on multilevel models. Results show no impact of ANet on teachers' attitudes towards assessment and assessment data, or their confidence using data or making instructional plans after two years (table 4.10). In other words, the

average differences in these hypothesized teacher-level mediators between teachers in ANet schools and control schools were small (ranging from 0.06 to 0.15 *sd*) and not statistically significant. The point estimates are positive, indicating these conditions are higher among ANet teachers; however, the size of the differences make it less likely that these teacher-level conditions will mediate the relationship between ANet and teachers' data-based instructional practices.

Controlling for all other predictors, the pattern in teacher attitudes and confidence across districts and waves is similar to that of teacher practices seen in research question one. Across hypothesized teacher-level mediators and all else being equal, teachers in Chelsea consistently report more negative attitudes and lower levels of confidence than their Boston peers, though the differences were not statistically significant. In contrast, with few exceptions, average teacher attitudes and confidence levels were higher among teachers in other districts than in Boston. The difference is most marked in Chicago where teachers' adjusted mean attitudes are about a half a standard deviation higher than in Boston (0.52 *sd*, $p < 0.05$). In Chicago, Jefferson Parish, and Springfield, mean teacher confidence in using data ranged from one-quarter to one-third of a standard deviation higher than in Boston (all $p < 0.10$ or better). Finally, teachers in Jefferson Parish reported higher confidence in instructional planning than their Boston peers (0.35 *sd*, $p < 0.01$) (table 4.10).

Table 4.10. Teacher Mediator Impact Results

Variable	Model 1 Assessment/ Data Attitudes	Model 2 Data Use Confidence	Model 3 Instructional Planning Confidence
Fixed effect			
Assigned to treatment: school	0.15	0.11	0.06
	0.109	*0.087*	*0.090*
District			
Chelsea	-0.18	-0.21	-0.05
	0.286	*0.220*	*0.234*
Chicago	0.52 **	0.35 *	0.34
	0.248	*0.209*	*0.219*
Jefferson Parish	-0.06	0.27 **	0.35 ‡
	0.157	*0.130*	*0.135*
Springfield	0.19	0.24 *	0.20
	0.164	*0.129*	*0.136*
Data collection wave two: school	-0.10	-0.04	-0.07
	0.156	*0.123*	*0.129*
Unbalanced pair dummy: school	-0.29	-0.29	-0.20
	0.334	*0.248*	*0.267*
Years of teaching experience	0.02 ‡	0.02 ‡	0.02 ‡
(total): teacher	*0.005*	*0.004*	*0.004*
Highest degree: teacher			
Master's	0.19 *	0.16	0.25 **
	0.104	*0.103*	*0.102*
Doctorate	-0.43	-0.84 **	0.38
	0.374	*0.373*	*0.378*
Random effect			
School (intercept)	-0.27	-0.45 **	-0.53 ‡
	0.216	*0.183*	*0.188*
Model statistics			
n	572	576	602
Number of groups	67	67	67
Wald χ^2	37.09 ‡	68.06 ‡	47.05 ‡

Notes: Outcome scales were standardized within the teacher sample; results are reported in standard deviation units. Estimates are reported on the top row for each predictor. Standard errors are reported below, in italics. Omitted district = Boston; omitted degree = bachelor's; and data collection wave one = 1, wave two = 2.

‡ $p < 0.01$; ** $p < 0.05$, $p < 0.10$.

Each year of additional teaching experience is associated with a small, but statistically significant increase in attitudes and confidence (0.02 *sd*, $p < 0.01$). Teachers who hold a master's degree report, on average, more positive attitudes and greater confidence in instructional planning than their peers with a bachelor's degree (both $p < 0.10$). However, teachers with a doctorate reported lower mean confidence using data in various ways compared to teachers with a bachelor's degree (-0.84 *sd*, $p < 0.05$). This may be a factor of the role teachers with doctorate degrees hold within a school and potentially less frequency interactions with interim assessment data (model 2, table 4.10).

These models were re-estimated including the interaction between treatment assignment and district. In two of the three models, the omnibus F-test show that the interaction effect is statistically significant at the $p < 0.10$ level or better: models predicting attitudes and confidence in instructional planning (table 4.11). The results show a significant negative interaction in Chelsea for the data confidence and instructional confidence models. This indicates a lower, or less positive, effect of ANet on these measures in Chelsea compared to Boston (both $p < 0.05$ or better). However, none of the effects of ANet on teachers' attitudes and confidence within districts are statistically significant.

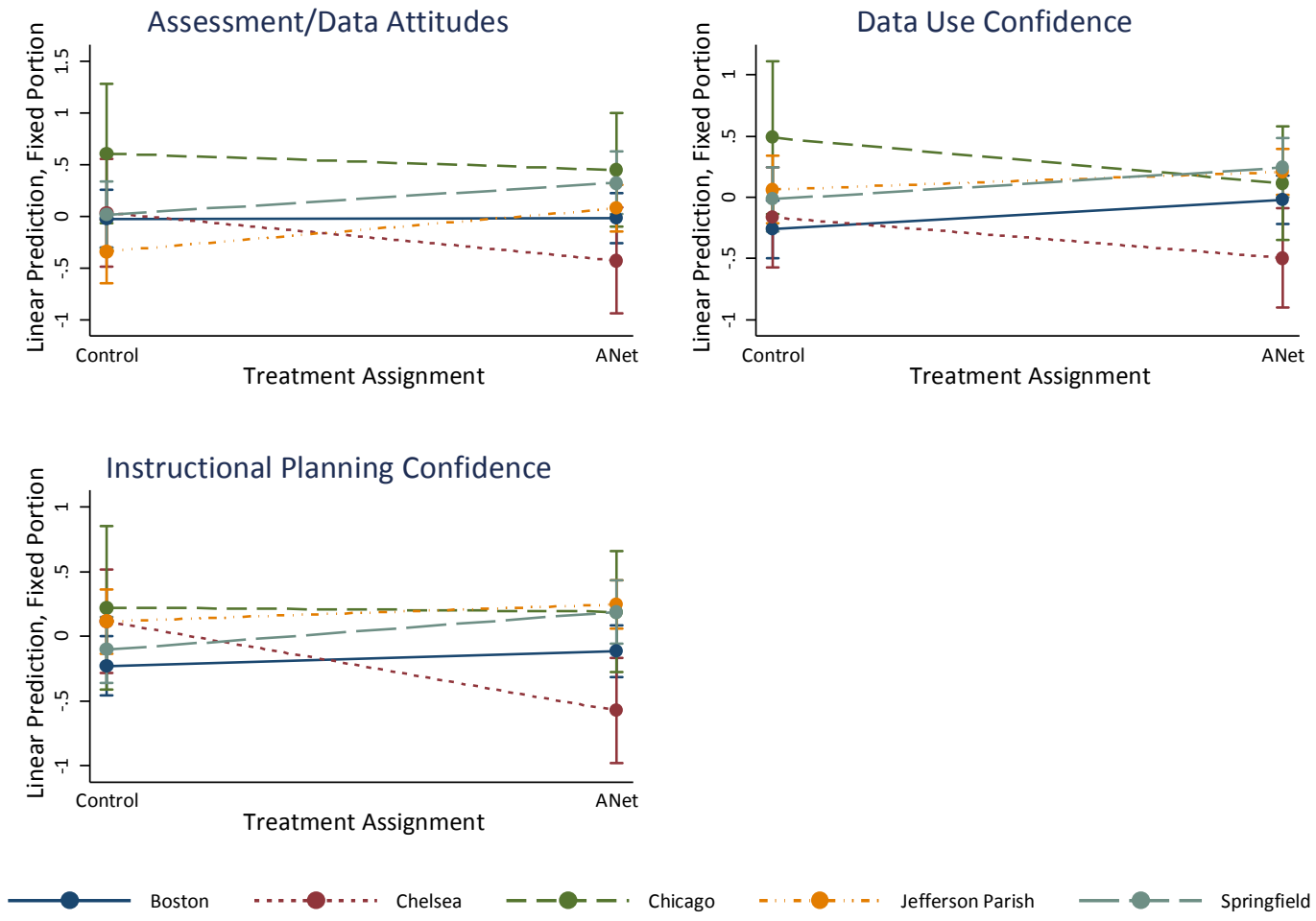
Table 4.11. Teacher Mediator Impact Results with Treatment by District Interaction

Variable	Model 1 Assessment/ Data Attitudes	Model 2 Data Use Confidence	Model 3 Instructional Planning Confidence
Fixed effect			
Assigned to treatment: school	0.01	0.24	0.12
	0.177	*0.147*	*0.145*
District			
Chelsea	0.06	0.10	0.34
	0.311	*0.248*	*0.243*
Chicago	0.63 *	0.75 **	0.45
	0.368	*0.337*	*0.340*
Jefferson Parish	-0.32	0.32 *	0.34 *
	0.214	*0.189*	*0.178*
Springfield	0.04	0.24	0.12
	0.215	*0.179*	*0.174*
Treatment x District Interaction			
Treatment*Chelsea	-0.47	-0.57 **	-0.81 ‡
	0.339	*0.261*	*0.260*
Treatment*Chicago	-0.16	-0.61	-0.15
	0.475	*0.421*	*0.425*
Treatment*Jefferson Parish	0.41	-0.09	0.02
	0.251	*0.213*	*0.204*
Treatment*Springfield	0.30	0.02	0.18
	0.281	*0.229*	*0.227*
Data collection wave two: school	-0.09	-0.01	-0.06
	0.142	*0.117*	*0.115*
Unbalanced pair dummy: school	-0.19	-0.23	-0.09
	0.299	*0.231*	*0.231*
Years of teaching experience	0.01 ‡	0.02 ‡	0.02 ‡
(total): teacher	*0.005*	*0.004*	*0.004*
Highest degree: teacher			
Master's	0.18 *	0.15	0.24 **
	0.104	*0.102*	*0.102*
Doctorate	-0.44	-0.87 **	0.40
	0.375	*0.374*	*0.378*
Random effect			
School (intercept)	-0.22	-0.54 ‡	-0.58 ‡
	0.216	*0.189*	*0.185*
Model statistics			
n	572	576	602
Number of groups	67	67	67
Wald χ^2	50.12 ‡	84.41 ‡	70.10 ‡
F (interaction)	8.34 *	7.23	14.32 ‡

Notes: Outcome scales were standardized within the teacher sample; results are reported in standard deviation units. Estimates are reported on the top row for each predictor. Standard errors are reported below, in italics. Omitted district = Boston; omitted degree = bachelor's; and data collection wave one = 1, wave two = 2.
‡ $p < 0.01$; ** $p < 0.05$, $p < 0.10$.

As a visual example, see the model for instructional planning confidence (model 3, table 4.11) and the plot of the interaction between treatment assignment and district for that model (bottom, figure 4.2). Comparing the line for Chelsea (in red, dotted line) with the line for Boston (in blue, solid line) in figure 4.2 (bottom), the main effect of ANet on teachers' confidence in various instructional planning strategies in Chelsea appears similar to Boston (i.e., "averaging" the treatment and control estimates within each district, Chelsea and Boston are very similar). This corresponds to the near-zero main effect for Chelsea teachers' instructional planning confidence of 0.34 (*ns*) in table 4.11 (model 3). However, the slope of line for Chelsea is negative (ANet teachers in Chelsea have lower instructional planning confidence than control-school teachers in Chelsea) in figure 4.2 (bottom) corresponding to the negative treatment effect of -0.69 (*ns*).

Figure 4.2. Interaction Between District and Treatment Assignment, by Hypothesized Teacher Mediator



Research Question Three: Teacher Practice Mediation Models

Do the hypothesized mediators account for differences in ANet and control-school teachers' data use and instructional practices?

The goal of research question three is to determine whether hypothesized school- and teacher-level mediators explain any of the statistically significant program impacts on teacher practices. First, the relationships between each of teachers' four data-based instructional practices are regressed separately on each of the school- and teacher-level mediators. Estimates were generated from two-level models with the same specifications in research question one. School-level covariates included treatment assignment, district, data collection wave, and unbalanced "pair" dummy, and teacher-level covariates included total teaching experience and highest degree. Table 4.12 reports only the coefficient on each hypothesized mediator; coefficients for all covariates are omitted.

Across all comparisons, there is evidence of relatively strong, positive relationships among mediator and outcomes measures. For school-level mediators, a one standard deviation change is related to a change in teacher practice outcomes generally ranging between one-tenth and one-third of a standard deviation (all $p < 0.05$) (table 4.12). On average, the frequency of collegial discussions during common planning time shows the strongest, positive relationship with teacher practice outcomes. For example, controlling for district-, school-, and teacher-level covariates, a one standard deviation change in the frequency of collegial discussions during common planning time is related to:

- a 0.32 standard deviation change in the frequency teachers review data ($p < 0.01$) and

- a 0.35 standard deviation change in the frequency teachers use various instructional planning strategies ($p < 0.01$).

This former finding should not be surprising given this is likely the most concrete connection between this study's mediators of interest and the ANet logic model. ANet expressly encourages teachers and leaders to review data during data meetings and other collegial settings (i.e., team planning time).

Table 4.12. Estimates from the Regression of Each Teacher Practice Outcome on Each Hypothesized School- and Teacher-Level Mediator

Survey Scale	Data Review	Data Use	Instructional Planning	Instructional Differentiation
School-level mediators				
Instructional leaders' abilities	0.19 ‡	0.15 ‡	0.21 ‡	0.10 **
Professional culture				
CPT discussions	0.32 ‡	0.23 ‡	0.35 ‡	0.18 ‡
General collegiality	0.16 **	0.12 **	0.22 ‡	0.11 **
Achievement culture	0.19 ‡	0.16 ‡	0.22 ‡	0.15 ‡
Teacher-level mediators				
Assessment/data attitudes	0.28 ‡	0.34 ‡	0.30 ‡	0.17 ‡
Confidence				
Data use	0.28 ‡	0.39 ‡	0.37 ‡	0.25 ‡
Instructional planning	0.23 ‡	0.35 ‡	0.53 ‡	0.36 ‡

Notes: Outcome scales were standardized within the teacher sample; results are reported in standard deviation units. Estimates are generated from two-level multilevel models where each teacher practice outcome is regressed separately on each mediator. Covariates in each model (but not shown) include: treatment assignment, district, data collection wave, and unbalanced "pair" dummy at the school level, and total teaching experience and highest degree at the teacher level.
‡ $p < 0.01$; ** $p < 0.05$.

For teacher-level mediators, a one standard deviation change is related to a change in teacher practice outcomes generally in the range of one-sixth to one-half of a standard deviation (all $p < 0.01$) (table 4.12). On average, teacher confidence in instructional planning shows the strongest, positive relationship with teacher practice

outcomes. As examples, controlling for district-, school-, and teacher-level covariates, a one standard deviation change in:

- confidence in using data in various ways is related to a 0.39 standard deviation change in the frequency teachers use data ($p < 0.01$)
- confidence in various instructional planning techniques is related to a 0.53 standard deviation change in the frequency teachers use various instructional planning strategies ($p < 0.01$).

Although these results are correlational, they suggest that a focus on these school- and teacher-level mediators is a potentially useful strategy for improving teachers' data-based instructional practices. Furthermore, it supports the links between school-level conditions and teacher practices in this study's conceptual model and the ANet logic model. However, evidence of generally larger correlations between teacher-level mediators and outcomes (compared to school-level mediators and outcomes) suggest that a greater focus by ANet on teacher characteristics – particularly confidence – could prove fruitful for improving teachers' data-based instructional practices.

Based on the results in table 4.12, research question three focuses on the teacher practice outcomes of data review and data use because they were the two outcomes for which ANet had a statistically significant, positive impact after two years. While it is possible that off-setting directional impacts between 1) ANet and the mediators and 2) the mediators and instructional practice outcomes could explain the overall null direct effects, results show that the impacts of ANet on the hypothesized mediators were all positive (tables 4.9-4.11) and that the relationship between mediators and instructional practice outcomes were also positive (table 4.12), making this scenario unlikely.

School Mediation Models. Recall from the results of research question one that, after controlling for all covariates, the frequency with which teachers' in ANet schools reported reviewing interim assessment data alone or with others was nearly a half a standard deviation higher than in control schools ($p < 0.01$) (table 4.7, model 1). These results are repeated in table 4.13 (model 1). In model 2, the block of hypothesized school-level mediators is added. Controlling for treatment-group assignment, all level-one and level-two covariates, and the other school-level mediators, only the frequency of collegial discussion during common planning time is significantly related to the frequency of teachers' data review (0.31 *sd*, $p < 0.01$) (table 4.13, model 2).

The inclusion of the block of hypothesized school-level mediators reduces the estimated impact of ANet on the frequency with which teachers' review data by about 30 percent: from 0.45 to 0.32 (table 4.13, model 1 versus 2). However, the impact is still large, positive, and statistically significant ($p < 0.01$), suggesting that, at best, the impact of ANet on teachers' frequency of reviewing data is only partially mediated by instructional leadership and school professional and achievement culture. Including the block of school-level (L2) mediators in model 2 explains an additional 65 percent of the between-school variance in teachers' frequency of data review as compared to model 1 (table 4.13).

In research question one, the frequency with which teachers used data in various ways was about one-quarter of a standard deviation higher in ANet schools after controlling for all covariates ($p < 0.01$) (table 4.13, model 5). In the school-mediation model, only the frequency of collegial discussion during common planning time is a

significant predictor of teachers' data use (0.20 *sd*, $p < 0.01$) all else being equal (table 4.13, model 6). The school-mediation model also suggests that the block of hypothesized school mediators explains, at most, one-third of the impact of ANet on data use, which declines from 0.25 to 0.17 *sd* (table 4.13, model 5 versus 6). The overall impact of ANet on the frequency with which teachers use data is still positive and statistically significant (0.17 *sd*, $p < 0.05$), suggesting that the impact of ANet is, at best, only partially mediated by school leadership and cultural conditions. Including the block of school-level (L2) mediators in model 6 appears to explain the remaining between-school variance in teachers' data use practices as compared to model 5 (table 4.13).

Teacher Mediation Models. Next, the block of hypothesized school-level mediators are removed and the block of hypothesized teacher-level mediators are added to the initial impact models from research question one. Results show that teachers' individual attitudes toward interim assessments and assessment data, as well as their confidence in using data in various ways are each positively related to both the frequency with which they review and use data (all $p < 0.01$) (table 4.13, models 3 and 7). Although teachers' confidence in various instructional planning activities is not predictive of the frequency they reported reviewing data, it is a positive predictor of the frequency they report using data ($p < 0.01$) (table 4.13, model 7). This makes sense if teachers' instructional confidence is not a prerequisite for reviewing data, but is for using that data for instructional improvement. The magnitude of the relationships between teacher characteristics and data practices are non-negligible, ranging between 0.13 and 0.24 *sd*.

Table 4.13. Teacher Practice Mediation Results

	Data Review				Data Use			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Variable	Impact Model	School Mediation Model	Teacher Mediation Model	Combined Mediation Model	Impact Model	School Mediation Model	Teacher Mediation Model	Combined Mediation Model
Fixed effect								
Assigned to treatment: school	0.45 ‡	0.32 ‡	0.37 ‡	0.29 ‡	0.25 ‡	0.17 **	0.18 **	0.15 **
	0.112	*0.094*	*0.102*	*0.092*	*0.089*	*0.082*	*0.078*	*0.074*
School-level mediators								
Instructional leaders' abilities		0.03		-0.05		0.03		-0.08
		0.084		*0.083*		*0.074*		*0.068*
CPT discussions		0.31 ‡		0.26 ‡		0.20 ‡		0.13 **
		0.071		*0.069*		*0.064*		*0.059*
General collegiality		-0.09		-0.07		-0.10		-0.07
		0.093		*0.090*		*0.082*		*0.075*
Achievement culture		0.08		0.08		0.11		0.12
		0.089		*0.087*		*0.080*		*0.072*
Teacher-level mediators								
Assessment/data attitudes			0.22 ‡	0.20 ‡			0.24 ‡	0.23 ‡
			0.039	*0.039*			*0.037*	*0.037*
Data use confidence			0.18 ‡	0.19 ‡			0.24 ‡	0.24 ‡
			0.048	*0.048*			*0.046*	*0.046*
Instructional planning confidence			0.07	0.06			0.14 ‡	0.13 ‡
			0.047	*0.047*			*0.045*	*0.045*
District								
Chelsea	-0.19	-0.20	-0.11	-0.11	-0.14	-0.17	-0.08	-0.10
	0.300	*0.245*	*0.272*	*0.240*	*0.230*	*0.210*	*0.202*	*0.191*
Chicago	0.70 ‡	0.44 **	0.51 **	0.33	0.27	0.15	0.01	-0.01
	0.245	*0.221*	*0.229*	*0.217*	*0.217*	*0.208*	*0.195*	*0.191*
Jefferson Parish	-0.14	-0.45 ‡	-0.20	-0.44 ‡	-0.11	-0.31 **	-0.22 *	-0.30 **
	0.160	*0.159*	*0.147*	*0.155*	*0.132*	*0.144*	*0.117*	*0.131*
Springfield	0.30 *	0.05	0.22	0.05	0.29 **	0.10	0.14	0.07
	0.169	*0.147*	*0.154*	*0.143*	*0.132*	*0.128*	*0.117*	*0.116*
Data collection wave two: school	0.40 **	0.37 ‡	0.41 ‡	0.38 ‡	0.38 ‡	0.38 ‡	0.41 ‡	0.42 ‡
	0.161	*0.138*	*0.147*	*0.134*	*0.126*	*0.118*	*0.111*	*0.108*
Unbalanced pair dummy: school	-0.35	-0.12	-0.22	-0.11	-0.61 **	-0.37	-0.41 *	-0.32
	0.353	*0.294*	*0.319*	*0.290*	*0.260*	*0.246*	*0.229*	*0.225*
Years of teaching experience (total): teacher	0.01 ‡	0.01 ‡	0.01	0.01	0.01 ‡	0.01 ‡	0.00	0.00
	0.004	*0.004*	*0.004*	*0.004*	*0.004*	*0.004*	*0.004*	*0.004*
Highest degree: teacher								
Master's	0.21 **	0.21 **	0.12	0.12	0.06	0.07	-0.06	-0.05
	0.099	*0.098*	*0.093*	*0.092*	*0.101*	*0.099*	*0.089*	*0.088*
Doctorate	-0.18	-0.18	0.01	0.01	-0.43	-0.38	-0.15	-0.14
	0.350	*0.347*	*0.330*	*0.328*	*0.363*	*0.359*	*0.321*	*0.320*
Random effect								
School (intercept)	-1.05 ‡	-0.77 ‡	-0.84 ‡	-0.64 ‡	-0.77 ‡	-0.62 ‡	-0.51 ‡	-0.47 ‡
	0.220	*0.206*	*0.202*	*0.199*	*0.185*	*0.187*	*0.163*	*0.169*
Variance Components								
L1	0.748	0.744	0.646	0.644	0.824	0.811	0.625	0.624
L2	0.094	0.033	0.073	0.038	0.020	0.000	0.016	0.004
Additional Variance Explained (%)								
L1		1%	14%	14%		2%	24%	24%
L2		65%	22%	60%		100%	109%	79%
Model statistics								
n	559	559	555	555	569	569	564	564
Number of groups	67	67	67	67	67	67	67	67
Wald χ^2	70.09	129.62 ‡	175.79 ‡	219.2 ‡	81.76	131.34 ‡	283.19 ‡	319.04 ‡

Notes: Outcome scales were standardized within the teacher sample; results are reported in standard deviation units. Estimates are reported on the top row for each predictor. Standard errors are reported below, in italics. Omitted district = Boston; omitted degree = bachelor's; and data collection wave one = 1, wave two = 2. Additional variance explained for models 2-4 is in comparison to model 1, and for models 6-8 is in comparison to model 5.

‡ $p < 0.01$; ** $p < 0.05$. * $p < 0.10$.

When included as their own block (i.e., excluding school-level mediators), the three teacher-level mediators appear reduce the impact of ANet on teachers' use of data by a slightly greater amount than their review of data – about 28 percent of the impact of ANet on the frequency teachers use data and 17 percent of the frequency teachers review data (table 4.13, models 7 and 3, respectively) – but account for less of the ANet impact on teachers' data review and use than did school-level factors. However, it is unlikely that differences in the proportion of the ANet impact that is accounted for 1) in models with the same mediators, but different outcomes or 2) between models within outcomes are statistically significant. Lastly, including the block of teacher-level (L1) mediators in models 3 and 7 explains an additional 14 percent of the within-school variance in teachers' data review practices and 24 percent of the within-school variance in teachers' data use practices as compared to models 1 and 4, respectively (table 4.13).

Combined Mediation Models. In models including both school- and teacher-level mediators, the frequency of teachers' discussions during common planning time remains a strong, positive predictor of the frequency with which teachers review ($p < 0.01$) (model 4) and use interim assessment data ($p < 0.05$) (model 8) (table 4.13). Specifically, a one standard deviation change in the frequency of teachers' collegial discussions is associated with a change in the frequency with which teachers review data of about one-quarter of a standard deviation ($p < 0.01$) (table 4.13, column 4) and a change of about one-eighth of a standard deviation in frequency teachers use data ($p < 0.05$) (table 4.13, column 8).

Likewise, teachers' attitudes toward interim assessments and assessment data, as well as their confidence in using data in various ways, remain positive predictors of the frequency with which they review and use data (all $p < 0.01$) (table 4.13, models 4 and 8). Specifically, controlling for all other mediators and covariates, a one standard deviation change in teachers' attitudes towards and confidence in using interim assessment data are associated with a change in the frequency of teachers' data review of about one-fifth of a standard deviation (all $p < 0.01$, table 4.13, model 4) and a change in the frequency of teachers' data use of about one-quarter of a standard deviation (all $p < 0.01$, table 4.13, model 8). As before, teachers' confidence in various instructional planning tasks remains a positive predictor of only the frequency of their data use ($p < 0.01$) (table 4.13, model 8).

The addition of the block of hypothesized teacher-level mediators reduces the size of the estimated impact of ANet on teachers' data-based practices over and above that of the school mediators by a relatively small amount: a 29 to 34 percent reduction of the impact of ANet on the frequency teachers review data and a 33 to 39 percent reduction of the impact of ANet on the frequency teachers use data (table 4.13, models 2 versus 4, and 6 versus 8, respectively). This is not surprising given the small program impacts on teacher-level mediators seen in table 4.11. Including both blocks of mediators in model 4 explains an additional 60 percent of the between-school variance and an additional 14 percent of the within-school variance in teachers' data review practices as compared to model 1 (table 4.13). Including both blocks of mediators in model 8 explains an additional 79 percent of the between-school variance and an additional 24 percent of the

within-school variance in teachers' frequency of data use as compared to model 4 (table 4.13).

All else being equal, teachers in Jefferson Parish appear to review and use data less frequently than their Boston peers. This pattern was seen in the analyses from research question one and continues to hold after adjusting for hypothesized school- and teacher-level mediators. In models controlling for all hypothesized mediators, teachers in Jefferson Parish review data less often than teachers in Boston by a magnitude of nearly one-half of a standard deviation (-0.44 *sd*, $p < 0.01$, model 4) and use data less often by a magnitude of about one-third of a standard deviation (-0.30 *sd*, $p < 0.05$, model 8). Teachers in Chicago and Springfield review and use data more frequently than their Boston counterparts; however, after controlling for hypothesized mediators and other covariates, the differences are not statistically significant (models 4 and 8, table 4.13).

Teachers in wave two review data (0.38 *sd*, $p < 0.01$) and use data (0.42 *sd*, $p < 0.01$) more frequently than wave-one teachers after controlling for all hypothesized mediators. The difference is larger than the treatment-control group difference in the frequency of teachers' data review and data use in the fully specified models (models 4 and 8, table 4.13). Finally, higher levels of teacher experience and education are not associated with more frequent review and use of data in the models controlling for all hypothesized mediators (models 4 and 8, table 4.13).

The models in table 4.13 were re-run including an interaction between school treatment assignment and district. The omnibus F-test indicates that the interaction is statistically significant in the school mediation model (model 2) and teacher mediation

model (model 3) predicting teachers' frequency of data review ($p < 0.05$) (table 4.14). The results also show a significant negative interaction in Chelsea and Chicago for most data review models indicating a lower, or less positive, effect of ANet on data review in these districts compared to Boston (most $p < 0.05$ or better). The effect of ANet on the frequency teachers review data in Boston remains significant in the school-, teacher-, and combined-mediation models (all $p < 0.01$). The impact of ANet on the frequency teachers review data in Jefferson Parish remains significant in the teacher-mediation model ($p < 0.05$).

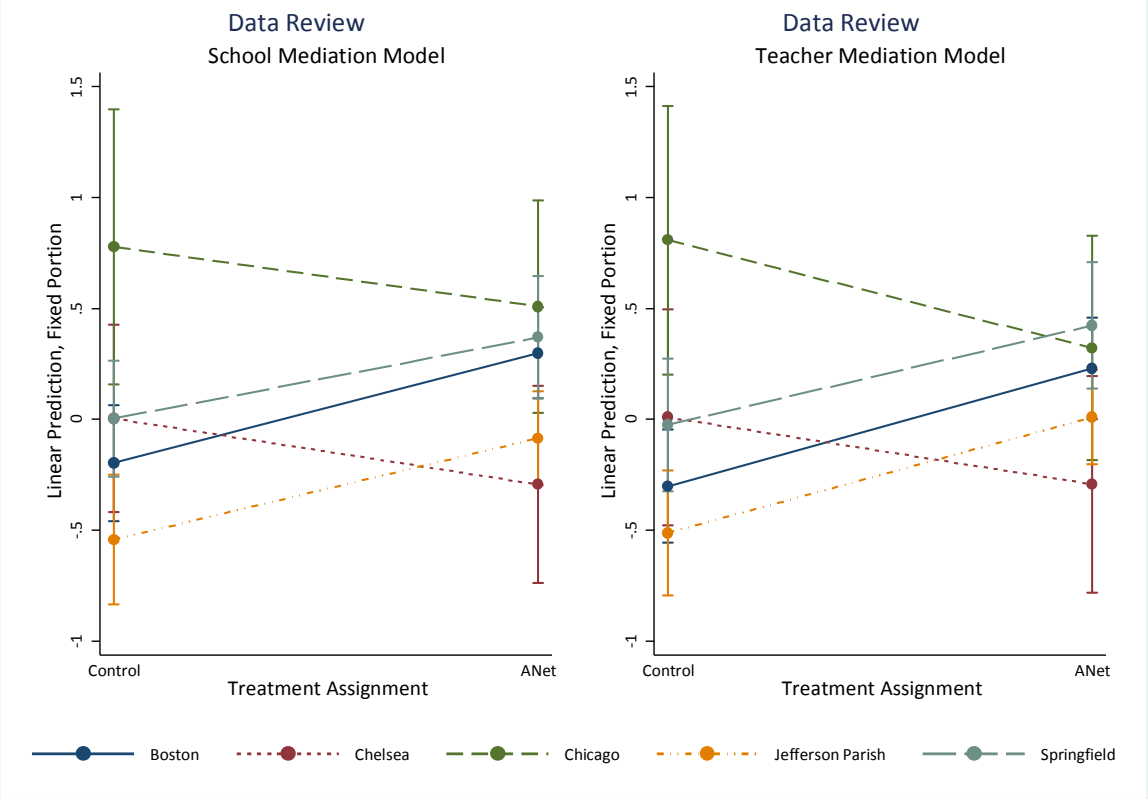
In this example, consider the teacher mediation model for data review (model 3, table 4. 14) and the plot of the interaction between treatment assignment and district for that model (right, figure 4.3). Comparing the line for Chelsea (in red, dotted line) with the line for Boston (in blue, solid line) in figure 4.3 (right panel), the main effect of ANet on teachers' confidence in various instructional planning strategies appears similar to Boston (i.e., "averaging" the treatment and control estimates within each district, Chelsea and Boston are very similar). This corresponds to the near-zero main effect for Chelsea teachers' instructional planning confidence of 0.31 (*ns*) in table 4.14 (model 3). However, the slope of line for Chelsea is negative (ANet teachers in Chelsea have a lower frequency of reviewing data than control-school teachers in Chelsea) in figure 4.3 (right) indicating a negative treatment effect of -0.30 (*ns*).

Table 4.14. Teacher Practice Mediation Results with Treatment by District Interaction

	Data Review				Data Use			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Variable	Impact Model	School Mediation Model	Teacher Mediation Model	Combined Mediation Model	Impact Model	School Mediation Model	Teacher Mediation Model	Combined Mediation Model
Fixed effect								
Assigned to treatment: school	0.59 ‡	0.50 ‡	0.53 ‡	0.47 ‡	0.27 *	0.22	0.21	0.22
	0.168	*0.152*	*0.164*	*0.155*	*0.147*	*0.147*	*0.137*	*0.133*
School-level mediators								
Instructional leaders' abilities		0.01		-0.06		0.03		-0.07
		0.077		*0.079*		*0.075*		*0.068*
CPT discussions		0.23 ‡		0.20 ‡		0.16 **		0.11 *
		0.070		*0.072*		*0.069*		*0.062*
General collegiality		-0.08		-0.06		-0.07		-0.05
		0.086		*0.087*		*0.084*		*0.075*
Achievement culture		0.10		0.10		0.10		0.11
		0.085		*0.087*		*0.083*		*0.075*
Teacher-level mediators								
Assessment/data attitudes			0.21 ‡	0.19 ‡			0.24 ‡	0.23 ‡
			0.039	*0.039*			*0.037*	*0.037*
Data use confidence			0.18 ‡	0.18 ‡			0.24 ‡	0.24 ‡
			0.048	*0.048*			*0.046*	*0.046*
Instructional planning confidence			0.06	0.05			0.14 ‡	0.13 ‡
			0.047	*0.047*			*0.046*	*0.046*
District								
Chelsea	0.36	0.20	0.31	0.21	0.13	0.00	0.03	-0.02
	0.295	*0.259*	*0.291*	*0.268*	*0.249*	*0.250*	*0.234*	*0.225*
Chicago	1.42 ‡	0.97 ‡	1.11 ‡	0.76 **	0.77 **	0.46	0.36	0.21
	0.346	*0.351*	*0.332*	*0.344*	*0.369*	*0.385*	*0.331*	*0.343*
Jefferson Parish	-0.19	-0.35	-0.21	-0.31	-0.29	-0.37 *	-0.32 *	-0.31
	0.203	*0.216*	*0.196*	*0.217*	*0.186*	*0.212*	*0.171*	*0.192*
Springfield	0.34 *	0.20	0.28	0.18	0.32 *	0.20	0.22	0.20
	0.204	*0.186*	*0.199*	*0.189*	*0.178*	*0.181*	*0.165*	*0.163*
Treatment x District Interaction								
Treatment*Chelsea	-1.10 ‡	-0.79 ‡	-0.83 ‡	-0.64 **	-0.53 **	-0.31	-0.21	-0.14
	0.322	*0.282*	*0.321*	*0.296*	*0.264*	*0.268*	*0.252*	*0.242*
Treatment*Chicago	-1.19 ‡	-0.77 *	-1.02 **	-0.64	-0.70	-0.43	-0.51	-0.32
	0.444	*0.436*	*0.432*	*0.435*	*0.446*	*0.460*	*0.404*	*0.413*
Treatment*Jefferson Parish	0.07	-0.04	-0.01	-0.10	0.26	0.15	0.16	0.04
	0.239	*0.221*	*0.233*	*0.225*	*0.211*	*0.215*	*0.196*	*0.194*
Treatment*Springfield	-0.01	-0.13	-0.08	-0.14	-0.02	-0.12	-0.14	-0.20
	0.267	*0.238*	*0.261*	*0.244*	*0.229*	*0.229*	*0.214*	*0.206*
Data collection wave two: school	0.42 ‡	0.39 ‡	0.42 ‡	0.40 ‡	0.39 ‡	0.38 ‡	0.41 ‡ ‡	
	0.136	*0.125*	*0.133*	*0.127*	*0.117*	*0.119*	*0.109*	*0.108*
Unbalanced pair dummy: school	-0.21	-0.04	-0.12	-0.04	-0.53 **	-0.35	-0.38 *	-0.31
	0.286	*0.258*	*0.284*	*0.269*	*0.236*	*0.247*	*0.223*	*0.221*
Years of teaching experience	0.01 ‡	0.01 **	0.00	0.00	0.01 ‡	0.01 ‡	0.00	0.00
(total): teacher	*0.004*	*0.004*	*0.004*	*0.004*	*0.004*	*0.004*	*0.004*	*0.004*
Highest degree: teacher								
Master's	0.21 **	0.22 **	0.12	0.13	0.06	0.07	-0.05	-0.04
	0.099	*0.098*	*0.093*	*0.092*	*0.100*	*0.099*	*0.089*	*0.088*
Doctorate	-0.25	-0.22	-0.04	-0.01	-0.48	-0.41	-0.20	-0.17
	0.351	*0.349*	*0.331*	*0.330*	*0.364*	*0.361*	*0.323*	*0.322*
Random effect								
School (intercept)	-1.13 ‡	-0.94 ‡	-0.92 ‡	-0.79 ‡	-0.79 ‡	-0.66 ‡	-0.51 ‡	-0.50 ‡
	0.206	*0.212*	*0.200*	*0.213*	*0.187*	*0.208*	*0.172*	*0.187*
Variance Components								
L1	0.754	0.749	0.647	0.648	0.822	0.805	0.623	0.624
L2	0.039	0.010	0.045	0.023	0.004	0.000	0.012	0.002
Additional Variance Explained (%)								
L1		1%	14%	14%		2%	24%	24%
L2		75%	-16%	40%		100%	-218%	56%
Model statistics								
n	559	559	555	555	569	569	564	564
Number of groups	67	67	67	67	67	67	67	67
Wald χ^2	112.56 ‡	166.04 ‡	204.57 ‡	240.83 ‡	112.25 ‡	136.18 ‡	296.11 ‡	329.02 ‡
F (interaction)	20.92 ‡	10.43 **	12.34 **	5.86	11.22 **	3.94	4.54	2.11

Notes: Outcome scales were standardized within the teacher sample; results are reported in standard deviation units. Estimates are reported on the top row for each predictor. Standard errors are reported below, in italics. Omitted district = Boston; omitted degree = bachelor's; and data collection wave one = 1, wave two = 2. Additional variance explained for models 2-4 is in comparison to model 1, and for models 6-8 is in comparison to model 5.
‡ $p < 0.01$; ** $p < 0.05$. * $p < 0.10$.

Figure 4.3. Interaction Between District and Treatment Assignment for School and Teacher Meditator Models with Data Review as Outcome



Having controlled for this interaction, teachers in Chelsea now report higher frequencies of data review and use (main effect of district) compared to Boston teachers (table 4.14). Teachers in Chicago still report higher frequency of each teacher practice outcome in comparison to Boston teachers, but the magnitudes have roughly doubled for data review (table 4.14) compared to the models without the interaction term. District patterns in Jefferson Parish and Springfield remain largely the same as the previous models (table 4.13) that did not include the treatment by district interactions.

Research Question Four: Exploratory Pre-Conditions Model

Does the effect of ANet on teachers' data use and instructional practices vary by the schools' baseline implementation "readiness" rating?

In the final research question, the role of school-level conditions such as leadership and culture move from that of mediator, to one of potential moderators. Recall that during the school recruitment process ANet administered a school readiness survey. The survey included questions that fell under nine broad categories and were thought to be linked to a school's readiness to successfully implement the ANet program. Across the nine categories, schools were given a score of 1-3 based on how their responses aligned to a predetermined scoring rubric.

This analysis capitalizes on the scores in the five survey categories thought to be most related to the hypothesized mediators of interest in this study. These categories are: school opt-in to the program; school's prioritization and organization of the program; dedication of school leadership; rigorous, complete, and aligned standards and curriculum; and a commitment to making time in the school's schedule for implementing ANet practices (see appendix C for descriptions of each measure and rubric level). Scores in these categories were used to place matched pairs of schools into two groups: pairs with an average score in the top versus bottom of the distribution. Specifically, each school's total score was first calculated. Next, an average was generated for each matched pair of schools. Finally, pairs were assigned to the top and bottom readiness groups based on this average in order to maintain the treatment assignment balance within readiness groups.

While the potential range of pair-average scores is 5-15, actual scores ranged from 6.5-15. The pair-average scores were negatively skewed with nearly one-third of the pairs reporting a score of 13.5 (modal value). Because of the small number of pairs and nonnormality of the data, two groups were constructed: 1) pairs with average scores of 13.5 or above were assigned to the "higher" readiness group (schools = 34; teachers = 296), and 2) pairs with average scores below 13.5 were assigned to the "lower" readiness group (schools = 33; teachers = 320).

With these groups established, the four models in research question one were re-estimated separately by readiness group in order to determine whether baseline readiness was a useful predictor of ANet's impact on teachers' data-based instructional practices. Specifically, the results show the estimated impact of ANet on each of the teacher practice outcomes when models were run separately for teachers in schools in the higher- versus lower-readiness groups. As before, estimates were generated from two-level models that included indicators for treatment assignment, district, data collection wave, and the Chelsea unbalanced "pair," and teacher-level covariates for total teaching experience and highest degree. Table 4.15 reports only the coefficient on the treatment indicator; coefficients for all covariates are omitted. For comparison, the "all schools" column provides the impact estimates from research question one (table 4.7).

Compared to the estimates in the full sample ("all schools"), impacts of ANet on data review and data use are larger and more positive in higher-readiness schools and smaller in lower-readiness schools (table 4.15). Impacts on instructional planning are

similar in the higher- and lower-readiness schools. In contrast, impacts on differentiation are larger and more negative in lower-readiness schools.

Table 4.15. Teacher Practice Impact Results by Baseline School Readiness Rating

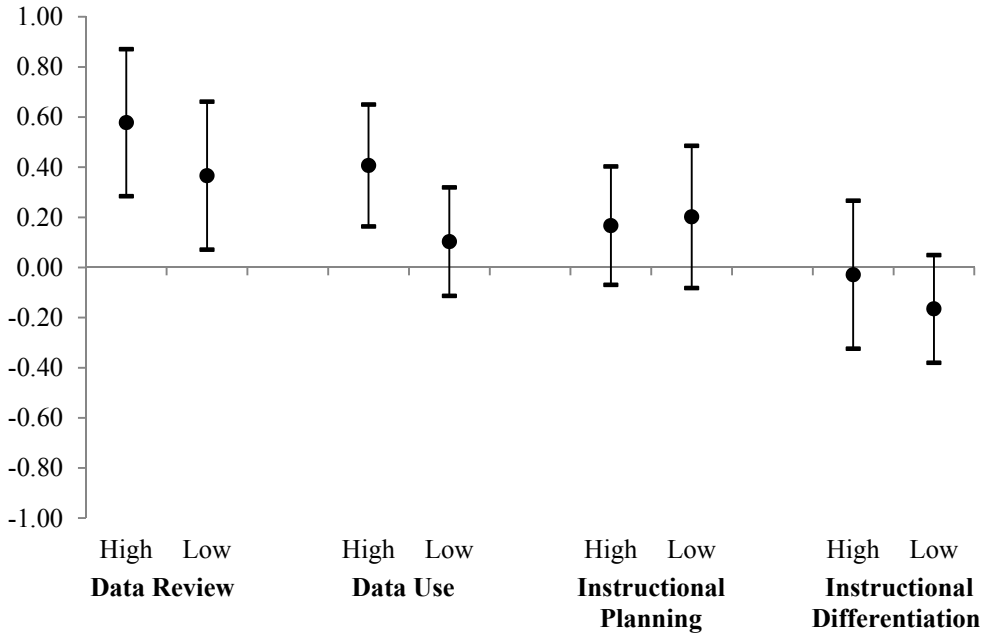
Teacher Practice Outcome	All Schools	Pair-Average Baseline School Readiness Group		Difference between groups
		High	Low	
Data review	0.45 ‡	0.58 ‡	0.37 **	0.21
	0.112	*0.150*	*0.151*	
Data use	0.25 ‡	0.41 ‡	0.10	0.30
	0.089	*0.124*	*0.110*	
Instructional planning	0.16	0.17	0.20	-0.04
	0.096	*0.121*	*0.145*	
Instructional differentiation	-0.10	-0.03	-0.17	0.14
	0.088	*0.151*	*0.110*	

Notes: Outcome scales were standardized within the teacher sample; results are reported in standard deviation units. Estimates are reported on the top row for each outcome. Standard errors are reported below, in italics. Estimates are generated from two-level multilevel models. Covariates in each model (but not shown) include: treatment assignment, district, data collection wave, and unbalanced "pair" dummy at the school level, and total teaching experience and highest degree at the teacher level. Omitted district = Boston; omitted degree = bachelor's; and data collection wave one = 1, wave two = 2.
‡ $p < 0.01$; ** $p < 0.05$.

The off-setting difference between readiness groups makes mathematical sense. However, accounting for uncertainty in estimates, it cannot be said that the differences in the estimated impacts of ANet in higher- versus lower-readiness schools on any of the four teachers practice outcomes are significantly different (figure 4.4). Though the differences between readiness groups are not statistically significant, results indicate that, in schools that were rated as more ready to partner with ANet, ANet's estimated impact on data-related practices after two years is slightly larger and more positive. Overall, this

suggests that these five baseline readiness items may predict where ANet will have a larger impact on some data-based practices. It also suggests that the overall negative impact (*ns*) of ANet on instructional differentiation is driven more by lower readiness schools. Though these results may not be considered statistically "important," the substantive implication are that lower readiness schools may require additional support to ensure that data are used effectively to target students' learning individual needs.

Figure 4.4. Estimates of the Impact of ANet on Each Teacher Practice Outcome, by School Readiness Group



SUMMARY

The quantitative results suggest several patterns that are both interesting in their own right and served to focus the qualitative analysis presented in the following chapter. Initially, the goal of the qualitative analysis was to add explanatory value and context for the quantitative findings. However, the findings suggest that a large portion of the

variance in teacher practices is not explained by the focal school- and teacher-level mediators. Therefore, a second purpose for the qualitative analysis emerged: to seek evidence of other conditions that may be associated with teachers' adoption of data-based instructional practices. The following framing questions were used to guide the qualitative analysis and serve as segue into those results.

Teacher Practice Outcomes. Based on survey analyses, there is strong evidence to suggest that, after two years, ANet had an impact on the frequency with which teachers reviewed and used data, but that this did not translate to larger impacts on instructional planning or differentiation of instruction. With more frequent use of whole-class instruction by ANet teachers, it might be that teachers are identifying skills on which the entire class is struggling. However, with detailed interim assessment data made available at the student level, one would expect to see this translate into greater instructional differentiation according to students' individual needs by year two. Qualitative analysis focused on the following questions.

- Did teachers describe data and instructional practices that prior research has shown to be effective?
- Is there evidence of a focus, order, or hierarchy of the implementation of these outcomes by the ANet program?

Mediator Impact Models. The quantitative results also suggest that, after two years, ANet had a larger impact on school-level mediators as compared to the teacher-level mediators of interest in this study. This pattern in findings is not surprising given ANet's logic model. ANet does not have an explicit goal of changing teacher attitudes toward interim assessment and assessment data, nor does its coaching model focus

directly on building teachers' data or instructional skills and confidence. However, its logic model does explicitly identify key intermediate outcomes related to building school culture and school leaders' instructional leadership capacity. These results provided the basis for several key questions guiding the qualitative analysis:

- How did leaders and teachers talk about instructional leadership and various facets of school culture?
- What did teachers say about their own attitudes and abilities, and is it consistent with characteristics shown to foster effective data-based practices?

Teacher Practice Mediation Models. Results from the survey analysis suggest that, compared to teacher-level characteristics, hypothesized school-level mediators likely explain more of the impact of ANet on teachers' data practices. In models that include both blocks of mediators, the school- and teacher-level mediators reduce the estimated direct impact of ANet on each data-based practice by just over one-third. Controlling for treatment assignment and all other mediators and covariates, the frequency of teachers' collegial discussions during common planning time, as well as their attitudes towards and confidence in using data remain important predictors of teachers' data practices. Again, these results suggest several key areas of qualitative analysis:

- When talking about data use or instructional practice, what school-level facilitators or barriers were discussed?
- When talking about data use or instructional practice, what did teachers cite as individual (teacher-level) facilitators or barriers?
- Are there school- or teacher-level conditions that offer context for the impact of ANet on some teacher practice outcomes, but not others?
- Are there other notable conditions that might account for the unexplained variation in teachers' data practices?

Several other findings emerged that are less likely to be illuminated through qualitative analysis of the year-two treatment-school leader and teacher interviews. First, the results of research question four point to the potential importance of identifying markers for successful implementation and measuring schools on these markers at the outset. Much like ANet's focus on tailoring instruction to students' needs and, to the extent that these baseline markers predict successful uptake of practices, some schools may require greater support in implementing the program.

Additionally, it is clear from the results of research questions one and three that there was variation in the impacts of ANet on teachers' data-based instructional practices across districts (both in magnitude and direction). This finding is important for two reasons. First, it speaks to the need for sufficiently powered subgroup analyses as a means to unpacking overall null impacts on instructional practices. Second, district variation in outcomes points to important interactions between the treatment and contextual differences such as characteristics of schools in which the intervention was implemented or the teachers who were exposed to the treatment (Weiss, Bloom, and Brock, 2013). Ignoring the additional noise introduced by a treatment-by-district interaction can lead to more type I errors which contribute to a lack of replicability in research (Benjamini, 2015).

CHAPTER FIVE: QUALITATIVE ANALYSES & RESULTS

The quantitative results in chapter four suggest that ANet teachers used various data-related practices with greater frequency than their control-school counterparts, but no difference was found in the frequency with which they used various instructional planning strategies or differentiated their instruction. Results also suggest that school- and teacher-level mediators play only a modest role in explaining ANet teachers' more frequent data-related practices. The purpose of this chapter is to provide a context for these quantitative results and a different perspective on the conceptual framework underpinning this study. Specifically, the findings in this chapter describe the types of data-based instructional practices reported by a subset of ANet teachers and the roles played by instructional leadership, school culture, and teacher attitudes and confidence in their adoption of these practices. Furthermore, given the evidence of substantial unexplained variance in teacher' data-based practices, these results also describe other factors – not accounted for in the conceptual model – that may explain the pattern of results seen in the previous chapter.

This chapter capitalizes on the larger evaluation's mixed methods design by analyzing teacher and leader interview data from 12 year-two ANet schools: 3 from each of the 4 geographic networks in the i3 evaluation. In total, interviews were conducted with 19 leaders (e.g., principals, assistant principals, and instructional or data leaders) and 16 teachers. Additionally, focus groups with teachers in ANet schools were held in each of the 5 participating districts.[1] Since the interviews are a secondary data source, the

[1] Teachers from all ANet schools in the i3 study were invited to participate in the focus group in their district.

transcripts were fully recoded and reanalyzed with this study's purposes in mind. Due to the secondary nature of the analysis, findings are often summarized from interviewee comments that arose spontaneously. Exceptions are noted where the protocol question explicitly asked about a topic.

This chapter first presents the results from teacher interviews, followed by the results from leader interviews. Because this chapter reviews findings related to each outcome and mediator and are, therefore, quite dense, it closes with full summary of findings. A glossary of ANet terms is provided in appendix D for the many references to ANet-specific practices or program components made throughout this chapter.

ANET TEACHERS' DATA-BASED PRACTICES

In chapter two, instructional data use was defined as the dual process of (1) analyzing data (2) and using the results to inform instructional planning and remediation (Faria, et al., 2012). Recall that the quantitative results for research question one showed no differences in instructional planning or differentiation practices between ANet and control-school teachers, but greater frequency in ANet teachers' review and use of data. For context on these quantitative results, this section provides evidence of the types of data-based instructional practices ANet teachers were using in year two of the evaluation. This section concludes with a discussion of some of the barriers teachers faced in making effective changes in practice.

Data Practices

Across the majority of interviews, teachers in ANet site visit schools had generally positive opinions of the usefulness of the interim assessments for helping them identify gaps in student learning. Several teachers also pointed out that having student results returned to them quickly was particularly useful for instructional decision making (T4, T6, T11, T14, T41, CHLS-FG SPS-FG)[2]. Their comments also suggested that this may have been a feature of ANet that was different from other interim assessment programs they were using or had used previously. Some focus group teachers said that they were able to identify students' gaps in knowledge sooner because they received detailed results so quickly (BPS-FG, CHLS-FG). Before ANet, one teacher explained that she and her peers would review the prior year's state summative assessment data to inform the instructional units they would teach during the current year. However, she felt that wasn't enough (T4). Another teacher appreciated the "ongoing" nature of the ANet assessment cycle and the opportunity this provided to immediately inform instruction compared to state summative assessment data that were received too late (T6).

Additionally, several teachers stated their appreciation of the detailed information that the ANet assessment results provided on student learning (T1, T4, T6). They noted that, because the ANet data could be disaggregated to show student performance by assessment item, items under the same standard, and for individual students, they no longer had to guess whether their students had mastered a skill.

[2] Throughout the chapter, this notation references the specific interviews and focus groups cited in making a particular claim. Although the notation does not allow readers to know the identity of cited interviewees, it provides context on the frequency of supporting evidence for a given statement.

> [I]nstead of just feeling like they can't or they weren't yet able to infer, I had more of an understanding of, like, is it informational text, literary text, what kind of questions, what are the stems that are difficult for them. So I think it… allowed me to dive a little bit deeper into the standards. (T1)

> I think it's easy to have surface understanding and kind of get by, and until you really have that data and dig into it, you're pretty much guessing as a teacher what the kids are understanding and taking away. (T6)

Another teacher explained that the ANet results opened her eyes to the number of students who hadn't yet mastered a skill, students whose lack of mastery might have previously been overlooked (T29).

As a tool for data analysis, ANet provides teachers with a misconceptions guide. There were mixed feelings among the teachers who referenced this resource. Three teachers found the distractor information useful for identifying the aspect of the skill students were missing (T14, T24, T33). However, several others felt it was difficult to use item-level distractor information in planning if there wasn't a single, obvious misconception students held (i.e., the distribution of wrong answers was split across the various options) (T19) or that students had no reason for selecting particular distractors, they simply guessed. Citing both possibilities, one teacher explained "trying to predict why the kid picked that, because unless you have a conversation with each kid, you really don't know what they were thinking or if they were thinking at all." (CHI-FG)

Instructional Planning & Practices

Teacher interviews also suggest that the detail provided by students' ANet results informed the planning of instruction. When prompted to share how ANet had influenced

their teaching practices, teachers generally mentioned two ways students' ANet assessment results were used: planning a reteach of past content to address students' learning gaps and backward planning of future instruction. One teacher explained how she incorporated both strategies,

> I'm using the feedback from the data a lot when I teach on a regular basis. It's not only, it gives me a picture of what my kids do, but it really helps me figure out what I'm going to do, like how I'm going to plan the next week's, you know, should I go back to things that I've probably taught too quickly or not, I don't know, not deeply enough. And also just to see how I'm going to plan a new topic. (T23)

However, where year-one of implementation appeared to focus on getting teachers to use data to inform their reteaching plan, backward planning of instruction seemed to be something ANet introduced during year two (T3, T4, T14, T41). As a result, planning a reteach appears to have been a more common practice than backward planning.

Reteaching. When asked about how their reteaching had changed with the implementation of ANet, the majority of teachers said that the ANet assessment data helped them "focus" ("drive," or "inform") the planning of their reteach. Several others said that the ANet assessment results helped them "pinpoint" where their students were struggling: for example, the specific skill or type of text with which students struggled, not just the broader standard. Teachers in one focus group felt that the ANet process forced them to acknowledge the need to reteach by exposing the students who hadn't mastered a skill. It seems that having these learning gaps made evident by students' ANet results made it harder for teachers to say "Okay, we've done that, check it off, move on." (JPS-FG)

Across interviews, several teachers mentioned ways they selected content for reteaching. In most instances, they described a skill or standard on which most of the class performed poorly on the ANet interim assessment (T14, T23, T19, T33, CHI-FG). Although it was unclear whether teachers were commonly using thresholds described by Goertz, Oláh, and Riggan (2009a), several teachers hinted at such a strategy. One explained,

> [I]f I'm looking at the data and I see that 80 percent of my class had an issue on a question on a GLE or regarding interpreting data, well, that means that I didn't do a good job, obviously. I mean, if 80 percent of the kids missed the question, it's not their fault, it's my fault. (T23)

Other teachers who described their reteach planning strategies mentioned that they tried to identify ways to teach a skill differently (T4, T33, T41, CHI-FG, SPS-FG).

> [ANet t]ook focus off "They [students] didn't get it," to us really looking at what we can do differently. How can I reteach it? (T4)

> [W]hen I look at my data, and especially with a skill that I see so many children did poorly with, then it makes me sit down and rethink. Maybe I taught it a way that the children weren't able to get it, so now I need to go in and see what I need to do as a teacher to get it across to the children so that they can understand it better. (T29)

However, several teachers acknowledged that they often didn't know another way to teach the same subject matter or skill (T33, BPS-FG). For example, one teacher felt that her reteaching was often no different than the initial instruction because she didn't know how to go about "doing this any different" and teachers in her school were on their own to figure out a new approach to reteaching a skill or standard (T33).

Reflecting on the research presented in chapter two, the types of planning strategies that teachers shared indicate that they may not have always been using – or

have been encouraged to use – practices that were particularly effective at improving student learning. For example, three teachers mentioned that there were times when their students performed poorly on a standard or skill that was measured by only one assessment item. Basing their reteach on these limited results made them uncomfortable, but was something they were encouraged to do by ANet or their school leader (T3, T6, T41). Specifically, they expressed concern as to whether a single item was sufficient evidence of lack of mastery.

> I question if there's only, like, one question for a standard, if that could really say that that child was struggling with summarizing because they didn't get one question on it. I feel like really, when you think about data, you should have a large population of data in order for you to make any firm decisions. So sometimes I feel like their [ANet's] messages … didn't really align with what I had always known about data. (T3)

Another teacher expressed a similar sentiment when items under the same standard or skill were structured very differently. She worried that students might have gotten the items wrong not because they lacked mastery of the common skill across the set, but rather some secondary – and uncommon – skill needed to answer each item correctly. (T30-1) In both situations, teachers questioned whether their reteaching time was well spent focusing on these skills or standards.

These and other teachers mentioned that they either opted to or were encouraged to develop a single grade-level reteaching plan for the same standard or skill. Some felt that this worked because the data showed that all students were struggling on the same skill and (or) it was a skill that teachers identified as being a priority standard on the state summative test. For example,

> [W]hen we look at the thing to reteach, we look to see which area that's the largest chunk for LEAP testing, to see where our kids did the—I don't want to say the poorest, but where we can move the most. We [the teacher team] kind of talk about it and which one fits the needs of all [students], which I know that's not necessarily how it's supposed to be done. (T30-2)

While this may have been appropriate in some classrooms (e.g., where the great majority of students hadn't mastered a skill), two BPS focus group teachers who were encouraged to generate grade-level reteaching plans felt that their students could benefit from reteaching a different skill. One teacher described her experience:

> They wanted us to have a team action plan. And the team is—you know, the other two classes did really bad on main idea. And I said at the time, "This is a waste of my time. This is a waste of my students' time. I don't need to have an action plan for the reteach on something they did well on. Why can't I look at author's purpose, or identifying details?" (BPS-FG)

The other teacher explained that she had switched schools that year and her new school allowed her to design a reteaching plan tailored to her students' results.

Backward Planning. While some teachers were using backward planning as a strategy even before their schools partnered with ANet, it seems that ANet formally introduced it in year two of implementation. One teacher described the shift:

> I feel like last year was really focused on helping us to identify priority standards for reteach, whereas this year I think it's more like prep work before you teach, like looking at your text before you teach it and then looking at your data after and making a plan for doing it again, whereas last year I feel like it was more a response to data. (T3)

From examples of backward planning given across interviews, there seemed to be variation in how teachers approached the strategy. In some cases, standards, and presumably the district's pacing guide, were used to align lessons over the course of the

year. In other cases, backward planning seemed to mean a process where teachers used the content of upcoming ANet assessments to plan lessons to cover those standards and skills before the next assessment. Based on interviews, it appears that some teachers tried to combine both approaches, planning much of their instruction according to the standards and curriculum, but attempting to slot in quick lessons if their original plans didn't include a skill or standard that was going to be tested on the next ANet interim assessment. One teacher explained her process:

> [W]e get an outline on what's going to be on the four ANet tests, the skills that are going to be covered. And I do make sure that I'm going to cover those skills in the period before the test, at some point in time. I mean, I don't want them to go into the test cold. ...I mean, I don't gear my lessons specifically around the ANet test, because there's so many other things we need to gear our lessons around, like the LEAP test, and the curriculum guide, comprehensive curriculum, but I do look at what's going to be on the ANet test and make sure that it's going to be covered at some point. (T22)

It is possible that teachers who were attempting to balance multiple – and especially misaligned – instructional resources (i.e., ANet assessments and the district pacing guide) were unable to effectively backward plan, taught skills too quickly (T41), or shifted topics and skills around in unproductive ways that disrupted student learning.

For teachers balancing multiple sources, the success with which they effectively backward planned their instruction may have related to the alignment of the ANet assessments and their district's curricular pacing guide or additional sources of instructional guidance. Teachers in one district mentioned that they had no pacing guide in math and, therefore, were able to use the ANet schedule of assessed standards to plan their math instruction. Others raised issues with managing the alignment of the district

pacing guide and ANet interim assessment content, a topic explored in detail later in this chapter.

Instructional Strategies. When asked to explain how ANet had changed their instructional practices, teachers described several strategies they were using. Using the ANet data to group students by their instructional needs was commonly mentioned (T4, T6, T10, T16, T29, T30, T40, SPS-FG, CHLS-FG); for example, one teacher explained that ANet had "reinforced the importance of grouping, of finding students that need the skills and teaching those specifically." (T10) Across interviews, teachers mentioned a variety of other strategies: co-teaching (n = 3, two schools); providing one-on-one instruction to individual students (n = 3); conferencing with students (T19); and differentiating instruction more generally (n = 2). Several teachers recounted ways of reviewing reading or math skills on which students were struggling during the other lessons in the same content area (but not during the designated reteach) or during lessons in other content areas (BPS-FG, CHLS-FG, CHI-FG).

In one school, teachers discussed how a shift to station teaching meant that different student learning needs – often identified by ANet results – could be simultaneously addressed. In this school, station teaching meant that two teachers could each work with a small group of students on a specific skill while other small groups worked independently on tasks related to the skill. It's worth noting, however, that this school also had a co-teaching model leading one of the interviewees to wonder whether such targeted reteaching could happen without it:

> I think if we didn't have the technology we have as well in this building, if we didn't have smart boards, if we didn't have iPads, I think the co-

> teaching model also lends to being able to do a lot of station teaching. So I think if we didn't have all of those other external factors, I'm not sure, if I was one teacher in a classroom, how that would work. (T4)

In a similar vein, a special education teacher from another school compared his class size to that of his regular education peer and the implications it had for individualized attention: "[S]ince I have such a low student-teacher ratio, I can actually have a conversation with each kid…. [Y]ou can't do that in almost any classroom. I'm lucky." (CHI-FG) As Goertz, Oláh, and Riggan (2009a) suggested, this level of differentiation might be possible only when there are sufficient classroom resources and staff to carry it out.

Barriers to Practice. Teachers were not directly asked to identify barriers to reteaching; however, patterns of common hindrances arose during the course of interviews. Several teachers remarked that, due to the rigor of the ANet assessments and the policy of only testing kids at their grade level, they had to focus on their relatively higher-performing students because the data were not telling them anything instructionally useful for their lowest performing students (T7, T30). "There was no information. And I think that's true for some of my low ones. I can't use the data meaningfully to reflect on what I need to change." (T30) In the BPS focus group, one teacher explained that she taught struggling readers, so her students' ANet scores were very low. She often focused her reteaching plan on only those students who were performing a little below grade level. A more detailed discussion of the implications of the rigor of the ANet assessments is discussed below.

Teachers also talked about finding enough time in the school day to implement the reteach (T7, T11). One teacher from Jefferson Parish explained:

> It's just a matter of finding time, you know, on top of the regular day to add reteaching to everything. Sometimes I find that it—you know, really force myself to get to it, because there's so much other stuff. You know, LEAP writing, and it's 20 days until LEAP writing test, so we're like gearing up for LEAP writing. And there's just always something, something, something, and to squeeze in something else. (T22)

For another teacher, the lack of sufficient instructional time made teaching for mastery more difficult:

> I feel like I'm always in a rush to get things done, to have the standards met before the test date. I feel like the tests are only spread out by four or five weeks. … So we have a month to get all this stuff taught, and it's a lot of information, and so I feel like I'm cramming everything in, and I'm not really able to dig in deep into the standards. I'm really just going the baseline—okay, I taught it for a day, move on, get through the next standard. (T41)

A few teachers explained that finding time to conduct the reteach put them behind in starting the next unit or lesson (T41, T35, SPS-FG) and that this could be exacerbated by an ambitious district pacing guide (T4).

Some teachers addressed the time limitation by working their reteach into new lessons as they moved on to new skills. However, this may have been easier to accomplish in ELA compared to math (CHI-FG). For the former, one teacher explained that an ELA skill could be worked into any story even if she had moved on to new ELA skills (CHLS-FG). Several teachers noted that ELA skills could also be incorporated into other subject areas easier than math skills (T41, CHLS-FG).

Summary

Patterns in ANet teachers' data-based instructional practices were consistent with prior research on data use from interim assessments. Teachers consistently described using the ANet data to uncover gaps in their students' learning, something that had been harder to do before. The majority of teachers described using their students' ANet results to plan their reteaching by focusing on the gaps found during data analysis. However, it is unclear if the frequency of data analysis and instructional planning translated into effective instructional practices. Some teachers said that they attempted to reteach the content in a different way, taking responsibility for their students' lack of mastery the first time. However, teachers' discussions of instructional practices provided few examples of teaching strategies such as differentiation and individualization beyond general references to grouping students based on their ANet results. Instead, some teachers described how they were encouraged to create a common reteaching plan across their grade level team even if it was targeted to a skill on which their students performed well. These findings seem to fit with the quantitative results.

Several barriers to practice also emerged from the interviews. Some teachers felt that the ANet results provided too little information on which to base a reteach; e.g., only one item was linked to a specific skill or, since the assessments were created to be grade-level appropriate, they were too difficult for their lagging learners. Others simply found it hard to make time to fit the reteach into their schedule. Those teachers who described differentiating their instruction tended to have classroom conditions that made it more feasible (e.g., co-teachers or small classes). In all, the evidence seems to suggest that

teachers were unable to take full advantage of the richness of the ANet data to shift instruction toward greater differentiation.

ANET TEACHERS' PERCEPTIONS OF INSTRUCTIONAL LEADERSHIP & SCHOOL CULTURE

In the conceptual framework for this study, instructional leadership, professional culture, and achievement culture are hypothesized to play a role in teachers' implementation of data-based instructional practices. The quantitative results suggest that ANet had a modest effect on these school conditions, but they appear to explain only a little over one-third of the variation in teachers' data practices. In this section, teachers' perceptions of these school-level factors are explored with a goal of understanding how ANet may have shaped these school-level conditions and the role they played in affecting teacher practices (research questions two and three).

Leadership

Teachers were not asked about the various roles their leaders played in the implementation of ANet or their abilities in these roles, but examples did arise during the course of interviews. In prior studies, researchers have suggested that school leaders play an important role in setting norms and expectations for data use.[3] One ANet teacher described a very concrete example of this: mandatory weekly grade-level meetings in which teachers reviewed data were just "part of what we do." (T19) For these weekly

[3] Heritage & Yeagley, 2005; Marsh, Pane, & Hamilton, 2006; Datnow, Park, & Wohlstetter, 2007; Goertz, Oláh, & Riggan, 2009a; Blanc, et al., 2010; Coburn & Turner, 2011; Datnow & Park, 2014; Gerzon, 2015

meetings, the principal expected teachers to bring their data so that they could share what they were going to do to improve student learning on a particular skill or standard. Despite being mandatory, the teacher thought the collaborative approach worked well. (T19)

Teachers in two focus groups felt that leaders had a responsibility for setting expectations or "parameters" for the reteach, but that these expectations either varied across schools or were not present.

> I think that the message is, you give a test, here is the data, now you need to look at where the kids fell down and do a reteach plan, and then you need to teach it, and then you need to reassess it. But what that comes out like in every school is a lot different. And I think that that's part of the messaging that needs to be told to the school leaders. (SPS-FG)

Without realistic expectations for the reteach, some teachers were concerned that the reteach could get out of control, be counterproductive, or focus on "stuff that you don't need." (CPS-FG, SPS-FG)

Additionally, the role of the principal as instructional leader was mentioned by several teachers. One teacher attributed the improvement in implementation of ANet in year two to the support of their administrator.

> I think the administration—there was sort of a realization that it wasn't going to be sort of a magic thing that happened. There was going to have to be more leadership in what to do with this and how to make it useful, and that's been extremely helpful. As I said, there's sort of a school-level understanding of, like, "Here are the big things we're working on. Now, what does that look like for you?" We're getting a lot of guidance within our groups from administration, you know, and then our teams. So I think that's probably been the biggest change. (CHLS-FG)

The handful of other specific examples of instructional support come from teachers describing how an instructional specialist, master teacher, or other lead teacher was the person to sit with them, review their data, and help identify the resources they may need to carry out their instructional plans (T14, T29). For example, one teacher was asked what resources had helped her change her reteaching:

> I think a lot of it has to do with the master teacher that we have this year…, because she has a wealth of knowledge, and she brings it to the table, and she shares it with us. She gives us the support that we need in our classrooms. If we're lacking in something, and we mention it to her, she makes sure we get it as best as she can. (T29)

To be clear, the interview protocol did not directly ask about the types and quality of support teachers received from school leaders. However, the interview data suggest that, besides setting expectations for data use, school leaders were not a significant source of instructional support for teachers' implementation of data-based instructional practices: e.g., offering feedback on reteaching plans such as guidance on appropriate content, delivery, and student groupings. Furthermore, a few teachers felt that they and their team were on their own to figure out how to reteach skill gaps identified in their students' ANet data (T33, BPS-FG). Several others explicitly stated a need for more feedback on their reteaching plans including a desire for periodic observations of their reteach (T4, T10).

Several teachers felt that their school leader was actually not helpful in their feedback or support. Related to the discussion of potentially ineffective instructional practices, above, two teachers felt their school leader was pushing them to focus on the

wrong skill during their reteach, skills that weren't supported by their analysis of their students' data (T4, T30-1). One stated,

> When I go to the data meeting, I sort of feel like I'm being forced through some steps and sometimes those steps aren't the things that I think would be best for my kids, and so I'm almost having to just comply and do something that isn't what I actually want to do based on my data.
>
> [Later…] I feel like when I've done a lot of work, putting thought into how I'm going to reteach something, and then I'm told, "You actually need to come up with something else...." Whoever's leading those data meetings can really make or break what is done with the data. (T30-1)

Several teachers in the Springfield focus group expressed feelings that feedback from their leaders on their reteaching plans was highly critical. In response to her reteaching plan feedback, one teacher said: "[I]t's going to be criticized, then why not come in and model the lesson?"

Rather than a source of instructional support, some teachers' commented on their leaders' focus on accountability. In particular, accountability came up with regard to carrying out reteaching plans, as interviewees suggested that school leaders often collected teachers' plans after data meetings (T14, T41, SPS-FG, P26, P28, P31). Teachers' reactions to this sense of accountability varied. One teacher suggested that having to hand in their reteaching plans seemed to ensure they followed through with the reteach and reflection; she explained that the "administration is very good about having us, you know, turn in plans to make sure that we are being reflective. It was easier to get around it sometimes before." (T14) Others felt the "mandate" to create a formal reteaching plans was unnecessary and certain aspects of the process were there only so leaders could tell if teachers were doing their job (i.e., reteaching appropriately) (SPS-

FG). A teacher in the Springfield focus group explained that her reteaching plan specified the day on which it would take place so that the school leader could check, "[I]f you said you're going to do it on this day, and they walk in, you should be doing it." However, she acknowledged that she had never actually had her reteaching observed.

These points of view highlight an important question: without ANet's formalized process and school leaders' accountability checks, would teachers put the effort into developing a plan for reteaching *and* implement it? Prior research points to the collection of reteaching plans as a way to hold teachers accountable to their implementation (Datnow, Park, & Wohlstetter, 2007; Goertz, Oláh, & Riggan, 2009a). In interviews, some leaders and teachers acknowledged that, prior to ANet, reteaching was easy to get around doing (T14) and now, ANet helps keep them on track instructionally (T41).

However, teachers from schools in one district expressed a concern that they were being held accountable to more than just reteaching; they felt accountable for students' ANet interim assessment results because they were now considered part of their evaluation (BPS-FG). One teacher explained how this affected the culture in her previous school:

> There were a lot of people who didn't want to speak up in data meetings. There wasn't, like, the collaboration, because people were feeling like they didn't want to talk about—they felt like they needed to defend their data. (BPS-FG)

If students' ANet interim assessment results were being included in teachers' evaluations in Boston it is a practice that ANet would have strongly discouraged.[4]

[4] Site visit data suggest that the use of students' ANet results in teachers' performance evaluations was not a practice that districts mandated but, rather, the decision of individual school leaders.

Professional Culture

In response to a question about discussions of ANet data with leaders and peers, over half of the teachers remarked that collaborating with peers over student data – be it as part of data meetings, reflection meetings, team meetings, or in discussions with their co-teacher – provided an opportunity to collectively unpack student results, understand which skills to target for student mastery, and share ideas on potentially effective instructional strategies (T4, T9, T10, T11, T19, T16, T29, T14, T33, T41, CHLS-FG).

> I think sitting down with my team is really helpful. I don't necessarily find it helpful sitting down with the other grades as much. I think it's helpful just to sit down with my fourth grade team and say, "Okay, what do we need to do as a team?" And I can kind of get their ideas of how they do their reteaching. (T41)

While most teachers referred to their grade-level teams, one teacher described how her school was doing team planning across disciplines, even including science and social studies teachers (CHLS-FG). In two examples, teachers expressly felt that these opportunities to collaborate with peers helped facilitate the implementation of ANet practices (T4, T11).

However, analysis of statements around professional culture produced some unexpected findings. Some schools were small enough that there was only one teacher for a grade or subject, making collaboration more challenging (CHI-FG). Several teachers spoke about feeling isolated from their peers or explained that their peers' cautiousness or competitiveness around sharing data and instructional strategies inhibited discussions and collaboration (T7, T14, T22, BPS-FG). For one teacher, this manifested as fear among her peers to give unsolicited advice to others. "I feel like it's just the culture right now of

most of the schools that we're afraid if you ever say, 'Did you ever think of doing this…?', you just get, 'Who do you think you are?'" (BPS-FG) While it's unclear how that culture developed, several teachers felt that their school had (or could) overcome a similar environment in time (T7, T14).

> As a team,… we can eventually get to the point where I can say, "Oh, your kids did really well with this, mine struggled with this. What did you do?" Once again, I think that takes time to build that relationship so it's not like I was sneaking and looking at your scores. (T14)

Prior research has shown that a sense of trust among teachers is important for healthy collaboration (Datnow & Park, 2014). Therefore, trust is likely an important condition for teachers to feel comfortable sharing student results, and giving or receiving instructional feedback. In one district, this sense of trust may be subverted by using students' ANet results as part of teachers' evaluations and putting teachers on the defensive with regard to student results, something teachers reported was happening in some schools.

> At my current school, it may be that it's a safer environment and it's an environment where people feel comfortable sharing ideas. … Even though it's a safe school, the data comes up and instantly, I'm like, "Please just let me be equal or a couple points –" and I love the other third-grade teachers. I think they're great. But I feel like it's caused this sense of competition. I want us to equally perform or do better, because if we do worse, I feel like it is going to be in my evaluation. (BPS-FG)

Since this use of student ANet results runs counter to the program's philosophy, it may have undermined the program's effectiveness in some schools in this district; for example, by fostering ineffective data cultures – e.g., characterized by less frequent or open collaboration – and negative attitudes towards ANet.

Achievement Culture

Among comments related to achievement culture, teachers expressed a desire for their students to be successful and held high expectations for student performance. In response to a question regarding ANet's impact on students, two themes arose. The first relates to the previous discussions of ownership over students' interim assessment results and internal accountability. Several teachers expressed a personal responsibility for their students' performance and success (T6, T29, T14). One shared her perspective: "[A]s a teacher, you want to move them along, because that's a big, like, 'Aha,' you know?" (T14) Another teacher expressed something seen in prior research: an acknowledgement that a student's poor performance may not be a product of the student, e.g., a student's disability or primary language. She explained a shift from:

> "Okay, the students just didn't get it and it's their problem," to, "Okay, what do I need to do differently?" I think that's the main focus—the problem's not the kid and their disability or lack of whatever. But honestly, it makes you think, "Okay, they didn't get it, and why they didn't get it, and I have to do something about that." So I think that is how it's shifted our conversations. (T4)

The second change that some teachers attributed to ANet was the inclusion of students in discussions of data. For one teacher, including her students was a practical response to keeping them motivated in the face of many tests over the course of the school year (T11). However, for others it was an attempt to foster an achievement culture among the students. One teacher remarked that her students weren't allow to "wimp out," because she was looking forward to seeing the same progress in her class as her peers' (T19). Many teachers reported that their students became more engaged when their

interim assessment results came back, especially when their performance improved. They explained that the results bred excitement, a desire to improve and succeed, and even bolstered students' inquisitiveness (T6, T14, T23, T41, T19, CHLS-FG).

> We've built a culture in our classroom of where the kids – they developed goals as a homeroom, and they kind of know where they stand as individuals. [When ANet results arrive] the kids can… see how they've done, and they've started… shouting each other out for improved performance or just great overall performance. And I think that's helped build a supportive, positive atmosphere in each one of the home rooms. (CHLS-FG)

Summary

Among the variety of roles leaders played in fostering instructional data use, ANet teachers discussed three in particular: setting expectations, providing instructional feedback, and checking for implementation fidelity. In some schools, teachers were aware of leaders' expectations for data analysis and instruction. Since much of the interview protocol focused on reteaching, teachers who provided examples of instructional feedback spoke to their school leaders' feedback on their reteaching plans. In these instances, the person providing feedback was not always the principal, but sometimes a data leader or instructional coach. Ultimately, though, some teachers expressed a need for greater support. A number of teachers described interactions with their leaders that were more consistent with accountability: checking their reteaching plans only to ensure teachers were complying with expectations.

When it came to examples of the professional culture in ANet schools, most teachers described examples of collaborating over student data: e.g., reviewing and

analyzing data with their peers, particularly their grade-level teams. This was generally presented as a positive experience: a way to unpack results together and share strategies for addressing students' learning gaps. However, it appears that, at least in some schools, there was work to be done in fostering a collaborative rather than competitive culture among teachers. In one district, the latter may have been exacerbated by the inclusion of student results in some teachers' annual evaluations, something ANet strongly discouraged. Finally, interviews revealed that many teachers saw ANet as a way to set and hold students to high expectations. This was manifested as teachers taking responsibility for student success and involving students in the process as a way to invest them in their own achievement.

Although these results are not easy to reconcile with the modest, positive impacts of ANet on teachers' school-mean perceptions of leader abilities and school culture (chapter 4), they do point to additional work that may need to be done to improve school conditions. For example, if school leadership is meant to affect teachers' data-based instructional practices, leaders' roles may need to be balanced between holding teachers accountable for implementation and providing them with instructional support. Leaders who do not have the expertise to provide instructional support to their teachers likely need training, resources, or support of their own. Finally, while quantitative results suggest that ANet teachers review student data more frequently, some schools may need to foster a culture of trust that allows deeper collaboration without fear of judgment.

ANET TEACHERS' ATTITUDES AND CONFIDENCE RELATED TO DATA-BASED INSTRUCTION

Teachers' attitudes toward assessments and confidence or skill with assessment data were not an explicit focus of the larger evaluation because they were not a defined target of the ANet program. It is, therefore, not surprising that the quantitative results showed no impact of ANet on these teacher characteristics. However, quantitative results did show that, controlling for treatment assignment and other covariates, teachers' attitudes and confidence were positively related to each of the four data-based instructional outcomes in this study. In seeking evidence to supplement the quantitative findings, coding focused on interview questions asking teachers to share their satisfaction with the ANet program and how the implementation of ANet had changed over time. Though they are specific to the ANet program and related practices, several patterns emerged from these interview questions.

Attitudes

Across interviews, many teachers held favorable views of the ANet program and its resources, but some teachers acknowledged that they didn't initially have such views. Several teachers talked about themselves or their peers being on board with or buying into ANet practices. Comments were often framed as a change in beliefs between year one and year two of implementation. Specifically, many teachers talked about having an initially unfavorable response to ANet (e.g., being overwhelmed by the amount of data or the ANet process), but that something made them come to support ANet over time: students' results on the first set of ANet results, the available resources on the MyANet

website, an enthusiastic or supportive leader or coach, or seeing improvements in their students' state summative assessment results that they attributed to their implementation of ANet practices (T4, T6, T19, T23, CHLS-FG). One teacher explained, "It's been a complete turnaround in usage and enjoyment, honestly. I did not even want to see an ANet test last year, whereas this year, it's a lot better because I know I'm going to get something useful out of it." (CHLS-FG) Another teacher equated ANet to learning to use any new type of classroom technology, stating the first time "you're like, 'I'm never going to be able to use this thing,' and after a while, you're like, 'I can't teach without it!' … It's the same thing with ANet. I'd probably be a little lost without it. It definitely simplifies my life now." (T23)

Not all teachers came to hold positive opinions of ANet. Several teachers raised concerns that ANet practices encouraged teaching to the test (T16, T29, T41, CHLS-FG, SPS-FG). For example, one veteran teacher of 30 years expressed her wish to "get back to teaching" rather than being told to teach certain standards because they would be on the test (T29). Some teachers lamented the amount of time spent on testing or the amount of instructional time lost to testing (T30-1, T41).

Confidence & Skill

Like shifts in teachers' attitudes toward ANet, teachers also expressed a change in their confidence and skills over time. Many teachers felt that year one of implementing ANet was challenging, either because they were overwhelmed by the process or amount of information (e.g., data), or they needed time to understand the program and what was

expected of them (T3, T4, T10, T23, JPS-FG, SPS-FG). In Chelsea, one focus group teacher felt that there had been a shift in year two toward making data more accessible to teachers who weren't comfortable "picking numbers apart." Other teachers mentioned that they were more comfortable with or better understood the ANet process in year two than they had in year one. Specifically, some teachers felt that, after going through the process in year one, their understanding of student assessment data had improved and they were quicker at analyzing it, were better at generating standard-specific quizzes through the quiz builder tool, or they understood the purpose of the reteach and ways to focus it effectively (T7, T3, T14, T10, T23, T41, SPS-FG).

Teachers in the Chelsea focus group also felt that their instructional skills improved from year one to year two; specifically, their effectiveness in backward planning and general lesson planning improved. However, as mentioned above, a few teachers across the sample felt that, while they understood the rationale for the reteach, they simply didn't know another way to teach the same skill or standard (T33, T6).

Summary

There was clearly a range in teachers' attitudes towards ANet and confidence in its implementation. Though many teachers had positive perceptions of ANet, some initially disliked the program and tended to come around when they saw improvement in their students' learning that they attributed to using ANet strategies. Likewise, confidence levels varied, with some teachers still feeling overwhelmed by the work involved in implementing the ANet data cycle, even in year two.

This variation in teachers' in attitudes and aptitudes may offer one explanation for why the quantitative results showed no impact of ANet on these outcomes Acknowledging and increasing teachers' attitudes, confidence, and skill likely requires directly meeting teachers' learning needs and providing instructional training, tools, or resources that increase the breadth of effective instructional strategies. It may also require a more explicit focus on teachers' beliefs and skills around data-based instructional practices by ANet and programs like it. For as one teacher said, "[I]t's not just the kids learning, but it's us learning from it as well, and most especially the teachers." (T35)

ANET TEACHERS' FEEDBACK ON THE INTERVENTION

Initially, the coding schema for the qualitative analysis did not include dedicated codes for the various ANet program components as they were outside the conceptual framework for this study. However, after discovering that the hypothesized mediators explained only a little more than one-third of the variation in teachers' data practices, the second round of coding included new codes aimed at identifying other factors that teachers felt helped or hindered their implementation of ANet and data-based instructional practices. In the vast majority of cases, these factors related to specific ANet program components. In this section, the most common themes in teacher responses are discussed.

ANet Coaches

Because of the structure of the program and ANet's focus on building leader capacity, teachers had less direct interaction with the ANet coach than their school leaders. However, teachers who provided examples of interactions with their ANet coach generally spoke favorably about them. Coaches were perceived as being helpful and available to answer questions (e.g., navigating the MyANet website, offering guidance on data analysis). Still, several teachers did express a wish that the coaches were more available or provided more support (T30, BPS-FG, JPS-FG). Two teachers from the same school remarked that their ANet coach had been a good resource the year before, but that "no one's really come from ANet this year." They recognized that the ANet coach was likely to visit less in year two, and even though they trusted their school leaders, they missed having the ANet coach around because of their familiarity with the program and their outside perspective on student data (T30-1, 30-2).

Focus group feedback tended to be more critical. Teachers in Boston and Springfield remarked that they didn't see their ANet coach enough and would have liked more support from them. Some teachers in the Jefferson Parish focus group expressed dissatisfaction with their coach and others didn't even know who their coach was. Focus group teachers also understood the gradual release of support, but at least some teachers felt that part of ANet's service was to provide a coach as a resource for analyzing data and sharing instructional ideas. In combination with earlier findings on school leader support, it seems that teachers were looking for additional support in implementing data-based practices.

ANet Website & Resources

In coding teachers' comments on the MyANet website, feedback on its ease of use was favorable compared to previously having no online resource or a similar platform provided by another assessment program. Two additional themes arose. First, teachers often talked about the student data reports. They appreciated the quickness with which results were uploaded after an assessment, the ease of retrieving results, the clarity of data reports, the ability to view students' results in multiple ways, and the ability to compare their class' results to other classes within their school or other schools within their network. One teacher said, "I think the data's very useful. It's easy to read, easy to understand, easy for me to see… how my children have grown. …I also really like how quickly the data is available. That is amazing." (T14)

Second, teachers appreciated having access to various instructional materials and assessment resources through the MyANet website. One teacher explained, "ANet just helps me to move toward success by providing me with the information and all the strategies I need on one website." (T19) The quiz tool seemed to be one of the more widely used resources on the website for assisting in reteaching and reassessment. Teachers generally appreciated how quickly and easily they could use the tool to identify items associated with specific standards and skills and build short quizzes to identify where their students continued to struggle. However, comments indicated that it may have been used as a form of test prep in at least some cases. One teacher explained, "[I]t's really helpful at just getting students used to the types of passages they're going to read on standardized tests." (CHI-FG, T30)

Many teachers pointed out, however, that both the pool of available quiz items and other instructional materials (e.g., sample lesson plans) were limited or missing entirely for some standards and skills (T4, T6, T14, T30-2, JPS-FG, CHI-FG, BPS-FG). In particular, teachers noted that the availability of materials for the new Common Core standards was extremely limited (T41, CHI-FG, JPS-FG, SPS-FG). In cases such as these, some teachers made their own lessons or created their own quiz items (T6, T30-2). Among the skills and standards for which instructional materials were available, multiple teachers commented that they often weren't specific or written clearly enough to carry out (T10, T14).

ANet Assessments

In the first year of the larger evaluation, the research team heard in interviews, and found in the survey results, clear patterns in ANet teachers' perceptions of the rigor and alignment of the ANet assessments to the standards, curriculum and curricular pacing guide, and the state summative assessment. To explore further, some questions on the year-two interview protocols focused on finding out more about these perceptions. As a result, the interviews include frequent examples of teachers' perceptions of the ANet assessments and resulting data, especially in relation to their utility in informing instruction.

Rigor. The majority of teachers found the rigor of the ANet interim assessments to be higher than their standards, curriculum, or state assessment. However, teacher opinion varied in how they viewed – and defined – the high level of rigor. This seemed

related to whether they spoke directly about assessment rigor or whether they were using rigor as a proxy for assessment difficulty. Rigor implies deeper, more complex or critical thinking. Difficulty implies tasks that are above a students' ability; e.g., an item that might be above their grade-level or level of mastery.

Among those teachers who viewed rigor favorably, some appreciated that the ANet assessments focused on high standards and felt they prepared students for meeting the demands of the Common Core standards (T6, T10). One teacher remarked that the ANet assessments exposed her students to something "authentic and with high rigor." (T10) Although the unstated assumption is that the rigor of the assessments influenced the rigor of instruction, it was impossible to tell whether teachers were, in fact, increasing the rigor of their instruction more broadly.

Much more often, teachers seemed to appreciate the rigor of the ANet assessments because they believed that it was preparing their students for the state summative assessment. These teachers thought that the ANet assessments allowed their students to experience, practice or prepare for, get used to, more comfortable with, or less overwhelmed by a rigorous interim assessment before sitting for the state's summative assessment (T4, T10, T14, T30-2, T22, T29, T33, SPS-FG, JPS-FG). Some of these teachers' comments indicated that they viewed the administration of the ANet interims as test preparation, a way to build their students' test-taking confidence and stamina before the state test. In comparing the rigor of the ANet tests to the state summative assessment, one teacher said: "I definitely feel like it [ANet] prepares them. Even the practice of taking that type of test. Like, the test-taking skills." (T4). This use of ANet as a way to

prepare students for the rigor, style, or format of the state test was echoed numerous times across schools and districts.

> I think in comparison to other district assessments that have been administered in the past, this is definitely the most rigorous test that the students have been exposed to. In some ways I think it's harder than the MCAS…. I'm okay with it being harder from MCAS, in my perspective, because I think that'll just prepare them more. …I think that's preparing them for MCAS. That's something for me to look at; how am I going to build their stamina to get through what they need to get through. (T33)

> So after the test is over, then I go over it with them, I model for them what it is they should be doing, I give them the strategies that they need in order to get a correct answer. (T29)

It is impossible to know whether these teachers used the ANet assessments *only* for test preparation. However, if the ANet tests were seen as more rigorous than the curriculum and standards, and, therefore, seen only as a chance to expose students to rigorous tests so that they would be accustomed to that rigor by the time they took the state test, the results of the ANet interims might have played a smaller part in shaping the rigor of some teachers' instruction.

One theme that supports this hypothesis are remarks by teachers who spoke about rigor of the ANet interims in the context of their being simply too difficult for their students (BPS-FG, JPS-FG). Some teachers judged the Lexile levels of reading passages to be at the high end of a grade level or a higher grade level (CHLS-FG, T41, SPS-FG). Opinions that the ANet assessments were overly rigorous – i.e., difficult – were particularly prevalent among teachers assigned to teach ELL and special education students, or whose regular-education students were performing far below grade level (T7, T30-1, CHI-FG, CHLS-FG). Several ELL teachers explained that their students could

master a skill, but not be able to apply it to a grade-level text in order to demonstrate mastery (BPS-FG, CHLS-FG). For example, one teacher explained, "[I]t's a test assessing literacy skills, but they don't have the language to access the reading." (CHLS-FG) A second teacher echoed this sentiment and also explained that she had to teach her class of ELLs at a slower pace and, thus, often hadn't covered all the material on the ANet interim (CHLS-FG).

For these teachers and teachers whose class was largely below grade level in reading or math, students often scored so poorly that teachers felt the data were of little instructional value. Some of these teachers felt that their students were often guessing. In response, a few teachers said they used the online quiz tool to create quizzes at lower grade levels, meeting students closer to where they thought they were academically (CHLS-FG,). Several teachers wondered why ANet didn't offer tests that were more appropriately matched to students' abilities or why they didn't allow off-grade-level testing (CHLS-FG, CHI-FG). One teacher explained, "[I]t would be wonderful… to delve into data for something that's meaningful for my kids. [Our ESL team] started piecing together things from the quizzes, but why should we be doing this? Shouldn't ANet provide something for us?" (CHLS-FG) Another teacher expressed a similar desire for flexibility to test kids at their ability level rather than grade level:

> Sometimes I feel like when I'm giving them those interim assessments, they're just guessing a lot of the time. And then the ones that I make up do pretty well because they're usually at their grade level, so it gives me a pretty good idea of where they are. And that does leave the question of—they have to take the grade-level test in the state test, and that's important, but I feel like it's more important for me to know where my kids are than

> it is for me to just kind of have them try their best at high-level stuff they're not going to do well on. (CHI-FG)

Teachers frequently remarked that the rigor or difficulty of the ANet interim assessments caused some students great frustration or anxiety (T3, T7, T16, T41, BPS-FG, CHI-FG, SPS-FG). One teacher explained that she sometimes saw dips in students' scores because students were "tired and they get exhausted at the idea of taking another test." (T3) Many teachers also felt that exposing students to overly difficult interim assessments or too many assessments of various types meant they might not take them seriously or may experience a sort of lack of motivation or burnout, thus making the results less valid or useful for informing instruction (T3, T22, T41, SPS-FG, BPS-FG, JPS-FG).

Alignment. Prior to the start of each school year, ANet works with districts and schools to develop interim assessments that are aligned with their curriculum and curricular scope and sequence. Despite this, another frequent frustration regarding the ANet interim assessments was their misalignment with the curriculum, standards, or, most commonly, the district's instructional pacing guide. Alignment seemed to vary by year of implementation, district, and content area. However, one thing was consistent: teachers often described how misalignment made backward planning and the use of students' results for reteaching frustrating and difficult (T4, T7, T3, T41). In particular, many teachers were frustrated when the content of an ANet interim assessment included skills that they had not yet taught because the skill or standard had not yet come up in the sequence of their curriculum (T29, T33, T41, BPS-FG, CHLS-FG). Some of these teachers felt it put them in the position of deciding between sticking to the pacing guide,

aligning with the ANet interims, or attempting to balance the two (CHLS-FG, SPS-FG, JPS-FG).

> [ANet] wasn't aligned with our scope and sequence last year. So…it impacted my instruction because I was trying to figure out how to fit it in… what I should be working on. So I didn't know how to do it. I kept flip-flopping back and forth, like, should I just go with what the city wants me to do, what the district wants me to do, or should I go with ANet? …[S]o it impacted me, I guess, to be more frustrated as to what to do. (CHLS-FG)

In response, some teachers tried to compensate for misalignment by fitting in unplanned lessons that were going to be on the ANet assessment (SPS-FG, BPS-FG).

> [W]e get an outline on what's going to be on the four ANet tests, the skills that are going to be covered. And I do make sure that I'm going to cover those skills in the period before the test, at some point in time. I mean, I don't want them to go into the test cold. .... I mean, I don't gear my lessons specifically around the ANet test, because there's so many other things we need to gear our lessons around, like the LEAP test, and the curriculum guide, comprehensive curriculum, but I do look at what's going to be on the ANet test and make sure that it's going to be covered at some point. (T22)

However, teaching something out of sequence meant that teachers didn't always have a lesson already prepared and had to go outside the curriculum to develop one (BPS-FG). This misalignment had the potential to create additional work or require additional skill.

> Everything's just really confusing, and then the pacing guide will throw in one standard that's not on ANet, but we need to teach it—so I think having it all really line up would be beneficial in that sense, because it's just a lot of—I feel like I'm doing so much extra work." (T41)

> And that's the teachers'—that's what the teachers do. That's the skill. That's where it comes in, that's where the talent comes in—teachers that know how to add in a little bit of this and a little bit of that and move

> things around so that everything gets taught by the end of the year. (SPS-FG)

Teachers in Springfield may not have had the option to ignore the district pacing guides; they often expressed a lack of freedom to realign instruction to the ANet interims. For example, when exposing misalignment between an ANet interim and the order of the curriculum for a particular math unit, teachers in one school were given a directive "from above" to stick to the district pacing guide (T33, T35). Whether teachers chose not to realign their instruction to upcoming ANet assessment content (T2) or weren't permitted to, the presence of misalignment was another instance in which the results were not as instructionally useful as they could have been.

Summary

ANet coaching focused on building the capacity of school leaders to support teachers' data-based instructional practices. Generally, it seemed that teachers understood this, but some felt that they could benefit from additional coach support. This is especially relevant since there was limited evidence that school leaders were a frequent instructional resource for teachers. Whether ANet coaching of teachers is something that can be worked into the program model is unclear, however. In terms of resources, the data reports were almost universally praised, but teachers saw gaps in the available instructional resources for certain standards. Tools such as the quiz builder were widely used; however, they were used not only to assess students' mastery, but as a way to expose and prepare students for the state summative assessment.

One of the most consistent themes that emerged across all teacher interviews and all study foci was teachers' perceptions of the rigor and alignment of the ANet assessments compared to the standards, curriculum, and state summative assessment. For many ANet teachers who were interviewed, the rigor and misalignment of the ANet assessments meant that the resulting data were not always able to inform their instruction. However, some teachers did feel that the tests were useful in preparing their students for the rigor and structure of the state summative assessment. In all, these results call into question whether teachers were provided with sufficient support, resources, and viable data to improve their instruction and effectively meet students' learning needs.

ANET SCHOOL LEADERS' PERCEPTIONS OF INSTRUCTIONAL LEADERSHIP & SCHOOL CULTURE

This chapter now turns to the results of the analysis of the leader interviews. Greatest attention is devoted to leaders' perceptions of their own leadership skills and the culture in their school. Leaders' perceptions of coach support and the ANet assessments are also discussed. Priority is generally given to teachers' self-reported attitudes, confidence, and practices. However, attention is also given to school leaders' perceptions. Although only possible in aggregate, patterns across leader and teacher interviews can offer evidence that either corroborates or contradicts teacher self-reports.

Leadership

The quantitative results suggest that ANet had a modest effect on leader abilities as measured by school-mean teacher perceptions of their leaders' abilities. The school leader interview data offer a view of a subset of treatment-school leaders' self-reported roles in fostering data-based instructional practices in their schools. This provides important context for the quantitative results.

The research reviewed in chapter two found that leaders play a key role in fostering data-based instructional practices by setting expectations among their staff.[5] This was a role several teachers also acknowledged during interviews. Over the course of interviews, four leaders discussed their expectations for how teachers should prepare for data meetings, review and use data, and plan and carry out their reteaching (P1, P15, P28, P31). For example, leaders expected their teachers to come to the data meeting having reviewed their students' data, to turn in completed reteaching plans for review, to carry out the reteaching with fidelity, or to reflect on the success of the reteach. However, it wasn't always clear whether or how these expectation were made known to teachers.

More frequently than expectation setting, leaders spoke about their attempts to build staff capacity to lead the work associated with ANet implementation, and the ways they provided feedback or monitored the work of ANet in their schools. In discussions of feedback and monitoring, two themes arose: the familiar roles of providing teachers with support and ensuring teachers carried out their reteaching plans. First, leaders' goals for

[5] Heritage & Yeagley, 2005; Marsh, Pane, & Hamilton, 2006; Datnow, Park, & Wohlstetter, 2007; Goertz, Oláh, & Riggan, 2009a; Blanc, et al., 2010; Coburn & Turner, 2011; Mandinach & Jackson, 2012; Datnow & Park, 2014; Gerzon, 2015.

capacity building are discussed before turning to leaders' roles in overseeing the work of ANet.

Capacity Building. About a quarter of the school leaders that were interviewed mentioned that ANet helped build their capacity as a data leader or instructional leader (P1, P5, P15, P27). For example, one leader explained that he and his assistant principal were pretty data savvy before partnering with ANet, but that working with ANet had improved their data analysis skills. (P15) As often, school leaders described to interviewers their desire to build the capacity of *other* leaders in the school – e.g., an assistant principal or teacher leader – so that ANet practices would continue if they were unavailable to lead it themselves. Specifically, one leader explained how this would affect data meetings:

> I'd like to see more people on our leadership team taking a role in the data meeting so there'd be some shared in that whole practice. So when I have to be absent, I don't feel like we're dropping the ball. More leaders have the capacity the better. (P28)

Another school leader admitted she was often too busy to provide feedback to teachers on their reteaching plans (P1). She felt that distributing some of this work would improve the feedback process for teachers' reteaching plans: "Being the only administrator, I need other people to step up and take things on."

However, distributed leadership was difficult to achieve for some of these leaders. The leader who admitted she often didn't have time to provide teachers with feedback on their reteaching plans wanted to hand off some of the responsibility to her literacy leadership team, but was met with some resistance because they didn't want the act of providing feedback to be perceived as criticizing their peers. During the interview, the

school leader explained that she was searching for ways to involve the literacy leadership team in providing feedback without their peers feeling as though their toes were being stepped on (P1). Another school leader felt it made sense for her to be the school's data leader. Although she thought there were teachers who had the capacity, she didn't feel they had the time to be away from their classrooms to do the work related to ANet (P5). Both of these leaders felt that, because their teachers were becoming more independent at reviewing and analyzing data, they could spend less time supporting that work and more time providing feedback on reteaching plans.

Building the capacity of the entire staff was the focus of leaders who wanted their teachers' to have the necessary skills to implement – and own – the data-based instructional practices promoted by ANet (P1, P5, D37). One data leader explained that she had initially done some data analysis for teachers, but realized that teachers needed to do that analysis themselves so that the process was "real" and they owned it (D37). The leader of another school had seen progress in this regard, "I honestly firmly believe that in a few years, they're not going to need me at all, in terms of being the data facilitator. They'll be able to facilitate their own [teams]." (P5)

Feedback & Support. Discussions of instructional support were almost exclusively in the context of the reteaching that was expected to take place after each ANet assessment administration. Specifically, leaders talked about the ways they provided feedback to teachers on their reteaching plans. Most often, leaders said that their feedback focused on whether teachers were addressing an appropriate skill or planning to use an appropriate instructional strategy. Consistent with what teachers shared, some

school leaders and data leaders mentioned conversations with teachers about considering a different way of reteaching a skill when students had shown a lack of mastery on the ANet assessment (P5, P15, P12, D22, D26, D35). Getting teachers to consider their student groupings during the reteach was also a goal of some leaders (P12, P18, D35). An assistant school leader described a conversation she had with her principal after they realized that teachers were relying heavily on whole-group reteaching.

> We said, we've got to change that. We've got to get more individualized, like, you know, start pulling small groups. So we wanted, we hoped that it would be a tool that would enable us to use data with the teachers so that they could see the impact they could make by doing small group, and it has. (P28)

As part of the reteaching process, several leaders required teachers to submit their completed reteaching plans. For some leaders, this was an opportunity to provide individual feedback on some of the aforementioned strategies (P1, P12, P34). For example, one leader explained how she reviewed each plan and, with teachers' strengths and weaknesses in mind, provided them with individual feedback:

> [I]f it's a teacher that I know is strong on content, knows her students very well, yet he or she may be afraid to incorporate technology to make it more engaging, I push them to get to that point. Or sometimes I have teachers who tend to want to teach whole group all the time. I push them toward differentiation and grouping. So I think the reteaching allows me to give them some one-on-one that they may need and kind of push them to look at how they taught the standard before, and then reassess, "Okay, it didn't work this time, so let's do this again, and this time let's try something different." So I think it has really pushed their level of thinking and really made them think about how they're teaching different standards. (P12)

However, other leaders seemed to require reteaching plans be handed in more as an act of accountability in order to ensure that reteach planning and, presumably, the reteach itself took place. One school leader felt that teachers' plans may not be purposeful or specific every time, nor would reteaching always be carried out "unless someone demands they do it, which is what ANet is doing by building this habit of, 'Okay, you're going to have to give this to me.'" (P1)

In addition to collecting teachers reteaching plans, a few leaders visited classrooms to observe the reteaching, although the regularity with which this occurred was unclear (P12, P18). Like collecting reteaching plans, there were two purposes that observations seemed to serve: feedback and accountability. For a few school leaders, observing the reteach provided an opportunity to provide teachers with real-time feedback. One school leader explained how, while observing a reteach, she had an opportunity to model lessons with teachers and to give immediate feedback on what and how they were teaching with respect to addressing students' skill gaps (P18). Other leaders seemed to use classroom visits as another accountability check (P34). A data leader from one school described how teachers turned in their plans to her and she tried to "pop in" to classroom to check that teachers were reteaching (D26). There was no direct mention of feedback being provided to teachers on the reteaching plans.

Generally, leaders tended not to view requirements such as handing in completed reteaching plans or observing the reteach as accountability in a negative sense (i.e., as accountability to *them*) and leaders didn't necessarily use only one approach – providing instructional support versus accountability. For example, several leaders explained that

they wanted to help teachers improve their reteaching so students would succeed. However, they felt that, without some oversight, teachers would focus their reteach inappropriately (P28) or might not follow through with reteach at all (P17). One school leader used the data meeting to review student data with each teacher and probe them to explain how and why they were going to "move this kid." However, he explained that he planned to follow up later to check that teachers were implementing what they laid out in their reteaching plans (P27).

Barriers to Practice. Across interviews a few leaders described barriers to leading the activities of the ANet data cycle. Although these are single examples, they seem potentially illustrative of barriers other ANet leaders would have cited had this been asked directly in the interviews. For one school leader, school size played a role in overseeing reteaching in his school. He admitted that he wasn't entirely sure how reteaching was really going. The year before, he had been at a smaller school where it was possible to observe each reteach. Having moved to a much larger school, it wasn't clear to him whether teachers were implementing their reteaching plans with fidelity (P27). This was partly because he didn't have time to "monitor" all teachers' reteaching and partly because he wasn't confident in his teachers' commitment to ANet more generally. These same challenges – providing feedback to a large staff of teachers and ensuring reteaching is happening with fidelity – were expressed by at least one other data leader (D35).

For another school leader, balancing competing demands was a barrier to participating in teachers' reflection meetings, specifically, and ANet implementation

more generally. She described what she called a "pretty broken week" and the need for structures to protect ANet implementation that weren't compromised by scheduling disruptions. She went on to explain that working with ANet was her choice and was not a district-wide initiative, so the ANet assessment cycle schedule and district schedule for meetings and professional development, etc., were sometimes in conflict.

> So you really have to have it [ANet] where it's something that is fully supported, so you don't have those conflicts coming in where you're trying to say, "Is it ANet or this?" And that's the frustrating part for me, because I value what ANet has, but I have to give sometimes preference to whatever [district] directive I may have at the time. (P18)

She expressed concern that ANet would not be successful unless it was supported at the district level or schools had control over putting structures in place to facilitate ANet implementation.

Culture

Professional Culture. Evidence from interviews suggests that school leaders saw ANet as a way to foster teacher collaboration over student data during data meetings, reflection meetings, and teacher team meetings. Many leaders described positive experiences of collaboration: their staff were more comfortable discussing data and what content to reteach, sharing and borrowing instructional ideas on reteaching specific skills or standards, and using common language and practices to do so (P1, P5, P12, P31, P40, P36, P34, D38). When asked to describe the culture around data in her school, a leader described how teachers interacted during data meetings, "For the most part, the teachers

are open, they can talk to each other, they share. So they're not like, 'I don't want you to see my data….'" (P12)

However, consistent with teacher interviews, this sharing atmosphere wasn't universal and a few leaders admitted there was room for improvement in teacher collaboration around student data and instructional strategies. One leader explained how her teachers tended to make excuses for their students' results, rather than recognize possible deficiencies in instruction and collaborate over solutions (P24). Another leader described what her teachers considered collaboration: breaking up the task of developing lesson plans by subject, working independently on their assigned subject, then trading plans with their team members. She explained,

> That's not my perspective on collaborative planning. To me, collaborative planning means we all come together and say, "Okay, are we all going to be doing this piece of lit for this week? Okay, how are you going to address it with your kids? I have this deficit with my kids. It worked for you last time, tell me what…" That kind of thing. (P28)

Another leader acknowledged an issue that some teachers had also mentioned: needing to tackle the closed-off nature of their staff when it came to sharing student data.

> [D]iscussions have come a long way, because in the first two cycles,… we had to use talking chips to get [teachers] to talk about things among themselves. Because it was almost like, "These are my results, I don't want to talk about it. I don't care what you did in your class." And now there's more sharing. Is it where we need it to be? No. But are some of the walls coming down? Yeah. (P28)

Achievement Culture. As it wasn't a focus of the larger evaluation, there were very few comments from leaders that could be coded as examples of achievement culture. One leader did remark that teachers in her school wanted to do a good job and, when

seeing that students were struggling, they were thinking about how to reteach a skill differently (P5). The data leader of another school expressed a similar sentiment, explaining that even though the school historically scored well in ELA and had an experienced teaching staff, teachers were still open to trying new things to continue improving (D38). Finally, two school leaders commented on a noticeable shift from teachers blaming students for not performing well to teachers taking responsibility and focusing on improving their instruction (P15, P18). One elaborated:

> You're hearing less and less of, "The students can't get it, the students are low," and more of, "Let me try it this way. What did you do? How did you get students to understand main idea? How did you get students to understand author's purpose?" When they're having conversations with me, that's what the conversations sound like more. (P18)

Summary

In coding interviews for examples of instructional leadership, leaders tended to speak about their ongoing attempts to build the capacity of their staff to do the work associated with the ANet program and their own efforts to lead the ANet work. With respect to the latter, they most frequently shared their methods for providing feedback to teachers on their reteaching plans. While some feedback efforts focused on improving reteaching, other leader comments seemed to describe measures to ensure that teachers were simply writing reteaching plans and carrying them out. What cannot be answered with the available data is whether one approach was better than the other, or whether both might play a part in the effectiveness of ANet. As one leader described, the goal was to

improve teaching and learning, but that some amount of accountability seemed necessary to ensure teachers' implemented ANet practices.

Some teachers' comments, however, seem to indicate that rather than ensuring the implementation of ANet practices, leaders collection of reteaching plans or observations of reteaching lesson was sometimes interpreted as external accountability for student learning (i.e., ensuring teachers were planning and implementing their reteach). This could be because, in some districts, leaders made mention of how writing a reteaching plan that targeted gaps based on student data addressed a component of a new teacher evaluation system (P18, P34). Recall that teachers in one district also reported that their students' ANet results were included as part of their evaluation.

Related to professional culture, leaders provided a number of examples of collegial conversations their teachers were having over student data. However, some leaders felt there was room for improvement, acknowledging that their staff was not open to sharing results, resistant to change, or had flawed conceptions of collaboration. Though examples of achievement culture were limited, a few leaders felt their teachers were taking responsibility for student learning and held a broader desire to see their students succeed.

ANET SCHOOL LEADERS' FEEDBACK ON THE INTERVENTION

To the extent that ANet coaching of school leaders played a role in leaders' subsequent support of teacher practice, this section focuses briefly on leaders' feedback on coach support. More attention is given to reviewing leaders' comments on the ANet

assessment, particularly, their perceptions of their rigor and alignment in order to compare leader and teacher perceptions on what may be a key factor in teacher implementation of data-based instructional practices.

ANet Coaches

Leader comments on their ANet coaches were mostly positive. Leaders stated that their ANet coach helped them implement ANet by preparing for – or facilitating – the data meetings, helping them have focused conversations with teachers, and honing their own' data analysis skills and knowledge of the Common Core standards. Leaders also said that, at various times, their ANet coaches offered feedback, support, and encouragement, discussed problems and next steps, assuaged teachers' fears and concerns, and answered their questions (P1, P5, P15, P12, P13, P18, P24, P40, P34). Two data leaders also discussed how their ANet coach helped them break down, understand, and implement ANet, especially in the first year (D13, D35).

Still, several leaders felt that the coach support was not sufficient or that it shouldn't have tapered off quite so quickly in year two (P5, P12, P31). One leader said that she had been told that her school was functioning like a year one school, but the coach offered no elaboration on what this meant. She expressed a desire for more support from ANet and said, "It's kind of one of those things where you don't know what you don't know. I feel like there's so much we don't know that we could be doing." (P31) Another leader expressed why she felt the train-the-trainer model may not work well:

> I think you always need that person. … I think having them as the lead person, the expert facilitating, preparing the presentation, being that

> outside data expert consultant, is key. Just from a logistical standpoint, it would be a nightmare if I had to take on that along with all the other initiatives and all the other things I have to do at this school. I think that's just too much. (P12)

She went on to suggest that coach support should be differentiated to teachers and leaders in the same way that ANet expected teachers to differentiate the instruction of their students.

ANet Assessments

Rigor. Many leaders commented on the high level of rigor of the ANet assessments. Several leaders felt that the rigor of the ANet assessments was good for instruction. For example, one believed that teachers were challenged by the rigor of the assessments and, therefore, worked "smarter and harder to get [students to] the level they need to be." (P34) A data leader from another school felt similarly; she explained that the ANet assessments were more rigorous than the state test and were more aligned to the Common Core standards, and that the rigor pushed teachers further and gave them a "better understanding that this is… where we're going." (D25) A school leader in Jefferson Parish concurred. Knowing the ANet assessments would be rigorous, she felt that teachers had shifted instruction to keep instruction and assessment "on the same plane." (P31) It may be that the rigor of the ANet assessments opened teachers' eyes to the increasing rigor of standards, curriculum, and instruction (e.g., due to the implementation of the Common Core), but from teacher interviews, it is difficult to find evidence that it also generated greater rigor in instruction.

Some leaders also tended to use rigor as a synonym for difficulty. Paralleling teachers comments on test preparation, one leader appreciated the ANet assessments because they mirrored the complexity and length of reading passages on the state test, so in terms of building endurance, "it's a great practice for kids." (P40) Overall, fewer leaders than teachers made an explicit mention of the ANet assessments being useful as preparation for the state summative test, though it was mentioned by some (D38, P40, D25). One data leader in Jefferson Parish noted that the rigor (i.e., difficulty) affected the utility of the data for informing instruction. Given the perceived level of difficulty of one the ANet reading assessments, she believed students were largely guessing, thus making the data invalid for instructional purposes (D26). She also remarked on the length of the reading passages, having received feedback from teachers that their students were exhausted at the end of the ANet assessment (D26).

Alignment. When asked about the alignment of the ANet assessments to the standards or curriculum, there were few comments by leaders in Jefferson Parish, Chicago, and Boston that indicated misalignment was a major problem. Leaders in these districts were generally happy with ANet's shift to align with the Common Core standards (P2, P18, D26, P31). In Jefferson Parish, several leader comments indicated that teachers had the freedom to create their own pacing plan based on the ANet schedule of assessed standards, curriculum, and state or Common Core standards (D26, P28).

However, leaders in Springfield consistently shared issues of misalignment in year two (D37, D35, P34, P36, P40). One Springfield principal felt that the ANet assessments were reasonably well aligned with the Common Core standards, but not the

district pacing guide. In particular, one data leader noted that the district's math pacing guide and the ANet interim assessments were "about a month's difference." (P37). Even with this misalignment, leaders explained that teachers were generally expected to follow the pacing guide. One data leader described the tension this appeared to create:

> [T]he teachers are… stuck in this place where they're torn between the pacing guide and ANet, so, I think it has a huge impact on what happens in the classroom. As much as they don't want to quote-unquote teach to the test, many times it ends up that that's exactly what's happening because… [teachers] can go in and they can see what standards are coming along, so that it's kind of driving their instruction, and not necessarily driving it to the standards we should be teaching at this moment. So it does have an impact where the teachers are feeling that pressure…. [W]e don't want to be driven by ANet, we want to be driven by the standards, and we want to be driven by what the city says. (T35)

As mentioned above, some teachers added lessons to their plans when something on an upcoming ANet assessment hadn't yet been covered in the curriculum. However, a data leader explained that she didn't want to race through teaching the planned lesson just to squeeze in something that hadn't yet been taught: "I'm teaching for understanding and if I just teach to the standard [on the ANet assessment], I don't think I'm developing [students'] understanding." (D37). Similar to the leader who felt that ANet partnerships were something that needed to be supported at the district level, this data leader felt that alignment was another factor that needed to be resolved with the involvement of the district. An interview with the school leader of another Springfield school indicated that ANet and district leaders were working together to improve alignment (P40).

Like teachers who experienced misalignment between the ANet assessments and the curricular pacing guide, two Springfield leaders noted that the misalignment meant

that student data weren't particularly useful instructionally (D35, P36). One leader felt that the misalignment meant data couldn't be trusted as reliable (D35). Another explained that poor alignment with the pacing guide meant little actionable data for moving instruction forward (P36).

Summary

Leaders generally spoke positively about their ANet coaches and the support they were given to lead ANet implementation. However, though they understood the gradual release model, some leaders did express a wish that the withdrawal of coach support had not been so swift in year two. When comments about the assessments arose, rigor and alignment were again the most common themes. Leaders' perceptions of rigor were relatively consistent with teachers, both in how they compared the ANet assessments to the curriculum, standards, or state test, and the positive impact they thought it had on instruction. However, there was little qualitative evidence of the latter.

In Springfield, perceptions of poor alignment were common, but that there were relatively few comments on misalignment in other districts is notable. It's possible that leaders in these other districts were more distant from instruction and, therefore, less likely than their teachers to be exposed to issues with alignment of the ANet assessments to the curriculum and curricular pacing guide. In fact, some of the leaders interviewed in Springfield were instructional leadership specialists who, by definition of the role, likely were more involved in instructional support. It's also possible that leaders in other districts, particularly Jefferson Parish, gave teachers more freedom to backward plan their

instruction using the ANet schedule of assessed standards. While leaders in Springfield recognized the position in which this misalignment put teachers, there seemed to be less flexibility for teachers to ignore or alter the curricular scope and sequence. They acknowledged that the high level of rigor (i.e., difficulty) and poor alignment meant data were less useful for some teachers' instructional planning purposes.

ANET SCHOOL LEADERS' PERCEPTIONS OF TEACHERS' ATTITUDES & CONFIDENCE

Some leaders offered their perceptions of teacher confidence using data and various instructional strategies, as well as perceptions of teachers' attitudes towards assessments and assessment data. On the former, several leaders felt that their teachers' abilities to navigate and analyze the ANet data reports improved from year one to year two (P1, P34, P36, D38). However, several others felt there was work to do to get teachers to a point where they could analyze and process their students' data effectively and independently (i.e., outside of data meetings), understand and plan what to reteach and how to execute it, and break standards into smaller skills (P1, P15, P24, 27, P28, P34, P18). A data leader from one school explained that much of the upcoming data meeting would be spent on planning for the reteach, as that was an area in which her teachers needed to improve (D37).

Interviews with leaders suggest that teachers' attitudes were also changing over time, generally becoming more positive in year two. Overall, leaders who addressed this change in attitudes seemed to attribute it to teachers' realizations that ANet was a

valuable tool, not "just another test." (P31) Speaking about the changes in year two, one leader said:

> You know, [there is] definitely more receptivity, more acceptance of, this is what we do. And that's it. And while that may seem small, it's huge, because it opens doors for what we can do with the assessment. Once a mind is closed, it's really hard. (P31)

She attributed this change in teachers' perspectives to the fact that, in the prior school year, ANet projections of student performance on the state test were pretty accurate, essentially validating the tests and process (e.g., analysis and reteaching) for teachers (P31). This sense that teachers' perceived value of ANet increased over time, and was bolstered by student success and validation of the process, was echoed by other school leaders (P5, P40, P36, P28) and some teachers (T4, T6). However, like confidence levels, some leaders felt there was work to be done demonstrating to their teachers that ANet was useful tool generally, or for their students in particular (P12, P28, D26, P27).

ANET SCHOOL LEADERS' PERCEPTIONS OF TEACHERS' PRACTICES

In this section, leaders' comments on their teachers' data use and instructional practices are reviewed in order to provide comparison with teachers' self-reported practices. Some leaders commented that ANet was just a different way of doing some of things they already did around instructional data use. However, leaders also felt that ANet had given teachers a structure for these practices (e.g., templates and processes for reviewing data and identifying the key areas students were struggling) and added some key component that had been previously missing (e.g., formalizing planning of the reteach and reflecting on its success).

Data Review and Use

Across interviews, leaders described many of the same uses of the ANet data as teachers self-reported. They explained that teachers were using data to pinpoint trends in results and gaps in instruction, identify skills (or standards) on which students were struggling, and uncover possible misconceptions that were driving incorrect student responses to items (P5, P12, P24, P17, P18, D38, D13). For example, one leader explained that working with ANet meant every conversation about student achievement was grounded in actual data, "We don't have to ever wonder. We actually talk about the specific strands, the skills, the domains that… are challenging kids." (P5) Several leaders talked about their teachers' using a process of item analysis, a practice ANet encouraged as a way of identifying the key skills involved in answering a specific assessment item correctly (P15, P18, P24, P27). Other leaders talked about ways teachers used the data to focus – and maximize the impact of – their instruction. For example, one leader explained how her teachers used ANet data to identify the "neediest" students, students whom teachers can "move" by reteaching (P1). Following similar logic, other leaders explained how teachers identified the priority standards on which students struggled and focused their reteach on them (P36, P18, P34).

Instructional Planning & Practices

One of the most consistent comments from leaders regarding teachers' instructional planning strategies was that there was an increased focus on the standards due to ANet, particularly the Common Core standards, and how to align their instruction

(P5, P15, P18, P40). However, rather than aligning directly with the standards, many other school leaders discussed trying to engage teachers in backwards mapping instruction to the skills and standards on upcoming ANet assessments (P1, P24, P28, P36, D37, D26). Two of these leaders described how their teachers' planning strategies transformed over time. Initially, teachers were not using backward planning with fidelity; however, those who responded to their leaders' encouragement to adopt the practice tended to see improvements in their students' performance on the ANet assessments which reinforced its utility as a planning strategy (P24, P28).

As discussed earlier, several leaders felt their teachers had improved in their data analysis skills; these improvements were perceived by a few leaders as having instructional benefits. Specifically, three school leaders felt that teachers were able to shift focus from learning data analysis skills to implementing ANet practices related to instructional planning and reteaching in year two (P36, P40, P15). One of them explained how they had focused much of their first year of implementation on improving teachers' data analysis skills,

> [I]t wasn't until kind of the second half of the year that we moved into really developing action plans. So that was another thing where this year, we could start right away, because we kind of, we know what to do with the data and really are spending much more time on the planning. And we've seen some great results from reteaching. (P40)

Also as discussed earlier, several leaders felt that reteaching was going well (D38, P40) with some teachers starting to shift to individualized or small group instruction in addition to or in place of whole-group instruction (P1, P18, P27, P28, D35, P31). However, as with other skills, attitudes, and practices, this was an area where leaders saw

room for improvement (P28, P31, D35). When asked about their expectations for partnering with ANet, two school leaders said that they hoped that teachers would utilize more individualized instruction, but that they weren't there yet. One of these leaders explained that teachers had initially been generating a single grade-level reteaching plan. However, she wanted that to change:

> What I want them to look at, from here on out, is their individual students, and putting them into groups of three or four, or however many kids missed that particular standard, and do a differentiated instruction type of reteaching, rather than, we're going to go back, we're going to give up math for a whole week, and we're just going to go back and we're going to reteach this whole unit. (D35)

As one of the leaders pointed out, the expectation of providing teachers with student-level assessment results was that teachers would take advantage and develop instructional plans for each student (P28).

Summary

Overall, leaders seemed to have a good sense of the types of data-based instructional practices their teachers were implementing. Leaders acknowledged that teachers were using data to better pinpoint students' learning needs, but they still saw areas in which their teachers could improve instructional planning (i.e., the reteaching plan) and practice. For example, leaders acknowledged that teachers were not always using the detail provided by the ANet data to implement small-group or individualized instruction based on student need.

QUALITATIVE RESULTS SUMMARY

Teachers' seemed to appreciate the speed with which students' results were returned after the assessment administration. Coupled with the level of detail of ANet's data reports, many teachers explained that they were not only better able to identify students' gaps in learning, but also catch them more quickly than they had previously. For many, this took the guess work out of determining whether students had mastered a skill. Many teachers also commented on the usefulness of the detailed student data for planning their reteach; i.e., providing clarity on what to focus. Among those teachers who described the way they selected a skill for reteaching, most seemed to focus on a skill or standard on which most of their class performed poorly. Interviews with leaders corroborate teachers' statements about their data analysis and planning practices.

What is less clear is how these analysis and planning practices translated into instruction. Grouping of students during the reteach seemed to be a practice mentioned by many teachers, but it wasn't always clear what instruction of these groups looked like or how differentiated it was. Overall, few teachers expressly mentioned individualizing instruction based on student performance on the ANet assessments. Those teachers who said that they were differentiating instruction based on their students' ANet results seemed to have the benefit of a co-teacher or small classes. Some practices may have actually been counterproductive to student learning. A few teachers described being encouraged to focus their reteach on a skill that was measured by only one or two items. Others were encouraged to develop a grade-level reteaching plan; sometimes the focus of the reteach was a skill their students seemed to have mastered, but their peers' students

hadn't. Many teachers said that they tried to reteach a skill differently than they did the first time; however, a few teachers explained that they didn't always know another way.

It seems that ANet first focused on teachers' planning and implementation of the reteach, but in year two, had begun encouraging teachers to backward plan their instruction. As a result, teachers more frequently commented on their reteaching practices than their backward planning. Likewise, reteaching seemed to be the most common way in which the analysis of student data affected instruction. Many leaders also seemed to think that reteaching was going well, but some thought that teachers could be individualizing instruction and backward planning more often. Overall, it appears that ANet had an order in introducing skills to teachers: data analysis, then reteach planning, then backward planning.[6]

Teachers were not directly asked about their principal's or other leaders' role as instructional leader. However, in interviews, teachers described the expectations their school leader set for data use, data meetings, and the reteaching. When they offered examples of how they interacted with leaders in their school, most comments focused on the reteach, though there were not many examples of their principal providing them with detailed instructional feedback. In some cases, teachers felt that they were encouraged by leaders to focus their reteach on the wrong skill. The few examples of supportive feedback seemed to come from teachers working with a master teacher or instructional coach. In the end, a few teachers expressed a need for more instructional feedback.

[6] Correspondence with ANet staff (May 2016) indicates that this is consistent with the introduction of skills in most schools.

A consistent finding from teacher interviews was that leaders tended to collect teachers' reteaching plans and hold them accountable to other steps in the ANet data cycle. This is a role that has been validated in the literature, but teachers had a range of reactions. Some explained that being held accountable made them more likely to follow through with the implementation of their reteach. Others seemed to resent the oversight and felt that it was unnecessary. Leaders, too, acknowledged their dual role of providing feedback and holding teachers accountable for the implementation of ANet practices. Some described using the reteaching plan and lesson as opportunities to provide teachers with feedback, but other leaders used the lesson as a way to confirm teachers were implementing their reteaching plan. This is not to say that each leader followed only one approach. There may be a need for both perspectives; i.e., providing instructional feedback as opportunity to help teachers improve practice and accountability as a way to ensure they actually carry it out.

Collaboration came up almost exclusively in the context of an interview question which asked teachers to describe their discussions of data with peers. As a result, most teachers mentioned using data meetings or team meetings as a chance to analyze data together and brainstorm instructional strategies. Likewise, most leaders observed these opportunities as a way to foster collaboration among teachers in their schools. Both teachers and leaders tended to have a positive view of these opportunities; however, one leader provided an example that might call into question exactly what teachers meant by collaboration.

In addition, a collaborative, sharing atmosphere wasn't present in all schools. A few respondents – both teachers and leaders – provided examples of teaching staff that were unwilling or afraid to share student results or provide instructional feedback. It isn't clear how this culture arose, but in one district, it could have to do with some school leaders apparently including students' ANet results in their teachers' evaluations. Placing higher stakes on students' ANet results could inhibit a trusting, collaborative culture that would make teachers more willing to openly share student results and instructional strategies.

Establishing teachers' attitudes toward interim assessment and assessment data was challenging as this was not a focus of the interview protocols. At various points in the interviews, teachers spoke about their opinions of ANet which were generally favorable in year two. Some teachers, however, noted that they had come to like ANet because of the valuable resources, the support of a coach or leader, or only after seeing improvement in their students' test results which they attributed to ANet practices. There were teachers who still maintained opposition to ANet for reasons such as the amount of time assessments took from teaching or a belief that it encouraged teaching to the test. Leader interview data were consistent, explaining how teachers' attitudes toward ANet had trended more positively over the two years, but some teachers had still not been convinced of the value of the program.

Teachers' comments on their skill in implementing aspects of the ANet program were much like their comments on attitudes toward ANet. They generally felt that their skills had improved over time and, in year two, they were more comfortable with the

process overall. Specifically, teachers said that they better understood their students' data, were quicker at analyzing the data, and felt that they understood how to focus their reteach more effectively. If teachers alluded to any gap in their skills, it was their knowledge of alternative ways to teach a skill during the reteach. Leaders might have been slightly less sure that teachers' skills in data analysis, reteach planning, or differentiating instruction were where they thought they should be. However, they too acknowledged that teachers' skills had improved over time.

One of the most consistent and salient points made across interviews was teachers' opinions of the ANet assessments themselves. This may be in part because it was one of the only topics that was coded for this study that also was a focus of the larger evaluation. However, it also points to a likely source of the unexplained variance in teachers' data-based instructional practices and why no impact was seen on the frequency of differentiation of instruction. Teachers universally felt that the ANet assessments were more rigorous than their curriculum or state test. A few leaders and teachers thought that this had improved the rigor of instruction, but there is little evidence in this study or the larger evaluation to suggest this was true in many classrooms. Other teachers who looked favorably on the level of rigor of the ANet assessments thought they were preparing their students for the state test, Common Core, or Common Core-aligned tests to be rolled out in the near future. Some of these teachers used the ANet resources to further prepare their students for the types of items they would see in the new curricula and tests.

In contrast, many teachers had an unfavorable view of the level of rigor; though, they seemed to use rigor as a synonym for difficulty. They noted that the ANet interim

assessments were often too difficult for their students – ELL, special education, or particularly low-performing regular-education students – to perform well and, therefore, results were not always instructionally useful for the reteach. The lack of utility of results was also a common complaint among teachers who cited poor misalignment of the ANet assessments to their curricular pacing guide. Some teachers expressed a choice when faced with misalignment. Some teachers carried out their lesson plans as scheduled knowing that when students took the next ANet interim assessment, some of the items would relate to skills and standards they had not yet covered. Alternatively, some teachers opted to fit in lessons on a skill if it was on the next ANet assessment, but they'd not yet come to it in the curriculum. For a few teachers, this raised concerns that they were not teaching to mastery or disrupting the planned lesson sequence.

CHAPTER SIX: SUMMARY & CONCLUSIONS

This chapter begins with an integrated discussion of the quantitative survey results and the qualitative interview results. The quantitative results provide estimates of the impact of the Achievement Network (ANet) on teachers' data-based instructional practices and the mediating or moderating roles played by school- and teacher-level conditions. The qualitative results provide context for patterns in quantitative findings. Specifically, they provide potential explanations for why ANet had an impact on some outcomes and mediators, but not others. They also offer evidence of the mechanisms by which the school- and teacher-level conditions mediate the relationship between ANet implementation and teachers' uptake of data-based instructional practices.

Overall, the alignment of quantitative and qualitative findings is high. Some correspondence could be expected simply because of the methods that went into preparing for the study. Specifically, the same thorough review of research pertaining to the constructs of interest to this study served as the basis for informing the survey scale revisions and the coding framework for qualitative analysis. However, it is the degree of correspondence in the findings that is notable. Teachers and leaders in ANet schools reported conditions and practices that, in aggregate, were generally consistent across the two data sources, as well as with the ANet program model and findings from prior research.

After the discussion of key findings, an overview is provided on the steps taken to validate the methods, measures, and results of this study. The chapter includes a review

of the study limitations and opportunities for future research, and concludes with a final summary of lessons learned.

SUMMARY OF KEY FINDINGS

Research Question One

Research question one was concerned with differences in teachers' data-based practices in ANet and control schools. Based on survey analyses, there is evidence to suggest that, after two years, ANet had a moderate impact on the frequency with which teachers reviewed data alone or with others (data review) and smaller impact on the frequency they used data in various ways (data use). However, the more frequent focus on students' interim assessment data did not translate to more frequent use of various instructional planning strategies or differentiation of instruction in ANet schools compared to control schools. In fact, unreported analyses indicated that teachers in ANet schools were using whole-class instruction more frequently than their control-school counterparts.

Teachers in ANet schools described the many settings in which they reviewed data – notably, in the ANet data meetings and team meetings. They also appreciated the speed and ease with which they could use their students' ANet results to identify gaps in learning. Differences in the frequency of instructional planning favoring ANet teachers may have been marginally significant in survey analyses because a few items referred specifically to planning the reteach, making the construct overly aligned with terminology and practices that would be familiar to ANet teachers. In contrast, the rest of

the instructional planning survey index items referred to practices such as backward planning or creating differentiated instructional plans. While evidence from interviews suggested that most ANet teachers were planning and implementing the reteach component of ANet, they were less regularly utilizing backward planning or instructional differentiation.

Finding that ANet teachers, on average, reported a greater use of whole-class instruction in year two also fits with data gathered from interviews. Teachers described focusing their reteach on skills on which the majority of their class had performed poorly. A few even mentioned being encouraged to develop a grade-level reteaching plan. While these situations do not necessarily require whole-class instruction, reteaching a single skill would tend to favor the practice over identifying and grouping kids based on different learning needs.

These results may also indicate a shift in ANet's implementation focus from year to year which parallels the impacts seen on teachers' data-based instructional practices: i.e., data analysis and reteach planning in year one, shifting to the inclusion of backward planning in year two. This is consistent with the frequency these practices were seen across the research reviewed in chapter two (Clune & White, 2008; Christman, et al, 2009; Goertz, Oláh, & Riggan, 2009a). In studies of programs like ANet, researchers described teachers' review and use of student data to plan a reteach more frequently than backward planning and differentiating instruction.

Ultimately, changes in instructional planning and practice might require more time and skill to implement in comparison to reviewing and using data. Creating the

defined space of the data meeting in treatment schools, something control teachers might not have experienced, might explain the impact of ANet on teachers' data-based practices. ANet teachers might have also increased their focus on data during team meetings as a result of ANet. However, affecting change in teachers' instructional practices through professional development is notoriously difficult (Richardson, 1990; Guskey, 2002). This might be especially true in the context of ANet where minimum implementation could be as limited as reteaching lessons around four quarterly interim assessments.

There has also been an identified gap in professional development around instructional data use; specifically, a gap in helping teachers translate their analysis of student data into effective instructional plans (Clune & White, 2008; Goertz, et al., 2009a). This, too, could explain ANet's impact on teachers' data practices, but not on instructional practices. Given that ANet coaches do not focus directly on training teachers and instructional support from school leaders may be insufficient, some teachers in ANet schools might lack this skill.

Survey results also show that when compared to their counterparts in Boston, teachers in Springfield reported a greater frequency of all four practice outcomes, teachers in Chicago reported a greater frequency of reviewing data and using various instructional planning strategies, and teachers in Jefferson Parish reported a greater frequency of using various planning strategies. Models that include the treatment-by-district interaction terms show that the impact of ANet on teachers' data-based instructional practices was not the same in every district. The impact of ANet on the

frequency teachers review data in Boston (0.59 *sd*) and Jefferson Parish (0.58 *sd*) is significant ($p < 0.05$). The impact of ANet on teachers' data use in Jefferson Parish is also significant (0.52 *sd*, $p < 0.05$), as is the impact on teachers' instructional planning in Springfield (0.57 *sd*, $p < 0.05$). However, the impact of ANet on teachers' data-based practices in Chelsea and Chicago is often negative, though not statistically significant. Exploring this district variation further was beyond the scope of this study, but the implications are discussed in later sections.

Research Question Two

Models estimating the impact of ANet on proposed school-level mediators were likely underpowered. However, standard deviation differences in school-mean ANet- and control-school teachers' perceptions of their leaders' instructional leadership abilities, as well as school professional and achievement cultures, were moderate and positive indicating higher levels of these school-level conditions in ANet schools. The differences in teachers' individual attitudes and confidence levels were much smaller and not statistically significant. This apparent difference in impacts of ANet on school conditions versus teacher characteristics is not surprising given ANet's theory of action and program focus. These findings also have implications for their potential to mediate the impact of ANet on teachers' data-based practices, discussed below.

Taking a closer look at the survey items asked of teachers regarding their leaders' instructional leadership abilities, the majority focus on setting and monitoring goals, standards for learning, and practices around data use. Three of the nine items focus more

directly on supporting teachers and providing feedback. The fact that school-mean teacher ratings of their leader abilities' was higher, on average, in ANet schools is not surprising especially in light of the types of leader practices ANet leaders and teachers described in interviews. Establishing norms and expectations and monitoring implementation were commonly mentioned practices and prior research has found that these activities are an important part of leaders' roles in supporting instructional data use in their schools (Halverson, Grigg, Pritchet, & Thomas, 2005; Heritage & Yeagley, 2005; Marsh, Pane, & Hamilton, 2006; Datnow, Park, & Wohlstetter, 2007; Goertz, Oláh, & Riggan, 2009a; Blanc, et al., 2010; Coburn & Turner, 2011; Datnow & Park, 2014; Gerzon, 2015). However, interviews also uncovered a need for effective instructional feedback.

In terms of culture, conditions described by teachers during site visit interviews are also consistent with the types of items asked of teachers on the year-two survey. For example, achievement culture came up in interviews in the context of high standards and expectations for student learning and teachers' acknowledgment of their responsibility for student success. However, recall that a few ANet teachers acknowledged their peers were not open to sharing student results and instructional remediations.

Finally, survey data support a conclusion that ANet had little impact on teachers' attitudes toward assessment and assessment data. To the extent that teachers' attitudes changed, they likely only changed in respect to ANet specifically, not the more general construct measured by the survey scale. Interview data indicate that teachers' attitudes toward ANet improved from year one to year two, often as a result of implementing ANet

practices and experiencing improvements learning. However, some teachers still held neutral to negative opinions of ANet in year two. Some school leaders hoped that these teachers' opinions of ANet would also improve as they experienced a connection between ANet practices and student learning.

Likewise, evidence of the impact on teachers' confidence in using data and various instructional planning strategies is sparse. Teachers generally felt they had improved in certain skills – e.g., data analysis, focusing their reteach – but leaders tended to see room for growth. Like frequency of practices, many of the items asked on the year-two survey referenced confidence in activities such as differentiation, backward planning, and assessing gaps in student learning or the school's curriculum, activities that likely represent more advanced levels of implementation not yet achieved by most ANet teachers.

Research Question Three

Results from the survey analysis showed that, controlling for treatment-group assignment and all covariates – each of the school- and teacher-level conditions of interest (i.e., mediators) were strong predictors of teachers' data-based instructional practices. This provides evidence that the conceptual model underpinning this study is sound; the focal school- and teacher-level conditions in this study are related to greater frequency of teachers' data-based practices. The results also showed that controlling for treatment-group assignment, all covariates, and all other mediators, the frequency of teachers' collegial discussions during common planning time, as well as their own

attitudes towards and confidence in using data remain important predictors of teachers' data practices.

However, evidence is less conclusive that these school-level conditions and teacher characteristics can explain the positive impact of ANet on teachers' data-specific practices. When sets of school- and teacher-level mediators were included in models individually, the results suggest that, compared to teacher-level characteristics, hypothesized school-level mediators may explain more of the impact of ANet on teachers' data practices. This is not surprising given ANet's modest impact on instructional leadership and school culture, but lack of impact on teachers' attitudes and confidence after two years.

In models that simultaneously included both sets of mediators, the school- and teacher-level mediators reduce the estimated impact of ANet on both data-based practices by just over one-third. This supports the study's hypotheses that school leadership and culture – and to a lesser extent, teacher confidence and attitudes – play a role in the adoption of teachers' data-based practices; however, these results are not causal and only suggest that these conditions partially mediate the relationship. Furthermore, it is evidence that there likely are other important factors that explain teachers' implementation of data-based practices.

The qualitative results provide some context for the patterns in findings from survey analyses. Specifically, they point to potential reasons for an impact of ANet on data practices, but not instructional practices, as well as the role played by the proposed mediators. Recall earlier evidence of the numerous occasions set aside for teachers to

meet with one another to discuss data. ANet not only expected schools to hold quarterly data meetings, but teachers described meeting in grade-level or subject-area teams to review student data and discuss instructional strategies. However, there were relatively few descriptions of instructional practices such as differentiation. Among the small number of teachers who talked explicitly about differentiation, they citied co-teaching and small classes as a key facilitators.

What context do school- and teacher-level characteristics provide for these patterns in findings?

Leadership. In considering how leadership may have helped or hindered teachers' data-based practices, a few findings from interviews stand out. First, it is notable that there was as frequent reference to leaders' management of ANet implementation and setting of expectations for teachers' data-based practices as there was to their provision of instructional support. A few teachers did provide specific examples of leaders who worked closely with them to plan their reteach, but as many expressed frustration over being encouraged to focus on the same lesson as their grade-level peers even when it wasn't a skill on which their students struggled. This may reflect leaders' judgments of their teachers' proficiency in planning a reteach that required differentiation. However, it seems counter to ANet's purpose and the fact that the program provides teachers with close to real-time, student-level assessment data.

With regard to managing implementation, leaders frequently described collecting teachers' reteaching plans. While some leaders used this opportunity to provide teachers with feedback, they also cited a perceived need to hold teachers accountable to

developing and carrying out their reteaching plan. Teachers clearly picked up on this accountability and some even acknowledged they may not have been as likely to implement the reteach without it. However, in addition to ensuring implementation, teachers likely need to be provided with meaningful feedback and support; especially when a reteaching plan requires new instructional techniques. A key to ensuring teachers receive quality feedback might be what some school leaders alluded to in interviews: continuing to build the capacity of other leaders in the school so that the work of ANet implementation can be distributed. Beyond that, some leaders might need training on effective instructional leadership skills. For ANet to have an impact on teaching and learning, leaders need to support teachers in using the richness of available student data to differentiate instruction according to students' needs and provide structures that allow them to use differentiation in more than just a handful of lessons.

Culture. It makes sense that the frequency of teachers' common planning time (CPT) discussions would be positively related to reviewing data more often; as mentioned earlier, interviews suggested that teachers had frequent opportunities to meet with peers and used those meetings to review student data. What is important to note is that results also showed that the frequency of CPT discussions predicted the frequency they used data in various ways. Both relationships persist in models that control for treatment-group assignment, other mediators, and covariates.

Interview data on teacher professional culture provided evidence of collaboration in the context of discussions of data and instructional practices. Teachers and leaders generally felt that these were positive and useful experiences; however, two exceptions

stand out. First, a small number of leaders shared examples of what their teachers considered collaboration, but fell short of their expectations of true collaboration. Even though some of the examples of working together in data meetings and teacher team meetings likely fall under what Hargreaves (1994) called *contrived collegiality* rather than true collaboration, recall that contrived collegiality has the potential to turn into collaboration (Datnow, 2011) and could still effect change when coupled with a strong achievement culture (Hargreaves, Morton, Braun, & Gurn, 2014; Little, 1999). Though examples of achievement culture in teacher and leader interviews were few, they were consistent. Teachers held high expectations for student achievement and seemed personally committed to helping student attain them. Therefore, mandated settings like data meetings might still be useful for facilitating effective collaboration.

The second point is that some teachers and leaders described cultures in their schools that were characterized by a lack of trust or collaboration around student data. Specifically, teachers did not want to share their students' data or feared being perceived as overstepping bounds by offering advice to others. Trust has been shown in other studies to be an important factor facilitating collaboration around data-based strategies, especially when the data culture was non-punitive (Datnow, 2011). Though the evidence is extremely limited, teachers in one district did make a connection between a lack of collaboration and their school leaders' inclusion of ANet data in their performance evaluation. This use of ANet data was counter to the program's purpose and may have undermined collegial trust and program effectiveness in these schools.

Skills and Confidence. Interviews suggested that positive changes in teachers' attitudes toward ANet could come as a result of seeing its utility as a tool for improving teaching and learning. Teachers' attitudes might also improve with better alignment between the assessments and the curricular scope and sequence. Prior research supports a connection between teachers' attitudes and assessment utility and validity, as well as improving perceptions and use of data through training (Chen, Heritage, & Lee, 2005; Kerr, et al., 2006; Marsh, Pane, & Hamilton, 2006). It is difficult to argue that ANet should make an explicit attempt to change teachers' attitudes around assessment and data use more broadly. However, ensuring sufficient and appropriate training (e.g., in data literacy), as well as the validity of their assessments and assessment data for informing instruction, is likely critical for teachers' adoption of data-based instructional practices.

In contrast, an explicit focus on improving teachers' skill and confidence could be a missed opportunity for ANet given the strength of the quantitative relationship between teachers' data and instructional confidence and their frequency of data-based instructional practices. In interviews, teachers generally felt they better understood ANet and their students' data going into year two. Teachers also generally said they were better at analyzing data and focusing their reteach. If they acknowledged any gap, several teachers mentioned needing support to identify new ways to reteach topic on which their students struggled. Leaders also felt that teachers' skills had improved over time. However, they tended to acknowledge room for improvement in their teachers' abilities to analyze data effectively and independently and improve the planning and execution of their reteach.

Overall, interview data suggest that there was a range in ANet teachers' attitudes, confidence, and skills in year two.

Other Factors. The results of research question three indicate that the school- and teacher-level mediators of interest in this study may explain some of the impact of ANet on teachers' data-based practices, but that there are likely other important mediators not accounted for in the conceptual and statistical models. The interview data provided insight into what some of these mediators might be. One of the most important seems to be the rigor and alignment of the ANet interim assessments and the role this played in providing teachers with valid, reliable, and useful information on their students' learning on which to base instructional decisions.

In interviews, ANet teachers universally stated that the rigor of the ANet interim assessments was high; though, some teachers equated rigor with difficulty. A large proportion of teachers also noted the misalignment between the assessments and their curricular scope and sequence. While teachers varied in whether they viewed rigor positively or negatively, misalignment was consistently viewed as problematic. Even with high rigor and poor alignment, teachers likely still *reviewed* data in the ways captured by the survey. However, it was clear that rigor and misalignment affected teachers' ability to *use* data to plan and differentiate instruction, thus potentially contributing to the pattern in results in research question one.

In interviews, a few leaders and teachers felt the rigor of the ANet assessments raised the rigor of instruction. However, there is little additional evidence to support this. More often, teachers appreciated the rigor as a way to introduce students to what they

would encounter on the state summative test. Interview comments on the ANet assessments and the use of ANet resources – e.g., the quiz tool – raise concerns that some teachers' saw ANet as a form of test preparation.

In contrast, some teachers viewed the level of rigor negatively. These teachers tended to equate rigor with difficulty, however. They also tended to be teachers of particularly low-performing students, English-language learners, or special education students. They explained that, because of the level of difficulty, students performed so poorly that results were not always useful for informing or focusing instruction.

Teachers who expressed dissatisfaction with the alignment of the ANet assessments with their curricula, or with scope and sequence, raised similar concerns. When alignment was poor – e.g., the ANet tests covered skills that hadn't yet been taught – and teachers either were not permitted to realign their lessons or opted not to, resulting data were less useful for instructional decision making. When teachers attempted to realign their lessons to upcoming ANet test content, some had to prepare or fit new lessons into existing plans. Not only did this have the potential to disrupt planned lessons but also, as one teacher pointed out, meant teachers may not have had time to teach the skills for mastery. In sum, planning and differentiating instruction was likely challenging when faced with assessments that were too difficult for their students or poorly aligned with the curriculum, resulting in data that provided limited evidence of gaps in student learning.

District Variation. Before moving to the final research question, it's worth noting a few district patterns that stand out. First, controlling for treatment-group assignment,

school- and teacher-level mediators, and all other covariates (i.e., the "fully adjusted" model), the treatment effect in Chelsea remains lower, or less positive, than in Boston, but only in the frequency with which teachers' review data (table 4.14). The effect of ANet on the frequency teachers review data in Boston remains significant in the school-, teacher-, and combined-mediation models (all $p < 0.05$). The impact of ANet on the frequency teachers review data in Jefferson Parish remains significant in the teacher-mediation model ($p < 0.05$).

Second, in "fully adjusted" models, teachers in Jefferson Parish reported less frequent review and use of data than their peers in Boston (table 4.13). "Partially adjusted" models (i.e., containing no school- or teacher-level mediators, but other covariates) showed no difference in the frequency of data-based practices in these two districts. In Chicago, the partially-adjusted model showed that, on average, teachers reviewed data more frequently than their Boston peers. In Springfield, partially-adjusted results showed more frequent review and use of data compared to Boston. However, the fully-adjusted models showed no differences in the frequency of data review or data use in these districts.

Ignoring relatively high frequencies of some key school- and teacher level mediators of interest, teachers in Chicago, Jefferson Parish, and Springfield reviewed and used data as often (Jefferson Parish) or more often (Chicago, Springfield) than their Boston counterparts. However, accounting for differences in key mediators between teachers in these districts and Boston, the story changes; Boston teachers reviewed and used data as often or more often. This provides additional evidence of a positive role

played by school and teacher characteristics in teachers' implementation of data practices. Should Boston focus on improving these conditions, their teachers might report at last as frequent data review and use as other districts.

Research Question Four

Research question four examined whether a subset of measures on the ANet baseline school screener acted as moderators of the impact of ANet on teachers' data-based practices and could identify schools as ready to implement the program at the outset. Though differences in ANet and control-school teachers' data-based practices in higher readiness groups were generally larger than in lower readiness groups, the differences in impacts between readiness groups were small and not statistically significant.

However, when schools' total baseline readiness screener scores were used, there were greater differences in the impact of ANet on teachers' data-based instructional practices – data review, data use, and instructional differentiation – in the highest and lowest readiness groups (West, Morton, & Herlihy, 2016).[7] Specifically, the larger evaluation used all nine baseline readiness subcategory scores to create three groups of schools. The impact of ANet on teachers' practices in the highest readiness group were more positive than in the lowest readiness group (West, Morton, & Herlihy, 2016). Compared to the five screener subcategory scores used in this study, the total school

[7] Differences in model specification and estimation result in slight differences in the full sample estimates of teachers' data-based practices in this study versus the larger evaluation.

screener score could be a better resource in helping ANet identify schools with conditions in place that facilitate more successful implementation.

Additionally, analyses reported in appendix B showed that school- and teacher-level mediators – instructional leadership, school culture, and teacher attitudes and confidence – explained more of ANet's impact on teachers' data-based practices in lower readiness schools than in higher readiness school. This may be because higher readiness schools start out with a more supportive environment. For schools rated lower on readiness, it appears more important to foster supportive instructional leadership, school culture, more positive attitudes, and greater teacher confidence in order to encourage teachers' adoption of data-based practices. In sum, much like ANet's focus on tailoring instruction to students' needs and, to the extent that baseline school readiness predicted differences in teachers' uptake of data-based instructional practices, some schools may require more or different types of support in implementing the program from the outset.

Subgroup Variation and Replicability

Evidence of variation in the impact of ANet on teachers' data-based practices across districts and, to a lesser extent, readiness groups, speaks to concerns over the replication of research. Replicability is a hallmark of scientific research; it provides evidence of the strength and validity of scientific claims. The topic was made mainstream in the fall of 2015 with the release of a report from the Reproducibility Project, a collaboration of researchers at the Center for Open Science (the Open Science Collaboration, OSP). These researchers found that in their replication of 100 published

psychological studies, results were generally much weaker than the original studies as judged by comparisons of effect sizes and the frequency of statistically significant results (Open Science Collaboration, 2015).

There have been responses to the OSP report; specifically, the accuracy of *their* conclusions about replicability. Gilbert, King, Pettigrew, and Wilson (2016) found that when they corrected for flaws in the OSP replications – i.e., the comparability of methods used in the original studies and replications – they reached an entirely opposite conclusion. They found that replicated results were generally consistent with initial results and the number that failed was in line with the number expected to fail by chance.

Despite this, replication of research continues to spark concerns across a wide range of fields. Benjamini (2015) raises concerns about variation in study results across sites (e.g., laboratories). He cautions that when treatment-by-site interactions exist, standard estimates of variability used in calculating *t* statistics fail to account for additional noise in estimation and can lead to a greater likelihood of type I errors. By accounting for the treatment-by-site interaction, one can reduce false positives and better identify results that are more likely to be replicated. In the context of this study, the interaction between ANet implementation and district context is important to acknowledge. It may be impossible to know whether or how ANet will "work" in a new district, but substantial district variation in the program's effectiveness is likely and, thus, generalizability of these results to new districts may be difficult.

On a related note, recent efforts to explore null and site-specific results in RCTs has generated several frameworks. For example, the results of Jacob, Jones, Hill, and

Kim (2015) might also suggest that variation in results across districts suggests implementation and contextual differences – e.g., school- and teacher-level mediators – might play a role. Similarly, a framework by Weiss, Bloom, and Brock (2013) considers the roles of the contrast, clients, and context under which an intervention takes place. Contrast refers to differences in the conditions groups experienced; for example, the ANet treatment versus business-as-usual in the control group. This contrast explains program impacts or lack thereof. Characteristics of the clients (i.e., study participants) or context are used to explain variation in program effects; they typically moderate the size or direction of effects.

VALIDATION OF RESULTS

The validity of this study's design and results have been discussed throughout previous chapters. Here, the separate discussions are summarized in an effort to assure readers of the strength and merit of the findings.

Quantitative Results

First and foremost, the matched-pair school randomized design of the larger i3 evaluation of ANet provided high confidence in the findings for research questions one and two even in the face of school attrition. Randomized designs have the benefit of ruling out plausible threats to internal validity including selection bias. The matched-pair design helped to maintain treatment- and control-group equivalence when schools left the study after randomization. Still, sample recruitment and attrition do play a role in the

external validity of findings. This study's results are likely generalizable to other urban, low-performing schools with the financial resources to pay for a program like ANet and the motivation to maintain the program for multiple years.

In refining the year-two surveys that served as the basis for the quantitative analyses, attention was paid to content and construct validity. A thorough search of the literature related to each of the school- and teacher-level mediators and teacher practice outcomes was performed. This helped ensure that scales and indices would have high construct representativeness; new scales and items were added or removed, as needed. Analyses of baseline and year-one survey data were undertaken to explore whether any existing scale items were a poor statistical fit (i.e., having low item-scale correlations and component loadings). Finally, the correlations between year-two scales were calculated to demonstrate that measures within each level – school and teacher – were positively correlated, but not so highly correlated as to raise concerns of multicollinearity.

Lastly, the quantitative analyses were guided by standards of statistical conclusion validity. The methods used to ensure content and construct validity also sought to reduce measurement error. Power was determined to be sufficient for the detection of an effect of ANet on teacher practices that would be consistent with what is likely required to have a subsequent impact on student outcomes. School matching and covariate-adjusted models likely contributed to the sufficiency of power. When necessary, analyses were conducted using multilevel models so that the clustering of teachers within schools was taken into account. This allowed for the proper estimation of standard errors and the reduced likelihood of committing a type I error.

Qualitative Results

With regard to the validation of the qualitative findings, three former members of the i3 evaluation team at Harvard reviewed the results as a way to ensure high descriptive and interpretative validity. All three judged the qualitative findings to be consistent with their recollection of the site visits in which they participated and the site visit data analysis that was completed for the larger evaluation. The alignment of the qualitative results with the quantitative results, conceptual framework, and prior research all provide evidence of the validity of the findings; however, it is worth noting that few disconfirming examples of conditions were found in the interview data. This alignment of findings also supports the internal consistency of the qualitative findings and their external validity. The purposive sampling of schools that participated in the site visit data collection also provides a measure of external validity. Schools were purposively selected in an effort to understand the adoption of ANet practices in schools with varying levels of fidelity of implementation and coach support. However, external validity is likely limited by the fact that only about one-third of the treatment schools were visited.

LIMITATIONS

Despite this study's contributions to the field, there are some limitations that bear discussion. Most of these limitations were discussed at various points in previous chapters. This section recounts these limitations and raises several new ones. Naturally, some pose greater threats to the findings in this study than others.

Design

The Counterfactual. Because data-based instructional programs, including interim assessments, are widespread in American schools and because they are intended to improve teaching and learning, no restrictions were placed on control schools' use of similar practices. They were only prohibited from participating in the ANet program or adopting their practices until the end of the two-year treatment period. This means that the conditions in treatment schools and the practices of treatment teachers are compared to business-as-usual in control schools. Data collected from surveys and conversations with each district's central office staff indicate that business-as-usual included the administration of interim assessments in some grades and subjects, and varying types of associated supports (e.g., coaching, professional development).

This contrast in treatment- and control-group conditions means it is possible that effects of ANet were smaller than they would have been relative to the absence of any data-based instructional practices in control schools. Conducting a randomized-controlled experiment in an educational setting where the control-group condition is the absence of the treatment can be incredibly difficult, but with a treatment such as ANet, it was simply unrealistic. Therefore, this is less of a limitation than a context in which to view the results of this study. Where the survey results detected statistically significant differences, those differences can be claimed as the unique effect of ANet compared to other data-based instructional initiatives.

Internal Validity. One main threat to validity can be attrition of the sample after randomization. Before attrition, randomized treatment and control groups are said to be

similar in expectation on all observable and unobservable measures. However, the differential loss of sample in treatment and control groups can threaten initial group equivalence and reduce statistical power (What Works Clearinghouse, 2014). Concerns about internal validity and equivalence are minimized in this study because the matched-pair school-randomized design of the larger evaluation allowed for schools that left the study between years one and two and their pairs to both be dropped from the analysis. The analysis of remaining matched pairs in year two means estimates remain internally valid.

Results in table 4.7 indicate that, despite attrition, the study is sufficiently powered to detect treatment effects on teachers' data-based instructional practices ranging from about 0.18 to 0.22. Ultimately, the sound design of the larger evaluation means that although these present as limitations, there are few concerns regarding internal validity in this study. While baseline survey data would have allowed the inclusion of covariates to both adjust for observed treatment- and control-group differences and improve statistical power, the design of the larger study – and relatively small differences in the results of research question one when models included a baseline measure of the outcome (appendix B) – makes the absence of baseline data less of a concern.

External Validity. Attrition can also affect external validity if schools that left the evaluation between years one and two differ systematically from those that remained. The analyses in appendix A indicate there were few year-one differences on measures relevant to this study between teachers in schools that attritted and those that stayed in the

evaluation. However, results of this study are generalizable only to schools that would remain in the program for two years when given the opportunity to do so.

There are other sample considerations that affect generalizability. The larger evaluation recruited schools from medium to large urban districts with high numbers of low-performing students. Therefore, the results are likely only generalizable to similarly low-performing urban elementary and middle schools. This is because these types of schools are likely to differ in systematic ways from higher performing or suburban or rural schools that might also affect outcomes. The schools that were recruited into the evaluation were motivated to work with ANet; they were a convenience sample recruited from districts in which ANet already had a relationship. They also had the means to allocate about $3,000 in funding, annually, to cover the remainder of the i3-subsidized annual fee for partnering with ANet. Therefore, results are likely generalizable only to other similarly motivated and financially able schools.

ANet's own leadership team was also concerned that the i3 sample represented a very different sample than the Boston charter schools in which they developed their program. Specifically, they hypothesized that the i3 sample had lower overall readiness to partner with the program than the schools in which the program was developed (West, Morton, & Herlihy, 2016). Although the baseline readiness measure in this study is shown to have a relatively weak association with teachers' data-based instructional practices, the readiness measure used in the larger evaluation was more strongly related to differences in teacher practices – data review, data use, and instructional differentiation – and student math and reading achievement. As a result, findings in this study and the

larger evaluation may not represent the findings that might be expected in higher-readiness schools (West, Morton, & Herlihy, 2016). This is partially supported by differences in the results of research question three when analyzed separately for higher- and lower-readiness schools (table B.3 and B.4).

Violations of Key Assumptions of Mediation. As described in chapter three, estimates of the mediating effect of school and teacher characteristics that are calculated from the difference method are valid and unbiased as long as certain assumptions can be met: specifically, that there is 1) no unmeasured confounding of the relationship between the treatment and outcome, 2) no unmeasured confounding of the relationship between the mediator and outcome, 3) no treatment-mediator interaction (treatment-mediator confounding), and 4) no mediator-outcome confounder that is affected by the treatment. There is also an assumption that the treatment, mediator, and outcome are temporally ordered (Valeri & VanderWeele, 2013). Although the design of the i3 evaluation makes meeting the first assumption plausible, scenarios can be imagined in which the others might plausibly be violated. This study's conclusions regarding mediation are observational; causal links cannot be draw between ANet, the proposed mediators, and the outcomes of interest. In addition, there is no way to test the assumptions or know how weak – or subject to bias – these results are.

More recent methods propose a counterfactual approach to estimating causal meditation effects using the potential outcomes framework to decompose the average treatment effect into the average direct effect (ADE) and average causal mediation effect (ACME, or indirect effect). Imai, Keele, Tingley, and Yamamoto (2011) developed one

such alternative (see also Imai & Yamamoto, 2013). Their estimation of mediating effects is also assumption based and results are non-causal; however, their framework provides a sensitivity analysis of results to violations of key assumptions that cannot be tested directly. Specifically, their sensitivity analyses provide a measure of the degree to which violations of the assumptions alter the conclusions regarding a single mediating mechanism (Imai, et al., 2011) or cases in which there are multiple mediating mechanisms, particularly when they are thought to be causally dependent (Imai & Yamamoto, 2013). The main reason this method was not used in this study was that the framework – and statistical program – had not been developed for multilevel data.

Data

Survey Nonresponse. Nonresponse may be of concern for several reasons. First, school leader survey nonresponse was the reason that aggregate, school-mean teacher responses were used for measures of school-level mediators. Chapter three reviewed the substantive implications of this choice: we cannot assume that aggregate teacher responses and school leader responses are measuring the same construct. Therefore, discussions of results took care to remind the reader that these are teacher perceptions of school leaders' abilities to perform various instructional leadership tasks, not a self-report and certainly not an objective measure of the quality of instructional leadership. Of less concern are aggregate teacher perceptions of school culture. These may be more appropriate than the school leaders' perceptions, particularly with respect to how they relate to teachers' practices.

Survey nonresponse at the teacher level is of concern in terms of the possible introduction of bias in the results. For example, nonresponding teachers may differ in important ways from responding teachers. Whether and how they differ is impossible to know and, to the extent that they do differ, biased estimates of the outcomes may be introduced. Of particular concern are low response rates in Chicago and Jefferson Parish control schools. Within-school survey nonresponse also plays into the creation of aggregate measures of school-level mediators. School-level response rates are unknown, but assumed in many cases to be less than 100 percent. To say that school-mean measures are representative of all teachers in the school would require responding teachers be a random sample of all teachers, including non-responders. This is unlikely to be true and, therefore, may be another source of bias in the results.

Measurement Error. Measurement error in the survey scales and indices in this study can attenuate bivariate correlations with the outcome, introduce bias in multiple regression estimates, and reduce statistical power. Year-two survey revisions were meant to improve the measures used in this study and minimize measurement error: i.e., improve validity and reliability. However, the reliability of measures of some predictors was unusually high. This was likely a product of teachers' straight-lining responses across all items in a set which could compromise their validity. The reliability of only one scale showed cause for concern: teachers' attitudes toward assessment and assessment data. Estimates from analyses utilizing this measure may be subject to some bias from measurement error.

Teacher Self-Reported Practice. All of the conclusions about teachers' data-based instructional practices are based on survey self-reports. Historically, there have been concerns about the accuracy and reliability of teacher self-reported practices. This is evidenced by early research showing that observations of teachers' classroom instruction and self-reported teacher practices are poorly correlated (Desimone, 2009). This may be because self-reports typically ask teachers to recount their practices after significant time has passed or encourage socially-desirable responses, e.g., teachers may over-represent their use of desirable instructional practices (Mayer, 1999; Muijs, 2006). Specifically, knowing some of the expected outcomes of partnering with ANet, teachers in treatment schools might over-reported their practices in order to appear more competent or effective. Therefore, not having some external or more objective measure of teacher practices – or even a social desirability scale to include as a covariate in the statistical models – could be considered a weakness of this study.

However, evidence of bias in self-reports in recent research has been limited. Desimone (2009) argues that many of the early studies that led to concerns in using self-reported teacher practices had methodological flaws: observations that were too few or too short, comparisons of self-reported average practices with observations of specific ones, or lack of comparable teacher and observer instruments. Instead, Desimone points to recent studies that either included multiple observations of behavioral measures or where teachers and observers used the same data collection protocol; these studies found moderate to high correlations between observations and self-reports (2009, p. 189; Mayer, 1999). Additionally, the relatively low-stakes nature of the ANet survey

compared to, for example, teacher evaluations would suggest teachers felt less inclined to misrepresent the frequency of the data-based instructional practice in this study (Mayer, 1999). This lends some confidence to the use of teacher self-reported outcomes in this study, especially since no impact was found on the frequency with which teachers reported differentiating instruction (a potentially desirable practice in today's classroom).

Interview Data. In some respects, the study was limited by the secondary use of interview data. Since the interview protocols were designed with the larger evaluation's purposes in mind, some measures of interest to this study were less well represented than they would have been had the protocols been developed to provide context for the quantitative findings in this study. It also meant that much of the evidence related to any measure of interest to this study arose spontaneously and not in response to a specific question, therefore limiting the ability to say what proportion of teachers held certain beliefs or implemented certain practices. This also made it impossible to reliably compare beliefs and practices across districts or between leaders and teachers within schools.

RESEARCH IMPLICATIONS & FUTURE DIRECTIONS

This study adds to our understanding of the roles of commonly cited mediators in the context of a specific data-based instructional program. However, the design still does not allow for the calculation of unbiased estimates of the causal mechanisms. Other design and data improvements in future research could continue to close this gap.

Design

Other Settings. Results from this study indicate that the impact of ANet on teachers' data-based instructional practices varied by district. As with most interventions, future research would benefit from continuing to study the use of interim assessments and data-based instructional practices under rigorous designs and in other settings. Efforts should be undertaken to define and collect information on the varying contexts – e.g., in implementation, educators' uptake, or the complex interactions of the two – in which ANet and other similar interventions unfold so that these can be explored as factors affecting outcomes across sites.

Alternative Designs for Assessing Causal Mediation. As is well-known, there are often ways to correct for design problems during analysis, but a better approach is to start with sound design. For example, the framework for estimating causal mediation effects that was introduced in the prior section (Imai, Keele, Tingley, & Yamamoto, 2011) is a post-design option for attempting to test the sensitivity of the estimation of average causal mediation effects – or indirect effects – to violations of assumptions when the design makes these violations plausible. However, the authors also propose designs that are less susceptible to violations of identification assumptions. They explain that the key is to design studies where the hypothesized mediator(s) can be directly or indirectly manipulated (Imai, Keele, Tingley, & Yamamoto, 2011). Details of these designs are provided; however, while they may provide more rigorous tests of causal mechanisms, they might also be difficult to put into practice.

One example of a stronger design is the parallel design in which two randomized experiments are conducted in parallel (Imai & Yamamoto, 2013). In the first experiment, only the treatment is randomized and at the conclusion of the study, levels of both the mediator(s) and outcome are measured. This is similar to ANet evaluation design. In the second experiment, both the treatment and mediator(s) are randomized and, at the conclusion, the outcome is measured. Here, all confounders are controlled for, including causally-dependent mediators. In both experiments, a participant's potential outcomes are assumed to be the same: the outcomes depend on the treatment and mediator values, not how the values came to be. The first experiment allows for the estimation of the impact of the treatment on the mediator(s) and outcome. The second estimates the indirect effect assuming there is no interaction between the treatment and the mediator for each unit *i*; something the authors acknowledge is unverifiable and unrealistic in most settings (Imai & Yamamoto, 2013, p. 164). They propose a sensitivity analysis for violations of this assumption alone.

Their second example of a stronger design for causal mediation is called the parallel encouragement design. It is proposed in situations where manipulation of the mediator isn't possible. The set-up is the same – two experiments are run in parallel – however, in the second, a random subset of participants are encouraged to take on a range in values of the mediator through some type of manipulation. For example, the manipulation might be some task meant to elicit specific levels of the mediator of interest. The randomized encouragement is considered the instrument that elicits exogenous variation in the mediator (Imai, Keele, Tingley, & Yamamoto, 2011, p. 781).

For those participants who comply with the encouragement, this design provides an estimate of the complier average causal mediation effect (CACME). Like the parallel design, this design provides more information about the causal mechanisms at work. With perfect manipulation, the parallel encouragement design reduces to the parallel design. Both designs assume that the indirect and direct manipulation of the mediator does not influence the outcome, participants must behave as though they chose the mediator value (Imai, Tingly, & Yamamoto, 2013).

Data

Measuring Mediators. Alternative designs such as those discussed above would provide more rigorous evidence of the conditions that mediate outcomes of data-based instructional programs; however, validated measures of these constructs are also important. For example, one of the key mediators in this study was hypothesized to be teacher data literacy. Data literacy continues to be a topic of interest in teacher preparation and professional development. This study had no direct measure of data literacy, but did show that teachers' confidence in using data was a relatively strong predictor of the frequency teachers' reviewed and used data controlling for treatment group assignment, all other mediators, and all other covariates.

Future research should continue to explore whether these findings related to data *confidence* are also true when measured more directly as data *literacy.* The content validation undertaken during the revision of year-two survey measures uncovered some validated scales of assessment literacy and statistical literacy, but work to define,

develop, and measure the complex construct of data literacy should be pursued. While this could be argued for any one of the constructs in this study, it is particularly important for the educational community to understand the role of teachers' data literacy given the call to increase time, attention, and funding for developing this skill and the prevalence of data-based strategies in schools.

Measuring Outcomes. Obtaining an objective measure of the frequency with which teachers take part in specific instructional practices is difficult; however, a study like this one could endeavor to measure instructional differentiation by asking teachers to keep instructional logs. Instructional logs have the benefit of being collected immediately after instruction and at multiple points, reducing the recall bias of surveys and allowing for more complete sampling of practices in comparison to classroom observations. As part of their Indiana benchmark assessment study, Konstantopoulos and colleagues asked 2nd- and 5th-grade teachers to complete 16 instructional logs for a random sample of 8 of their students over the course of the school year. Logs captured the type and level of content that was taught to which students; differentiation was identified when content was taught to some, but not all students during the same lesson.

The authors concluded that teacher logs provided a reliable representation of differentiation at the log- and teacher-levels. However, they acknowledged that school-level reliability was slightly lower. Therefore, studies that measure the impact of a program on school-level estimates of instructional differentiation would need to account for this in power calculations of the number of teachers required to complete logs (Williams, et al., 2014).

CONCLUSIONS

Despite the methodological limitations, this study makes an important contribution to our knowledge of data-based instructional practices and the conditions under which they may or may not be implemented effectively. Not only does this study provide internally valid estimates of the impact of ANet on teachers' data-based practices, but exploratory analyses of quantitative and qualitative data provide some of the most rigorous and complete insights to date into the mediating roles played by oft-cited school and teacher characteristics. In particular, despite earlier noted limitation, the fit between research questions in the larger evaluation and this study was quite high and, therefore, provide an illustration of how researchers can use qualitative data that have been collected for a different purpose to answer unique questions and support quantitative results.

In addition to the empirical contributions of this study, these results have important practical implications for educators looking to implement data-based instructional practices, yet they also challenge the larger research and practitioner community to explore some important unanswered questions. In this concluding section, the study's findings are framed around key implications for policy and practice. The hope is that these might help educators focus their efforts when implementing ANet or other data-based instructional programs and practices. That said, the results of this study are an important illustration of how program effects can differ across districts and speak to the importance of knowing the context in which programs like ANet unfold. While this might

make generalizing the findings to new districts more difficult, the following represent a key set of conclusions for any district or school leader, or program provider, to consider.

What Impact on Teacher Practices Can We Expect from Data-Based Instructional Programs?

Among the four practice outcomes – reviewing data, using data, planning instruction, and differentiating instruction – the data-related outcomes appear not only be *easier* for teachers to adopt and implement, but teacher interviews indicated ANet put a stronger focus on data analysis and planning the reteach in year one and backward planning and carrying out the reteach with fidelity in year two. Since teachers likely had more time to put data-related skills into practice, seeing impacts in data-related, but not instructional-related, outcomes after two years is neither surprising nor inconsistent with prior research (Cavalluzzo, et al., 2014).

The results indicate that teachers need time and support to implement effective data-based instructional strategies. Instructional practices such as differentiation might represent a major shift in practice and the introduction of ANet itself may require an adjustment period. District and school leaders should give programs time to take hold and ensure that teachers are provided with support to bridge the gap between analyzing data and using it to adjust their instruction. On the latter, researchers must continue to uncover the characteristics of programs that help teachers make this transition successfully and the types of support they need to make data-based instruction a part of their regular practice.

In addition to supporting teachers, patterns in impacts of ANet on teachers' data-based instructional practices by baseline school readiness indicate that programs might

need to differentiate support at the school level. Though not conclusive, results indicate that impacts might be larger or more positive in schools judged more ready to implement ANet than in schools judged less ready. This highlights the importance of districts, schools, and program providers collecting information about school capacity before implementing new data-based instructional programs (Datnow, 2005). Schools with lower initial readiness should be monitored for progress; they may require greater program support to see similar impacts on teacher practices and student achievement (West, Morton, & Herlihy, 2016). Much like ANet encourages teachers to do with their students, programs like ANet should consider principled program variations that adapt constructively to school and teacher needs.

How Can We Support Conditions that Make Data-Based Practices More Likely to Take Hold?

Results from this study also show that school-mean teacher reported perceptions of school leaders' instructional leadership abilities were meaningfully, if not statistically, higher in ANet schools. Proficiency in setting expectations for instructional data use and monitoring whether teachers meet these expectations might account these findings; however, teacher interviews indicate that instructional leaders might better balance these accountability-focused roles with providing instructional support to teachers around reviewing data and planning a reteach. Leaders also need to ensure that the instructional support they provide is based on effective practice and takes full consideration of the richness of ANet's data reports.

It wasn't always clear whether the broader professional culture in ANet schools more closely resembled true collaboration or contrived collegiality. In some schools, analyzing data with peers and sharing instructional strategies may have taken place regularly. However, conditions in other schools were likely more contrived; that is, focused on ANet implementation (e.g., the reteach) and timed around quarterly data meetings. In light of qualitative evidence that ANet schools were characterized by positive achievement cultures, contrived collegiality may still be effective in fostering data-based instructional practices. However, leaders need to ensure not only positive achievement cultures, but that their school culture engenders trust among staff members. ANet or other data-based programs might be more likely to succeed in schools where teachers collaborate without fear of being judged on their students' data or perceived as offering unsolicited advice to peers.

Models that explored whether school- and teacher-level characteristics explained the impact of ANet on teachers' data-based practices showed the comparatively small role of teacher attitudes and confidence. However, these teacher characteristics showed strong, positive associations with teachers' data-based instructional practices. These results raise questions about a potential gap in the program's focus. During this study, ANet's program model focused on training leaders to support teachers' data-based instructional practice. With evidence that leaders may need more support to be effective instructional leaders, these results suggest that ANet could leverage direct support to teachers as a way to move them from reviewing and using data in various ways to adopting data-based instructional practices. Providing support to leaders *and* teachers in

low-performing or high-poverty schools could be even more necessary given relatively higher turnover and loss of human capital and more limited internal transfer of knowledge (Clotfelter, Ladd, Vigdor, & Wheeler, 2006; Simon & Johnson, 2015).

ANet's value-add over many other data-based instructional programs is their coaching. However, results suggest that the gradual release model might not be gradual enough for leaders in ANet schools to effectively implement the program and foster the desired teacher practices on their own. Furthermore, the program could see more effective implementation of practices if teachers were provided more frequent, direct coaching on skills such as planning instruction, analyzing their students' results, and identifying and implementing effective instructional responses to gaps in student learning. More frequent coaching around data use has been shown to be associated with teachers' self-reported changes in instructional practices (Marsh, McCombs, & Martorell, 2010). District- or school-based coaches build teacher capacity around data use when they model data practices and help teachers bridge the gap between data and instruction (Huguet, Marsh, & Farrell, 2014).

Another potentially promising approach to providing teachers with greater instructional support and for maintaining knowledge around ANet practices was one used by ANet leaders in some schools: building the capacity of other school leaders and distributing ANet responsibilities to better provide teachers with feedback (Copland, 2003; Lachat & Smith, 2005; Wayman & Stringfield, 2006; Knapp, Copland, & Swinnerton, 2007; Gerzon, 2015). While ANet asks schools to form a data leadership team, interviews with some school leaders indicated that they either found it difficult to

engage others in the work or felt they couldn't ask them to be away from their teaching responsibilities.

The findings from this study suggest that teachers need support at all points in the adoption of data-based instructional practices: e.g., analyzing data, forming a reteaching plan, and backward planning instruction. Additionally, teachers seemed to adopt these practices at different speeds or with different levels of confidence. Building leader capacity is important, but should not be the sole strategy for improving teacher practice. Leaders themselves are often poorly trained and turnover can be high in the schools ANet tends to serve. A greater focus on direct and differentiated teacher support is likely needed to ensure that they are able to translate student data into effective instructional actions.

The question that remains is whether we can expect program-based coaches like those working for ANet or district- or school-based coaches to be the sole line of support for teachers' data-based practices. It seems logical to ask whether teacher preparation programs are doing enough to provide the necessary preservice support for teachers to effectively use student data to inform classroom instruction. The answer seems to be, no. Research has shown that schools of education do offer assessment literacy coursework (i.e., design, implementation, and analysis of assessments); however, fewer had courses that were devoted to the broader skill of data literacy or data-based decision making (Mandinach, Friedman, & Gummer, 2015). Schools of Education report not having flexibility or expertise to provide coursework in data literacy, or failing to see it as a "sufficiently important" for allocating a faculty member to teach those courses

(Mandinach, Friedman, & Gummer, 2015, p. 34). However, given the prevalence of student data and expectations for teachers to use those data to inform instruction, data literacy is an important skill in today's classrooms. Thus, it may rest on districts to put pressure on Schools of Education to make this a priority in their curriculum.

What Other Factors Should We Consider?

Recall that the school conditions and teacher characteristics of focus in this study explained, at best, only about one-third of the impact of ANet on teachers' data practices. Therefore, it is important to identify other factors that affect the implementation of data-based programs and practices in schools. The site visit interviews pointed to one important factor: teachers' perceptions of the ANet assessments. Misalignment seems to have played a part in making student results less actionable for some teachers. If we accept that interim assessments can affect positive changes in teacher instruction and student learning, then the results of this study speak to the need to ensure curricular alignment so that student results are likely to inform instruction. Results from the larger study also showed that teachers' confidence in fitting the reteach into curricular scope and sequence, frequency of backward planning, and belief that policies limit curricular flexibility all predicted their perceptions of alignment between their math interim assessments and the math scope and sequence (West, Morton, & Herlihy, 2016).

In the spectrum of assessment types, interim assessments appear to dominate teachers' access to and work with student data (Datnow & Hubbard, 2015). However, in an age where policymakers, parents, and educators are concerned about over-testing, this

raises concerns about the dual purpose of interim assessments. As it relates to rigor (i.e., difficulty), ANet interim assessments are given to students at their grade level so that leaders and teachers can gauge students' progress toward grade-level standards. However, the organization often partners with schools that serve a low-performing student population, students who aren't performing at grade level. For many teachers, this meant the results were not always as useful for informing classroom instruction. Even if the difficulty of the tests were better matched to students' ability levels, the multiple-choice focus does not lend itself to true rigor, nor does it provide clear information on student reasoning that would best inform instructional decisions. This suggests the need to supplement interim assessment programs with formative and performance assessments (Hoffman, Goodwin, & Kahl, 2015) and systems of continuous improvement in order to truly attend to individual students' instructional needs.

It would not be fair or appropriate to put the responsibility of ensuring alignment between a district's curriculum and a program's interim assessments solely on the shoulders of program providers. These findings argue not only for district support for program implementation, but also district cooperation in designing assessments that match the content standards and curricular sequence. Furthermore, since misalignment is likely to be inevitable at some point, it reinforces the notion that teachers need the skills, flexibility, and support to manage misalignment: i.e., preservice training, coaches who can provide instructional support, and policies that allow for curricular flexibility when appropriate. Finally, a program like ANet should not be the only source of data teachers have as a resource for informing instruction. Districts, schools, and coaches should

ensure teachers have a portfolio of student-level information on which to base instructional decisions and are supported in using these various forms of information effectively.

In these concluding pages, some important questions were raised for the field; questions about what support teachers need to move from data analysis to instructional improvement and the roles of outside partners – Schools of Education and districts – in supporting these efforts. These questions are important because data-based instructional programs like ANet are prevalent in American schools. At the same time, there is a call to ensure that testing – a key component of these programs – provides information that is instructionally useful while limiting intrusions on classroom time. This study fills an empirical need for more rigorous evidence of the effectiveness of one such data-based instructional program and the roles played by instructional leadership, culture, and teachers' attitudes and confidence in teachers' adoption of data-based instructional practices. The results are also substantively important as they provides practitioners and policymakers with valuable insights on conditions that might increase the likelihood that programs like ANet lead to effective teacher practices.

REFERENCES

Achievement Network. (2012). *Program logic model for i3 evaluation*. Unpublished internal document.

Achievement Network. (2015). Website homepage and school leader support page. Retrieved on August 10, 2015 at: http://www.achievementnetwork.org/ and on September 27, 2015 at: http://www.achievementnetwork.org/school-leaders/.

Bamberger, J. M., Rugh, J., & Mabry, L. S. (2006). *RealWorld evaluation: Working under budget, time, data, and political constraints*. Thousand Oaks, CA: Sage Publications, Inc.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-82.

Benjamini, Y. (2015, October). *The replicability crisis in science: It's not the p-values' fault.* Center for the Study of Testing, Evaluation, and Education Policy fall lecture series. Chestnut Hill, MA: Boston College.

Bennett, R. E., & Gitomer, D. H. (2008). Transforming K–12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J.J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43-61). Netherlands: Springer.

Bickel, R. (2007). *Multilevel analysis for applied research: it's just regression!* New York, NY: The Guilford Press.

Bickman, L. (2000). Summing up program theory. *New directions for evaluation*, 87, 103-112.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 92*(1), 81-90.

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education*, *85*(2), 205-225.

Blank, R. K., Smithson, J., Porter, A., Nunnaley, D., & Osthoff, E. (2006). Improving instruction through schoolwide professional development: Effects of the data-on-enacted-curriculum model. *ERS spectrum, 24*(2), 9-23.

Bloom, H. S. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reforms. Evaluation Review, 27(1), 3–49.

Bloom, H. S. (2006). *The core analytics of randomized experiments for social research.* MDRC Working Papers on Research Methodology. New York: MDRC.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas Accountability System. *American Educational Research Journal, 42*(2), 231-268.

Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (1997). Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education, 13*(3), 259-278.

Borman, K. M., et al. (2005). Meaningful urban education reform: Confronting the learning crisis in mathematics and science. Albany, NY: SUNY Press.

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.

Bryk, A. S., Sebring, P. B., Allensworth, E., Easton, J. Q., & Luppescu, S. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: University of Chicago Press.

Bulkley, K. E., Oláh, L. N., & Blanc, S. (2010). Introduction to the special issue on benchmarks for success? Interim assessments as a strategy for educational improvement. *Peabody Journal of Education*, *85*(2), 115-124.

Campbell, P. F., & Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal, 111*(3), 430-454.

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. *Educational Evaluation and Policy Analysis*, *33*(3), 378-398.

Cavalluzzo, L., Geraghty, T. M., Steele, J. L., Holian, L., Jenkins, F., Alexander, J. M., & Yamasaki, K. Y. (2014, March). '*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.* Arlington, VA: CNA Corporation.

Chen, E., Heritage, M., & Lee, J. (2005). Identifying and monitoring students' learning needs with technology. *Journal of Education for Students Placed at Risk*, *10*(3), 309-332.

Cho, V. & Wayman, J. C. (2013, November). *District leadership for computer data systems: Technical, social, and organizational challenges in implementation*. Paper presented at the Annual Convention of the University Council for Educational Administration, Indianapolis, IN.

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, PA: Research for Action.

Clotfelter, C., Ladd, H. F., Vigdor, J., & Wheeler, J. (2006). High-poverty schools and the distribution of teachers and principals. *North Carolina Law Review, 85*, 1345-1379.

Clune, W. H., & White, P. A. (2008*). Policy effectiveness of interim assessments in Providence public schools.* WCER Working Paper No. 2008-10. Madison, WI: Wisconsin Center for Education Research, School of Education, University of Wisconsin-Madison.

Coalition for Evidenced-Based Policy. (2013, July). *Randomized controlled trials commissioned by the Institute of Education Sciences since 2002: How many found positive versus weak or no effects.* Washington, DC.

Coburn, C. E., & Turner, E. O. (2012). The practice of data use: An introduction. *American Journal of Education, 118*(2), 99-111.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research & Perspective*, *9*(4), 173-206.

Cohen D. K. & Hill, H. C. (2000) Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record, 102*, 296-345.

Copland, M. A. (2003). Leadership of inquiry: Building and sustaining capacity for school improvement. *Educational Evaluation and Policy Analysis, 25*(4), 375-395.

Cordray, D., Pion, G., Brandt, C., Molefe, A., & Toby, M. (2012). *The impact of the Measures of Academic Progress (MAP) program on student reading achievement*. (NCEE 2013–4000). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Couper, M. P. (2000). Review: Web surveys: A review of issues and approaches. The Public Opinion Quarterly, 64(4), 464-494.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research.* Thousand Oaks, CA: Sage publications.

Crocker, L. & Algina, J. (2008). *Introduction to Classical and Modern Test Theory*. Mason, OH: Cengage Learning.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Daly, A. J. (2012). Data, dyads, and dynamics: Exploring data use and social networks in educational improvement. *Teachers College Record, 114*(11), 1-38.

Daniel, L. G., & King, D. A. (1998). Knowledge and use of testing and measurement literacy of elementary and secondary teachers. *The Journal of Educational Research, 91*(6), 331-344.

Datnow, A. (2011). Collaboration and contrived collegiality: Revisiting Hargreaves in the age of accountability. *Journal of Educational Change, 12*(2), 147-158.

Datnow, A. (2005). The sustainability of comprehensive school reform models in changing district and state contexts. *Educational Administration Quarterly, 41*(1), 121-153.

Datnow, A. & Hubbard, L. (2015). Teachers' use of assessment data to inform instruction: Lessons from the past and prospects for the future. *Teachers College Record, 117*(4), 1-26.

Datnow, A. & Park, V. (2014). *Data-driven leadership.* San Francisco, CA: Jossey-Bass.

Datnow, A., & Park, V. (2009). School system strategies for supporting data use. In T. J. Kowalski & T. J. Lasley II (Eds.), *Handbook of data-based decision making in education* (pp. 191-206). New York, NY: Routledge.

Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving with data: How high-performing school systems use data to improve instruction for elementary students.* Los Angeles: Center on Educational Governance, Rossier School of Education, University of Southern California.

David, J. L. (2008, October). What research says about pacing guides. *Educational Leadership, 66*(2), 87-88.

Dembosky, J. W., Pane, J. F., Barney, H., & Christina, R. (2005). *Data driven decisionmaking in southwestern Pennsylvania school districts.* (WR-326-HE/GF). Santa Monica, CA: RAND Corporation.

Deming, W. E. (1993). *The new economics for industry, education, government*. Cambridge, MA: MIT Center for Advanced Engineering Study.

Denton, C. A., Swanson, E. A., & Mathes, P. G. (2007). Assessment-based instructional coaching provided to reading intervention teachers. *Reading and Writing, 20*(6), 569-590.

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

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. Educational evaluation and policy analysis, *24*(2), 81-112.

DeVellis, R. F. (2003). *Scale development: Theory and applications*. Thousand Oaks, CA: Sage Publications.

Diamond, J. B., & Cooper, K. (2007). The uses of testing data in urban elementary schools: some lessons from Chicago. *Yearbook of the National Society for the Study of Education*, *106*(1), 241-263.

Dobbie, W., & Fryer Jr, R. G. (2011). *Getting beneath the veil of effective schools: Evidence from New York City* (Working Paper 17632). Cambridge, MA: National Bureau of Economic Research.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*(1), 1-17.

Donner, A., & Klar, N. (2004). Pitfalls of and controversies in cluster randomization trials. *Public Health Matters*, *94*(3), 416-422.

DuFour, R., & Eaker, R. (1998). *Professional learning communities at work: Best practices for enhancing student achievement*. Bloomington, IN: National Educational Service.

DuFour, R., Eaker, R., & DuFour, R. (2005). Recurring themes of professional learning communities and the assumptions the challenge. In R. DuFour, R Eaker, & R. DuFour (Eds.), *On common ground: the power of professional learning communities* (pp. 7-29). Bloomington, IN: Solution Tree.

Every Student Succeeds Act (ESSA) of 2015, Pub. L. No. 114-95 (2015).

Faria, A. M., Heppen, J., Li, Y., Stachel, S., Jones, W., Sawyer, K., ... & Palacios, M. (2012). *Charting success: Data use and student achievement in urban schools.* Washington, DC: Council of the Great City Schools.

Flumerfelt, S., & Green, G. (2013). Using lean in the flipped classroom for at risk students. *Educational Technology & Society, 16*(1), 356–366.

Fowler, Jr., F. J. (2009). Survey research methods (4th ed.). Thousand Oaks, CA: Sage.

Fullan, M. (2007). *The new meaning of educational change*. New York, NY: Teachers College Press.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American educational research journal*, *38*(4), 915-945.

Gerzon, N. (2015). Structuring professional learning to develop a culture of data use: Aligning knowledge from the field and research findings. *Teachers College Record, 117*(4).

Gilbert, D., King, G., Pettigrew, S., & Wilson, T. (2016). Comment on 'Estimating the reproducibility of psychological science', *Science*, *351*(6277), 1037a-1037b.

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

Goertz, M. E., Oláh, L. N., & Riggan, M. (2009b). *Can interim assessments be used for instructional change?* CPRE Policy Brief #RB-51. Philadelphia, PA: Consortium for Policy Research in Education.

Greene, J. C. (2008). Is mixed methods social inquiry a distinctive methodology? *Journal of mixed methods research*, *2*(1), 7-22.

Grubb, W. N., & Flessa, J. J. (2006). "A job too big for one": Multiple principals and other nontraditional approaches to school leadership. *Educational administration quarterly, 42*(4), 518-550.

Guskey, T. R. (2002). Professional development and teacher change. *Teachers and Teaching: theory and practice, 8*(3), 381-391.

Guskey, T. R. (1988). Teacher efficacy, self-concept, and attitudes toward the implementation of instructional innovation. *Teaching and teacher education, 4*(1), 63-69.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (Vol. 7). Upper Saddle River, NJ: Pearson Prentice Hall.

Hallinger, P., & Murphy, J. (1985). Assessing the instructional management behavior of principals. *The elementary school journal, 86*(2), 217-247.

Halverson, R. (2010). School formative feedback systems. *Peabody Journal of Education, 85*(2), 130-146.

Halverson, R., Grigg, J., Prichett, R., & Thomas, C. (2007). The new instructional leadership: Creating data-driven instructional systems in schools. *Journal of School Leadership, 17*, 159-194.

Halverson, R., Prichett, R., Grigg, J., & Thomas, C. (2005). The new instructional leadership: Creating data-driven instructional systems in schools. WCER Working Paper No. 2005-9. Madison, WI: Wisconsin Center for Education Research, University of Wisconsin.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., & Wayman, J. C. (2009). *Using student achievement data to support instructional decision making*. (NCEE 2009-4067). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (2002). *Making sense of test-based accountability in education*. Santa Monica, CA: RAND Corporation.

Hannaway, J. (2005). Poverty and student achievement: a hopeful review. In J. Flood and P. Anders (Eds.), Literacy development of students in urban schools (pp. 3–21). Newark, DE: International Reading Association.

Hargreaves, A. (1994). *Changing teachers, changing times: teachers' work and culture in the postmodern age.* New York, NY: Teachers College Press.

Hargreaves, A., & Braun, H. (2012). *Leading for all: final report of the review of the development of essential for some, good for all–Ontario's strategy for special education reform devised by the Council of Directors of Education*. Toronto, CA: Council of Directors of Education.

Hargreaves, A., & Braun, H. (2013). *Data-driven improvement and accountability.* Boulder, CO: National Education Policy Center.

Hargreaves, A., & Dawe, R. (1990). Paths of professional development: Contrived collegiality, collaborative culture, and the case of peer coaching. *Teaching and teacher education, 6*(3), 227-241.

Hargreaves, A., Morton, B., Braun, H. & Gurn, A. (2014). The Changing dynamic of educational judgment and decision making in a data-driven world, In S. Chitpin & C. Eves (Eds.), *Decision-making in educational leadership: Principle, policies and practice* (pp. 3-20). New York, NY: Routledge.

Hargreaves, D. H. (1995). School culture, school effectiveness and school improvement. *School effectiveness and school improvement*, *6*(1), 23-46.

Hart, R., Casserly, M. Uzzell, R., Palacios, M., Corcoran, A., and Spurgeon, L. (2015, October). Student testing in America's Great City Schools: An inventory and preliminary analysis. Washington, DC: Council of the Great City Schools.

Heaton, J. (2008). Secondary analysis of qualitative data: an overview. *Historical Social Research, 33*(3), 33-45.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement* (Issues & Answers Report, REL 2007–No. 039). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2008). *A second follow-up year for "Measuring how benchmark assessments affect student achievement."* (REL Technical Brief, REL Northeast and Islands 2007–No. 002). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands.

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: a seamless process in formative assessment? *Educational measurement: issues and practice, 28*(3), 24-31.

Heritage, M., & Yeagley, R. (2005). Data use and school improvement: Challenges and prospects. *Yearbook of the National Society for the Study of Education, 104*(2), 320-339.

Herman, J. L., & Baker, E. L. (2005). Making benchmark testing work. *Educational Leadership*, *63*(3), 48-54.

Hesse-Biber, S. N. (2010). *Mixed methods research: Merging theory with practice*. New York, NY: Guilford Press.

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research consensus, crossroads, and challenges. *Educational Researcher, 42*(9), 476-487.

Hochberg, E. D., & Desimone, L. M. (2010). Professional development in the accountability context: Building capacity to achieve standards. *Educational Psychologist, 45*(2), 89-106.

Hofman, P., Goodwin, B., & Kahl, S. (2015). *Re-balancing assessment: Placing formative and performance assessment at the heart of learning and accountability.* Denver, CO: McREL International.

Honig, M. I., & Coburn, C. (2007). Evidence-based decision making in school district central offices: Toward a policy and research agenda. *Educational policy, 22*(4), 578-608.

Howard, A. (2012). *Data for the public good*. Sebastopol, CA: O'Reilly Media, Inc.

Huguet, A., Marsh, J. A., & Farrell, C. C. (2014). Building teachers' data-use capacity: Insights from strong and developing coaches. *Education Policy Analysis Archives, 22*(52).

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105*(04), 765-789.

Imai, K., King, G., & Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statistical Science, 24*(1), 29-53.

Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 5-51.

Imai, K., & Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis, 21*(2), 141-171.

Ingram, D., Seashore Louis, K., & Schroeder, R. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *The Teachers College Record, 106*(6), 1258-1287.

Jacobs, J., Gregory, A., Hoppey, D., & Yendol-Hoppey, D. (2009). Data literacy: Understanding teachers' data use in a context of accountability and response to intervention. *Action in Teacher Education, 31*(3), 41-55.

Jacob, R. T., Jones, S. M., Hill, H. C., & Kim, J. (2015). *Randomized trial meets read world: A conference to explore the nature and consequences of null effects in educational research*. May 7, 2015. Arlington, VA.

Jencks, C., & Phillips, M. (Eds.). (1998). The black-white test score gap. Washington, DC: Brookings Institution.

Jimerson, J. B., & Wayman, J. C. (2015). Professional learning for using data: Examining teacher needs & supports. *Teachers College Record*, *117*(4).

Jimerson, J. B. & Wayman, J. C. (2011). *Approaches to data-related professional learning in three Texas school districts.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Johnson, R. B. (1997). Examining the validity structure of qualitative research. Education, *118*(2), 282-292.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher, 33*(7), 14-26.

Johnson, S. M., Berg, J. H., & Donaldson, M. L. (2005). *Who stays in teaching and why?: A review of the literature on teacher retention*. Cambridge, MA: Project on the Next Generation of Teachers, Harvard Graduate School of Education.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus, and Giroux.

Kenny, D. A. (2014, October). *Mediation.* Retrieved from: http://davidakenny.net/cm/mediate.htm.

Kerr, K. A., Marsh, J. A., Ikemoto, G. S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American journal of education, 112*(4), 496-520.

Kisa, Z., & Correnti, R. (2015). Examining implementation fidelity in America's choice schools a longitudinal analysis of changes in professional development associated with changes in teacher practice. *Educational Evaluation and Policy Analysis, 37*(4), 437-457.

Knapp, M. S., Copland, M. A., & Swinnerton, J. A. (2007). Understanding the promise and dynamics of data-informed leadership. *Yearbook of the National Society for the Study of Education, 106*(1), 74-104.

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.

Konstantopoulos, S., Miller, S. R., van der Ploeg, A., & Li, W. (2016). Effects of interim assessments on student achievement: Evidence from a large-scale experiment. *Journal of Research on Educational Effectiveness*. [Accepted author version.] Retrieved February 21, 2016 from http://www.tandfonline.com/doi/abs/10.1080/19345747.2015.1116031.

Konstantopoulos, S., Miller, S., van der Ploeg, A., & Li, W. (2014). *Combining evidence from two RCTs about diagnostic assessments*. Paper presented at the 2014 spring conference of the Society for the Research on Educational Effectiveness. March 8, 2014. Washington, DC.

Lachat, M. A., & Smith, S. (2005). Practices that support data use in urban high schools. *Journal of Education for Students Placed at Risk, 10*(3), 333-349.

Lazarin, M. (2014). *Testing overload in America's schools*. Washington, DC: Center for American Progress.

Lee, V. E., & Smith, J. B. (1996). Collective responsibility for learning and its effects on gains in achievement for early secondary school students. *American Journal of Education, 104*(2), 103-147.

Leech, N. L., & Onwuegbuzie, A. J. (2008). Qualitative data analysis: A compendium of techniques and a framework for selection for school psychology research and beyond. *School Psychology Quarterly*, *23*(4), 587-604.

Leithwood, K. A., & Montgomery, D. J. (1982). The role of the elementary school principal in program improvement. *Review of educational research, 52*(3), 309-339.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry (Vol. 75). SAGE Publications: Newbury Park, CA.

Linn, R. L. (2000). Assessments and accountability. *Educational researcher*, *29*(2), 4-16.

Little, J. W. (1999). Organizing schools for teacher learning. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice*, (pp. 233-262). San Francisco, CA: Jossey-Bass.

Little, J. W. (2012). Understanding data use practice among teachers: The contribution of micro-process studies. *American Journal of Education*, *118*(2), 143-166.

Love, N. (2008). 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.

Luo, M. (2008). Structural equation modeling for high school principals' data-driven decision making: an analysis of information use environments. *Educational administration quarterly, 44*(5), 603-634.

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, *47*(2), 71-85.

Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational researcher, 42*(1), 30-37.

Mandinach, E. B., Gummer, E. S., & Muller, R.D. (2011). *The complexities of integrating data-driven decision making into professional preparation in schools of education: It's harder than you think.* Alexandria, VA, Portland, OR, and Washington, DC: CNA Education, Education Northwest, and WestEd.

Mandinach, E. B., Friedman, J. M., & Gummer, E. S. (2015). How can schools of education help to build educators' capacity to use data: A systemic view of the issue. *Teachers College Record, 117*(4).

Mandinach, E. B., & Honey, M. (2008). Data-driven decision making: An introduction. In E. B. Mandinach & M. Honey (Eds.), *Data-driven school improvement: Linking data and learning* (pp. 1-9). New York, NY: Teacher's College Press.

Mandinach, E. B., & Jackson, S. S. (2012). *Transforming teaching and learning through data-driven decision making*. Thousand Oaks, CA: Corwin.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. New York, NY: McKinsey.

Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record, 114*(11), 1-48.

Marsh, J. A., McCombs, J. S., & Martorell, P. (2010). How instructional coaches support data-driven decision making: Policy implementation and effects in Florida middle schools. *Educational Policy*, *20*(10), 1-37.

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

Mason, S. (2002). Turning data into knowledge: Lessons from six Milwaukee public schools. WCER Working Paper No. 2002-3. Madison, WI: Wisconsin Center for Education Research, University of Wisconsin—Madison.

Massell, D. (2001). The theory and practice of using data to build capacity: State and local strategies and their effects. In S. H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states* (pp. 148–169). Chicago: University of Chicago Press.

Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review, 62*(3), 279-301.

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33*(2), 3-11.

Maxwell, J. A. (2012). The importance of qualitative research for causal explanation in education. *Qualitative Inquiry*, *18*(8), 655-661.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data?. *Educational evaluation and policy analysis, 21*(1), 29-45.

McLaughlin, J. A., & Jordan, G. B. (1999). Logic models: a tool for telling your programs performance story. *Evaluation and program planning, 22*(1), 65-72.

Means, B., Padilla, C., DeBarger, A., & Bakia, M. (2009). *Implementing data-informed decision making in schools: Teacher access, supports and use*. Washington, DC: U.S. Department of Education.

Means, B., Padilla, C., & Gallagher, L. (2010). *Use of education data at the local level: From accountability to instructional improvement*. Washington, DC: U.S. Department of Education.

Messick, S. (1990). Validity of test interpretation and use. Princeton, NJ: Educational Testing Service.

Moore, R. T., & Schnakenberg, K. (2013). *Package 'blockTools'* [for R software].

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation, 12*(1), 53-74.

Murnane, R. J., Sharkey, N. S., & Boudett, K. P. (2005). Using student-assessment results to improve instruction: Lessons from a workshop. *Journal of Education for Students Placed at Risk, 10*(3), 269-280.

National Research Council (NRC). (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.), Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Newmann, F. M., King, M. B., & Rigdon, M. (1997). Accountability and school performance: Implications from restructuring schools. *Harvard Educational Review, 67*(1), 41-75.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Notz, P. (2005). Secondary qualitative analysis of interviews. A method used for gaining insight into the work/life balance of middle managers in Germany. *Forum: Qualitative Social Research, 6*(1). Retrieved March 1, 2015 from http://www.qualitative-research.net/index.php/fqs/article/view/506.

O'Cathain, A. (2010). Assessing the quality of mixed methods research: Toward a comprehensive framework. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social & behavioral research* (2nd ed.) (p. 531-555). Thousand Oaks, CA: SAGE Publications, Inc.

Oláh, L. N., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education, 85*(2), 226-245.

Onwuegbuzie, A. J., & Johnson, R. B. (2004). Mixed research. In R. B. Johnson & L. B. Christensen (Eds.), *Educational research: Quantitative, qualitative, and mixed approaches* (2nd ed.) (p. 408-431). Needham Heights, MA: Allyn & Bacon.

Onwuegbuzie, A. J., & Teddlie, C. (2003). A framework for analyzing data in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (p. 351-383). Thousand Oaks, CA: SAGE Publications, Inc.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Page, R. (1987). Teachers' perceptions of students: A link between classrooms, school cultures, and the social order. *Anthropology & education quarterly, 18*(2), 77-99.

PARCC Consortium. (2010). Website: The PARCC Assessment, Assessment System, grades 3-8. Retrieved on August 11, 2014 at http://www.parcconline.org/3-8.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). United States: Wadsworth-Thompson Learning.

Pellegrino, J. W. (2006). *Rethinking and redesigning curriculum, instruction and assessment: What contemporary research and theory suggests.* Washington, DC: Report commissioned by the National Center on Education and the Economy for the New Commission on the Skills of the American Workforce.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, *28*(3), 5-13.

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.

Purkey, S. C., & Smith, M. S. (1983). Effective schools: A review. *The elementary school journal, 83*(4), 426-452.

Quint, J. C., Sepanik, S., & Smith, J. K. (2008). *Using student data to improve teaching and learning: Findings from an evaluation of the Formative Assessments of Students Thinking in Reading (FAST-R) program in Boston elementary schools*. New York, NY: MDRC.

Randel, B., Beesley, A. D., Apthorp, H., Clark, T. F., Wang, X., Cicchinelli, L. F., & Williams, J. M. (2011). *Classroom assessment for student learning: Impact on elementary school mathematics in the central region* (NCEE 2011-4005). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173-185.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd Ed.). Thousand Oaks, CA: Sage Publications, Inc.

Ree, M. J., & Carretta, T. R. (2006). The role of measurement error in familiar statistics. *Organizational Research Methods, 9*(1), 99-112.

Richardson, V. (1990). Significant and worthwhile change in teaching practice. *Educational researcher, 19*(7), 10-18.

Riggan, M., & Oláh, L. N. (2011). Locating interim assessments within teachers' assessment practice. *Educational Assessment*, *16*(1), 1-14.

Rogers, P. J., Petrosino, A., Huebner, T. A., & Hacsi, T. A. (2000). Program theory evaluation: Practice, promise, and problems. *New directions for evaluation, 2000*(87), 5-13.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics, 29*(1): 159-183.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688-701.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, *26*(1), 20-36.

Saldaña, J. (2012). *The coding manual for qualitative researchers*. Thousand Oaks, CA: Sage Publications, Inc.

Sammons, P. (2010). The contribution of mixed methods to recent research on educational effectiveness. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 697-723). Thousand Oaks, CA: Sage Publications, Inc.

Sarason, S. (1996). *Revisiting "The culture of the school and the problem of change."* New York: Teachers College Press.

Schafer, W. D., & Lissitz, R. W. (1987). Measurement training for school personnel: Recommendations and reality. *Journal of Teacher Education, 38*(3), 57-63.

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis, 36*(3), 259-280.

Senge, P. M. (1990). *The fifth discipline: The art and practice of the learning organization*. New York: Doubleday/Currency.

Shadish, Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Independence, KY: Wadsworth Cengage Learning.

Shepard, L. A. (2010). What the marketplace has brought us: Item-by-item teaching with little instructional insight. *Peabody Journal of Education*, *85*(2), 246-257.

Shirley, D., & Hargreaves, A. (2006). Data-driven to distraction. *Education Week*, *26*(6), 32-33.

Simon, N. S., & Johnson, S. M. (2015). Teacher turnover in high-poverty schools: What we know and can do. *Teachers College Record, 117*(3), 1-36.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research, *Review of Educational Research, 75*(3), 417–453.

Slavin, R. E., Cheung, A., Holmes, G., Madden, N. A., & Chamberlain, A. (2013). Effects of a data-driven district reform model on state assessment outcomes. *American Educational Research Journal*, *50*(2), 371-396.

SMARTER Balanced Consortium. (2014). Website: Smarter Balanced Assessments. Retrieved on August 11, 2014 at: http://www.smarterbalanced.org/smarter-balanced-assessments/.

Spillane, J. P. (2012). Data in practice: Conceptualizing the data-based decision-making phenomena. *American Journal of Education*, *118*(2), 113-141.

Spillane, J. P., Pareja, A. S., Dorner, L., Barnes, C., May, H., Huff, J., & Camburn, E. (2010). Mixing methods in randomized controlled trials (RCTs): Validation, contextualization, triangulation, and control. *Educational Assessment, Evaluation and Accountability, 22*(1), 5-28.

Stecher, B. M., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A., Sloan McCombs, J., ... Naftel, S. (2008). *Pain and gain: Implementing No Child Left Behind in three states, 2004-2006* (Vol. 784). Santa Monica, CA: RAND Corporation.

Stecher, B. M., & Hamilton, L. S. (2006). *Using test-score data in the classroom* (Working Paper WR-375-EDU). Santa Monica, CA: RAND Corporation.

Stecher, B. M., Kirby, S. N., Barney, H., Pearson, M. L., & Chow, M. (2004). Organizational improvement and accountability: lessons for education from other sectors. Santa Monica, CA: RAND Corporation.

Stiggins, R. (2005). From formative assessment to assessment for learning: a path to success in standards-based schools. *Phi Delta Kappan, 87*(4), 324-328.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758-765.

Stoll, L. (1998). School culture. *School improvement network's bulletin, 9*, 9-14.

Supovitz, J. (2009). Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. *Journal of Educational Change*, *10*(2-3), 211-227.

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

Teddlie, C., & Tashakkori, A. (2003). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In C. Teddlie & A. Tashakkori (Eds.), *Handbook of mixed methods in social & behavioral research*, (pp. 3-50).Thousand Oaks, CA: Sage Publications, Inc.

Turner, E. O. & Coburn, C. E. (2012). Interventions to promote data use: An introduction. *Teachers College Record, 114*(11), 1-13.

Tyack, D. B. (1995). *Tinkering toward utopia*. Cambridge, MA: Harvard University Press.

U.S. Department of Education. (2009). Race to the top program: Executive summary. Retrieved on August 11, 2014 at: http://www2.ed.gov/programs/racetothetop/executive-summary.pdf.

U.S. Department of Education. (2014). Website: ESEA Flexibility. Retrieved on August 11, 2014 at http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html.

U.S. Department of Education. (2014). Website: Race to the Top Assessment Program. Retrieved on August 11, 2014 at: http://www2.ed.gov/programs/racetothetop-assessment/index.html.

U.S. Department of Education. (2015). Website: Fact Sheet – Testing Action Plan. Retrieved on October 30, 2015 at: http://www.ed.gov/news/press-releases/fact-sheet-testing-action-plan.

Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods, 18*(2), 137.

Van den Berg, H. (2005). Reanalyzing qualitative interviews from different angles: the risk of decontextualization and other problems of sharing qualitative data. *Forum: Qualitative Social Research, 6*(1). Retrieved on March 1, 2015 from http://www.qualitative-research.net/index.php/fqs/article/view/499/1075.

Watson-Gegeo, K. A. (1988). Ethnography in ESL: Defining the essentials. *TESOL Quarterly, 22*(4), 575-592.

Wayman, J. C. (2005). Involving teachers in data-driven decision making: Using computer data systems to support teacher inquiry and reflection. *Journal of Education for Students Placed at Risk, 10*(3), 295-308.

Wayman, J. C., & Cho, V. (2009). Preparing educators to effectively use student data systems. In T. J. Kowalski & T. J. Lasley II (Eds.), *Handbook of data-based decision making in education* (pp. 89-104). New York, NY: Routledge.

Wayman, J. C., Cho, V., Jimerson, J. B., & Spikes, D. D. (2012). District-Wide Effects on Data Use in the Classroom. *Education policy analysis archives*, *20*(25).

Wayman, J. C., Jimerson, J. B., & Cho, V. (2012). Organizational considerations in establishing the Data-Informed District. *School Effectiveness and School Improvement*, *23*(2), 159-178.

Wayman, J. C., Snodgrass Rangel, V. W., Jimerson, J. B., & Cho, V. (2010). Improving data use in NISD: Becoming a data-informed district. Austin: The University of Texas.

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.

Webb, N. (2002, April). *Assessment literacy in a standards-based urban education setting*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Weiss, M. J., Bloom, H. S., & Brock, T. (2013). *A conceptual framework for studying the sources of variation in program effects*. New York, NY: MDRC.

West, M. R., Morton, B. A., & Herlihy, C. (2016). *Achievement Network's Investing in Innovation Expansion: Impacts on Educator Practice and Student Achievement.* Cambridge, MA: Center for Education Policy Research, Harvard Graduate School of Education.

What Works Clearinghouse. (2014). *WWC procedure and standards handbook* (v 3.0). Washington, DC: Retrieved November 11, 2014 from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf.

White, H. (2013). The use of mixed methods in randomized control trials. *New Directions for Evaluation, 138,* 61–73.

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design* (2nd Ed.). Alexandria, VA: Association for Supervision and Curricular Development.

Williams, R. T., Swanlund, A., Miller, S., Konstantopoulos, S., Eno, J., van der Ploeg, A., & Meyers, C. (2014). Measuring instructional differentiation in a large-scale experiment. *Educational and Psychological Measurement, 74*(2), 263-279.

Yin, R. K. (2013). *Case study research: Design and methods*. Thousand Oaks, CA: Sage Publications, Inc.

Young, V. M. (2008). Supporting teachers' use of data: The role of organization and policy. In E. B. Mandinach & M. Honey (Eds.), *Linking data and learning* (pp. 87–106). New York: Teachers College Press.

Young, V. M. (2006). Teachers' use of data: Loose coupling, agenda setting, and team norms. *American Journal of Education, 112*(4), 521-548.

Young, V. M., & Kim, D. H. (2010). Using Assessments for Instructional Improvement: A Literature Review. *Education Policy Analysis Archives*, *18*(19).

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research, 37*(2), 197-206.

Zigarelli, M. A. (1996). An empirical test of conclusions from effective schools research. *The journal of educational research, 90*(2), 103-110.

APPENDIX A: SAMPLE COMPARISONS

Across both data collection waves, a total of 119 schools were recruited for the larger evaluation. Of those, 67 schools are in the year-two survey impact sample used in this study. In this appendix, comparisons on available measures are made between three groups: schools that were recruited and randomized, but (1) dropped out prior to year one of the study, (2) dropped out after year one of the study, and (3) schools that remain in the year-two analysis sample for this study.

Of the schools that are not included in the analyses in this study, 13 dropped out of the study after randomization, but before any interaction with the Achievement Network (ANet) began. Additionally, one school closed, another refused to participate in survey data collection, and a third is excluded due its alternative student population. These schools and their matched pairs (n = 32) comprise group "A": schools that "left" the study prior to year one. Group "B" includes the 10 treatment schools that declined to continue their partnership with ANet after year one and their matched pairs (n = 20).

The loss of schools has implications for internal validity. Internal validity is addressed by the matched-pair method of randomization. Because the matched pair of any attritted school can also be excluded, the analyses in this study provide internally valid estimates of program impacts under the assumption that the decision to remain in the study is uncorrelated with the outcomes of interest within school pairs. As assurance, these supplemental analyses explore whether this balance is maintained on a key set of observable school, teacher, and student characteristics.

Attrition at Baseline

The first set of comparisons explore whether there is something systematically different – and observable – about schools in the year-two survey impact sample (n = 67) and schools from group "A": those schools that left the study after randomization, but prior to implementation or that refused to take part in survey data collection (n = 32). These analyses capitalize on existing school-level data collected by the National Center for Education Statistics' Common Core of Data, as well as school-level state summative assessment performance reports to test for differences in the two groups on measures of student demographics and performance at baseline (2010-11 for wave one and 2011-12 for wave two).

Among all schools, those that remain in the year-two sample are about 16 percent smaller than the schools that were excluded from the study prior to the start of year one ($p < 0.05$) (table A.1). In addition, a slightly lower proportion of students in the schools that remain in the year-two sample are proficient in math (-8 percentage points, $p < 0.10$). Among the treatment schools, those that remain in the year-two sample have a slightly lower proportion of students who are of a racial or ethnic minority (-8 percentage points, $p < 0.10$). These difference are relatively small and raise no major concerns regarding the attrition of schools after randomization, but prior to implementation.

Table A.1. Baseline Comparison of School Characteristics for Schools that Attritted Prior to Year One and Schools that Remain in Year-Two Sample, by Full Sample and Treatment Sample

	Full Sample		Treatment Sample	
Baseline Measure	Schools that Attritted	Schools in Y2 Sample	Schools that Attritted	Schools in Y2 Sample
Mean School Readiness Rating	13.1	12.9	13.4	13.0
Total Enrollment (n)	501	422 **	510	462
Percentage of Student Enrollment (%)				
FRPL eligible	88.0	87.6	90.2	87.0
Minority	90.8	86.2	93.5	86.3 *
Proficient in reading	50.8	48.2	51.8	48.0
Proficient in math	55.7	47.7 *	55.5	47.4

Note: The number of schools that dropped out after randomization but prior to implementation was 32. The number of schools in the year-two sample is 67. One treatment school is missing FRPL data. One treatment school is missing readiness data.
Source: Readiness data come from Achievement Network administrative files. All other data come from the Common Core of Data: 2010-11 for wave one schools and 2011-12 for wave two schools.
** $p < 0.05$; * $p < 0.10$.

Attrition at Year One

The loss of schools between year one and two also has implications for internal validity. Since many schools left the study after year one due to a change in leadership or a leadership decision, it is plausible that a new leader or an existing leader facing budgetary constraints might look to cut what they perceive to be their least effective programs and partnerships, thereby leaving only those treatment schools that are the most motivated or that perceived ANet was having a positive impact. These supplemental analyses test for differences between schools in the year-two survey impact sample (n = 67) and schools in group "B": those schools that closed or attritted from the sample between year-one and year-two (n = 20). Comparison are made on year-one teacher scale measures comparable to those used in the main study (where available), as well as student math and reading achievement (from state or district administrative data files). A

positive coefficient signifies that teachers or students in schools that dropped out between years one and two scored higher on a given measure (on average).

Table A.2. Year-One Comparison of Teachers Survey Scales and Student Achievement Scores for Schools that Attritted After Year One and Schools that Remain in Year-Two Sample, by Full Sample and Treatment Sample

	Full Sample			Treatment Sample		
Year One Measure	Standardized Difference	SE	*p*-value	Standardized Difference	SE	*p*-value
Teacher Scales and Indices						
Professional culture: CPT discussions	0.16	0.160	0.308	0.11	0.170	0.522
Confidence: instructional practice	0.20	0.136	0.134	0.05	0.160	0.738
Data review	0.29	0.191	0.123	-0.01	0.175	0.963
Data use	0.23	0.158	0.148	0.04	0.177	0.818
Instructional planning	0.27 **	0.133	0.042	0.13	0.158	0.419
Instructional differentiation	0.19	0.164	0.256	-0.05	0.201	0.788
Student Achievement						
Math achievement	0.02	0.198	0.921	-0.33 ‡	0.017	0.000
Reading achievement	0.08	0.142	0.591	-0.15 ‡	0.014	0.000

NOTE: Comparisons are made between the 67 "stayer" schools and 20 "leavers." Because some schools had no in-scope teachers, survey comparisons include 66 stayers and 17 leavers in the "Full Sample" models. The "Treatment Sample" models compare the 34 treatment schools that remain in the year-two sample with the 10 "leavers." Differences in teacher scale and index scores are generated from two-level models that include controls for district, data collection wave and Chelsea triad, as well as total teaching experience and highest degree at the teacher level. Student achievement models are calculated using cluster-adjusted OLS models and the same model specifications as the larger i3 evaluation (see West, Morton, & Herlihy, 2016).
Source: The year-one teacher survey, year-one (2011-12 or 2012-13) state or district administrative student data.
‡ $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

In the full sample, teachers in schools that left the study after year one reported a slightly higher frequency of instructional planning practices ($sd = 0.27$, $p < 0.05$) as compared to teachers in the year-two analysis sample schools (table A.2). Among treatment schools, students in those that left the study after year one had lower achievement ($p < 0.01$). This suggests that the decision to leave the study could have been based on the program's perceived effects on student performance, but likely not teacher performance. Though this may be of concern for the larger study, no observable

differences in these samples raise concerns regarding the effects of school attrition on the analyses in this study.

APPENDIX B: ADDITIONAL MODELS

RESEARCH QUESTION ONE: TEACHER PRACTICE IMPACT MODELS WITH BASELINE COVARIATE

Due to survey nonresponse and non-administration, there was a high percentage of missing baseline data for the year-two teacher sample. As a result, primary analyses in this study do not include a baseline measure of the outcome of interest. With randomization, teachers in the treatment and control schools will be equivalent on all observable and unobservable characteristics, except for the receipt of treatment, up to statistical sampling error. Even so, analyses of data from randomized studies often include baseline measures of the outcome of interest in order to correct for chance differences in that outcome between treatment and control groups. Additionally, to the extent that the baseline measure is correlated with the outcome, its inclusion can improve power by explaining a portion of the variance in the outcome not due to the treatment, thus, making the treatment effect easier to detect.

The analyses in table B.1 report the results for research question one after including a baseline measure of the outcome of interest.[1] Compared to the main models in chapter four, there is a potential trade-off in statistical power given the specifications and sample in these models. On one hand, the inclusion of the baseline covariate should improve power by explaining a portion of the variance in the outcome not due to

[1] These baseline covariate models are only shown for research question one because many of the composite measures of hypothesized school- and teacher-level mediators were new or revised for the year-two teacher survey and have no equivalent baseline measure.

treatment. On the other hand, only 35 to 39 percent of the year-two teacher sample has a non-missing value of the baseline covariate. A loss in sample typically decreases power.

Overall, the results from models that control for the baseline measure of the outcome are very similar to the main models in chapter four. The magnitudes of the estimates change slightly, but the frequency that teachers in ANet schools report reviewing and using data is still significantly higher than that of their control-school peers (data review = 0.40 *sd,* data use = 0.25 *sd*, both $p < 0.01$). Treatment-control differences in the frequency with which teachers report various instructional planning strategies and differentiate instruction remain indistinguishable from zero (table B.1).

Each baseline measure is positively correlated with its respective year-two outcome. A one standard deviation increase in the respective baseline measure is associated with about a one-quarter standard deviation increase in the frequency with which teachers review and use data ($p < 0.01$). A one standard deviation increase in the respective baseline measure is associated with about one-half a standard deviation increase in the frequency with which teachers use various instructional planning strategies and differentiate instruction. However, despite their strong correlation with the outcomes, the addition of the baseline covariates seems to have resulted in a slight reduction in power. Standard errors across estimates generally increased (table B.1 compared to table 4.7). This is likely due to the significant loss of sample.

Table B.1. Teacher Practice Impact Results with Baseline Covariate

Variable	Model 1 Data Review	Model 2 Data Use	Model 3 Instructional Planning	Model 4 Instructional Differentiation
Fixed effect				
Assigned to treatment: school	0.40 ‡	0.28 ‡	0.14	-0.15
	0.147	*0.110*	*0.111*	*0.128*
Baseline measure of outcome	0.26 ‡	0.24 ‡	0.46 ‡	0.51 ‡
	0.052	*0.059*	*0.054*	*0.059*
District				
Chelsea	-0.15	0.06	-0.24	0.13
	0.333	*0.238*	*0.232*	*0.265*
Chicago	0.40	-0.19	-0.05	-0.96 **
	0.387	*0.387*	*0.372*	*0.414*
Jefferson Parish	-0.01	-0.07	0.25	-0.15
	0.205	*0.178*	*0.172*	*0.193*
Springfield	0.24	0.30 *	0.20	0.12
	0.221	*0.162*	*0.162*	*0.187*
Unbalanced pair dummy: school	-0.25	-0.71 ‡	-0.10	-0.39
	0.397	*0.273*	*0.278*	*0.318*
Years of teaching experience	0.01	0.01 *	0.00	0.00
(total): teacher	*0.006*	*0.006*	*0.006*	*0.007*
Highest degree: teacher				
Master's	0.26 *	0.30 *	0.03	-0.06
	0.146	*0.154*	*0.145*	*0.161*
Doctorate	-0.69	-1.17 **	-0.13	-0.31
	0.505	*0.562*	*0.580*	*0.637*
Random effect				
School (intercept)	-0.68 ‡	-0.69 ‡	-0.33 *	0.11
	0.211	*0.195*	*0.188*	*0.209*
Variance Components				
L1	0.425	0.550	0.609	0.731
L2	0.119	0.000	0.012	0.023
Model statistics				
n	202	201	233	228
Number of groups	47	47	49	49
Wald χ^2	61.31 ‡	75.90 ‡	102.49 ‡	96.37 ‡

Notes: Outcome scales were standardized within the teacher sample; results are reported in standard deviation units. Estimates are reported on the top row for each predictor. Standard errors are reported below, in italics. Omitted district = Boston; omitted degree = bachelor's. Data collection wave is omitted since all teachers are from wave one.
‡ $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Finally, it is worth noting the varying changes in the magnitude and direction district effects across outcomes after including baseline measures of the outcome. While this may have implications for the interpretation of results in chapter four, it is impossible to tell whether the changes in the patterns of estimates are due to a lack of baseline equivalence in some districts or the fact that the estimates in table B.1 are generated from a different sample of teachers than the main models in chapter four.

RESEARCH QUESTION TWO: WEIGHTED SCHOOL MEDIATOR IMPACT MODELS

Recall that inverse weighting is proposed as a way to account for the unknown rate of nonresponse by teachers within schools. Inverse variance weighting makes an assumption that larger variance is due to fewer responding teachers and, therefore, higher nonresponse. However, larger variance could also indicate greater heterogeneity in teacher responses within school regardless of response rate. If this is the case, weighted models can also introduce bias when re-estimating the models from research question two. As a result, the "true" estimates may lie in the range between unweighted and weighted estimates.

Weighting appears to alter some point estimates by a sizable amount. The weighted and unweighted estimates in models predicting school-mean teacher-perceived leader abilities are very similar (table B.2, model 1 versus model 2). However, the treatment effect on the school-mean frequency of common planning time discussions nearly doubles in the weighted model (table B.2, model 3 versus model 4).

Table B.2. School Mediator Impact Results, Unweighted and Inverse Variance Weighted

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
			Professional Culture					
	Instructional Leaders' Abilities		CPT discussions		General Collegiality		Achievement Culture	
Variable	Unweighted	Weighted[1]	Unweighted	Weighted[1]	Unweighted	Weighted[1]	Unweighted	Weighted[1]
Fixed effect								
Assigned to treatment: school	0.41 *	0.43 *	0.29	0.52 **	0.35	0.48 *	0.45 *	0.52 **
	0.232	*0.236*	*0.209*	*0.206*	*0.239*	*0.262*	*0.230*	*0.225*
District								
Chelsea	0.39	0.58	0.46	0.31	-0.40	-0.06	0.25	0.58
	0.781	*0.998*	*0.703*	*0.967*	*0.804*	*1.126*	*0.771*	*0.824*
Chicago	0.39	0.50	0.84 **	-0.10	0.50	0.26	0.09	0.35
	0.456	*0.489*	*0.410*	*0.351*	*0.470*	*0.636*	*0.451*	*0.450*
Jefferson Parish	0.60	0.73	1.40 ‡	0.83 *	0.60	0.69	0.00	-0.13
	0.520	*0.462*	*0.468*	*0.426*	*0.535*	*0.517*	*0.514*	*0.432*
Springfield	0.63 *	0.69 *	0.65 *	0.16	0.63	0.82 *	0.50	0.49
	0.376	*0.398*	*0.338*	*0.310*	*0.387*	*0.438*	*0.371*	*0.385*
Data collection wave two: school	-0.19	-0.28	0.00	-0.22	-0.10	-0.60 *	-0.49	-0.71 **
	0.345	*0.345*	*0.310*	*0.292*	*0.355*	*0.358*	*0.341*	*0.308*
Unbalanced pair dummy: school	-1.49	-1.39	-0.64	-0.74	-0.53	-0.63	-1.58 *	-1.51
	0.893	*1.066*	*0.803*	*1.044*	*0.919*	*1.209*	*0.882*	*0.937*
School mean years of teaching	0.05	0.06	0.09 ‡	0.14 ‡	0.03	0.10 **	0.05	0.10 ‡
experience (total)	*0.035*	*0.042*	*0.032*	*0.032*	*0.036*	*0.042*	*0.035*	*0.036*
School mean highest degree								
Master's	0.39	0.42	1.22 **	1.28 **	1.08	1.77 **	0.70	0.28
	0.667	*0.653*	*0.600*	*0.587*	*0.686*	*0.732*	*0.659*	*0.577*
Doctorate	-4.40	-1.84	-5.23	-5.47	-1.97	-2.48	-2.55	-2.74
	4.478	*4.085*	*4.029*	*4.674*	*4.609*	*4.135*	*4.424*	*4.125*
School (intercept)	-1.23	-1.66 *	-2.78 ‡	-2.98 ‡	-1.53 *	-2.56 ‡	-0.85	-1.17
	0.797	*0.919*	*0.717*	*0.728*	*0.820*	*0.939*	*0.787*	*0.769*
Model statistics								
n	67	67	67	67	67	67	67	67
Adjusted R^2	0.13	0.12	0.29	0.31	0.08	0.17	0.15	0.27

[1] Weighted by the inverse of the variance in the school-mean scale score.

Notes: Outcome scales were standardized within the school sample; results are reported in standard deviation units. Estimates are reported on the top row for each predictor. Standard errors are reported below, in italics. Omitted district = Boston; omitted degree = bachelor's; and wave one = 1, wave two = 2.

‡ $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

District estimates vary widely in whether they are more strongly positive, more strongly negative, or closer to zero. Compared to Boston, district estimates of the school-mean frequency of common planning time discussions are generally larger and more positive; estimates of the school-mean frequency of general collegial conversations are similar (Jefferson Parish and Springfield) or larger (more negative in Chelsea, more positive in Chicago); and school-mean estimates of the achievement culture are generally smaller (closer to zero) in the unweighted models.

RESEARCH QUESTION THREE: TEACHER PRACTICE MEDIATION MODELS BY SCHOOL READINESS

In the analyses shown below, the models from research question three were re-estimated separately on schools in the higher readiness (N = 34) and lower readiness (N = 33) groups in order to explore whether the role of the hypothesized mediators differs in these schools. Recall that readiness was rated at baseline and based on school responses to items measuring school opt-in, program priority and organization, dedication of leadership, standards and alignment, and scheduling (see appendix C for descriptions of each measure and rubric level).

The results for higher readiness schools indicate that hypothesized school- and teacher-level mediators do little to reduce the overall impact of ANet on teachers' data review and data use practices. In models 4 and 8, where blocks of both hypothesized school- and teacher- level mediators are included, estimates of the impact of treatment assignment remain large and statistically significant (table B.3, $p < 0.01$). Separately, the block of hypothesized school-level mediators reduce the estimate of the treatment effect

slightly more than the block of hypothesized teacher-level mediators for both outcomes (17 versus 7 percent for data review, 24 versus 10 percent for data use).

Table B.3. Teacher Practice Mediation Results for Higher Readiness Schools

	Data Review				Data Use			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Variable	Impact Model	School Mediation Model	Teacher Mediation Model	Combined Mediation Model	Impact Model	School Mediation Model	Teacher Mediation Model	Combined Mediation Model
Fixed effect								
Assigned to treatment: school	0.58 ‡	0.48 ‡	0.53 ‡	0.47 ‡	0.41 ‡	0.31 ‡	0.37 ‡	0.30 ‡
	0.150	*0.125*	*0.147*	*0.125*	*0.124*	*0.120*	*0.114*	*0.106*
School-level mediators								
Instructional leaders' abilities		0.04		-0.08		-0.02		-0.14
		0.117		*0.119*		*0.112*		*0.100*
CPT discussions		0.39 ‡		0.34 ‡		0.17		0.15
		0.108		*0.107*		*0.106*		*0.093*
General collegiality		0.04		0.11		-0.04		0.02
		0.117		*0.117*		*0.111*		*0.098*
Achievement culture		-0.05		-0.04		0.17		0.18 *
		0.120		*0.120*		*0.115*		*0.101*
Teacher-level mediators								
Assessment/data attitudes			0.26 ‡	0.23 ‡			0.22 ‡	0.21 ‡
			0.058	*0.058*			*0.049*	*0.049*
Data use confidence			0.16 **	0.16 **			0.20 ‡	0.19 ‡
			0.071	*0.070*			*0.063*	*0.062*
Instructional planning confidence			0.03	0.01			0.18 ‡	0.17 ‡
			0.071	*0.070*			*0.063*	*0.062*
District								
Chicago	0.46	-0.07	0.33	-0.11	0.21	0.05	-0.02	-0.10
	0.282	*0.277*	*0.275*	*0.274*	*0.243*	*0.259*	*0.222*	*0.228*
Jefferson Parish	-0.20	-0.81 ‡	-0.32	-0.76 ‡	0.07	-0.04	-0.14	-0.14
	0.228	*0.290*	*0.227*	*0.289*	*0.185*	*0.279*	*0.173*	*0.248*
Springfield	0.49 *	0.07	0.32	-0.09	0.25	0.03	0.00	-0.21
	0.275	*0.251*	*0.274*	*0.253*	*0.221*	*0.242*	*0.206*	*0.214*
Data collection wave two: school	0.42 *	0.53 ‡	0.44 *	0.49 **	0.22	0.26	0.30 *	0.29 *
	0.239	*0.198*	*0.237*	*0.200*	*0.190*	*0.186*	*0.177*	*0.163*
Years of teaching experience	0.01 *	0.01	0.01	0.00		0.01	0.00	0.00
(total): teacher	*0.007*	*0.006*	*0.006*	*0.006*		*0.006*	*0.006*	*0.005*
Highest degree: teacher								
Master's	0.27 *	0.20	0.11	0.07	0.01 *	0.18	0.01	-0.01
	0.150	*0.149*	*0.142*	*0.143*	*0.006*	*0.141*	*0.125*	*0.125*
Doctorate	-1.94 **	-2.04 **	-0.95	-1.15	0.21	-2.34 ‡	-1.40 *	-1.53 *
	0.940	*0.924*	*0.917*	*0.909*	*0.141*	*0.888*	*0.821*	*0.811*
School (intercept)	-1.16 ‡	-0.87 ‡	-0.94 ‡	-0.70 **	-2.37 ‡	-0.73 ‡	-0.58 **	-0.48 **
	0.317	*0.281*	*0.312*	*0.280*	*0.902*	*0.269*	*0.244*	*0.238*
Variance Components								
L1	0.827	0.810	0.715	0.713	0.778	0.750	0.584	0.573
L2	0.0493963	0.000	0.058	0.011	0.0079875	0.000	0.015	0.000
Model statistics								
n	266	266	264	264	272	272	270	270
Number of groups	34	34	34	34	34	34	34	34
Wald χ^2	35.75 ‡	69.60 ‡	74.58 ‡	101.77 ‡	30.90 ‡	47.56 ‡	117.94 ‡	139.79 ‡

Notes: Outcome scales were standardized within the teacher sample; results are reported in standard deviation units. Estimates are reported on the top row for each predictor. Standard errors are reported below, in italics. Omitted district = Boston; omitted degree = bachelor's; and wave one = 1, wave two = 2. Dummy variables for Chelsea schools and the unbalanced Chelsea pair are omitted since no schools from Chelsea were in the higher-readiness group.

‡ $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

All else being equal, the patterns in the relationships among hypothesized mediators and outcomes are similar to the full-sample models in chapter four (table 4.13). Controlling for all other measures, greater frequency in teachers' common planning time conversations is associated with greater frequency of reviewing data (0.34, *sd,* $p < 0.01$) (model 4, table B.3). All else being equal, teachers' attitudes towards and confidence using data are positively associated with the frequency they review and use data (all $p < 0.05$) (models 4 and 8, table B.3). Finally, greater confidence in instructional planning is associated with more frequent use of data (0.17 *sd,* $p < 0.01$) (model 8, table B.3).

The patterns in the relationships among hypothesized mediators and outcomes are similar in lower-readiness schools compared to the full-sample models in chapter four (table 4.13). Controlling for all other measures, greater frequency of teachers' common planning time conversations is associated with greater frequency of reviewing data (0.30 *sd*, $p < 0.01$) (model 4, table B.4). All else being equal, teachers' attitudes towards and confidence using data are positively associated with the frequency they review and use data (all $p < 0.01$) (models 4 and 8, table B.4).

There are two notable deviations from the pattern of findings in these models compared to the full sample and higher-readiness schools. Adjusting for treatment assignment, other hypothesized mediators, and all other school- and teacher-level covariates, in lower-readiness schools (1) collegiality among the staff is negatively associated with the frequency they review data (-0.24 *sd*, $p < 0.10$) (model 4, table B.4) and (2) teachers' confidence in instructional planning is no longer positively associated with the frequency of their data use (model 8, table B.4). Though the former finding is

counterintuitive, one seemingly anomalous finding in a study that includes many model estimates is not surprising.

Table B.4. Teacher Practice Mediation Results for Lower Readiness Schools

	Data Review				Data Use			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Variable	Impact Model	School Mediation Model	Teacher Mediation Model	Combined Mediation Model	Impact Model	School Mediation Model	Teacher Mediation Model	Combined Mediation Model
Fixed effect								
Assigned to treatment: school	0.37 **	0.22 *	0.24 *	0.19 *	0.10	0.05	0.03	0.03
	0.151	*0.121*	*0.125*	*0.108*	*0.110*	*0.115*	*0.098*	*0.102*
School-level mediators								
Instructional leaders' abilities		0.08		-0.01		-0.08		-0.19 *
		0.120		*0.108*		*0.116*		*0.104*
CPT discussions		0.34 ‡		0.30 ‡		0.19 **		0.13
		0.092		*0.084*		*0.090*		*0.081*
General collegiality		-0.28 *		-0.24 *		0.01		0.05
		0.152		*0.136*		*0.146*		*0.129*
Achievement culture		0.19		0.13		0.06		0.00
		0.145		*0.130*		*0.141*		*0.125*
Teacher-level mediators								
Assessment/data attitudes			0.17 ‡	0.14 ‡			0.26 ‡	0.26 ‡
			0.052	*0.052*			*0.054*	*0.055*
Data use confidence			0.20 ‡	0.21 ‡			0.27 ‡	0.29 ‡
			0.067	*0.067*			*0.068*	*0.069*
Instructional planning confidence			0.12 *	0.10			0.10	0.09
			0.064	*0.063*			*0.066*	*0.066*
District								
Chelsea	-0.19	-0.28	-0.03	-0.13	-0.49 *	-0.46 *	-0.33	-0.31
	0.339	*0.280*	*0.279*	*0.250*	*0.253*	*0.272*	*0.222*	*0.240*
Chicago	1.02 **	0.99 ‡	0.89 **	0.85 **	0.12	0.05	-0.17	-0.25
	0.416	*0.354*	*0.380*	*0.346*	*0.387*	*0.385*	*0.368*	*0.368*
Jefferson Parish	-0.06	-0.44 *	-0.02	-0.33	-0.46 **	-0.66 ‡	-0.40 **	-0.50 **
	0.258	*0.233*	*0.220*	*0.212*	*0.217*	*0.233*	*0.190*	*0.207*
Springfield	0.19	0.08	0.22	0.16	-0.04	-0.14	-0.10	-0.06
	0.265	*0.215*	*0.221*	*0.193*	*0.206*	*0.211*	*0.180*	*0.186*
Data collection wave two: school	0.62 **	0.30	0.58 ‡	0.34 *	0.53 ‡	0.33	0.55 ‡	0.41 **
	0.273	*0.224*	*0.224*	*0.200*	*0.196*	*0.211*	*0.172*	*0.187*
Unbalanced pair dummy: school	-0.32	0.07	-0.19	-0.01	-0.57 **	-0.50 *	-0.38 *	-0.55 **
	0.345	*0.308*	*0.276*	*0.272*	*0.229*	*0.289*	*0.201*	*0.255*
Years of teaching experience	0.01 ‡	0.01 **	0.00	0.00	0.02 ‡	0.02 ‡	0.00	0.00
(total): teacher	*0.006*	*0.005*	*0.005*	*0.005*	*0.006*	*0.006*	*0.005*	*0.005*
Highest degree: teacher								
Master's	0.20	0.22 *	0.17	0.19	-0.06	-0.06	-0.08	-0.10
	0.133	*0.131*	*0.123*	*0.122*	*0.142*	*0.141*	*0.125*	*0.125*
Doctorate	0.07	0.08	0.35	0.18	-0.28	-0.23	-0.02	-0.03
	0.369	*0.365*	*0.624*	*0.343*	*0.405*	*0.400*	*0.359*	*0.356*
School (intercept)	-1.24 ‡	-0.74 **	-1.05 ‡	-0.68 **	-0.44	-0.11	-0.31	-0.09
	0.378	*0.336*	*0.320*	*0.304*	*0.312*	*0.333*	*0.273*	*0.293*
Variance Components								
L1	0.675	0.669	0.582	0.579	0.839	0.814	0.639	0.628
L2	0.091	0.021	0.047	0.011	0.000	0.000	0.000	0.000
Model statistics								
n	293	293	291	291	297	297	294	294
Number of groups	33	33	33	33	33	33	33	33
Wald χ^2	48.77 ‡	102.29 ‡	127.39 ‡	178.93 ‡	94.02 ‡	105.90 ‡	217.57 ‡	226.16 ‡

Notes: Outcome scales were standardized within the teacher sample; results are reported in standard deviation units. Estimates are reported on the top row for each predictor. Standard errors are reported below, in italics. Omitted district = Boston; omitted degree = bachelor's; and wave one = 1, wave two = 2.
‡ $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

The results also suggest that hypothesized school- and teacher-level mediators could play a greater role in explaining the impact of ANet on teachers' data review and data use practices in lower readiness schools. In the combined model predicting teachers' frequency of reviewing data, hypothesized school- and teacher-level mediators combine to reduce the treatment effect by about half: from 0.37 *sd* ($p < 0.05$) to 0.19 *sd* ($p < 0.10$) (models 1 versus 4, table B.4). Though the effect of ANet on the frequency with which teachers use data is already not statistically significant in the partially-adjusted model (model 5), the inclusion of both hypothesized school- and teacher-level mediators reduce the treatment coefficient by about 70 percent: from 0.10 *sd* (*ns*) to 0.03 *sd* (*ns*) (models 5 versus 8, table B.4).

Readiness was defined by (1) the level of school buy-in and a clear plan for how ANet can address specific school needs, (2) a priority on improving student performance and clarity on the role of ANet, (3) dedicated leaders and the allocation of time to implement ANet in the school, (4) alignment of rigorous content standards and quality curriculum, as well as the prioritization of measuring progress towards standards, and (5) allocated time for ANet implementation and professional development. These results suggest that, in schools where readiness was rated to be high at baseline, hypothesized mediators may do less to facilitate teachers' data practices because of an already supportive school environment. In contrast, where readiness was rated lower, it may be more important to ensure supportive instructional leadership, school culture, more positive attitudes, and greater teacher confidence in order to assist teachers' adoption of data practices.

APPENDIX C: ACHIEVEMENT NETWORK READINESS SCREENER

Exhibit C.1. Scoring Rubric for Baseline Screener of School Readiness

Conditions Included in Analysis			
	1 point	**2 points**	**3 points**
School Opt-in	The school is only applying because the district/CMO told it to and does not seem to understand the value of the work. They may want to apply, but are expecting something different than what we offer, such as coaching for their own assessments.	The school is applying because either they are interested in the program on their own or because their district/CMO or an external funder recommended it, but the schools still sees the general value in the program as a whole.	The school not only shows genuine interest in the program as a whole, but also can identify specific needs as well as a coherent explanation for how ANet will help address these needs.
Priority & Organization	The school has either explicitly said that ANet would not be one of its top priorities, or alternatively does not have the ability or capacity to set priorities. This could be evidenced by having essentially no strategic plan for the year, planning time that usually has no agenda and ends up dealing with the most pressing issues on that day, or focusing on more urgent problems such as student safety and classroom management.	The school has a set of pre-determined priorities, but is not always successful at continuing to focus on them throughout the year. The school has expressed interest in making ANet one if its top priority, but can't promise it will get full attention at each interaction as there may be more pressing issues going on.	The school has a clear set of priorities for the year that are defined in advance. The leadership team strongly believes in accountability and measuring progress to outcomes. Improving student achievement is one of their top 3 priorities for the year, and can describe how ANet fits into these priorities.
Dedication of Leadership	The leadership team has not been identified for next year, possibly because they expect significant leadership turnover before the next school year. The current leadership team is focused on many things at once and cannot necessarily commit to all meet at the same time for the ANet work.	The leadership team for next year has been identified but have not explicitly confirmed that each of them have the capacity for the work. The leadership team would like to be able to focus on the ANet work as much as possible, but has many other issues. They may not have time set aside yet for the work, but have expressed an interest in figuring out a way to do so.	The leadership team has been fully identified and each person has clarified that both they have the capacity for the work and that they are committed to it. They have already planned to set aside the required time to implement the ANet work.
Standards & Alignment	The school is unwilling to align to any standards. This could be because they do not see the value, because they work from a certain curriculum program that they don't want to deviate from, or other reasons. The school likely does not understand how well they are currently aligned to standards.	The school understands the value of standards and would like to align their curriculum to them. The school may either not know how well they are currently aligned, or alternatively does know they are not well aligned but does not know how become better aligned. They may use interim assessments, but don't find them that useful as they are not clear on what goals they are measuring progress towards.	The school believes in the value of standards and has aligned its curriculum to them as much as possible. If their current curriculum systems has gaps in standard coverage, they have a clear plan for how to fill those gaps. The school may potentially align to other more rigorous standards if they think the state standards are too low-level. The school has set student achievement goals and believes in regularly measuring progress to those goals.
Scheduling	The school is either unwilling or unable to commit to blocking out the required time for the ANet program.	The school has blocked off some or most of the required time, but has not yet found time for all of the scheduled meetings. They might ask to have shorter data meetings or space the meeting out over different times with different groups of teachers.	The school has fully committed to blocking off the required time for ANet programming in their curriculum and PD schedule before the school year begins. This time may already exist that they are re-purposing for ANet work and are asking ANet to help sharpen how they use that time.

Exhibit C.1. Baseline Program Screener Rubric, Continued

Conditions NOT Included in Analysis			
	1 point	**2 points**	**3 points**
School & District Payment	The school is either not able or not willing to find a path to pay for the program through any combination of school and district budgets or external funders.	The school believes the value of the work and would like to find a way to pay for it. The school believes it can find a way to pay the fee through a combination of school and district budgets or external funders. The school may not be able to pay for any part of the fee out of its own budget.	The school is willing to pay part of the annual fee out of its own budget because it sees the value of the work and recognizes the costs it takes to do this work. The school already has a clear plan for how to pay for the program for this year and is committed to finding the financing to pay for upcoming years as well.
District Support	The district/CMO may not see the value in the ANet work and may not support the school in scheduling PD time for it. The district/CMO has also explicitly mandated initiatives that conflict with ANet such as alternative PD or additional interim assessments.	The district/CMO general supports ANet's work with the school. The district/CMO also helps facilitate the application and opt-in process. The district/CMO may be tentative as to the value of the work and wants to see results before discussing any potential growth plan for other schools.	The district/CMO is completely in support of ANet work and would like to establish a plan to expand ANet services into as many schools as possible and to work with all levels of district management. They pave the way for the application and opt-in process and strongly encourage schools to apply. The district/CMO is willing to participate in ANet's escalation plan to assist any school that falls off track.
Data and Logistics	Someone who does not have the capacity has been assigned as the logistics lead, such as the principal, or no one has been assigned. The school has no plan to update their student roster information and cannot necessarily promise it's current accuracy. The school may also have misconceptions that all logistics will be taken care of for them.	The school has assigned some to be the logistics lead, but that person has not been told all that the role will entail and agreed that they have the capacity. That person might not be clear on what all is expected. The school would like to maintain a current roster, but admits there may be errors from time to time.	The school has a dedicated logistics lead who already has the responsibilities built into their job description. They have updated systems that ensure a smooth printing and scanning process. The school has an established plan to regularly update the student roster and continuously monitor it to make sure it is accurate.
Collaboration (Network)	The school does not want to share their interim results with other schools because they don't see the value or for other reasons. The school has not expressed interest in attending any network events.	The school would like to attend some or all of the network events. They're interested in finding out what they can learn from the network, although don't have a clear idea at this point. They may not be enthusiastic about sharing results, but understand that it's a required part of the program.	The school is excited about learning best practices from other schools and explicitly names the Network as one of the main reasons they are applying. The school has committed to attend all of the network events. The school agrees with the importance of measuring itself against other schools to better understand gaps in learning.

APPENDIX D: GLOSSARY OF ACHIEVEMENT NETWORK TERMS

Action Plan: see Reteaching plan

Backward Planning (or Mapping): a method of instructional planning promoted by the Achievement Network (ANet) that begins by focusing on the end goal of a unit or lesson: typically, mastery of a skill or standard. The first step in this process is to identify the target skill or standard. This target is translated into a learning objective. With this defined, teachers identify instructional strategies and develop a unit or lesson plan, instruction is carried out, and finally, student mastery is assessed. This process is usually guided by content standards, performance standards, and, in many i3 evaluation districts, a district pacing guide (see "Pacing Guide"). In some ANet schools, backward planning began from the skills or standards to be assessed on upcoming interim assessments.

Data Meeting: meetings that take place after each administration of the ANet interim assessments. Teachers and leaders come together for a block of time in which results can be shared, discussed, and analyzed. In some cases, the ANet coach or a school leader delivers some type of data-related professional development. There may also be time built in for teachers to begin to develop their reteaching plan (see "Reteaching Plan").

Distractor Guide: see Misconceptions Guide.

Misconceptions Guide: also called a distractor guide, this resource provides teachers with explanations for why students chose incorrect answers to multiple choice items (i.e., distractors).

MyANet: an integrated online platform that provides tools and resources meant to help teachers meet their students' needs. The website includes: student data reports, sample lesson plans, a quiz builder tool (see "Quiz Builder Tool") and question stems (see "Question Stems"), a misconceptions guide (see "Misconceptions Guide"), templates to organize data analysis and lesson planning, content standards guides, and a Schedule of Assessed Standards (see "Schedule of Assessed Standards").

Pacing guide: also called a curriculum map, or curriculum scope and sequence. These guides are often developed by districts to define the content and order of material to be covered in each grade and subject. The content to be covered is often timed so that specific content has been taught before the state summative assessment and, in some cases, prior to the administration of periodic benchmark assessments such as ANet. Pacing guides can vary in the level of specificity they provide on what must be taught, how it must be taught, and on what days or class periods. (David, 2008).

Priority standards: Foundational standards presumed necessary for mastering more advanced standards or standards that are deemed priorities in each grade and/or subject and, therefore, highly likely to be tested on the summative assessment.

Question stems: Phrases commonly used on summative assessments to frame questions for various content areas, standards, and skills. These typically are the introduction to an assessment item which teachers can use to craft their own items. For example, a vocabulary item might be framed as: "In the context of lines ___ and___ of the story, the best meaning of _____ is…?"

Quiz Builder tool: a tool on the MyANet platform (see "MyANet") that can be used to generate quizzes from a bank of sample multiple choice assessment items that are organized by standard.

Reteaching/Reteaching plan: After each administration of the ANet interim assessment, teachers analyze their students' data and identify a skill or standard on which students failed to achieve mastery. The expectation is that teachers will use the data to develop a reteaching plan that targets this skill or standard, implement the plan, and then reassess student learning to see if the reteaching was successful.

Reteaching template: a template provided by ANet for assisting teachers in developing their reteaching plan (see "Reteaching Plan").

Schedule of Assessed Standards: A breakdown of the standards and skills to be covered by each ANet interim assessment. This is available at the start of each year and intended to align with a district or school's curriculum.