Center for Education Policy Research
HARVARD UNIVERSITY

RESEARCH PAPER

Achievement Network's Investing in Innovation Expansion: Impacts on Educator Practice and Student Achievement

Martin R. West
Beth A. Morton
Corinne M. Herlihy

March 2016

ABOUT THE CENTER FOR EDUCATION POLICY RESEARCH AT HARVARD UNIVERSITY

The Center for Education Policy Research (CEPR) at Harvard University is a unique partnership among districts, states, foundations, and university-based researchers designed to leverage the overwhelming amount of newly available school-, teacher-, and student-level data to address previously intractable policy questions in education and improve educational outcomes for all students.

Today is a moment of great promise in American education. Pockets of innovation are forming across the country, yet the true breakthroughs are hard to distinguish from the false claims. To recognize the difference, the current exuberance must be matched with an equal focus on evidence. In recent years, taxpayers have invested hundreds of millions on education data systems. However, because public agencies are not positioned (or staffed) to evaluate their own programs, these data are an underused national resource. By deploying two of its greatest assets—analytic talent and national credibility—Harvard has an irreplaceable role to play in K–12 education reform. CEPR is building a network of relationships with school leaders around the country and engaging the best minds in social science to learn what's working—and what's not.

For more information, visit http://cepr.harvard.edu

This work has been supported by the U.S. Department of Education's Investing in Innovation (i3) program, through grant U396C100771.The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

ACKNOWLEDGEMENTS

The research described in this report was funded by a grant from the U.S. Department of Education's Investing in Innovation (i3) Program. However, the contents of the report, as well as any errors or omissions, are those of the authors, not the U.S. Department of Education. No endorsement by the federal government should be assumed.

This report is the culmination of nearly five years of work involving many people. We would first like to thank our current and former colleagues at CEPR, particularly project manager Hilary Bresnahan, who not only ensured that the evaluation ran smoothly and efficiently, but also supported our data collection and analysis efforts. Jon Fullerton and Katie Buckley made significant contributions throughout the evaluation process, including the drafting of this report. We would also like to thank the numerous CEPR staff members and graduate students who assisted our team on school site visits: Fallon Blossom, Zubair Butt, Olivia Chi, Sara Gussin, Sara Howard, Amal Kumar, Meg Nipson, Antoniya Owens, Ashley Snowdon, Sarah St. Germain, and Sylvia Zelaya.

We benefited from excellent advisors in the early stages of the evaluation. Hunter Gehlbach, associate professor of education at the University of California, Santa Barbara, was instrumental in the development of our educator surveys. John Tyler, professor of education and economics at Brown University, made numerous contributions to the study design, including providing feedback on our surveys and site visit protocols. Geoffrey Borman served as our i3 technical assistance liaison and offered valuable feedback at each phase of our study design process.

We would also like to thank those who provided us with the data we needed to answer our research questions and helped us understand it, including Carrie Conaway, Geraldine Stewart, and Paula Willis from the Massachusetts Department of Elementary and Secondary Education; Kylie Klein from Chicago Public Schools; and Susan Adam and Thomas Lambert from Jefferson Parish Public Schools.

Our evaluation would not have been possible without additional support on the ground. Achievement Network network executive directors facilitated data collection by connecting us to district and school leaders and provided context on the program's rollout in their districts. We are also indebted to the leaders and teachers at the schools participating in the study for completing

surveys and answering our questions during site visits. Finally, we thank the central office staff who answered our questions about the context for data-based instruction in their districts.

Finally, the leadership team at Achievement Network remained enthusiastic and engaged thought partners throughout the evaluation process. Together, we sought to ensure that our results would provide valuable evidence not only for their program, but also for other organizations working to implement data-based instructional programs in schools. In particular, we would like to thank past and present staff members who were close collaborators, including John Maycock, Mora Segal, Emma Dogget, Carter Romansky, Jennifer Appleyard, Chris Rupprecht, Cori Stott, and Jennifer Poulos.

TABLE OF CONTENTS

TABLES, FIGURES, AND EXHIBITS

Tables

Figures

Exhibits

EXECUTIVE SUMMARY

Achievement Network (ANet) was founded in 2005 as a school-level intervention to support the use of academic content standards and assessments to improve teaching and learning. Initially developed within the Boston charter school sector, it has expanded to serve over 500 schools in nine geographic networks across the United States. The program is based on the belief that if teachers are provided with timely data on student performance from interim assessments tied to state standards, if school leaders provide support and create structures that help them use that data to identify student weaknesses, and if teachers have knowledge of how to improve the performance of students who are falling behind, then they will become more effective at identifying and addressing gaps in student learning. This will, in turn, improve student performance, particularly for high-need students.

In 2010, ANet received a development grant from the U.S. Department of Education's Investing in Innovation (i3) Program. The grant funded both the expansion of the program to serve up to 60 additional schools in five school districts, as well as an external evaluation of the expansion. The Center for Education Policy Research (CEPR) at Harvard University partnered with ANet to design a matched-pair, school-randomized evaluation of their program's impact on educator practice and student achievement in schools participating in its i3-funded expansion.

Background

Data-based instructional programs like ANet are increasingly widespread in American schools. The theory of action behind these programs is that providing teachers with periodic assessment data and, in some cases, other supports (e.g., coaching on data use) will allow teachers to target instruction to areas where student mastery is lacking. The adoption of data-based instructional programs has been driven in part by pressures from accountability systems to improve student achievement, as well as by evidence suggesting that these practices are a distinguishing characteristic of high-performing schools, particularly within the charter sector (Angrist, Pathak, & Walters, 2013; Dobbie & Fryer, 2013).

Despite their prominence, there is limited research on the effects of data-based instructional programs on teacher practice and student outcomes. The few experimental evaluations that have been conducted to date show mixed findings. Some find an impact on teacher practice, but no impact on student outcomes (Cordray, Pion, Brandt, Molefe, & Toby, 2012; Randel et al., 2011). Others find effects on achievement in specific grades or subjects, but

these vary in magnitude and even in sign (Carlson, Borman, & Robinson, 2011; Konstantopoulos, Miller, & van der Ploeg, 2013; Slavin, Cheung, Holmes, Madden, & Chamberlain, 2013; Konstantopoulos, Miller, van der Ploeg, & Li, 2014). Yet these studies lack detailed process data that would provide descriptive insights into the factors that might explain the variation in impacts.

The Achievement Network Program

During our evaluation period, the ANet program had four key components: (1) quarterly interim assessments in math and English language arts (ELA) for Grades 3 through 8 that are aligned to state content standards; (2) data tools including reports on individual students' progress; (3) coaching of school leaders to support their teachers' use of assessment data to improve instruction; and (4) a network of peer schools in their geographic area that shares results and engages in joint professional development. Through these components, the program aims to embed the use of interim assessment data into educators' everyday routines in an effort to identify and close gaps in student achievement.

Upon partnering with a district or school, ANet works to align the interim assessment content and administration with the curriculum and curricular scope and sequence. For each school, they set a schedule for assessment administration and regular coach visits. Coaches work closely with school leaders to build their capacity to support the implementation of data-based instructional practices by teachers. Coach visits are an integral component of the ANet program and occur at key points in the data cycle. For example, ANet coaches meet with school leaders or the data leadership team to help them plan and prepare prior to the administration of each assessment cycle. This includes preparing them to support teachers' review and analysis of data and decisions about instruction. Initially, the goal is for the ANet coach and school leader to be partners. However, as implementation proceeds, school leaders are expected to lead the data cycle in their school and ensure structures are in place to support it.

After each interim assessment is administered, ANet returns students' results to leaders and teachers in roughly two business days. It is at this point that the ANet coach meets with the school leader, data team, and teachers in data meetings. In the first year, data meetings typically bring together all of the ELA and math teachers of Grades 3 through 8. Meetings are often scheduled for a single three-hour block after the regular school day. Coaches or school leaders typically open the data meeting by calling attention to and celebrating positive student results.

The bulk of the meeting is then devoted to group professional development (e.g., how to analyze an assessment item) and assisting teachers in reviewing data on their own students. Specifically, teachers use the time to analyze their data and design lesson plans to address identified gaps in student knowledge.

Study Design, Sample, and Data Collection

Our evaluation sought to estimate the impact of ANet's data-based instructional program on student achievement on state math and reading tests after two years and to inform the program's development during a period of rapid growth. Because ANet is a schoolwide initiative, schools—not individual principals or teachers—were recruited into the evaluation sample.

Recruitment and design. The evaluation exploited a planned expansion of the program's services to serve up to 60 additional schools in five school districts: Boston, Chelsea, and Springfield (MA), Jefferson Parish (LA), and Chicago (IL). We worked with ANet to recruit schools in those districts that were willing to participate in the study in exchange for receiving subsidized ANet services immediately (i.e., treatment schools) or at the conclusion of the two-year implementation period (i.e., control schools). In order to assess their readiness to implement the ANet program, each school completed an ANet-developed screener survey. All schools that expressed interest in participating in the study were determined to meet a minimum level of readiness to implement ANet and therefore none were screened out. In total, 119 schools were recruited to participate in the expansion of ANet's data-based instructional program in two waves beginning either in the 2011–12 (Wave 1) or 2012–13 school year (Wave 2). All of the recruited schools served high proportions of students who were eligible for subsidized lunch and of students who were not performing at proficient levels on state math and reading assessments.

After schools were accepted into the study, we created matched pairs of schools within each of the five districts based on grade span and school-level measures of student demographics and prior achievement. One school within each pair was randomly assigned to the treatment group and received subsidized ANet services for two school years (2011–13 for Wave 1, 2012–14 for Wave 2). Control schools became eligible to receive ANet services in 2013–14 (Wave 1) or 2014–15 (Wave 2) (Table ES.1).

Table ES.1. School Year (SY) and Study Year, by Data Collection Wave

Data collection wave	SY 2010–11	SY 2011–12	SY 2012–13	SY 2013–14
Wave 1 (W1)	Recruitment	Year 1 (Y1)	Year 2 (Y2)	
Wave 2 (W2)		Recruitment	Year 1 (Y1)	Year 2 (Y2)

Note. Shaded cells represent the sample for which we report Year 2 findings.

No restrictions were placed on control schools' use of other interim assessment programs or other data-based interventions during the evaluation period. In fact, all control schools administered interim assessments in some grades and subjects, and educators received some type of data-related support from their district or school (e.g., instructional leaders, data strategists, master teachers, etc.). Our estimates of the impact of ANet on educator practice and student achievement therefore capture its effects over and above those of district-led efforts to administer and use interim assessments to inform instruction.

Sample. The initial sample of 119 schools across both waves was reduced due to the withdrawal and reorganization of schools in several districts. Several treatment schools closed or withdrew from the study prior to any interaction with ANet. These schools, and their matched pairs, are not included in analyses in this report. Ten additional treatment schools elected not to continue receiving ANet services in the second study year. Although we include these schools and their matched pairs in our main analyses of impacts on student outcomes, we also report results for the reduced Year 2 sample of schools that worked with ANet for both years. Our three main analytic samples are:

- **Year 2 full student impact sample**: 89 schools (45 treatment, 44 control) serving just over 21,000 students in Grades 3–8.
- **Year 2 reduced student impact sample**: 69 schools (35 treatment, 34 control) serving nearly 16,000 students in Grades 3–8. The reduced sample includes schools that actively worked with ANet for two years (and their matched pairs).
- **Year 2 survey impact sample**: 67 schools (34 treatment, 33 control). Neither the treatment schools that ended their partnership with ANet between Year 1 and Year 2 nor their matched pairs were surveyed in Year 2. This sample is identical to the Year 2 reduced student impact sample with the exception of one additional school (and its matched pair) that declined to take part in the surveys.

Evaluation data sources. To address questions of the impact of ANet on student achievement, we obtained student-level demographic, enrollment, and performance data for the 2010–11 through 2013–14 school years from the relevant state or district. To address questions about ANet's impact on educator beliefs and practices, we designed and administered school leader and teacher surveys. Surveys were administered at the end of the first and second years of the evaluation. School leaders were asked about the culture of their school and their attitudes towards data use; the presence of interim assessments, 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 available supports for instructional data use; school culture; and teacher background.

Analyses. The Year 2 student and survey impact analyses are based on cluster-adjusted ordinary least squares (OLS) regressions of survey scales or students' state test scores on an indicator of treatment assignment. All outcomes were standardized prior to analysis (within the respondent sample for leader and teacher survey scales, and by subject and grade within the relevant state for student achievement). The impact models also include school-pair dummy variables to account for the matched-pair randomization design and improve the precision of treatment estimates. The student impact models additionally include baseline achievement and demographics and fixed effects for each grade level. Missing baseline student achievement and demographics have been imputed.

Findings

Impacts on educator beliefs and practice. After two years, school leaders and teachers in ANet schools reported that their math and ELA interim assessments were more rigorous than reported by their counterparts in control schools. They also reported higher satisfaction with the timeliness and clarity of data and with the support they received for various data-based practices. School leaders and teachers in ANet schools reported reviewing student data more frequently, and teachers in ANet schools reported using student data more frequently to inform their instruction. However, teachers in ANet schools reported that their math interim assessments were less well aligned to their state's math content standards and state test and their school or district's math curriculum and curricular scope and sequence. Teachers in ANet schools were also no more likely to report differentiating instruction in response to students' needs.

Impacts on student achievement. After two years, we found no overall impact of ANet on student achievement in math or reading in either the full or reduced student impact samples (Table ES.2). In the full Year 2 sample, positive impacts were seen in math in Springfield. However, negative impacts were seen in eastern Massachusetts and Chicago in both math and reading.

Table ES.2. Year 2 Impacts of ANet on Students' Summative State Test Scores, by Subject and Sample

Subject and sample	Impact estimate	*SE*	*p* value	*n*
Math				
Full sample	–0.04	0.037	0.300	21,335
Reduced sample	0.00	0.044	0.978	15,806
Reading				
Full sample	–0.05	0.028	0.099	21,258
Reduced sample	–0.01	0.032	0.751	15,746

Note. Analyses were run on the full intention-to-treat sample of 89 schools (45 treatment and 44 control), as well as the reduced survey impact sample of 67 schools (34 treatment and 33 control). Models include fixed-effects for grade level and a set of paired school dummy variables. All models include students' baseline math test score and student demographics. Baseline test score is interacted with Year 2 grade level and Year 2 state test score in all models. An additional baseline test score imputation flag and a third- and fourth-grade flag were also included (and their interaction term). The Year 2 test score (outcome) was standardized by subject and grade using state means and standard deviations. Models are cluster-adjusted. Dummy variable imputation was used to replace missing baseline math test scores and demographics. Source: Student-level district administrative files from baseline (2010–11 or 2011–12), Year 1 (2011–12 or 2012–13), and Year 2 (2012–13 or 2013–14).
‡ $p < 0.01$; ** $p < 0.05$.

Exploratory analyses. Compared to their control-school counterparts, educators in ANet schools generally held more positive opinions of particular components of interim assessment programs and the support they received. They also reported reviewing and using data more frequently. However, this did not translate to positive impacts on student achievement in ANet schools. We explored several potential explanations for this unexpected pattern of findings.

***ANet theory of change*.** To explore the credibility of the ANet theory of change, we first asked whether schools that rated higher on various indicators of instructional data use elicited larger gains in student achievement over the two-year study period. With a few exceptions, we find that each of these measures is positively correlated with students' math and reading

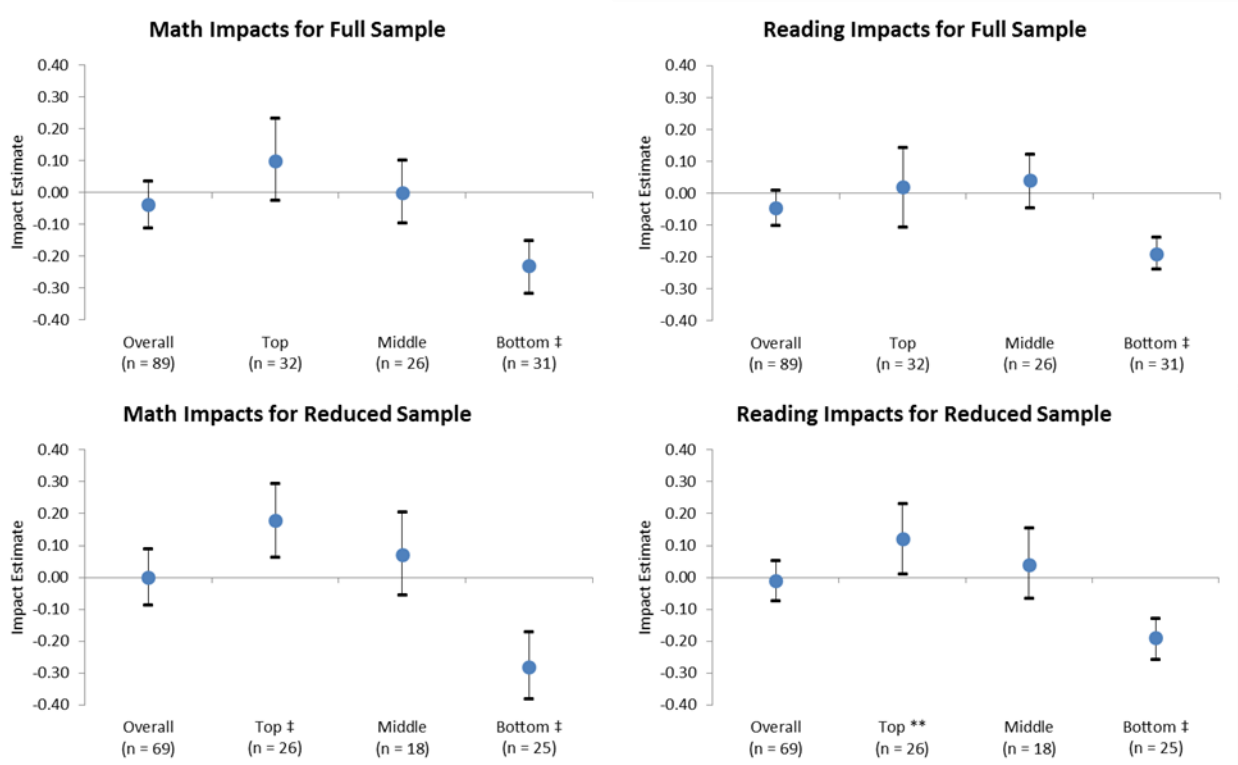
achievement. For example, in schools where teachers report more frequent use of interim assessment data, students had higher levels of achievement. In multivariate models, we see that, holding other scales constant, the frequency with which teachers use data to inform their instruction remains among the strongest predictors of student achievement in both subjects. These results cast doubt on the notion that the null impacts on student achievement we estimate stem from a flawed theory of change.

Teacher capacity and instructional flexibility. Within the same school, teachers generally share the same content standards, curriculum, and curricular scope and sequence. As a result, we might expect to see little within-school variation in the perceived alignment of interim assessments with standards and curricula. However, the vast majority of the variation in teacher-reported alignment of the interim assessments to the curriculum and the curricular scope and sequence was within rather than between schools. This suggests that perceived alignment is a proxy for something that varies from teacher to teacher.

Although teachers in ANet schools held poorer perceptions of the alignment of their math interim assessments, looking across all teachers we found that those who had more confidence fitting reteaching into the school or district's existing curricular scope and sequence and who more frequently used the content of upcoming interim assessments to plan instruction rated the alignment of their math interim assessments with the math curriculum and curricular scope and sequence more favorably. In contrast, those who felt that there was limited flexibility to adjust their instruction rated alignment less favorably.

School readiness. We also exploited the average scores on the school screener survey within matched pairs to divide the pairs into three "readiness" groups: top, middle, and bottom. Among schools in the highest readiness group, we found that the impact on student achievement in both math and reading after two years is positive but not statistically significant in the full sample and positive and statistically significant in the reduced sample (Figure ES.1). Among schools in the middle readiness group, there was no impact on student achievement. For the schools in the bottom readiness group, however, the impact on student achievement in math and reading after two years is negative and statistically significant in the full and reduced samples.

Figure ES.1. Year 2 Impacts on Math and Reading Test Scores, by Sample and Readiness Group



‡ $p < 0.01$; ** $p < 0.05$.

Conclusions

The lack of significant impacts of ANet on student achievement in the full evaluation sample is surprising given (1) emerging evidence that intensive data use is a distinguishing feature of many high-performing schools; (2) the fact that the program increased teachers' satisfaction with the available supports for data use and the extent to which they reported reviewing data and using it to inform their instruction; and (3) the fact that many indicators of program satisfaction and instructional data use are positively associated with schools' performance in raising student achievement within our study sample.

Our analysis suggests two potential explanations for these findings related to the capacity of participating teachers and schools. First, the program reduced teachers' perceptions of the extent to which their math interim assessments were aligned to the curriculum and the scope and sequence used in their schools. This result may reflect a genuine lack of alignment but appears to be mediated by teachers' capacity and the flexibility they have to align their instruction to the

content of interim assessments. Second, we find that negative impacts on student achievement were concentrated within a subset of treatment schools that were rated by program staff as having lower levels of readiness to engage in instructional data use. Among schools that received higher readiness ratings and participated in the program for two years, we estimate large positive treatment effects. In short, our results are consistent with the idea that intensive data use is an effective strategy for schools with the right structures and flexibilities in place to support teachers and where leadership is ready to prioritize the work.

This pattern suggests a choice facing ANet—and perhaps the broader field of data-based instructional programs—as it continues to refine its strategy for improving student achievement. First, ANet could narrow its focus to schools that are ready and able to implement an improvement strategy centered on data-based instruction. Our evidence indicates that the organization is able to identify in advance those settings in which it is most likely to improve student achievement, suggesting that exercising more discretion about the schools with which it chooses to partner is viable. Alternatively, ANet could explore making its services easier to be used productively by schools and teachers with less incoming capacity. The organization could also consider a hybrid approach, narrowing the range of schools that it serves with its existing program model while simultaneously experimenting with new models for working with schools with less initial readiness. This would allow it to continue to work with those schools in which it appears to be making a positive difference, while at the same time advancing the field's understanding of what is needed for a focus on instructional data use to translate into improvements in student achievement.

Finally, our results suggest that districts or schools considering an intensive data-based instructional program should work with providers to identify what is necessary for successful implementation. District and school leaders should also ensure that, if misalignment between interim assessments and curricula becomes a challenge, teachers are provided with support to address it through planning and the freedom to reorganize their instruction, so that the assessment data they receive is instructionally useful. As an evaluation of a specific program implemented in particular settings, our study cannot speak to the efficacy of data-based instructional programs generally. It seems likely, however, that the factors that appear to have generated positive results in a subset of schools within our evaluation sample would also increase the probability of success elsewhere.

1. INTRODUCTION

Data-based instructional programs are increasingly widespread in American schools. The theory of action behind these programs is that providing teachers with periodic assessment data and, in some cases, other supports (e.g., coaching on data use) will allow teachers to target instruction to areas where student mastery is lacking. The adoption of data-based instructional programs has been driven in part by evidence suggesting that these practices are a distinguishing characteristic of high-performing schools, particularly within the charter school sector (Angrist et al., 2013; Dobbie & Fryer 2013). Despite their popularity, however, there is limited research on the causal effects of data-based instructional programs on teacher and student outcomes. Only recently have quasi-experimental and experimental studies attempted to link the use of interim assessments and related supports with improved student achievement.[1]

In a 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. FAST-R consists of periodic, short reading assessments aligned to content standards and the state test, data coaching, and professional development aimed at helping teachers interpret and use the assessment results. The study produced mixed results that mostly were not statistically significant. Quint et al. hypothesized that the program's training and coaching element was not intensive enough, nor was it sufficiently different from professional development in the comparison schools, to have a measurable impact on teaching and learning.

Henderson, Petrosino, Guckenburg, and Hamilton (2007, 2008) used a similar non-experimental method to estimate the impact of quarterly benchmark math assessments on math achievement in a study that matched 22 treatment with 44 comparison middle schools. Their results showed positive but not statistically significant differences in student achievement after one and two years.

Studies with stronger empirical designs have also produced mixed findings. Some find an impact on some indicators of teacher practice related to data use but no impact on student outcomes. For example, Cordray et al. (2012) conducted a cluster-randomized evaluation with grade-level (Grade 4 or 5) assignment of the Northwest Evaluation Association's (NWEA)

[1] Portions of this literature review, as well as later discussions of the evaluation design and study, are adopted from Morton (2015).

Measures of Academic Progress (MAP) benchmark assessment program in Illinois. The MAP program consists of a series of computer-adaptive interim assessments, online instructional resources for educators, and on-site and on-demand training throughout the year. No evidence of an impact on student achievement in reading was found, as measured by the state test or the MAP composite score. 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-report surveys did suggest a positive impact on the extent of differentiation in Grade 5.

The Classroom Assessment for Student Learning (CASL) program is 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). Using a school-randomized design, Randel et al. (2011) found that the program improved teachers' knowledge of classroom assessment. However, they found no detectable effects on the quality of teachers' classroom assessment practices or on students' math achievement after two years.

The Using Data program, developed by TERC, is a professional development and technical assistance program that aims to help teachers use data in collaboration with peers to address students' learning needs (Cavalluzzo et al., 2014). A block-randomized evaluation of Using Data found a positive impact after one year on the frequency with which teachers used data as well as on their data literacy and attitudes about the value of data for improving instruction. Despite this, there was no detectable difference in overall math achievement between treatment- and control-school students after two years.[2]

Other studies find effects on student achievement in specific grades or subjects, but these vary in magnitude and even in sign. Carlson et al. (2011) examined the impact of a district-wide intervention 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. The study included 549 schools in 59 districts across seven states. Districts were randomly assigned to treatment and control groups within each state. The authors found small but statistically

[2] However, students in the lowest-performing block of schools (at baseline) did score higher than their counterparts in control schools after two years (Cavalluzzo et al., 2014).

significant positive effects on math achievement after the first year of implementation. Estimates of the program's impact on reading achievement were positive but not statistically significant.[3]

Konstantopoulos et al. (2013) examined the impact of Indiana's system of benchmark assessments: mCLASS and Acuity. Teachers in Grades K through 2 administer periodic mCLASS assessments in the form of face-to-face language tasks or short (one-minute) probes. Acuity provides online multiple-choice assessments in reading, math, science, and social studies for Grades 3 through 8. The system also provides teachers with item banks to construct on-demand assessments and access to instructional tools. A school-randomized trial found a significant positive effect on average math and reading achievement in Grades 3 through 8 but not in Grades K–2. The impact was largest in fifth- and sixth-grade math, where results showed impacts greater than one-quarter of a standard deviation, and in third- and fourth-grade reading, where results showed impacts of about one-seventh of a standard deviation (Konstantopoulos et al., 2013).

Finally, Konstantopoulos et al. (2014) conducted a second school-randomized trial of mCLASS and Acuity in a separate sample of schools as a replication of their prior study. The results for this sample showed no overall impact on K–8 achievement in either math or reading. However, a negative impact was found on student achievement in math and reading in Grades K–2. Based on combined estimates from their initial study and its replication, the authors conclude that 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).

In sum, some evaluations of data-based instructional programs find impacts on teacher, but not student outcomes. Others find impacts on student achievement in specific grades and subjects, but these impacts are not consistently positive. Unfortunately, those studies that do find impacts on student achievement provide very limited information on the changes in educator practice that may be responsible.

A key challenge in interpreting research on data-based instructional programs is that such programs typically comprise at least three distinct elements, each of which could plausibly impact student achievement. First is the administration of interim assessments, which could

[3] A follow-up study focused on fifth- and eighth-grade reading and math scores conducted two years after the two-year experiment concluded showed some signs of long-term impacts of CDDRE. The estimated impacts on math and reading achievement after four years were relatively large but not always statistically significant due to smaller sample sizes (Slavin et al., 2013).

improve student achievement directly as a result of practice effects. Second is the interim assessment data that is reported to teachers, which may on its own enable teachers to refocus their instruction in productive ways. Third is whatever coaching or professional development is provided to teachers in order to support them in identifying gaps in student learning and modifying their instruction in response. Although data-based instructional programs by definition include the administration of interim assessments, they vary in how they report the data from those assessments back to teachers and especially in the extent of professional development they offer. As important, the schools that serve as the comparison group in the evaluation of a given data-based instructional program may themselves have one or more of these elements in place. For example, the practice of administering regular interim assessments is now ubiquitous in American schools, regardless of whether the school is engaged in a focused attempt to promote instructional data use.

In this report, we present the results of a two-year evaluation of a data-based instructional program known as Achievement Network. Designed as a matched-paid, school-randomized trial, the study adds to the small body of experimental evidence on the impact of data-based instructional programs on educator practice and student achievement. As described below, the Achievement Network combines interim assessments and data reports with intensive coaching of school leaders designed to enhance their practices around instructional data use. However, no restrictions were placed on control schools' practices around data use, and all schools in the evaluation sample administered interim assessments. Our evaluation therefore estimates the impact of participation in Achievement Network over and above alternative approaches to interim assessment and support for data use.

Achievement Network

Achievement Network (ANet) was founded in 2005 as a school-level intervention based on the belief that if teachers are provided with timely data on student performance from interim assessments tied to state standards, if school leaders provide support and create structures that help them use that data to identify student weaknesses, and if teachers have knowledge of how to improve the performance of students who are falling behind, then they will become more effective at identifying and closing gaps in student learning. This will, in turn, improve student performance in their school, particularly for high-need students.

ANet was initially developed to serve the Boston charter school sector but has since expanded to serve over 500 schools in nine geographic networks across the United States. Most of its current partnerships are with schools in traditional school districts. During the period of our evaluation, the ANet program had four key components: (1) quarterly interim assessments in math and English language arts (ELA) for Grades 3 through 8 that are aligned to state content standards; (2) data tools including reports on individual students' progress delivered through an online platform called MyANet; (3) coaching of school leaders to support their teachers' use of assessment data to improve instruction; and (4) a network of peer schools in their geographic area that shares results and engages in joint professional development.

Through these components, the program aims to embed the use of interim assessment data into educators' everyday routines as a way to identify and close gaps in student achievement. Much of ANet's work with schools is organized around what in 2011 it called the data cycle. Prior to the school year, ANet works with individual schools or their district to align the content and administration of the ANet interim assessments with the existing curriculum and curricular scope and sequence. They schedule assessment administration and coach visits around what they consider to be the key leverage point in the data cycle: data review and action planning meetings. In addition, ANet coaches work closely with school leaders throughout the year to build their capacity to support teachers' implementation of data-based instructional practices. Specifically, ANet coaches meet with school leaders and a data leadership team to help them plan prior to the administration of each cycle of interim assessments. This includes preparing them to support teachers' review and analysis of data and decisions about instruction. Initially, the ANet coach and school leader work as partners. As the program becomes more embedded, however, school leaders are expected to take ownership of the data cycle in their school.

After each interim assessment is administered, ANet returns students' results to leaders and teachers within two business days and the ANet coach meets with the school leader, data team, and teachers in data meetings. In the first year, data meetings typically bring together all teachers who instruct students in ELA and math in Grades 3–8. Meetings are scheduled for a block—typically three hours—after the regular school day. In these meetings, coaches can play a variety of roles as agreed upon with the school leadership team, depending on the leadership's own capacity level: They may open by celebrating signs of student progress in specific areas, offer group professional development focused on data analysis strategies, and assist teachers in reviewing their students' data. The goal is for school leaders to support teachers in identifying

gaps in their students' learning and address them through targeted re-teaching plans, which teachers are expected to develop at the data meetings. The coach is again present to support this work, but the intention is that ultimately school leaders and data leadership team members will lead these efforts.

Below, we detail the program components and intermediate outcomes as well as provide a logic model illustrating how they are hypothesized to interact to produce improvements in student achievement. As a relatively young organization, ANet continues to refine its program logic model in response to contextual changes and its experiences with partner schools. The logic model presented in Exhibit 1.1 reflects the organization's approach at the start of the 2012–13 school year, the second year of implementation for most of the schools participating in this evaluation. The most notable change reflected in this logic model as compared with the first year of implementation is the inclusion of "school structures" as both a responsibility of leaders to put in place and a key mediator of student outcomes; although the tasks reflected under this heading had always been an implicit part of the program, they had not been formally incorporated into the organization's logic model.

Since that time, ANet has made additional changes to its program motivated in part by the adoption of the Common Core State Standards in the states in which it works and the resulting increase in expectations for student achievement. In particular, it has expanded the scope of support provide by its coaches to include assisting teachers to align their instruction to the skills covered by the next assessment cycle and develop quizzes to assess student mastery after re-teaching. To reflect this change, it renamed the "data cycle" the "teaching and learning cycle." ANet now also provides direct support to districts to coordinate the alignment of assessments, standards, and curricula.

Intervention inputs. Quarterly interim assessments are the core source of data to inform teachers' instructional decision making. ANet provides quick turnaround of student assessment results through an online platform. School leaders and teachers can use this platform to generate reports at the school, grade, classroom, and student levels and to easily compare their results to those of other schools in their geographic network. The online platform also provides teachers with other resources for instructional planning such as data analysis and re-teaching plan templates, schedules of standards to be covered on future interim assessments, curriculum guides, and tools for diagnosing student misconceptions. Each school is assigned an ANet coach

who visits the school roughly 20 times each academic year, usually for the first two to three years of partnership. Coaches primarily work to build school leaders' capacity to support instructional data use in their school. School leaders also have access to a network of other leaders in their geographic network. Two annual meetings allow leaders from across the network to come together to share practices with their peers.

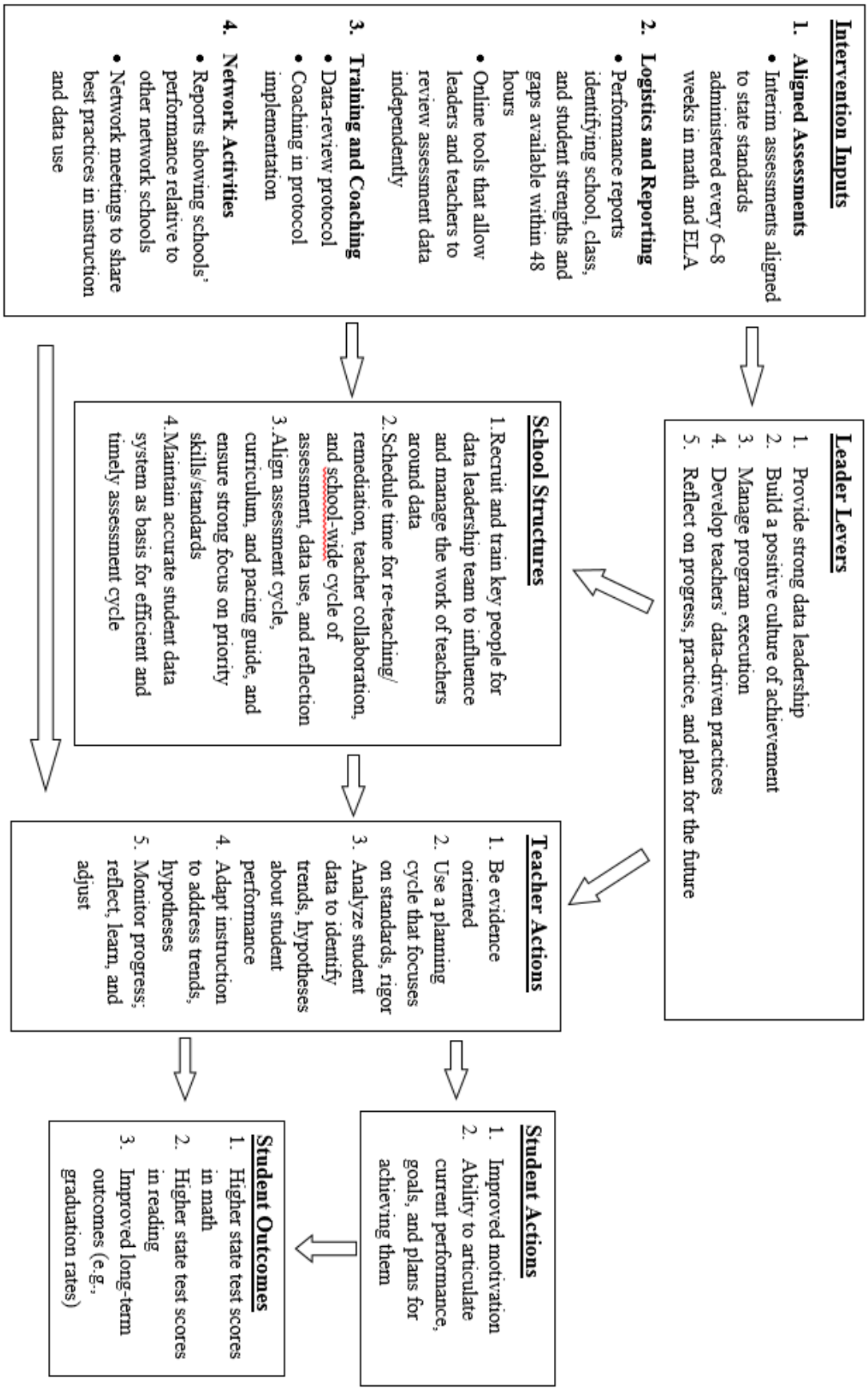
Leader actions. The program model focuses on strengthening leaders' engagement with interim assessment data and their prioritization of the use of data to inform instruction. Leaders are expected to build a school-wide culture of achievement, manage the implementation of the data cycle, develop teachers' skills and practices around instructional data use, and reflect on student and school progress, taking action where needed.

School structures. School leaders must also establish a data leadership team, typically including themselves, an assistant principal, and grade-level or content-area instructional leaders. With help from the ANet coach, school leaders provide the data leadership team with support in engaging teachers in data-based practices. School schedules are arranged to include time for activities such as data meetings, re-teaching and remediation, and key logistics of the data cycle. Leaders also ensure that the assessment cycle is aligned to curricular resources and planning. Finally, they must maintain a student data system that supports the implementation of the data cycle.

Teacher actions. Teachers are expected to use techniques such as backward planning to align their instruction to state content and performance standards. The goal is to develop a planning cycle that focuses on alignment and rigor. Teachers are also expected to analyze assessment results to determine students' progress toward standards and use this information to develop and implement action plans that address gaps in students' knowledge. Once they have implemented these action plans, teachers are expected to assess and reflect on their success.

Student actions and outcomes. The primary outcome of interest for the evaluation is student achievement as measured by state test scores in math and reading. However, through the sharing of interim assessment results with students, 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. It is also expected that short-term impacts on student achievement will translate into improvements in longer-term student outcomes, such as high school graduation and postsecondary success.

Exhibit 1.1. ANet Program Logic Model



The Evaluation

Achievement Network's expansion and external evaluation were funded by a 2010 grant from the federal Department of Education's Investing in Innovation (i3) program. The evaluation was designed to estimate the impact of the Achievement Network program on student achievement in schools served through its i3 expansion as well as to inform the program's development during a period of rapid growth. After discussing the evaluation design, sample, and data collection in section two, we provide final results from our two-year experimental evaluation of ANet's data-based instructional program (the "program") in five school districts across the United States. We first examine implementation fidelity in section three. In section four, we report on the impacts of ANet on educators' beliefs and practices around instructional data use as measured by surveys. Section five reports the impacts of ANet on student achievement in math and reading. Section six reports on exploratory analyses examining the roles of school readiness (to partner with ANet) and teacher capacity (to implement data-based instructional strategies) in explaining variation in program impacts within the evaluation sample. We close with a discussion of the results and their implications.

2. EVALUTION DESIGN, METHODS, AND DATA

In the fall of 2010, we partnered with Achievement Network (ANet) to design a matched-pair, school-randomized evaluation of their program's effectiveness in raising student achievement in participating schools. Because ANet is a school-wide initiative, schools—not individual principals or teachers—were recruited into the evaluation sample. The evaluation exploited a planned expansion of the program to serve up to 60 additional elementary, middle, or K–8 schools in five school districts with which ANet had pre-existing relationships: Boston, MA; Chelsea, MA; Springfield, MA; Jefferson Parish, LA; and Chicago, IL. All of the schools we recruited served high proportions of students eligible for subsidized lunch and students who were not performing at proficient levels on state math and reading assessments. In this section, we detail the recruitment efforts, resulting analysis samples, and estimation strategies.

School Recruitment

With memoranda of understanding in place at the district level, we worked with ANet to recruit schools that were willing to participate in the study in exchange for receiving subsidized ANet services immediately (i.e., treatment schools) or at the conclusion of the two-year implementation period (i.e., control schools). In order to assess their willingness and readiness to implement the ANet program, each school completed an ANet-developed screener survey. The screener survey included nine criteria related to the extent to which school leaders prioritized data use, the presence of or willingness to create a dedicated data leadership team, a standards-based curriculum and curricular scope and sequence, and dedicated teacher collaboration time. After schools completed the screener survey, ANet staff used a rubric to assign each school a score of one to three on each rating criterion (see Exhibit A.1 in Appendix A). All schools that expressed interest in participating in the study were determined to be at least minimally ready to implement ANet; none were screened out.

We set a goal of recruiting 120 schools in order to provide sufficient statistical power to detect a small effect of the program on student achievement. In total, 101 schools were recruited to participate in the expansion of ANet's data-based instructional program beginning in the 2011–12 school year. Because the initial recruitment efforts fell short of our target, we worked with ANet to recruit a second wave of schools in Jefferson Parish and Springfield with the primary purpose of improving statistical power (n = 18). Treatment schools in this second wave began receiving services one year later (2012–13) than the schools in the initial, Wave 1 sample.

All aspects of the design that were applied to the first wave of schools were applied to the second wave, including recruitment, screening, randomization, data collection, and data analysis. The only exception is that no site visits were made. Table 2.1 details the relationship between school years and the timing of study activities for each data collection wave.

Table 2.1. School Year (SY) and Study Year, by Data Collection Wave

Data collection wave	SY 2010–11	SY 2011–12	SY 2012–13	SY 2013–14
Wave 1 (W1)	Recruitment	Year 1 (Y1)	Year 2 (Y2)	
Wave 2 (W2)		Recruitment	Year 1 (Y1)	Year 2 (Y2)

Note. Shaded cells represent the sample for which we report Year 2 findings.

These recruitment and screening procedures may impact the generalizability of the evaluation findings. Because the evaluation sample consists of schools that were deemed ready to implement the program and were willing to commit a share of the school's discretionary budget to doing so, our findings generalize only to similarly equipped and motivated schools. At the same time, conversations with program staff suggest that the sample of i3 schools in the evaluation may differ systematically from the high-performing charter schools in which the program was initially developed and from most of the district schools in its portfolio. In particular, the rapid expansion of the program may have resulted in a sample of schools with lower levels of readiness to engage in data-based instructional reform, on average, than the schools with which ANet was accustomed to working. This may have implications for inferences about ANet's impact on schools with greater readiness, an issue we return to when discussing variation in treatment impacts in section six.

Evaluation Design and Random Assignment

Once schools were accepted into the study, we created matched pairs of schools within each of the five districts based on grade span and school-level variables likely to be correlated with the outcomes of interest: prior student proficiency rates in each subject, enrollment in Grades 3–8, percentage enrollment by race/ethnicity, and the percentages of students who are eligible for free or reduced-priced lunch, who are English language learners (ELL), and who have been identified for special education services. Schools within each pair were randomly assigned to treatment or control status. This matched-pair, school-randomized design ensures unbiased estimates of the treatment effect along with an improvement in the power to detect an

effect (Imai, King, & Nall, 2009). Matching and randomization were conducted in spring 2011 for Wave 1 and spring 2012 for Wave 2 using the blockTools package for the R statistical software program (Moore & Schnakenberg, 2011; Moore, 2012).

Treatment schools received subsidized ANet services for two school years (2011–12 and 2012–13 for Wave 1, 2012–13 and 2013–14 for Wave 2). Control schools became eligible to receive ANet services in 2013–14 (Wave 1) or 2014–15 (Wave 2). However, no restrictions were placed on their use of other interim assessment programs or other interventions during the evaluation period. In fact, district requirements meant that all control schools administered interim assessments in some grades and subjects. In most cases, teachers in control schools also received some type of data-related support from their district- or school-based coaches (i.e., instructional leaders, data strategists, master teachers, etc.). This implies that our estimates of ANet's impact on educator practice and student achievement capture its effects over and above those of district-led efforts to administer and use interim assessments to inform instruction, potentially making the detection of a treatment effect more difficult.

School Samples

The full sample of 119 schools across both waves was reduced due to the withdrawal and reorganization of schools in several districts. Several schools assigned to the treatment group closed or withdrew from the study prior to any interaction with ANet. We excluded these schools, and their matched pairs, from the evaluation sample. In addition, 10 treatment schools elected not to continue receiving ANet services in the second study year. Although we include these schools and their matched pairs in our main student impact analyses, we also report student impact results for the reduced sample of treatment schools that worked with ANet for both years. Details of the analysis samples used to estimate impacts on student achievement after two years of implementation, as well as the sample used to estimate impacts on educator practice in Year 2, are described below.

Year 2 full student (ITT) impact sample. In total, 28 schools withdrew, closed, or were dropped from the study after randomization, but prior to any implementation of the intervention, and are excluded from our student impact sample. Shortly after recruitment and randomization had been completed, Chicago Public Schools experienced a leadership transition and an internal reorganization. Because schools lost control over the discretionary budgets they had planned to use to pay for program services, only 10 of the original 18 schools that had been randomly

assigned to the treatment condition were able to remain in the study. The other eight treatment schools and their matched pairs were dropped from the student impact sample. Two schools in Boston and one school in Springfield also withdrew from the study prior to implementation. These schools and their pairs were dropped from the student impact sample. In Jefferson Parish, one school was excluded from the student impact sample because it was an alternative school serving a unique population of at-risk students, a second school dropped out of the study, and a third closed. These schools and their pairs were excluded from the student impact sample.

Between Year 1 and Year 2, one additional school in Jefferson Parish closed. Its pair was also dropped from the Year 2 student impact sample. The result after combining both recruitment waves is a Year 2 student impact sample that includes 89 schools serving students in Grades 3–8 (45 treatment and 44 control schools).[4]

Year 2 reduced student impact sample. Of the treatment schools in the Year 2 full student impact sample, 10 schools declined to continue their program participation in Year 2. For some analyses, we therefore present results from a "reduced" Year 2 student impact sample consisting of the sample of schools that worked with ANet for two years and their matched pairs (n = 69) (35 treatment and 34 control schools).

Year 2 survey impact sample. The schools that dropped out of the study after the first year could not be surveyed in Year 2; one additional control school also opted out of survey administration. We therefore exclude these schools and their matched pairs from our analysis of survey data. As a result, our Year 2 survey impact sample consists of 67 schools (34 treatment and 33 control). All analyses that link survey data with data on student achievement are based on this subset of the full evaluation sample.

Because both reduced samples exclude schools that decided not to work with ANet for a second year, results based on them are generalizable only to schools that would remain in the program for two years when provided the opportunity to do so. Due to the matched-pair structure 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 relevant outcomes (independent of treatment) within school pairs.

[4] The uneven number of schools is due to a "pair" in one district containing three schools, two of which were assigned to treatment status.

Note also that the Year 2 survey and student impact samples both include "joiners," or teachers and students who entered the schools during the evaluation period. This means that some participants in the Year 2 treatment school sample may have only experienced a single year of ANet exposure. Our results therefore capture cluster-level, not individual-level impacts (i.e., the impact of ANet on all teachers within a given school after two years, as opposed to the impact of two years of ANet participation on specific teachers).

Baseline equivalence. Table 2.2, reports on the baseline equivalence of student performance, demographics, and school readiness for schools in the final student impact sample (N = 89). Results for student performance in each subject are broken out by geographic network and grade level. For each row, we report the mean, student sample size, and proportion of missing data by group assignment. The regression adjusted treatment-control difference for each variable (row) is reported along with its level of statistical significance. For example, the first row shows the treatment-control difference in the standardized reading test scores for students in Grades 3–8. The mean for the control group is –0.48 standard deviations; the mean for the treatment group is –0.56 standard deviations; and the regression-adjusted difference is –0.04 standard deviations. While this difference is marginally statistically significant, it is quite small. In total, six comparisons of the baseline differences between treatment and control school achievement and demographics are statistically significant. However, the Year 2 analytic models include these baseline measures and therefore meet the baseline equivalence standards established by the What Works Clearinghouse (2014).[5]

[5] Our conclusions regarding baseline equivalence are based on What Works Clearinghouse (2014, p. 15), which states, "If the reported difference of any baseline characteristic is greater than 0.25 standard deviations in absolute value (based on the variation of that characteristic in the pooled sample), the intervention and comparison groups are judged to be not equivalent. . . . For differences in baseline characteristics that are between 0.05 and 0.25 standard deviations, the analysis must include a statistical adjustment for the baseline characteristics to meet the baseline equivalence requirement. Differences of less than or equal to 0.05 require no statistical adjustment."

Table 2.2. Baseline Student Achievement and Demographics for Schools in Year 2 Impact Sample, by Treatment Assignment

	Treatment group			Control group			Treatment-control difference		
	Mean	*n*	Sample with missing data (%)	Mean	*n*	Sample with missing data (%)	Difference	*SE*	*p* value
Baseline ELA score	–0.56	11,628	14.4	–0.48	9,841	10.8	–0.04	0.026	0.096
Network									
Eastern Massachusetts	–0.78	4,724	21.8	–0.69	3,073	14.5	–0.04	0.048	0.418
Springfield, MA	–0.48	982	17.6	–0.55	900	18.7	0.09	0.061	0.151
Chicago, IL	–0.71	2,652	12.2	–0.57	2,766	9.3	–0.16 ‡	0.031	0.000
Jefferson Parish, LA	–0.15	3,270	4.7	–0.16	3,102	6.2	0.03	0.050	0.541
Grade level									
Grades 3–5	–0.50	7,298	12.8	–0.44	5,910	10.1	–0.05	0.033	0.158
Grades 6–8	–0.66	4,330	17.2	–0.53	3,931	11.9	–0.05	0.038	0.165
Baseline math score	–0.52	11,678	14.0	–0.47	9,864	10.5	–0.04	0.031	0.226
Network									
Eastern Massachusetts	–0.69	4,769	20.6	–0.58	3,094	13.7	–0.08	0.059	0.184
Springfield, MA	–0.45	988	16.9	–0.50	904	18.1	0.08	0.084	0.383
Chicago, IL	–0.78	2,652	12.2	–0.65	2,763	9.4	–0.15 ‡	0.051	0.008
Jefferson Parish, LA	–0.10	3,269	4.8	–0.19	3,103	6.2	0.07	0.050	0.146
Grade Level									
Grades 3–5	–0.46	7,346	12.1	–0.42	5,936	9.6	–0.03	0.036	0.355
Grades 6–8	–0.63	4,332	17.1	–0.53	3,928	12.0	–0.10	0.049	0.058
School Readiness Rating	0.07	13,308	0.0	–0.09	10,734	1.6	0.27	0.176	0.132
Demographics									
Gender	0.52	13,286	0.2	0.51	10,881	0.2	0.01	0.007	0.246
Race/ethnicity	0.87	13,286	0.2	0.87	10,881	0.2	0.00	0.013	0.869
FRPL status	0.85	13,308	0.0	0.88	10,904	0.0	–0.02	0.009	0.085
IEP status	0.17	13,308	0.0	0.16	10,904	0.0	–0.01	0.008	0.529
ELL status	0.18	13,308	0.0	0.12	10,904	0.0	0.03	0.015	0.055

Note. The baseline school sample includes the 89 schools that remained in the impact sample in Year 2. Group means are unadjusted. The proportion of missing data is based on the total number of students in Grades 3 through 8 at baseline. The treatment–control difference represents the coefficient on treatment assignment when the baseline covariate is regressed on treatment assignment. Models are cluster-adjusted and include only grade-level and school-pair fixed effects. FRPL = free and reduced price lunch; IEP = individual education plan; ELL = English language learner. Source: Baseline district administrative data files from 2010–11 (Wave 1 schools) and 2011–12 (Wave 2 schools).

‡ $p < 0.01$; ** $p < 0.05$.

Data Collection and Samples

Student data. To study ANet's impact on student achievement, we obtained individual-level student demographic, enrollment, and performance data for the 2010–11 through 2013–14 school years from the relevant district (Jefferson Parish Public School System and Chicago Public Schools) or state (Massachusetts). We then assembled a single, longitudinal, student-level data file that allowed for cross-site analyses.

Student sample. Within each school in the samples described above, all students in Grades 3 through 8 with either a nonmissing math or reading test score are included in the Year 2 analyses. We describe our procedures for handling missing baseline demographic and performance variables below.

Surveys. To study the impact of ANet on educator beliefs and practices, we designed and administered school leader and teacher surveys.[6] School leaders were asked about the culture of their school and their attitudes towards data use; the presence and implementation of interim assessments and other elements of data-based instructional programs; and general information on school leadership. Teacher surveys focused on attitudes towards, and use of, data in the classroom; awareness and understanding of and satisfaction with data-based instructional programs and their implementation; their school's culture; and their background. A set of ANet-specific items was directed to the treatment group respondents to measure implementation fidelity as reported by educators and their satisfaction with specific ANet program components.

Survey sample. We aimed to survey the universe of eligible educators in all treatment and control schools after one and two years. In this report, we limit our analysis to the Year 2 sample and results.

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, the interim or assistant principal was surveyed instead. When estimating Year 2 impacts on school leaders, we use the "complete-pair" sample for the relevant item—that is, the sample of leaders for whom we have both leaders' survey responses within the matched pair. In Year 2, responses were received from 60 school leaders of the 67 survey impact sample schools (Table 2.3).

[6] Our Year 2 survey instruments are available at http://cepr.harvard.edu/achievement-network-evaluation-instruments

Table 2.3. Year 2 School Leader Survey Response Rates, Overall and by Treatment Assignment

	Year 2		
	Overall (%)	Treatment group (%)	Control group (%)
Overall response rate	89.6	97.1	81.8
Network response rate			
Eastern Massachusetts	92.0	100.0	83.3
Springfield, MA	91.7	83.3	100.0
Chicago, IL	66.7	100.0	33.3
Jefferson Parish, LA	91.7	100.0	83.3

Note. Leaders were considered a survey respondent if they were from a Year 2 impact sample school (*n* = 67), gave consent, and responded to at least some portion of the survey beyond the consent items. All schools leaders who responded are included in the response rate calculations. Source: Year 2 school leader surveys (treatment and control).

Teachers. All consenting teachers in a survey impact sample school were included if they reported a math, English language arts (including reading), or general elementary assignment in one or more of Grades 3 through 8. The overall Year 2 teacher survey response rate was 78%, including 70% of teachers in control schools and 85% in treatment schools (Table 2.4). When broken out by geographic network, the combined response rates ranged from a high of 88% in western Massachusetts to a low of 65% in Chicago. [7]

The Year 2 sample of 616 teachers consists of all eligible teachers in the 67 survey impact sample schools who completed a survey during the second year of the study, including teachers who moved to an eligible assignment in Year 2. The Year 2 sample includes at least one teacher from each of the 67 Year 2 survey sample schools. Overall, the number of eligible respondents per school ranges from one to 23 teachers, with an average of 9.2.

[7] Because the sampling frame did not contain complete teaching assignment information and because a portion of the survey respondents were found to be out of scope, these response rates use the known proportion of out-of-scope *respondents* in each geographic network to remove the unknown proportion of out-of-scope *nonrespondents* from the denominator of teacher response rate calculations. These adjusted response rates should be closer to the true in-scope teacher response rate than if the total number of nonrespondents were used.

Table 2.4. Year 2 Teacher Survey Response Rates, Overall and by Treatment Assignment

	Year 2		
	Overall (%)	Treatment group (%)	Control group (%)
Overall response rate	78.0	85.0	70.0
Network response rate			
Eastern Massachusetts	81.2	79.1	84.4
Springfield, MA	88.0	96.0	80.6
Chicago, IL	65.2	82.2	47.5
Jefferson Parish, LA	72.8	89.9	52.0

Note. Teachers were considered a survey respondent if they were from a Year 2 impact sample school (*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 ELA, math, or general elementary in at least one of Grades 3–8.) The denominator is adjusted to account for the likelihood that some nonrespondents were not in scope. Source: Year 2 teacher surveys (treatment and control).

Analytic Methods and Models

Estimates of the impacts of ANet on educator practice and student achievement after two years of participation were generated from a series of regressions of (1) Year 2 survey scales and (2) Year 2 math and reading test scores on a variable indicating treatment assignment.

Survey impact models. The Year 2 survey impact analyses are based on cluster-adjusted ordinary least squares (OLS) regressions of survey variables or scales on treatment assignment. Models were cluster-adjusted to account for the nesting of teachers within schools. The models include school-pair dummy variables to account for the matched-pair randomization design and improve the precision of treatment estimates.

Student impact models. Student impact models include all students with either a non-missing Year 2 reading or math test score. Because the student analyses include data from three states, test scores were standardized by grade and subject using statewide means and standard deviations. The student impact models estimate program impacts on reading and math achievement for all students who were enrolled in Grades 3–8 in a Year 2 impact sample school when state tests were administered (N = 89). The models are specified as follows:

$$Y_{igkdt} = \beta_1 Treat_{kdt} + \delta \boldsymbol{Ach}_{kdt-2} + \gamma \boldsymbol{X}_{igkdt-2} + \theta_g + \alpha_s + \varepsilon_{igkdt} \quad \text{(Eq. 1)}$$

where Y_{igkdt} represents the reading or math achievement of student *i* in grade *g*, school *k*, and district *d* at time *t*. The values of Y_{igkdt} have been standardized within state, grade, and subject to allow for comparability across state tests. $\boldsymbol{Ach}_{kdt-2}$ represents a vector of mean baseline achievement in math and reading in Grades 3–8,[8] and $\boldsymbol{X}_{igkdt-2}$ is a vector of student-level covariates measured at baseline (e.g., gender, race/ethnicity, special education, English language learner, grade progression). $\theta_g$and α_s represent fixed effects for grade and school pair. $Treat_{kdt}$ is an indicator of whether the school the student attends was assigned to receive the treatment.[9] β_1 therefore provides the intent-to-treat estimate of the causal effect of the intervention. Finally, ε_{igkdt} represents a standard mean-zero error term adjusted to account for the clustering of observations within schools. All models were cluster-adjusted to account for the nesting of students within schools.[10]

In addition to estimating ANet's impact on student achievement for the full sample, we also estimated program impacts for several student subgroups. More specifically, contrasts were run by geographic network and grade level (3–5, 6–8). Variations of the model in Equation 1 described below were also run to explore impacts by schools grouped according to the score they were assigned at baseline by ANet screeners.[11]

Because the student impact analyses employ district administrative data, there is little missing data in student demographic covariates; gender, race, free or reduced-price lunch status, and IEP or ELL status. The bottom panel of Table 2.5 shows the percentage of missing data for these variables, by analysis sample (rows) and treatment assignment (columns). Where there is missing data on baseline student-level demographic control variables, we created dummy variables identifying students for whom the relevant variable was missing.

---

[8] $\boldsymbol{Ach}_{kdt-2}$ includes a three-way interaction term between base-year test score, Year 2 grade level, and Year 2 state test to allow the relationship between baseline and Year 2 test scores to vary across grades and states. It also includes a two-way interaction term between the base-year test score imputation flag and a flag for students in Grades 3 and 4 to account for the impact of imputation.

[9] ANet decided that maintaining its partnership with one district required that it provide treatment to two schools that had initially been assigned to the control condition. The analyses presented here ignore this crossover between control and treatment conditions and therefore represent intent-to-treat models. However, all substantive conclusions regarding program effects are robust to the use of instrumental variable techniques to recover treatment-on-treated effects.

[10] All results reported below are qualitatively unchanged when estimated using various other model specifications, such as substituting network for school-pair fixed-effects, excluding baseline covariates, and using instrumental variable regression to account for treatment crossover.

[11] Because the two confirmatory impact analyses represent program impacts in different domains (math and ELA), and because all other reported contrasts are exploratory in nature, we do not adjust for multiple hypothesis testing.

Missing baseline test scores are more common, however (top two panels of Table 2.5). For example, in the overall model measuring impacts of ANet on student reading achievement after two years, 47% of the students in control schools and 53% of the students in treatment schools were missing a baseline measure of reading achievement. The high rate of missing baseline achievement data primarily reflects the inclusion of students in third and fourth grade, who were not in tested grades prior to the evaluation period. Students with missing baseline test scores were assigned a score of zero and a separate dummy variable identifying these students was created.[12] All student impact models reported below are based on this approach, but the results are qualitatively similar if we limit the analysis to students for whom baseline test scores are available.

[12] This method, called dummy variable imputation, has been shown to introduce an acceptably small amount of bias when baseline covariates are missing for some students within schools and when inferences are focused on the treatment indicator (Puma, Olsen, Bell, & Price, 2009).

Table 2.5. Proportion of Missing Baseline Data for Year 2 Analysis Models, by Analysis Sample and Treatment Assignment

	Treatment group (%)	Control group (%)
Missing ELA scores (overall)	53.1	47.1
Network		
Eastern Massachusetts	48.4	42.7
Springfield, MA	72.6	71.9
Chicago, IL	37.6	37.7
Jefferson Parish, LA	64.5	51.6
Grade level		
Grades 3–5	71.7	70.8
Grades 6–8	19.0	13.3
Missing math scores (overall)	53.4	47.2
Network		
Eastern Massachusetts	49.1	42.8
Springfield, MA	73.0	72.5
Chicago, IL	37.7	37.7
Jefferson Parish, LA	64.6	51.6
Grade Level		
Grades 3–5	71.9	70.8
Grades 6–8	19.6	13.5
Demographics		
Gender	0.0	0.0
Race/ethnicity	0.0	0.1
FRPL status	16.7	15.4
IEP status	16.7	15.4
ELL status	16.7	15.4

Note. The school sample includes the 89 schools that are in the Year 2 impact sample. The proportion of students with missing baseline ELA or math scores is based on the sample of students in the respective Year 2 impact models. The proportion of students with missing demographic data is based on the sample of students in either the ELA or math Year 2 impact models. Source: Year 2 district administrative data files from 2012–13 (Wave 1 schools) and 2013–14 (Wave 2 schools).

3. IMPLEMENTATION DATA ANALYSIS

In order to assess the fidelity of program implementation in treatment schools, we gathered administrative data from ANet on each of the four program components: assessment administration, logistical support, coaching support, and network events. We also studied fidelity of implementation from the perspective of treatment-school leaders and teachers by examining survey data on the usefulness of the data team meetings and the MyANet tool. We report fidelity of implementation for both Year 1 and Year 2 of the study for schools in the full student impact sample, which includes 46 treatment schools in Year 1 and 45 treatment schools in Year 2 (Table 3.1).[13]

Table 3.1. Number of Treatment Schools, by Study Year, School Year (SY), and Data Collection Wave

	Year 1 (Y1)	Year 2 (Y2)
Wave 1 (W1)	39 (SY 2012–13)	38 (SY 2013–14)
Wave 2 (W2)	7 (SY 2013–14)	7 (SY 2014–15)

Note. Shaded cells represent the sample of schools included in the Year 2 impact analyses.

The 10 treatment schools that worked with ANet in Year 1 but not in Year 2 are included in our Year 2 count of 45 schools because they are included in our Year 2 full student impact sample. However, no implementation data (ANet administrative data or survey data) were collected for these 10 schools in Year 2 of the study because they were no longer administering the program. In addition, some treatment schools are not represented because teachers or leaders from those schools failed to respond to the survey.[14]

Implementation Results from Administrative Data

Assessment administration. Assessment administration refers to whether assessments were administered within two days of their target date for each subject (ELA and math) and assessment cycle. For schools in Year 1 and Year 2 of the study with complete data, ANet assessments were administered within one to two days of their target date, on average, across subjects and data cycles. There were a few instances of assessments being administered several days after the target date (up to 16 days), but these were rare.

[13] The control-school pair of one treatment school closed between year 1 and year 2 of the study. As a result, this treatment school is excluded from both the implementation and impact analyses in year 2.
[14] Responses of leaders whose paired control school leader did not respond were removed from the results.

Logistical support. Logistical support refers to two factors: the degree to which interim assessment results were returned to schools within 48 hours (two business days) of each assessment administration and whether each school had access to the MyANet online platform. In Year 1 and Year 2 of the study, reports were returned to schools, on average, within the 48-hour window. Every school, in both years of the study, also had access to MyANet.

Coaching support. Coaching support measures the degree to which coaches completed planned school visits and led data meetings with teachers and principals. ANet guidelines indicate that coaches are expected to meet with leaders or teachers in each school approximately 20 times per year. While 20 coach visits is therefore a reasonable target to expect, the number of coach visits may be more or less for schools that require different levels of support or where particular visits were unusually long. This is particularly true for schools in their second year of participation, as ANet began to transition coach duties to school leaders in order to build school capacity. In Year 1 of the study, coaches visited with schools, on average, 17 times throughout the school year, slightly less than the 20-visit expectation. In Year 2 of the study, coaches visited with schools, on average, 15 times throughout the school year. The number of coach visits annually ranged across participating schools from as many as 29 visits to as few as eight visits over the course of the study.

Network events. Network events measures whether at least four network events were offered to schools in each network each year, including two major events, such as a spring and fall event. At least four, and often times more, network events were offered to ANet schools in each year of the study.

Implementation Results from Treatment-School Leader and Teacher Surveys

Teacher and leader perspectives on data meetings. In both years of the study, the majority of teachers and leaders who responded to the survey reported that they found data meetings somewhat or very useful, in terms of reviewing class and individual student results on ANet assessments. The majority of teachers and leaders also found the data meetings somewhat or very useful for developing lesson plans and sharing instructional strategies for re-teaching; however, time for planning lessons did not appear to be consistently provided according to teacher and leader reports. Only a small percentage of teachers in each year (2–6%) reported that the data meetings were not useful for reviewing data, developing lesson plans, or sharing instructional strategies for re-teaching.

Teacher perspectives on MyANet. In both years of the study, the majority of teachers (60–75%) reported that they used the MyANet tool at least once or twice a month in order to create quizzes, review and compare student work, examine state standards in math or ELA, and generate interim assessment materials. Fewer teachers reported using the MyANet tool to plan their lessons; 50% in Year 1 and 40% in Year 2 reported that they used it infrequently for this purpose. Survey responses indicate that teachers used the MyANet tool more frequently, for a variety of purposes, in Year 2 compared to Year 1.

4. IMPACTS ON EDUCATOR BELIEFS AND PRACTICE AFTER TWO YEARS

In this section, we report estimates of the impact of ANet participation on school leader and teacher beliefs and practices after two years within the survey impact sample. Each of the leader- and teacher-reported beliefs and practice outcomes are constructed from sets of survey items that are measured on five-point scales of satisfaction, frequency, or agreement.[15] A mean scale score was calculated for any leader or teacher responding to at least 50% of the items within a set. For school leaders and teachers, we provide descriptive statistics on the sample and relevant survey scales. Descriptive statistics for scale scores are reported in scale score points. Prior to analyzing program impacts, however, scores were standardized within the study sample. Impact estimates are therefore reported in standard deviation units (i.e., effect sizes).

School Leaders

Table 4.1 shows characteristics of school leaders in the complete-pair Year 2 impact sample.[16] On average, across both the treatment and control groups, leaders have seven years of experience as an administrator, with just over four years of experience in their current school. The sample is predominantly female (73%) and white (61%), and most hold a Master's degree (92%). The only characteristic on which the treatment and control school leaders differ by a statistically significant amount is total leadership experience, with control-school leaders having somewhat more experience ($p < 0.05$).

[15] Exhibit B-1 in Appendix B provides a list of the school leader and teacher survey items within each scale or index, along with the response scale.

[16] The complete-pair Year 2 leader sample is the sample of leaders for which we have both leaders' survey responses within the matched pair.

Table 4.1. Year 2 School Leader Demographics, Overall and by Treatment Assignment

	Overall			Treatment group			Control group		
	n	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
Leadership experience									
Total experience in current district	52	17.4	11.31	26	17.4	8.91	26	17.4	13.46
Total leadership experience	52	7.0	6.98	26	5.0	4.42	26	9.0	8.47
Leadership experience in current school	52	4.2	4.37	26	3.7	3.16	26	4.7	5.33
	Overall			Treatment group			Control group		
	n		%	*n*		%	*n*		%
Gender									
Female	38		73.1	18		69.2	20		76.9
Male	14		26.9	8		30.8	6		23.1
Race/ethnicity									
African American	16		31.4	8		32.0	8		30.8
White	31		60.8	15		60.0	16		61.5
Other	4		7.8	2		8.0	2		7.7
Highest degree									
Masters	47		92.2	23		92.0	24		92.3
Doctorate	4		7.8	2		8.0	2		7.7

Note. Descriptive statistics are reported for those leaders in the Year 2 survey impact sample school ($n = 67$) whose matched-pair school leader also responded to the survey. Treatment- and control-school leaders only differ with statistical significance on one measure: mean total leadership experience ($p < 0.05$). Source: Year 2 school leader surveys (treatment and control).

Table 4.2. Summary Statistics for Year 2 Leader Survey Scales

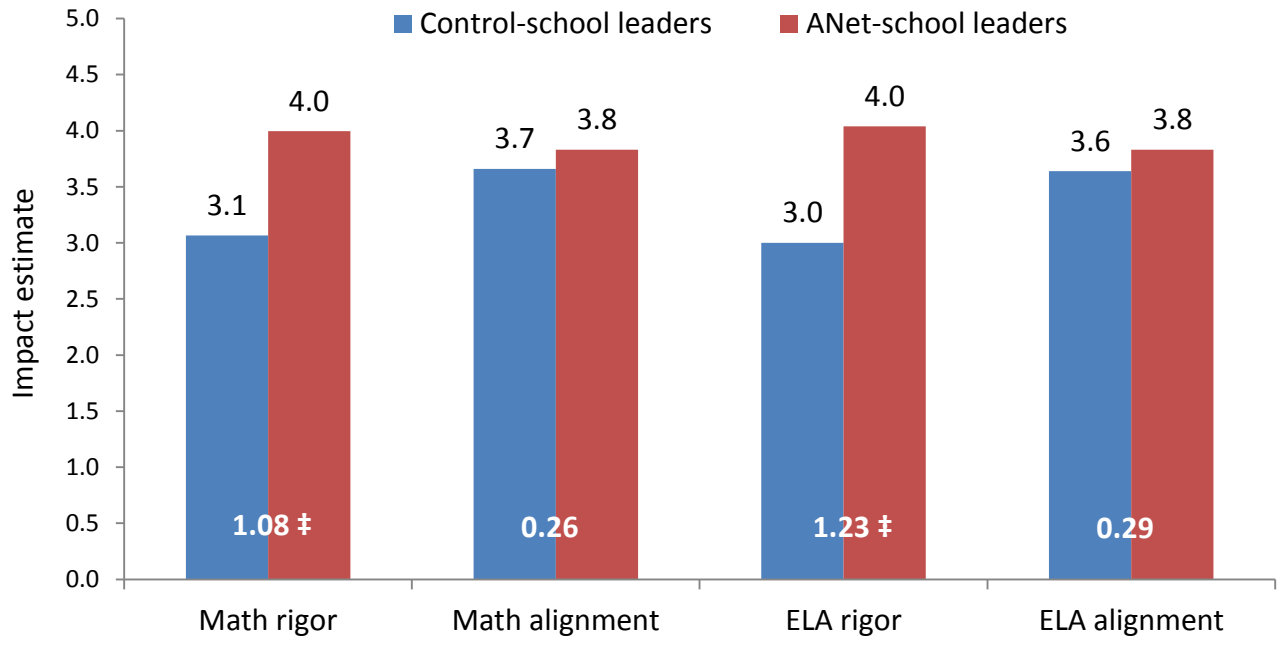
Survey Scale	Number of items	Mean	*SD*	*n*	Reliability
Math rigor	3	3.5	0.88	51	0.95
Math alignment	4	3.8	0.65	51	0.87
ELA rigor	3	3.5	0.87	51	0.95
ELA alignment	4	3.7	0.68	51	0.85
Data and reporting satisfaction	2	4.0	0.72	50	0.80
Support satisfaction	4	3.8	1.01	51	0.96
Leader instructional practices	9	4.1	0.56	53	0.81
Data review	3	3.8	0.66	51	0.76
Confidence using data	8	3.8	0.57	51	0.93

Note. Descriptive statistics are reported for those leaders in the Year 2 survey impact sample ($n = 67$) whose matched-pair school leader also responded to the survey. Scale statistics (mean, standard deviation, and counts) were calculated for any impact sample respondent who answered at least 50% of the scale items. Exceptions are for scales containing only two items where responses to all items were required. Scale reliability was computed within the impact sample on all available data (i.e.,pairwise calculations). Source: Year 2 school leader surveys (treatment and control).

Survey impacts. Figure 4.1.a summarizes the impact of ANet on school leaders' perceptions of the rigor and alignment of the interim assessments in use at their schools in both math and reading. For each outcome, we report the mean response for leaders in the control group (blue bars) and the adjusted mean response for the treatment group (red bars); the latter is calculated by adding the estimated treatment effect to the control group mean. Within each pair of bars we report the estimated treatment effect in standard deviation units (i.e., effect size).

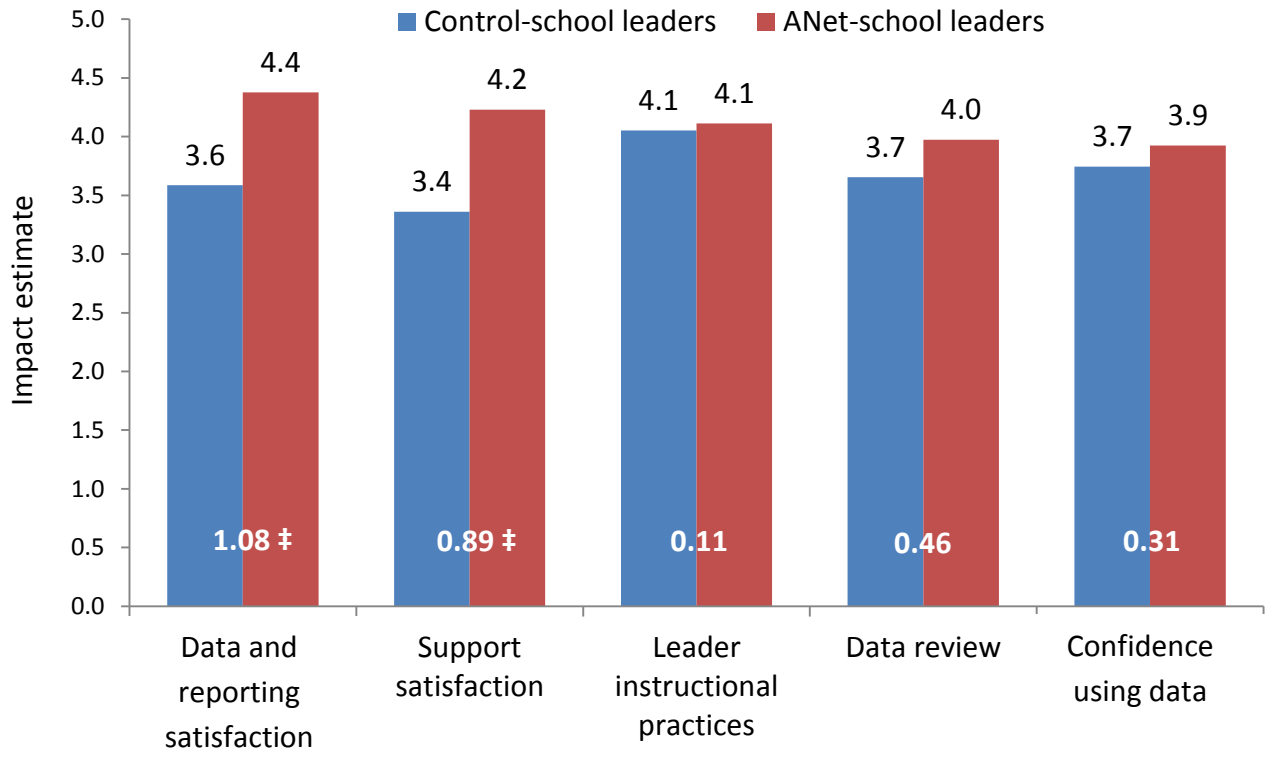
After two years, we find a large positive impact of ANet on school leaders' perceptions of the rigor of the ANet interim assessments compared to those used by control-school leaders. The difference is greater than one standard deviation in both subjects (math 1.08 *SD*, $p < 0.01$; ELA 1.23 *SD*, $p < 0.01$) (Figure 4.1.a). We see no statistically significant differences in school leaders' perceptions of the alignment of the math or ELA interim assessments with the respective state standards, the state summative assessments, and the curriculum and curricular scope and sequence in use in their school.

Figure 4.1.a. Year 2 Impacts on School Leaders' Perceptions of the Rigor and Alignment of Their Interim Assessments



Note. School leaders in ANet schools were asked to respond in regards to their ANet interim assessments. Data labels for columns report mean scale scores for the respective group (control or treatment). Standardized impact estimates are shown at the bottom of each set of columns.
‡ *p < 0.01*

Figure 4.1.b. Year 2 Impacts on School Leaders' Satisfaction with Program Components, Practices, and Confidence



Note. School leaders in ANet schools were asked to respond in regards to their ANet interim assessments. Data labels for columns report mean scale scores for the respective group (control or treatment). Standardized impact estimates are shown at the bottom of each set of columns.
‡ $p < 0.01$

Turning to measures of support, Figure 4.1.b confirms that leaders in ANet schools also reported much higher satisfaction with data and reporting (1.08 *SD*, $p < 0.01$), as well as with the support they receive around data use (0.89 *SD*, $p < 0.01$). However, no statistically significant differences were found between treatment- and control-school leaders on other measures: frequency of various instructional support practices and the review of interim assessment data, and their confidence in using various assessment and data practices.

Teachers

In the Year 2 impact sample, eligible 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 4.3). Since all teachers reported engaging in at least some amount of instruction in ELA, reading, or math, it is not surprising that the vast majority teach one or both of these subjects or in a general elementary assignment (43.7% in ELA/reading, 34.9% in math,

and 28.4% in general elementary—categories not mutually exclusive). Most teachers instruct students in Grades 3 through 5 only (77.1%), but 20.5% teach only middle grades (Grades 6 through 8) or at least one grade at both levels (2.4%). A large majority are female (88.4%), are white (73.2%), hold a Master's as their highest degree earned (68.2%), and entered teaching through a traditional certification route (79.7%). No differences are observed between treatment and control teachers on any of these variables.

Table 4.3. Year 2 Teacher Demographics, Overall and by Treatment Assignment

	Overall			Treatment group			Control group		
	n	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
Total teaching experience	609	12.6	9.08	367	12.7	9.28	242	12.4	8.79
Teaching experience in current district	608	10.9	8.68	367	10.8	8.72	241	11.0	8.63
Teaching experience in current school	608	7.2	6.82	369	7.2	7.16	239	7.1	6.27
Teaching experience in current grade and subject	611	6.8	6.73	369	7.0	7.10	242	6.4	6.11

	Overall		Treatment group		Control group	
	n	%	*n*	%	*n*	%
Primary subject area						
English language arts	269	43.7	161	43.2	108	44.4
Mathematics	215	34.9	132	35.4	83	34.2
General elementary	175	28.4	109	29.2	66	27.2
Special education	117	19.0	61	16.4	56	23.0
English as a second language/ English language development	63	10.2	36	9.7	27	11.1
Other	53	8.6	31	8.3	22	9.1
Grade level						
Elementary only	475	77.1	288	77.2	187	77.0
Middle 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						
Bachelor's	187	30.7	121	32.8	66	27.4
Master's	416	68.2	245	66.4	171	71.0
Doctorate	7	1.2	3	0.8	4	1.7
Alternative certification						
Yes	124	20.3	69	18.6	55	22.8
No	488	79.7	302	81.4	186	77.2

Note. Descriptive statistics are reported for in-scope teachers in the Year 2 survey impact sample (*n* = 67). There are no statistically significant differences between treatment- and control-school teachers on any of these measures. Primary subject categories are not mutually exclusive. Totals sum to greater than 616 teachers and percentages to greater than 100. Source: Year 2 teacher surveys (treatment and control).

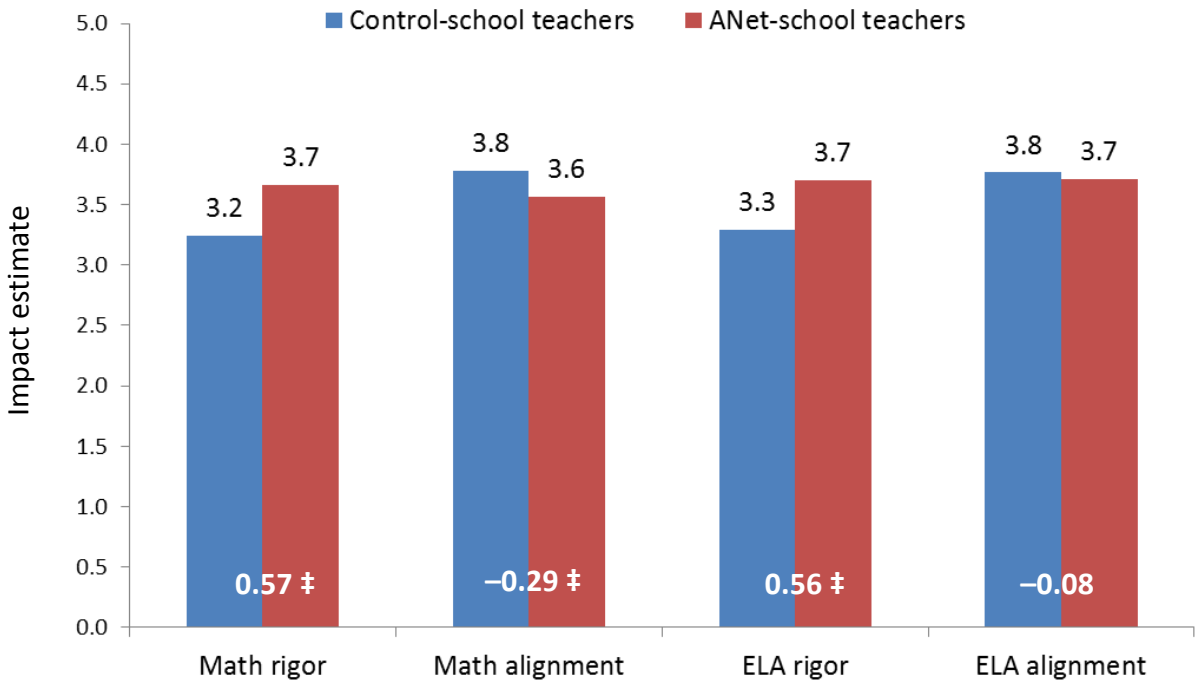
Table 4.4. Summary Statistics for Year 2 Teacher Survey Scales

Survey Scale	Number of items	Mean	*SD*	*n*	Reliability
Math rigor	3	3.5	0.73	488	0.93
Math alignment	4	3.7	0.75	491	0.90
ELA rigor	3	3.6	0.72	514	0.91
ELA alignment	4	3.7	0.66	518	0.90
Data & reporting satisfaction	2	4.0	0.78	580	0.84
Support satisfaction	4	3.6	0.87	586	0.95
Leader abilities	9	3.8	0.93	616	0.97
Data review	4	3.0	0.67	588	0.84
Data use	8	3.4	0.76	585	0.95
Confidence using data	10	4.0	0.63	588	0.96
Instructional differentiation	2	4.0	0.69	608	0.74

Note. Descriptive statistics are reported for in-scope teachers in the Year 2 survey impact sample (*n* = 67). Scale statistics (mean, standard deviation, and counts) were calculated for any impact sample respondent who answered at least 50% of the scale items. Exceptions are for scales containing only two items where responses to all items were required. Scale reliability was computed within the impact sample on all available data (i.e., pairwise calculations). Source: Year 2 teacher surveys (treatment and control).

Survey impacts. After two years, we observe wide variation in the impact of ANet on teachers' perceptions of their interim assessments, support for instructional data use, and several indicators of data use and related instructional practices. Compared to the interim assessment used by their control-school counterparts, teachers in treatment schools perceived ANet's math and ELA interim assessments to be substantially more rigorous (0.57 *SD* math, 0.56 *SD* ELA; both $p < 0.01$) (Figure 4.2.a). However, treatment-school teachers also perceived ANet's math interim assessments to be less well aligned than their control-school counterparts when comparing them to the state standards, the state test, and their school's curriculum and curricular scope and sequence (–0.29 *SD*, $p < 0.01$), a pattern we explore in detail in Section 6.

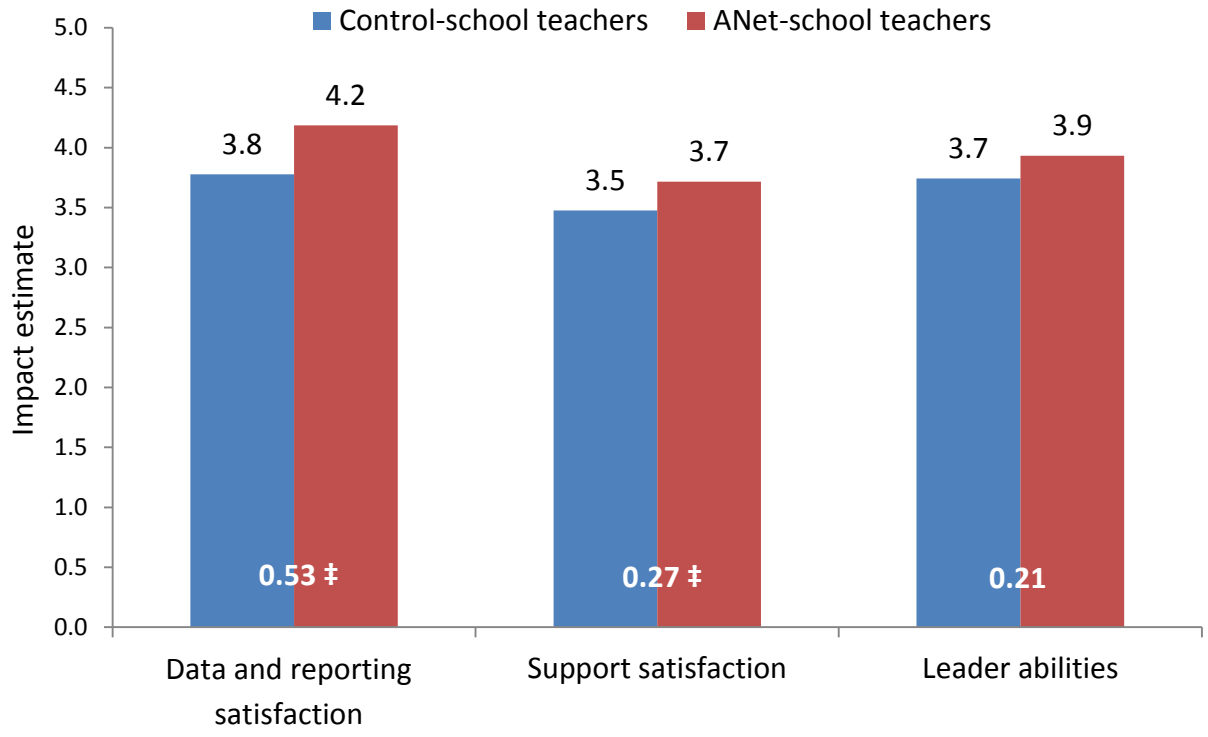
Figure 4.2.a. Year 2 Impacts on Teachers' Perceptions of the Rigor and Alignment of Their Interim Assessments[1]



Note: Data labels for columns report mean scale scores for the respective group (control or treatment). Standardized impact estimates are shown at the bottom of each set of columns.
‡ $p < 0.01$

Treatment-school teachers also reported higher satisfaction with interim assessment data and reporting (0.53 *SD*, $p < 0.01$) and with the level of support they received for various data-related tasks (0.27 *SD*, $p < 0.01$) (e.g., analysis of data, improving instructional practices); however, the magnitude of the latter impact is only half as large as that on satisfaction with data and reporting (Figure 4.2.b).

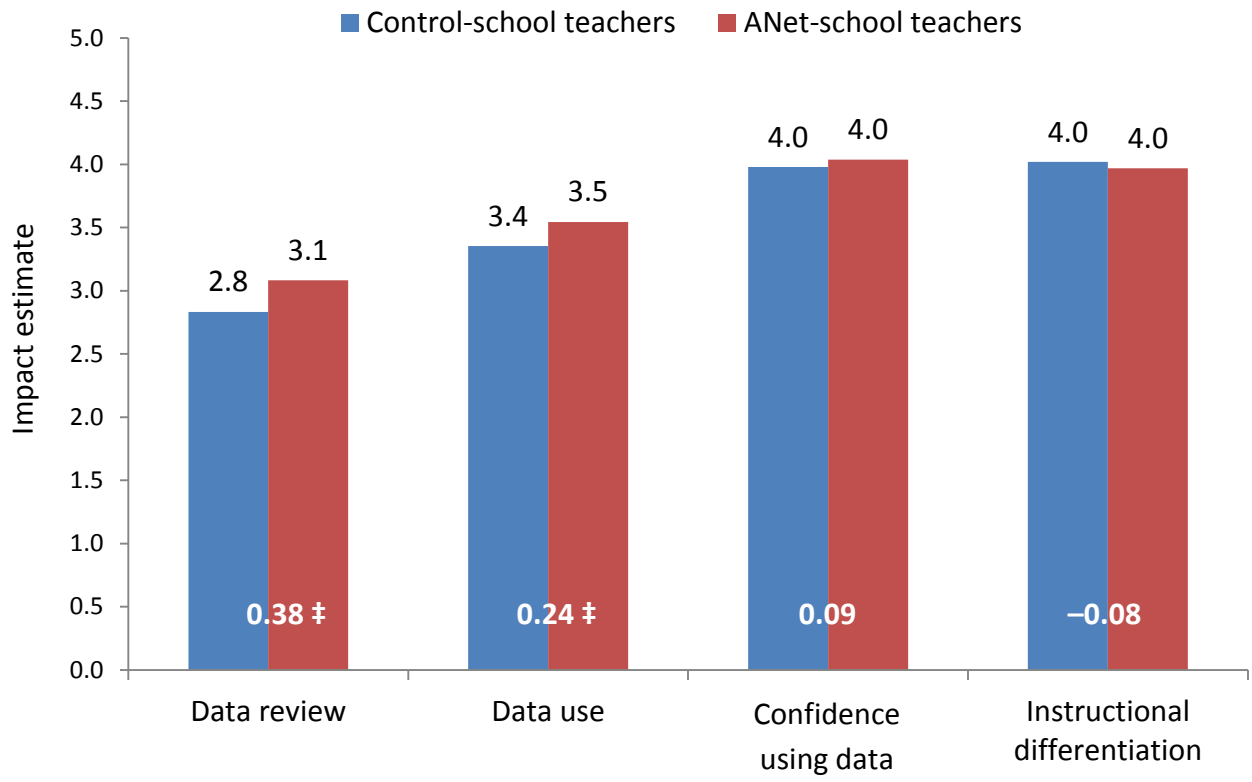
Figure 4.2.b. Year 2 Impacts on Teachers' Satisfaction with Various Forms of Support



Note. Data labels for columns report mean scale scores for the respective group (control or treatment). Standardized impact estimates are shown at the bottom of each set of columns.
‡ $p < 0.01$

In terms of practices, treatment-school teachers reported more frequent review (0.38 *SD*, $p < 0.01$) and use (0.24 *SD*, $p < 0.01$) of data (Figure 4.2.c). ANet had no impact, however, on teachers' perceptions of their instructional leaders' abilities, teachers' confidence in using various assessment and data practices, or on the frequency with which teachers reported differentiating instruction.

Figure 4.2.c. Year 2 Impacts on Teachers' Data-Related Confidence and Practices



Note. Data labels for columns report mean scale scores for the respective group (control or treatment). Standardized impact estimates are shown at the bottom of each set of columns.
‡ *p < 0.01*

5. IMPACTS ON STUDENT ACHIEVEMENT AFTER TWO YEARS

In this section, we report the impacts of ANet after two years on students' math and reading achievement, as measured by their performance on state tests. The overall model results for the full sample of schools represent our main confirmatory analyses, but we also present estimates for various subgroups of students and for the reduced sample of treatment schools that continued to work with ANet for both years. Recall that, in these analyses, student assessment scores were standardized within state, subject, and grade. Impacts on student achievement after two years are therefore reported in standard deviation units.

Student Achievement Impacts

The top portion of Tables 5.1 and 5.2 show the intent-to-treat results for the full sample of schools for which student data are available in math and reading, respectively (N = 89). After two years, ANet had no impact on the math achievement of students in Grades 3 through 8 as measured by their summative state test scores (–0.04 *SD*; p = 0.30). However, statistically significant impacts on math achievement were found in three geographic networks. Eastern Massachusetts showed a negative impact of 0.12 *SD* ($p < 0.05$) and Chicago showed a negative impact of 0.16 *SD* ($p < 0.01$). A positive impact on math achievement of 0.21 *SD* was found in Springfield ($p < 0.05$). When broken out by grade level, a negative impact on math achievement of 0.09 *SD* was found in the elementary grades ($p < 0.05$); however, the program had no impact on math achievement in the middle grades. No differences were found in the math achievement of students when broken out by free or reduced-price lunch eligibility or prior performance in Year 1.

Table 5.1. Year 2 Impacts on Math Test Scores, Overall and by Subgroup

	Impact estimate	*SE*	*p* value	*n*
FULL SAMPLE (*N* = 89)				
Overall	–0.04	0.037	0.300	21,335
Network				
Eastern Massachusetts	–0.12 **	0.056	0.040	7,908
Springfield, MA	0.21 **	0.082	0.028	1,907
Chicago, IL	–0.16 ‡	0.043	0.001	4,879
Jefferson Parish, LA	0.04	0.079	0.659	6,641
Grade level				
Grades 3–5	–0.09 **	0.040	0.027	13,233
Grades 6–8	–0.02	0.066	0.785	8,102
Free or reduced-price Lunch eligibility				
Not eligible	–0.03	0.070	0.705	2,686
Eligible	–0.04	0.034	0.246	18,637
Prior performance				
Top quartile	–0.01	0.045	0.891	3,276
Bottom quartile	–0.03	0.037	0.428	4,217
"Bubble" student	–0.05	0.032	0.105	4,511
REDUCED SAMPLE (*n* = 69)				
Overall	0.00	0.044	0.978	15,806
Network				
Eastern Massachusetts	–0.13	0.068	0.071	6,137
Springfield, MA	0.21 **	0.082	0.028	1,907
Chicago, IL	–0.03	0.069	0.686	1,121
Jefferson Parish, LA	0.04	0.079	0.659	6,641
Grade level				
Grades 3–5	–0.06	0.047	0.209	10,433
Grades 6–8	0.05	0.095	0.621	5,373
Free or reduced-price Lunch eligibility				
Not eligible	0.03	0.080	0.742	2,061
Eligible	–0.01	0.041	0.855	13,741
Prior performance				
Top quartile	0.03	0.053	0.566	2,395
Bottom quartile	0.02	0.045	0.686	2,960
"Bubble" student	–0.06	0.044	0.181	3,230

Note. Analyses were run on the full ITT sample of 89 schools (45 treatment and 44 control) as well as the reduced student impact sample of 69 schools (35 treatment and 34 control). Models include fixed-effects for grade level and a set of paired school dummy variables. All models include students' baseline math test score and student demographics. Baseline test score is interacted with Year 2 grade level and Year 2 state test in all models. An additional baseline test score imputation flag and a third- and fourth-grade flag were also included (and their interaction term). The Year 2 test score (outcome) was standardized by subject and grade using state means and standard deviations. Models are cluster-adjusted. Dummy variable imputation was used to replace missing baseline math test scores and demographics. Source: Student-level district administrative files from baseline (2010–11 or 2011–12) and Year 2 (2012–13 or 2013–14).

‡ $p < 0.01$; ** $p < 0.05$.

Table 5.2. Year 2 Impacts on Reading Test Scores, Overall and by Subgroup

	Impact estimate	*SE*	*p* value	*n*
FULL SAMPLE (*N* = 89)				
Overall	–0.05	0.028	0.099	21,258
Network				
Eastern Massachusetts	–0.08 **	0.035	0.037	7,840
Springfield, MA	0.04	0.068	0.561	1,889
Chicago, IL	–0.17 ‡	0.036	0.000	4,888
Jefferson Parish, LA	0.07	0.062	0.284	6,641
Grade level				
Grades 3–5	–0.07 **	0.034	0.031	13,188
Grades 6–8	–0.03	0.052	0.615	8,070
Free or reduced-price Lunch eligibility				
Not eligible	0.02	0.050	0.697	2,676
Eligible	–0.06 **	0.027	0.039	18,570
Prior performance				
Top quartile	–0.00	0.040	0.992	3,279
Bottom quartile	–0.06	0.031	0.078	4,213
"Bubble" student	–0.10 ‡	0.029	0.001	5,454
REDUCED SAMPLE (*n* = 69)				
Overall	–0.01	0.032	0.751	15,746
Network				
Eastern Massachusetts	–0.09 **	0.042	0.045	6,091
Springfield, MA	0.04	0.068	0.561	1,889
Chicago, IL	–0.11 **	0.041	0.045	1,125
Jefferson Parish, LA	0.07	0.062	0.284	6,641
Grade level				
Grades 3–5	–0.04	0.037	0.301	10,393
Grades 6–8	0.02	0.075	0.754	5,353
Free or reduced-price Lunch eligibility				
Not eligible	0.04	0.060	0.475	2,051
Eligible	–0.02	0.030	0.484	13,691
Prior performance				
Top quartile	0.06	0.038	0.118	2,394
Bottom quartile	–0.01	0.035	0.883	2,960
"Bubble" student	–0.10 ‡	0.033	0.003	3,759

Note. Analyses were run on the full ITT sample of 89 schools (45 treatment and 44 control) as well as the reduced student impact sample of 69 schools (35 treatment and 34 control). Models include fixed-effects for grade level and a set of paired school dummy variables. All models include students' baseline ELA test score and student demographics. Baseline test score is interacted with Year 2 grade level and Year 2 state test in all models. An additional baseline test score imputation flag and a third- and fourth-grade flag were also included (and their interaction term). The Year 2 test score (outcome) was standardized by subject and grade using state means and standard deviations. Models are cluster-adjusted. Dummy variable imputation was used to replace missing baseline ELA test scores and demographics. Source: Student-level district administrative files from baseline (2010–11 or 2011–12) and Year 2 (2012–13 or 2013–14).

‡ $p < 0.01$; ** $p < 0.05$.

After two years, ANet had no clear impact on the reading achievement of students in Grades 3 through 8 in the full sample; the point estimate of –0.05 *SD* is not statistically significant at conventional levels ($p = 0.099$). However, statistically significant impacts on reading achievement were found in two geographic networks. Eastern Massachusetts showed a negative impact of 0.08 *SD* ($p < 0.05$) and a negative impact of 0.17 *SD* was found in Chicago ($p < 0.01$). A negative impact on reading achievement of 0.07 *SD* was found in the elementary grades ($p < 0.05$); however, the program had no impact on reading achievement in the middle grades (Table 5.2). Statistically significant negative impacts on reading achievement were also found in the subset of students identified as eligible for free or reduced-price lunch (–0.06 *SD*, $p < 0.05$) and for students whose performance in Year 1 put them in the performance category below their states' designation of mastery (i.e., "bubble students": –0.10 *SD*, $p < 0.01$). [17]

Parallel student impact analyses were run for the reduced sample of 69 schools, which includes only those treatment schools that worked with ANet for two years and their pairs. The overall impacts in math (0.00 *SD*, $p = 0.98$) and reading (–0.01 *SD*, $p = 0.75$) are very close to zero and not statistically significant (bottom portion, Tables 5.1 and 5.2). The fact that the point estimates of program impacts are more favorable (although still statistically indistinguishable from zero) in this reduced sample is expected given that the reduced sample includes only those treatment schools that decided to continue with the program into a second year. It is likely that this sample represents a set of schools where the actual or perceived implementation of ANet was more successful than at the set of schools that decided not to continue participating.

[17] We calculated the Year 2 impacts separately for Wave 2 schools ($n = 14$). Recall, these schools were only recruited from Jefferson Parish, LA, and Springfield, MA. After two years, ANet had a positive impact on math achievement of 0.23 *SD* (0.077 *SE*, $p < 0.05$). The impact on reading achievement was 0.18 *SD* (0.062 *SE*, $p < 0.05$).

6. EXPLORATORY ANALYSES

Despite generally positive impacts of the ANet program on leaders' and teachers' perceptions and practices around instructional data use, our confirmatory analyses showed no clear impact on math or reading achievement in the full sample. Of course, it is not uncommon for randomized controlled trials (RCTs) to produce weak or null effects. A recent synthesis of education RCTs with no major design or methodological limitations indicated that only 9% found positive impacts (Coalition for Evidence-Based Policy, 2013). This figure may even overstate the share of RCTs with positive findings if studies with negative or null results are less likely to be published (Jacob, Jones, Hill, & Kim, 2015). Even so, our results with respect to student achievement are clearly disappointing and, in light of the positive effects on educator beliefs and practices, could be seen as casting doubt on the ANet logic model and the broader theory of change behind data-based instructional programs.

We therefore perform a series of exploratory analyses to unpack this puzzle in an attempt to distill as much learning for ANet and the field. In an effort to ensure that program providers, researchers, educators, and policymakers learn from all RCTs, even those with overall weak or null results, Jacob et al. (2015) propose three categories of factors that may contribute to null findings: design or methodological factors, contextual factors that act as barriers to implementation, and flaws in the program's theory of change. Although we conducted our exploratory analyses before we became aware of Jacob et al., their framework provides useful way to organize our findings. We first revisit design and methodological considerations before examining contextual factors that may have acted as barriers to implementation, as well as evidence related to ANet's theory of change.

Methodological Factors

Perhaps the most common methodological factor contributing to null effects in education RCTs is insufficient statistical power, an issue that is often aggravated due to attrition from the evaluation sample. Although we experienced modest attrition immediately after recruitment and between study Years 1 and 2, our matched-pair design allows for internally valid estimates of program impact in the presence of attrition and our evaluation remained sufficiently powered to detect a relatively small impact of ANet on student achievement had there been one. Specifically, the standard errors for our full student impact sample reported in Tables 5.1 and 5.2 imply that

the evaluation would have been able to detect impacts on student achievement as small as 0.07 *SD* in math and 0.06 *SD* in reading in the full sample.

There are nonetheless three aspects of the evaluation design that may help explain the null impacts of ANet on student achievement after two years. First, it is possible that transformative data-driven instructional practices take time to become embedded within schools before producing measurable improvements in student learning. For example, while we found that ANet teachers were more likely than their control-school counterparts to review and use data, there was no evidence that they differentiated their instruction more often. Measuring the main outcome of interest—student achievement—after two years therefore may not have allowed sufficient time for the program to take hold. Evidence from other studies suggests that data-driven practices can take as many as four years to impact student achievement (Slavin et al., 2013).

Second, the counterfactual conditions in control schools included the administration of interim assessments, as well as some amount of support for data use provided to leaders and teachers internally by their districts. Given the ubiquity of these elements of data-based instructional programs in American schools, comparing ANet's more intensive approach to alternatives is appropriate. Even so, it may have made it more difficult to detect an ANet treatment effect.

Finally, the sample of schools participating in the evaluation may have differed from those with which ANet typically works. In particular, the pressure to recruit a large number of partner schools quickly to meet the targeted scale of its i3-funded expansion and associated evaluation may have produced a sample of schools that was less equipped to engage in data-based instructional reform. Although this possibility is difficult to assess systematically, one potential indication of lower levels of readiness in the evaluation sample is the relatively high number of schools that elected not to continue their partnership with ANet into a second year. Ten of the 45 schools did not continue to participate for a second year, despite the fact that their participation was partially subsidized. In contrast, ANet reports that retention rates for non-i3 schools after their first year of partnership have consistently exceeded 90%. To the extent that a school's readiness to engage in data-based instructional reform is related to the strength of implementation, the results of our evaluation may therefore provide a misleading picture of

ANet's impact on educator practice and student outcomes across its entire portfolio of partner schools. Evidence from our next exploratory analysis is consistent with this interpretation.

Contextual Factors

School readiness. In their review of potential barriers to the implementation of educational interventions, Jacob et al. (2015) highlight the importance of school capacity to change (Lee, Shin, & Amo, 2013) and organizational structures that facilitate the implementation of instructional improvement strategies (Elmore, 1996; Olson, 2004). We focus here on schools' structures and their leaders' willingness to prioritize the implementation the ANet program, as reflected in their scores on the readiness screener survey that ANet administered to all schools during the evaluation's recruitment phase. Specifically, we examine whether program impacts vary systematically with program staff's assessment of their readiness to partner with ANet and engage in instructional data use. In focusing on school readiness, we distinguish the concept from more general notions of capacity, which is typically used to refer to generally high levels of skill and experience among educators in a school. We define readiness as the extent to which (1) structures are in place that enable a school to implement the program and (2) the school's leadership has prioritized this program as part of its overall improvement strategy.

We conducted this analysis in two ways. First, we grouped schools into thirds based on their average total screener score within each matched pair and estimated program impacts separately within each of these groups in both the full and reduced sample (Tables 6.1 and 6.2, Columns 1–3). Second, we re-ran our impact models for the full set of schools but included the main effect of the pair-average screener score on student achievement, as well as the interaction of the pair-average screener score with the treatment assignment indicator (Tables 6.1 and 6.2, Column 5). This provides a formal test of whether there is a statistically significant relationship between school readiness and the impact of program participation on student achievement.

For the schools in the top readiness group, the impact on student math achievement after two years is positive but not statistically significant in the full sample (0.10 *SD*, $p = 0.11$) and positive and statistically significant in the reduced sample (0.18 *SD*, $p < 0.01$). Students in treatment and control schools in the middle readiness group show no differences in achievement. For the schools in the bottom readiness group, however, the impact on student achievement in math after two years is negative and statistically significant in the full sample (-0.23 *SD*, $p < 0.01$) and in the reduced sample (–0.28 *SD*, $p < 0.01$). Column 5 confirms that there is a

statistically significant positive relationship between ANet's impact on student achievement and the average readiness of the schools within each pair to implement the program, as assessed by program staff prior to implementation. This is the case despite the fact that pair-average readiness on its own is not a statistically significant predictor of student achievement.

The same basic pattern is evident with respect to student achievement in reading. The program's estimated impact after two years is positive but not statistically significant in the full sample (0.02 *SD*, $p = 0.76$) and positive and significant in the reduced sample (0.12 *SD*, $p < 0.05$). In both samples, the program has a relatively large, negative impact on student achievement in reading in the bottom readiness group of schools ($p < 0.01$), i.e., those rated as having the lowest capacity to engage in instructional data use. The coefficient on the interaction between pair-average readiness scores and the ANet treatment is again positive and statistically significant, even though school readiness is not significantly related to student achievement.

Table 6.1. Year 2 Impacts on Math Test Scores, by Readiness Group and Interacting Readiness Score with Treatment Assignment

	Pair-average baseline readiness group			Overall math impact model	
	Top	Middle	Bottom	All schools	All schools with pair-average score interaction
FULL SAMPLE					
Treatment impact	0.10	0.00	–0.23 ‡	-0.04	–0.04
SE	0.063	0.048	0.041	0.037	0.031
p value	0.111	0.933	0.000	0.300	0.236
Pair-average score (main)	--	--	--	--	0.20
SE	--	--	--	--	0.186
p value	--	--	--	--	0.293
Treatment*Pair average score (int)	--	--	--	--	0.09 ‡
SE	--	--	--	--	0.017
p value	--	--	--	--	0.000
n (schools)	32	26	31	89	89
REDUCED SAMPLE					
Treatment impact	0.18 ‡	0.07	–0.28 ‡	0.00	–0.01
SE	0.056	0.061	0.051	0.044	0.029
p value	0.004	0.238	0.000	0.978	0.672
Pair-average score (main)	--	--	--	--	0.19
SE	--	--	--	--	0.146
p value	--	--	--	--	0.206
Treatment*Pair average score (int)	--	--	--	--	0.14 ‡
SE	--	--	--	--	0.016
p value	--	--	--	--	0.000
n (schools)	26	18	25	69	69

Note. These analyses were run on the full ITT sample of 89 schools (45 treatment and 44 control) as well as the reduced student impact sample of 69 schools (35 treatment and 34 control). Models include fixed-effects for grade level and a set of paired school dummy variables. All models include students' baseline math test score and student demographics. Baseline test score is interacted with Year 2 grade level and Year 2 state test in all models. An additional baseline test score imputation flag and a third- and fourth-grade flag were also included (and their interaction term). The Year 2 test score (outcome) was standardized by subject and grade using state means and standard deviations. Dummy variable imputation was used to replace missing baseline math test scores and demographics. Models are cluster-adjusted. Columns 1 through 3 report the student impact models for each of the three school readiness groups. Column 4 is a repeat of the student impacts shown in the first row of Table 5.1. Column 5 repeats the model but includes the main effect of the pair-average school readiness total score, as well as its interaction with the treatment indicator. Source: Student-level district administrative files from baseline (2010–11 or 2011–12) and Year 2 (2012–13 or 2013–14).

‡ $p < 0.01$; ** $p < 0.05$.

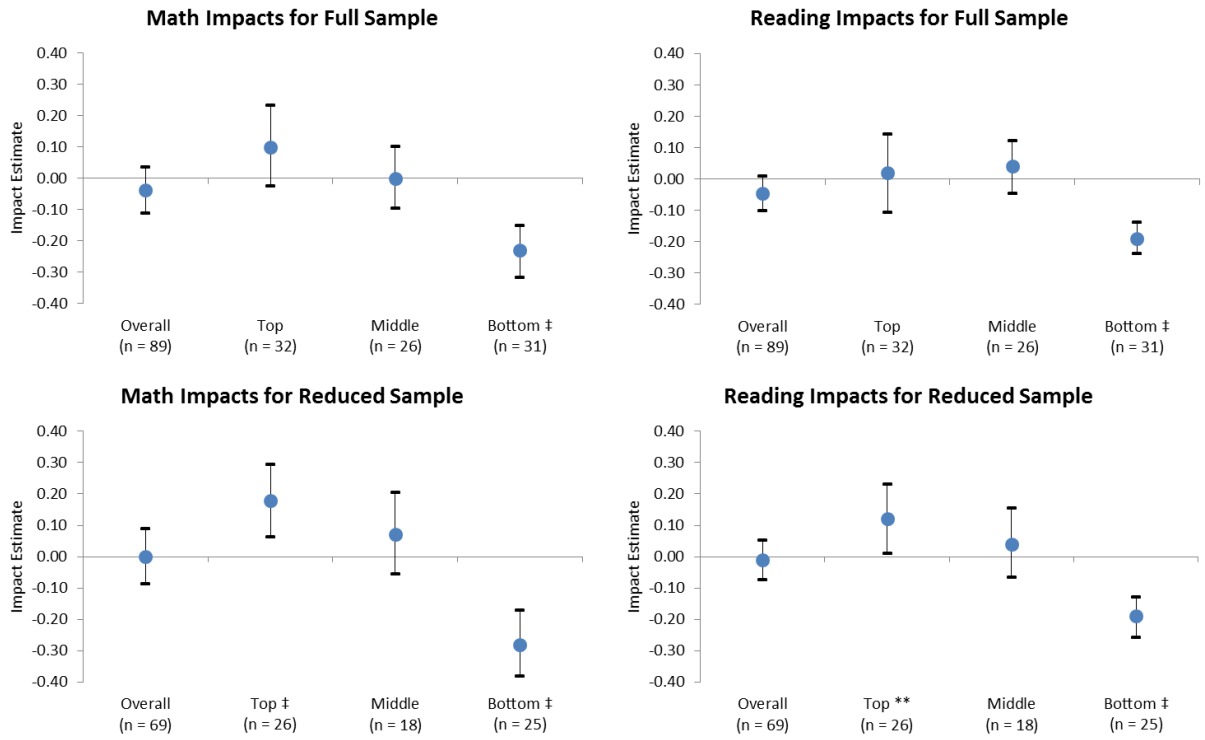
Table 6.2. Year 2 Impacts on Reading Test Scores, by Readiness Group and Interacting Readiness Score with Treatment Assignment

	Pair-average baseline school readiness group			Overall ELA impact model	
	Top	Middle	Bottom	All schools	All schools with pair-average score interaction
FULL SAMPLE					
Treatment impact	0.02	0.04	–0.19 ‡	–0.05	–0.05
SE	0.061	0.041	0.024	0.028	0.024
p value	0.763	0.360	0.000	0.099	0.061
Pair-average score (main)	--	--	--	--	0.16
SE	--	--	--	--	0.121
p value	--	--	--	--	0.181
Treatment*Pair average score (int)	--	--	--	--	0.06 ‡
SE	--	--	--	--	0.012
p value	--	--	--	--	0.000
n (schools)	32	26	31	89	89
REDUCED SAMPLE					
Treatment impact	0.12 **	0.04	–0.19 ‡	–0.01	–0.02
SE	0.054	0.052	0.031	0.032	0.023
p value	0.034	0.406	0.000	0.751	0.417
Pair-average score (main)	--	--	--	--	0.16
SE	--	--	--	--	0.111
p value	--	--	--	--	0.155
Treatment*Pair average score (int)	--	--	--	--	0.09 ‡
SE	--	--	--	--	0.013
p value	--	--	--	--	0.000
n (schools)	26	18	25	69	69

Note. These analyses were run on the full ITT sample of 89 schools (45 treatment and 44 control) as well as the reduced student impact sample of 69 schools (35 treatment and 34 control). Models include fixed-effects for grade level and a set of paired school dummy variables. All models include students' baseline ELA test score and student demographics. Baseline test score is interacted with Year 2 grade level and Year 2 state test in all models. An additional baseline test score imputation flag and a third- and fourth-grade flag were also included (and their interaction term). The Year 2 test score (outcome) was standardized by subject and grade using state means and standard deviations. Dummy variable imputation was used to replace missing baseline ELA test scores and demographics. Models are cluster-adjusted. Columns 1 through 3 report the student impact models for each of the three school readiness groups. Column 4 is a repeat of the student impacts shown in the first row of Table 5.2. Column 5 repeats the model but includes the main effect of the pair-average school readiness total score, as well as its interaction with the treatment indicator. Source: Student-level district administrative files from baseline (2010–11 or 2011–12) and Year 2 (2012–13 or 2013–14).

‡ $p < 0.01$; ** $p < 0.05$.

Figure 6.1. Year 2 Impacts on Math and Reading Test Scores, by Sample and Readiness Group



‡ $p < 0.01$; ** $p < 0.05$.

Figure 6.1, which displays estimated treatment effects and their standard errors for schools grouped by screener score, illustrates the relationship between readiness group and ANet's estimated impact on student achievement in both subjects.

We further exploit the school screener groupings to see if similar patterns are evident in the impacts of ANet on our survey-based measures of teacher beliefs and practices. For comparison, Column 1 of Table 6.3 displays the overall program impacts for the survey impact sample reported previously in Section 4. Columns 2 through 4 show the Year 2 impacts by baseline school readiness. While not as consistent as in the student impacts analysis, there are some clear differences in impacts on teacher beliefs and practices across schools of varying baseline readiness. First, while the impacts on teachers' perceptions of interim assessment rigor appear similar across levels of school readiness, the impacts on teachers' perceptions of alignment are consistently most positive in the top readiness group and most negative in the bottom readiness group. Impacts on satisfaction with support also appear similar across groups. However, the impacts on the frequency with which teachers' review and use data are largest in the top group and smallest in the bottom group.

Table 6.3. Year 2 Impacts on Teachers' Beliefs and Practices, by Readiness Group

		Pair-average baseline school readiness group		
Survey Scale	All Schools	Top	Middle	Bottom
Math rigor	0.57 ‡	0.49 ‡	0.78 ‡	0.57 ‡
SE	0.074	0.130	0.244	0.077
Math alignment	-0.29 ‡	0.05	-0.49 **	-0.54 ‡
SE	0.107	0.146	0.215	0.164
ELA rigor	0.56 ‡	0.54 ‡	0.66 **	0.54 ‡
SE	0.073	0.127	0.247	0.081
ELA alignment	-0.08	0.06	-0.12	-0.19
SE	0.079	0.084	0.244	0.129
Data and reporting satisfaction	0.53 ‡	0.62 ‡	0.97 ‡	0.28
SE	0.092	0.079	0.116	0.153
Support satisfaction	0.27 ‡	0.36 ‡	0.19	0.21
SE	0.088	0.127	0.203	0.147
Leader abilities	0.21	0.25	0.42	0.08
SE	0.110	0.134	0.279	0.186
Data review	0.38 ‡	0.76 ‡	0.38	0.01
SE	0.115	0.109	0.217	0.173
Data use	0.24 ‡	0.52 ‡	0.28	-0.04
SE	0.088	0.111	0.241	0.104
Confidence using data	0.09	0.30 ‡	-0.18	-0.01
SE	0.082	0.053	0.193	0.148
Instructional differentiation	-0.08	0.04	-0.19	-0.15 **
SE	0.065	0.125	0.147	0.070

Note. Models include the reduced survey impact sample of 67 schools (34 treatment and 33 control). Each scale is regressed on treatment assignment, school-pair dummies, and a data collection wave dummy. Models are cluster-adjusted. Column 1 repeats the standardized differences between the treatment and control groups shown in Figures 4.2.a through 4.2.c. Columns 2 through 4 repeat the survey impact model for each of the three school readiness groups. Source: Year 2 teacher surveys (treatment and control).
‡ $p < 0.01$; ** $p < 0.05$.

Collectively, these results are important for at least two reasons. First, they suggest that ANet may be more effective in improving student achievement in schools with better structures in place to support data-based instructional practices and where leadership prioritizes the work. In fact, the program appears to have had a negative impact on student achievement in schools that ANet rated as being less ready to engage in instructional data use. This pattern of results could reflect the fact that program model was initially developed in a group of charter schools where these preconditions were consistently in place. To the extent that the schools participating in the i3-funded expansion had lower levels of readiness than the schools typically served by the ANet, our evaluation may be not be representative of the organization's impact on teacher and student outcomes more generally. Second, the readiness preconditions identified by the program provider do appear to be useful for predicting where a partnership is most likely to yield positive results. This implies that the instrument the program developed to assess school readiness may be useful in identifying schools most likely to benefit from a partnership.

Teacher capacity and instructional flexibility. As shown in Figure 4.2.a, the one survey scale on which ANet was found to have a statistically significant negative impact measured teachers' perception of the alignment of their interim assessments in math; this survey scale is also positively correlated with student math achievement (see Table 6.6). To better understand the factors contributing to teachers' perceptions of assessment alignment, we first examined the extent to which those perceptions varied within and across schools. Within the same school, teachers generally have the same standards, curriculum, and curricular scope and sequence. As a result, we might expect to see little within-school (or even within-district) variation in perceived alignment. However, the vast majority of the variation in teacher-reported alignment of the interim assessments to the curriculum and the curricular scope and sequence was within schools (Table 6.4). This suggests that perceived alignment is a proxy for something that varies from teacher to teacher.

Table 6.4. Unconditional Variance Estimates of Teacher-Reported Alignment of Their Interim Assessments with the Curriculum and Curricular Scope and Sequence, by Subject, Overall and for Treatment Teachers

	All schools				Treatment schools			
	ELA		Math		ELA		Math	
Variance	Curriculum	Scope and sequence	Curriculum	Scope and sequence	Curriculum	Scope and sequence	Curriculum	Scope and sequence
Within school (σ^2)	94.7	92.1	86.7	82.8	92.1	90.4	86.1	83.7
Between school (τ_π)	0.4	0.0	0.3	0.0	0.0	0.0	1.4	3.3
Between district (τ_β)	5.0	7.9	13.0	17.2	7.9	9.6	12.5	13.0

Note. Variance decomposition from unconditional three-level model accounting for clustering of teachers within schools and districts where the dependent variable is teacher perceptions of the alignment of their interim assessments with their school's or district's (1) curriculum and (2) curricular scope and sequence. Models run by subject, for all reduced survey impact sample schools (n = 67) and treatment schools only (n = 34). Source: Year 2 teacher surveys (treatment and control).

What then explains variation in teachers' perceptions of the alignment of their interim assessments? To address this question, we regressed their ratings of the alignment of their math interim assessments with their curriculum and curricular scope and sequence separately on several individual survey items measuring factors that could influence those perceptions. These analyses indicate that teachers' confidence fitting re-teaching into the school's or district's existing curricular scope and sequence and the frequency that they use the schedule of standards to be assessed on upcoming interim assessments to plan instruction are all strong predictors of perceived alignment in math in the full sample of schools (all $p < 0.01$) (Table 6.5, Column 1). Teachers who reported that policies do not allow them flexibility to adjust their instruction based on student test results also had poorer perceptions of the alignment of the math interim assessments with curricular scope and sequence in math ($p < 0.05$). With a few exceptions, these patterns hold when examining treatment schools only (Table 6.5, Column 2).

Table 6.5. Relationship Between Teachers' Perceptions of the Alignment of Their Math Interim Assessments to the Math Curriculum or Curricular Scope and Sequence, and Measures of School Policy, Teacher Confidence, and Practices

	All schools	ANet schools
Alignment with math curriculum		
Lack of flexibility around instructional adjustments	–0.08	–0.09
Confidence fitting reteaching into scope and sequence	0.15 ‡	0.14 **
Use of upcoming interim assessment content to plan	0.14 ‡	0.15 **
Alignment with math scope/sequence		
Lack of flexibility around instructional adjustments	–0.09 **	–0.09
Confidence fitting reteaching into scope and sequence	0.13 ‡	0.11
Use of upcoming interim assessment content to plan	0.18 ‡	0.18 ‡

Note. Estimates were generated from three-level models predicting teacher reported alignment of the math interim assessments to the curriculum and curricular scope and sequence based on various policies, practices, and levels of confidence (multivariate models). Models run for all reduced survey impact sample schools (n = 67) and treatment schools only (n = 34). Source: Year 2 teacher surveys (treatment and control).
‡ $p < 0.01$; ** $p < 0.05$.

These results highlight the potential importance of aspects of teachers' confidence, as well as of organizational structures like flexibility of scope and sequence, as barriers to implementation at the individual level. Addressing these factors may help teachers manage real or perceived misalignment between their interim assessment cycle and the curriculum content and calendar. In turn, this may make the data and protocols from the ANet interim assessments more valuable instructionally if teachers are able to make changes in instructional practice and act on the data.

External context and program shifts. Finally, it is important to note that this study bridged one of the most important recent developments in American education: the launch of the Common Core State Standards (CCSS). Given the centrality of content standards to several components of ANet's model, we cannot discount the possibility that their introduction influenced the study results. As described earlier, the CCSS pushed ANet to adjust its program model by expanding its coaching support and tools to include backward planning from standards and by deepening its engagement with district leadership. The extent to which ANet was able to implement these adjustments across the different regions in this study could have influenced the study results.

ANet's Theory of Change

Our final set of exploratory analyses examines whether, within the evaluation sample as a whole, our measures of teacher perceptions and practices aligned to ANet's program logic model are related to schools' effectiveness in raising student achievement. In other words, we asked: Do schools that rate higher on various indicators of instructional data use elicit larger gains in student achievement over the two-year study period? To address this question, we first generated school-mean teacher responses for each of the survey scales reported in Section 4. We then examined whether schools where teachers reported higher levels of satisfaction with the assessments and related supports, as well as greater frequency of instructional data use, were more effective in raising student test scores (after adjusting for baseline test scores and student demographics). Although these correlational analyses cannot establish causal links between educator practices and school effectiveness, they shed some light on the underlying validity of the ANet logic model.

We find that most of the teacher survey measures are in fact predictive of student achievement gains. Each scale (on its own) is positively correlated with student math achievement with the exception of teachers' perceived rigor of their math interim assessments (no relationship) (Table 6.6, Column 1). Two of the strongest predictors of student achievement in math are the frequency with which their teachers review interim assessment data alone or with others and use data in various ways. For example, a 1 *SD* change in the frequency with which teachers use data is associated with an increase in student math achievement of 0.13 *SD* ($p < 0.01$).

Table 6.6. Bivariate and Multivariate Relationships Between Year 2 Teacher Survey Scale Scores and Math Test Scores

	Bivariate correlation of EACH survey scale with student math achievement			Student math impact model with ALL survey scales		
	Estimate	*SE*	*n*	Estimate	*SE*	*n*
Treatment				–0.10	0.058	15,357
Math rigor	–0.03	0.040	15,357	0.02	0.032	
Math alignment	0.11 ‡	0.023	15,357	0.05	0.033	
Data and reporting satisfaction	0.08 **	0.036	15,522	0.08 **	0.035	
Support satisfaction	0.10 ‡	0.029	15,522	–0.01	0.056	
Leader abilities	0.11 ‡	0.026	15,522	0.09 **	0.043	
Data review	0.12 ‡	0.025	15,522	–0.08	0.075	
Data use	0.13 ‡	0.031	15,522	0.12 **	0.025	
Confidence using data	0.09 ‡	0.031	15,522	–0.03	0.047	
Instructional differentiation	0.10 ‡	0.027	15,522	0.05	0.045	

Note. Analyses include the reduced student impact sample of 69 schools (35 treatment and 34 control). Columns 1–3 provide the bivariate estimate of the correlation between each scale and student math achievement after adding each scale to the student impact model. Columns 4–6 show the multivariate estimates of the correlation between each scale and student math achievement after adding all scales to the student impact model. Scales and indices are standardized within the student impact sample. Student impact models include fixed-effects for grade level and a set of paired-school dummy variables. All models include students' baseline math test score and student demographics. Baseline test score is interacted with Year 2 grade level and Year 2 state test in all models. An additional baseline test score imputation flag and a third- and fourth-grade flag were also included (and their interaction term). Dummy variable imputation was used to replace missing baseline math test scores and demographics. The Year 2 test score (outcome) was standardized by subject and grade using state means and standard deviations. Models are cluster-adjusted. Source: Student-level district administrative files from baseline (2010–11 or 2011–12) and Year 2 (2012–13 or 2013–14); Year 2 teacher surveys (treatment and control).
‡ $p < 0.01$; ** $p < 0.05$.

The bivariate relationships between the survey scales and student reading achievement are also generally positive but slightly smaller in magnitude. The exceptions are teachers' self-reported confidence in their ability to engage in various data use practices and, as in math, perceived rigor of their reading interim assessments (Table 6.7, Column 1). As with math achievement, one of the strongest predictors of student achievement in reading is the frequency with which their teachers use data in various ways. The frequency that teachers reported differentiating instruction (to small groups or individuals) was similarly related to reading achievement. In both cases, a 1 *SD* change in this frequency is associated with an increase in student math achievement of 0.08 *SD* ($p < 0.01$).

Table 6.7. Bivariate and Multivariate Relationships Between Year 2 Teacher Survey Scale Scores and Reading Test Scores

	Bivariate correlation of EACH survey scale with student ELA achievement			Student ELA impact model with ALL survey scales		
	Estimate	*SE*	*n*	Estimate	*SE*	*n*
Treatment				–0.15 ‡	0.040	15,462
ELA rigor	0.06	0.028	15,462	0.02	0.032	
ELA alignment	0.06 ‡	0.021	15,462	–0.02	0.020	
Data and reporting satisfaction	0.07 ‡	0.024	15,462	0.09 ‡	0.020	
Support satisfaction	0.05 **	0.023	15,462	–0.04	0.035	
Leader abilities	0.06 **	0.024	15,462	0.06	0.034	
Data review	0.07 ‡	0.021	15,462	–0.02	0.049	
Data use	0.08 ‡	0.029	15,462	0.11 ‡	0.033	
Confidence using data	0.03	0.022	15,462	–0.09 **	0.037	
Instructional differentiation	0.08 ‡	0.024	15,462	0.04	0.023	

Note. Analyses include the reduced student impact sample of 69 schools (35 treatment and 34 control). Columns 1–3 provide the bivariate estimate of the correlation between each scale and student ELA achievement after adding each scale to the student impact model. Columns 4–6 show the multivariate estimates of the correlation between each scale and student ELA achievement after adding all scales to the student impact model. Scales and indices are standardized within the student impact sample. Student impact models include fixed-effects for grade level and a set of paired-school dummy variables. All models include students' baseline ELA test score and student demographics. Baseline test score is interacted with Year 2 grade level and Year 2 state test in all models. An additional baseline test score imputation flag and a third- and fourth-grade flag were also included (and their interaction term). Dummy variable imputation was used to replace missing baseline ELA test scores and demographics. The Year 2 test score (outcome) was standardized by subject and grade using state means and standard deviations. Models are cluster-adjusted. Source: Student-level district administrative files from baseline (2010–11 or 2011–12) and Year 2 (2012–13 or 2013–14); Year 2 teacher surveys (treatment and control).
‡ $p < 0.01$; ** $p < 0.05$.

Simultaneously controlling for all survey scale measures as potential mediators in the student impact models results in negative impacts on student achievement in math (–0.10 *SD*, $p = 0.08$) and reading (–0.15 *SD*, $p < 0.01$) (Tables 6.6 and 6.7, Column 4). This is as expected given the pattern of program impacts on survey and student achievement outcomes. Because ANet generally had positive impact on survey scales related to student achievement gains, controlling for these impacts mechanically reduces the program's estimated impact on achievement.

More importantly, however, the results of this multivariate analysis also enable us to examine the relationship between each of the survey scales and student achievement gains while holding the others constant. We see that, holding other scales constant, the relationships between

student achievement in both subjects and the frequency with which teachers' review data ("Data review"), their satisfaction with support for data use ("Support satisfaction"), and their confidence around data use ("Confidence using data") take on negative signs although (with the exception of teachers' confidence around data use and reading achievement) they are not statistically significant. In addition, the relationship between student achievement in reading and perception of the alignment of interim assessments in that subject becomes negative in the multivariate model (Tables 6.6 and 6.7, Column 4). The frequency with which teachers use data in various ways remains one of the stronger predictors of student achievement in both subjects (all $p < 0.05$).

Taken as a whole, these results cast doubt on the notion that the null impacts on student achievement we estimate stem from a flawed theory of change. They also suggest several potential lessons for instructional data use that deserve examination in future research. First, conditional on the frequency teachers report using data to inform their instruction, we find that additional time spent reviewing data is not associated with student achievement. This implies that data review is only productive to the extent that it results in changes in instructional practice. Second, scales measuring their satisfaction with supports for and confidence around data use are, if anything, negatively related to achievement in the multivariate models. This suggests that these scales may be poor indicators of the actual quality of the supports teachers receive and their actual capacity to engage in instructional data use. This could be the case, for example, if programs that push teachers to use data in new ways make them aware of gaps in support or in their own capacity.

7. CONCLUSIONS

The lack of significant impacts of ANet on student achievement in schools participating in its i3-funded expansion is surprising given (1) emerging evidence that intensive data use is a distinguishing feature of many high-performing schools; (2) the fact that the program increased teachers' satisfaction with the available supports for data use and the extent to which they reported reviewing data and using it to inform their instruction; and (3) the fact that many indicators of program satisfaction and instructional data use are positively associated with schools' performance in raising student achievement within our study sample. On the other hand, our study is not alone in finding that efforts to promote instructional data use do not consistently translate into gains in student learning.

Our analysis points to at least two potential explanations for these findings related to the characteristics of participating teachers and schools. First, the program reduced teachers' perceptions of the extent to which their interim assessments were aligned to the curriculum and the scope and sequence used in their schools. This result may reflect a genuine lack of alignment but appears to be mediated by teachers' capacity to use a schedule of standards to be assessed on upcoming interim assessments to plan their instruction. In schools where teachers lack either the ability or the requisite flexibility to align their instruction to the content of interim assessments, efforts to promote the use of interim assessment data may increase perceptions of misalignment and render the data produced by the assessments less valuable.

Second, we find that negative impacts were concentrated within a subset of treatment schools that were rated by program staff as having lower levels of readiness to engage in instructional data use. Among schools that received higher readiness ratings and participated in ANet for two years, we estimate large positive treatment effects. In short, our results are consistent with the idea that intensive data use is an effective strategy for schools with the right structures in place to support teachers and where leadership is committed to prioritizing the work. In other settings, however, it may even be counterproductive. Such heterogeneity in impacts could in theory account for the variation in findings reported in the existing literature on data-based instructional programs.

This pattern suggests a choice facing ANet—and perhaps the broader field of data-based instructional programs—as it continues to refine its strategy for improving student achievement. First, ANet could narrow its focus to "ready" schools. ANet's current program may be analogous

to a high-end piece of athletic equipment that only produces results for individuals committed to daily practice and to making other changes to their lifestyle. Those less committed will find it hard to use productively and may even injure themselves; they would be better served by pursuing other strategies to improve their fitness. Our evidence indicates that the organization is able to identify in advance those settings in which it is most likely to improve student achievement, suggesting that exercising more discretion about the schools with which it chooses to partner is viable. Alternatively, ANet could explore making its services easier to be used productively by schools and teachers with less incoming readiness. This option is attractive in that it would allow the organization to work with a broader range of schools including those most in need of improvement. However, whether it is possible to identify and deliver in a cost-effective way the kinds of supports that would allow all schools to benefit from its data-based instructional program remains to be seen.

The organization could also consider a hybrid approach, narrowing the range of schools that it serves with its existing program model while simultaneously experimenting with new models for working with schools with less initial readiness. Such an approach would be a natural extension of its existing program model, under which coaches adjust the level of support provided to participating schools based on their success in implementing the program. This would allow ANet to continue to work with those schools in which it appears to be making a positive difference, while at the same time advancing the field's understanding of what is needed for a focus on instructional data use to translate into improvements in student achievement.

Finally, our results suggest that districts or schools considering whether to implement an intensive data-based instructional program should work with providers to identify what is necessary for successful implementation. District and school leaders should also ensure that, if misalignment between interim assessments and curricula becomes a challenge, teachers are provided with support to address it through planning and the freedom to reorganize their instruction so that the assessment data they receive is instructionally useful. As an evaluation of a specific program implemented in particular settings, our study cannot speak to the efficacy of data-based instructional programs generally. It seems likely, however, that the factors that appear to have generated positive results in a subset of schools within our evaluation sample would also increase the probability of success elsewhere.

REFERENCES

Angrist J. D., Pathak, P. A., & Walters, C. R. (2013). Explaining charter school effectiveness. *American Economic Journal: Applied Economics, 5*(4), 1–27.

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). *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.

Coalition for Evidenced-Based Policy. (2013). *Randomized controlled trials commissioned by the Institute of Education Sciences since 2002: How many found positive versus weak or no effects.* Washington, DC: Author.

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 Report No. 2013-4000). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Dobbie, W., & Fryer, R. G. (2013). Getting beneath the veil of effective schools: Evidence from New York City. *American Economic Journal: Applied Economics, 5*(4), 28–60.

Elmore, R. T. (1996). Getting to scale with good educational practice. *Harvard Educational Review, 66*(1), 1–26.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement* (REL Technical Brief No. 2007-039). 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 No. 2007-002). 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.

Ho, D., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199–236.

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.

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.

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., van der Ploeg, A., & Li, W. (2014, March). *Combining evidence from two RCTs about diagnostic assessments.* Paper presented at the spring conference of the Society for the Research on Educational Effectiveness, Washington, DC.

Lee, J., Shin, H., & Amo, L. C. (2013). Evaluating the impact of NCLB school interventions in New York state: Does one size fit all? *Education Policy Analysis Archives, 21*(67). Retrieved from http://epaa.asu.edu/ojs/article/view/1122

Moore, R. T. (2012). Multivariate continuous blocking to improve political science experiments. *Political Analysis*, *20*(4), 460–479.

Moore, R. T., & Schnakenberg. K. (2011). blockTools: Blocking, assignment, and diagnosing interference in randomized experiments [software and training manual] (Version 0.5-3).

Morton, B. A. (2015). *The effect of a "data-based instructional program" on teacher practices: the roles of school culture, instructional leadership, and teacher characteristics.* Manuscript in preparation.

Olson, D. (2004). The triumph of hope over experience in the search for "what works": A response to Slavin. *Educational Researcher*, *33*(1), 24–26.

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 Report No. 2009-0049). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

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 Student Thinking in Reading (FAST-R) in Boston elementary schools*. New York: 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: The impact on elementary school mathematics in the Central Region* (NCEE Report No. 2011-4005). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

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.

What Works Clearinghouse. (2014). *WWC procedure and standards handbook* (Version 3.0). Washington, DC: Author. Retrieved from: http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

Appendix A. School Screener Scoring Rubric

Exhibit A.1. Achievement Network School Screener Scoring Rubric

Developed by the Achievement Network. Reprinted for this report with permission.

	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.
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.
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.
Dedicated 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-Based	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 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.
Schedule	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.
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	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 B. Year 2 School Leader and Teacher Survey Scale Items

Exhibit B.1. Question Stem, Survey Items, and Response Scale for Each Leader-Reported Survey Scale or Index

Interim Assessment Alignment	**Scale**
Thinking about the math/ELA interim assessments you administered this year, please rate their alignment with:[1]	**Very Poor - Excellent**
a. Your state's math/ELA standards.	(5 point scale)
b. Your state's end-of-year math/ELA assessment.	
c. Your school or district's math/ELA curriculum.	
d. Your school or district's math/ELA curricular scope and sequence.	

Interim Assessment Rigor	**Scale**
Thinking about the math/ELA interim assessments you administered this year, please rate the level of rigor of test items compared to:[1]	**Much less rigorous - Much more rigorous**
a. Your state's math/ELA standards.	(5 point scale)
b. Your state's end-of-year math/ELA assessment.	
c. Your school or district's math/ELA curriculum.	

Data & Reporting Satisfaction	**Scale**
Thinking about the student interim assessment data you receive, please rate how satisfied you are with each of the following:	**Very Dissatisfied - Very Satisfied**
a. The time it takes to receive student scores on interim assessments.	(5 point scale)
b. The clarity of the data you receive.	

Support Satisfaction	**Scale**
How satisfied are you with the overall quality of the support you receive in the following:[1]	**Very Dissatisfied - Very Satisfied**
a. Analysis of interim assessment data.	(5 point scale)
b. Setting student performance goals.	
c. Timeliness of response to questions.	
d. Improving instructional practices to meet students' needs.	

Leader Abilities (Teacher-Reported)	**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**
a. Communicate a clear vision for teaching and learning for this school.	(5 point scale)
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.	

Exhibit B.1. Question Stem, Survey Items, and Response Scale for Each Leader-Reported Survey Scale or Index, Continued

Data Review	**Scale**
Over this past school year, how often have you reviewed interim assessment data:	**Never - More Than Once A Week**
a. Independently?	(5 point scale)
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?	

Data Use	**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 point scale)
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?	

IA/Data Use Confidence	**Scale**
When working with interim assessment data, how confident are you in your own ability to:	**I Don't Know How - Highly Confident**
a. Use data to set learning goals for individual students?	(5 point scale)
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 re-taught 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?	

Instructional Differentiation	**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**
a. Teach or re-teach content to small groups of students.	(5 point scale)
b. Teach or re-teach content to individual students.	

[1] The treatment survey specifically referenced the interim assessments and support provided by the Achievement Network.

Exhibit B.2. Question Stem, Survey Items, and Response Scale for Each Teacher-Reported Survey Scale or Index

Interim Assessment Alignment	Scale
Thinking about the math/ELA interim assessments you administered this year, please rate their alignment with:[1]	**Very Poor - Excellent**
a. Your state's math/ELA standards.	(5 point scale)
b. Your state's end-of-year math/ELA assessment.	
c. Your school or district's math/ELA curriculum.	
d. Your school or district's math/ELA curricular scope and sequence.	

Interim Assessment Rigor	Scale
Thinking about the math/ELA interim assessments you administered this year, please rate the level of rigor of test items compared to:[1]	**Much less rigorous - Much more rigorous**
a. Your state's math/ELA standards.	(5 point scale)
b. Your state's end-of-year math/ELA assessment.	
c. Your school or district's math/ELA curriculum.	

Data & Reporting Satisfaction	Scale
Thinking about the student interim assessment data you receive, please rate how satisfied you are with each of the following:	**Very Dissatisfied - Very Satisfied**
a. The time it takes to receive students' interim assessment scores.	(5 point scale)
b. The clarity of the data you receive.	

Support Satisfaction	Scale
How satisfied are you with the overall quality of the support you receive from your school or district in the following areas:[1]	**Very Dissatisfied - Very Satisfied**
a. Analysis of interim assessment data.	(5 point scale)
b. Setting student performance goals.	
c. Timeliness of response to questions.	
d. Improving instructional practices to meet students' needs.	

Exhibit B.2. Question Stem, Survey Items, and Response Scale for Each Teacher-Reported Survey Scale or Index, Continued

Leader Instructional Practices	Scale
How often did you or another instructional leader engage in the following activities during the school year?	**Never - Weekly or More**
a. Observe the instruction of individual teachers?	(5 point scale)
b. Participate in grade-level/content area team meetings?	
c. Implement activities that encourage teachers to reflect on their instructional practice?	
d. Meet with teachers to discuss students who are not meeting grade level expectations?	
e. Meet with teachers to discuss their unit and/or lesson plans?	
f. Coach teachers in planning backwards from standards?	
g. Coach teachers in using student assessment data to monitor student progress?	
h. Coach teachers in developing plans to re-teach content to struggling students?	
i. Coach teachers in evaluating the effectiveness of their re-teaching strategies?	

Data Review	Scale
Over this past school year, how often have you reviewed interim assessment data:	**Never - Weekly or More**
a. Independently?	(5 point scale)
b. With your leadership team?	
c. with teachers in your school?	

IA/Data Use Confidence	Scale
How confident are you in using student assessment data to perform each of the activities below:	**Not at All Confident - Highly Confident**
a. Set challenging yet attainable goals for student achievement school-wide	(5 point scale)
b. Set challenging yet attainable goals for student achievement at each grade level	
c. Compare your school's performance to that of other schools	
d. Examine trends in your school's performance over time	
e. Evaluate the performance of individual teachers in raising student achievement	
f. Identify struggling students in need of instructional support	
g. Evaluate the effectiveness of instructional programs	
h. Lead teachers in analyzing student assessment data	

[1] The treatment survey specifically referenced the interim assessments and support provided by the Achievement Network.

Appendix C. School Leader and Teacher Survey Impact Tables

Table C.1. Year 2 Impacts of ANet on Leaders' Perceptions and Practices

Survey scale	Impact estimate		*SE*	*p* value	*n*
Math rigor	1.08	‡	0.253	0.000	51
Math alignment	0.26		0.301	0.400	51
ELA rigor	1.23	‡	0.213	0.000	51
ELA alignment	0.29		0.305	0.349	51
Data and reporting satisfaction	1.08	‡	0.254	0.000	50
Support satisfaction	0.89	‡	0.258	0.002	51
Leader instructional practices	0.11		0.282	0.708	53
Data review	0.46		0.271	0.101	51
Confidence using data	0.31		0.309	0.318	51

Note. Models include the reduced survey impact sample of 67 schools (34 treatment and 33 control). Impact estimates include leaders who completed at least some portion of the survey, gave consent, and their matched-pair school leader also responded to the survey. Each scale is regressed on treatment assignment, school-pair dummies, and a data collection wave dummy. Scales were standardized within the leader sample; therefore estimates are in *SD* units.

‡ $p < 0.01$.

Table C.2. Year 2 Impacts of ANet on Teachers' Perceptions and Practices

Teacher survey scale	Impact estimate		*SE*	*p* value	*n*
Math rigor	0.57	‡	0.074	0.000	488
Math alignment	–0.29	‡	0.107	0.009	491
ELA rigor	0.56	‡	0.073	0.000	514
ELA alignment	–0.08		0.078	0.326	518
Data and reporting satisfaction	0.53	‡	0.092	0.000	580
Support satisfaction	0.27	‡	0.088	0.003	586
Leader abilities	0.21		0.110	0.062	616
Data review	0.38	‡	0.115	0.001	588
Data use	0.24	‡	0.088	0.007	585
Confidence using data	0.09		0.082	0.274	588
Instructional differentiation	–0.08		0.065	0.245	608

Note. Models include the reduced survey impact sample of 67 schools (34 treatment and 33 control). Each scale is regressed on treatment assignment, school-pair dummies, and a data collection wave dummy. Models are cluster-adjusted. Scales were standardized within the teacher sample; therefore estimates are in *SD* units. Source: Year 2 teacher surveys (treatment and control).

‡ $p < 0.01$.