All Together Now: The Impact of Team-Based Problem-Solving on Teacher Learning and Effectiveness

Robert Shand

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

All Together Now: The Impact of Team-Based Problem-Solving on Teacher Learning and Effectiveness

Robert Shand

Schools face a great challenge in recruiting and retaining quality teachers, given the documented importance of, variability in, and difficulty observing and predicting teacher quality. One option schools have is to identify what more effective teachers do and use that information to train less effective teachers to get better. Unfortunately, there is little empirical support for much traditional teacher training, as measured by gains in student test scores. Models of collaborative, team-based learning – such as Professional Learning Communities and Japanese lesson study – have been widely touted, and there is some evidence that they may be effective in certain contexts. Economic theory suggests this could be because of peer monitoring, peer pressure, specialization, knowledge-sharing, or market failure in pre-service training, particularly if learning to teach is primarily experiential. However, not all collaboration is good due to concerns about free-riding and substituting for more productive individual activity, so unbridled enthusiasm for collaborative professional development may need to be tempered.

This dissertation examines the effectiveness of a specific form of teacher collaboration in the form of inquiry teams, groups of teachers and administrators jointly engaged in action research projects with the aim of uncovering innovative instructional strategies and sharing effective approaches. It takes advantage of the phase-in of teams, eventually to all teachers in a large, urban school district in the northeastern United States from 2007-2010 to estimate the results of three natural experiments using difference-in-differences and instrumental variables approaches. The effects of teamwork on teacher value-added, teacher retention, and student test scores are small and sensitive to year, specification, and outcome, although results are mostly positive and occasionally statistically significant, suggesting that overall effects are potentially positive but modest at best. Further examination of heterogeneity and four qualitative case studies of teams suggest that small average effects mask considerable differences in team processes, and that under certain conditions, inquiry team work may be far more effective. A cost analysis reveals that, although it is costly to do inquiry work well, given the low-intensity of average treatment and the large number of students affected, the benefits of inquiry work could exceed the costs if the policy were more targeted. Overall, the policy recommendation is to temper unqualified enthusiasm about teacher collaboration, as without appropriate structures and supports it has little measurable effect on the outcomes examined here. As a policy lever, a universal mandate to participate on collaborative inquiry teams is unlikely to be effective or pass a cost-benefit test. Nonetheless, smaller-scale, higher intensity forms of collaboration that allow for more active leadership support and participation may be more promising, and more cost-effective than alternative forms of professional development, particularly for some sub-groups of teachers such as those in their first year of teaching.

CONTENTS

List of Tables……………………………………………………………………………………………….iv

List of Figures……………………………………………………………………………………………...vii

Chapter 1 Introduction 1

The Intervention 4

Research Questions 11

Conceptual Framework 13

Economic Model 13

Team Learning Model 18

Chapter 2 Literature Review 23

Defining and Measuring Teacher Effectiveness 23

Pre-service Training Programs..24

Value-added Measures and Incentives..26

Retention and Working Conditions...28

In-service Training 31

Experimental and Quasi-Experimental Studies...31

Non-experimental Studies 38

Returns to Experience and Diffusion of Expertise..42

Collaboration and teamwork in education and other sectors 44

Individual and Organizational Learning – Human and Social Capital Interactions 53

Chapter 3 Data and methods: Quantitative Strategy...57

Data 57

Simple Model 67

Model #1: Grade Switchers....71

Model #2: First-year Teachers 72

Model #3: Instrumental Variables....74

Descriptive Analysis of Heterogeneity 77

Chapter 4 Methods and Data: Qualitative and Cost Analyses....78

Qualitative Analysis....78

Cost Analysis 83

Chapter 5 Effects of Inquiry Team Participation on Measures of Teacher Productivity....86

Simple Model....86

Model #1: Grade Switchers....91

Model #2: First-year Teacher Model 98

Model #3: Instrumental Variables....106

Interpretation of Results....116

Chapter 6 Analysis of Mechanisms and Heterogeneity....118

Potential Mechanisms 118

Defining and Measuring Quality Teamwork 120

Descriptive Analysis of Heterogeneity and Quality....125

Incorporating Heterogeneity in Main Models....131

Chapter 7 Qualitative Case Studies of Teams in Action....144

Coding Scheme 145

Findings....146

Prototypical Case....146

School A 147

School B 152

School C 158

School D 163

Cross-Case Analysis 166

Chapter 8 Cost-benefit Analysis 173

Cost Analysis 173

Benefit-Cost Analysis 179

Sensitivity Analysis..181

Discussion 183

Chapter 9 Conclusion..184

Summary of findings..184

Policy implications...185

Limitations and suggestions for further research 186

Contribution 189

References..191

Appendix A. Additional Instrumental Variable Diagnostics and Results 203

Appendix B. Additional Heterogeneity Results...210

Instrumental Variables Quality Estimates..214

Appendix C. Qualitative Coding Scheme 217

List of Tables

Table 1-1 Team Learning Framework 20

Table 3-1 Outcomes by model and year 59

Table 3-2. Descriptive Statistics on Teachers, 2007-2010 60

Table 3-3. Descriptive Statistics on Value-Added, 2008-2010 63

Table 3-4. Summary Statistics on Test Score Outcomes, 2008-2010....66

Table 4-1. Description of Case Study Sample 82

Table 5-1 Summary Statistics on Covariates in Team-Switcher Model....87

Table 5-2 Ordinary Least Squares estimates of the association between Team Participation and value-added 89

Table 5-2 Fixed effects estimates of the effects of teamwork on value-added....91

Table 5-4 Difference-in-differences estimate of the effect of teamwork on value-added 93

Table 5-5 Difference-in-differences estimate of effect of teamwork on value-added, grade switchers 94

Table 5-6 OLS and difference-in-difference estimates of the effects of teamwork on other value-added measures...96

Table 5-7 Descriptive statistics on first-year teacher model, 2007-2008 99

Table 5-8 OLS estimates of association between team participation and retention, first-year teachers, 2007-2008 .100

Table 5-9 OLS estimates of association between team participation and value-added, first-year teachers, 2007-2008 101

Table 5-10 Team participation and retention of first-year teachers, 2008-2009 102

Table 5-11 Effects of team participation on value-added, first-year teachers, 2008-2009....103

Table 5-12 2009-2010 Retention 104

Table 5-13Effects of team participation on value-added, first-year teachers, 2009-2010....106

Table 5-14 OLS estimates of effects of Inquiry Team participation on math and ELA gain scores, K-8 schools, 2008-2009 108

Table 5-15 Instrumental Variables estimates of effects of Inquiry Team participation on math and ELA gain scores, K-8 schools, 2008-2009 110

Table 5-16 Quadratic Instrumental Variables estimates of effects of Inquiry Team participation on math and ELA gain scores, K-8 schools, 2008-2009 110

Table 5-17 Instrumental variables diagnostics, 2008-2009, K-8 Schools 111

Table 5-18 OLS estimates of association between teamwork and high school outcomes, 2008-2009. 112

Table 5-19 Instrumental Variables (2SLS) estimates of effects of Inquiry Team participation on high school outcomes, 2008-2009 114

Table 6-1 Rubric used to assess intensity of implementation of inquiry teams...123

Table 6-2 Descriptive statistics on team quality and intensity of implementation 127

Table 6-3 Correlations (2009-2010) 128

Table 6-4 Association Between team quality And retention, all teachers, 2007-2008..132

Table 6-5 Association Between team quality And retention, first-year teachers, 2007-2008 133

Table 6-6 Association Between team quality And value-added, all teachers, 2007-2008 133

Table 6-7 Association Between team quality And value-added, first-year teachers, 2007-2008..............................134

Table 6-8 OLS estimates of association between test score growth and team quality measures, K-8 135

Table 6-9 OLS estimates of association between test score growth and team quality measures, HS 136

Table 6-10 Association between indicators of team quality and retention, all teachers. 137

Table 6-11 Association between indicators of team quality and retention, first-year teachers. 138

Table 6-12 Relationship between team quality indicators and value-added, all teachers 139

Table 6-13 Relationship between team quality indicators and value-added, first-year teachers...............................139

Table 6-14 OLS estimates of association between test score growth and team quality measures, K-8 141

Table 6-15 OLS estimates of association between test score growth and team quality measures, HS 141

Table 7-1 Code application by school and co-occurrence 168

Table 8-1 Costs of teams by intensity 178

Table 8-2 Benefit-cost analysis of inquiry teams 181

Table A- 1 Instrumental variables diagnostics, high schools, 2008-2009 203

Table A- 2 Instrumental variables diagnostics, K-8 schools, 2008-2009 204

Table A- 3 Instrumental variables diagnostics, high school, 2009-2010...205

Table A- 4 OLS estimates of the association between teamwork and test score outcomes, K-8 schools, 2009-2010 206

Table A- 5 Instrumental variables results, effects of teamwork on gain scores, K-8 schools, 2009-2010................207

Table A- 6 OLS estimates of association between teamwork and high school outcomes, 2009-2010 208

Table A- 7 Instrumental variables estimates of association between teamwork and high school outcomes, 2009-2010 208

Table B- 1 Retention – 100 teams coded for quality, all teachers 210

Table B- 2 Retention 100 teams, first year teachers 210

Table B- 3, Value-Added, 100 hand-coded teams, all teachers 210

Table B- 4, Value-Added, 100 hand-coded teams, first-year teachers 211

Table B- 5, Retention, first-year teachers, hand-coded quality as covariate 211

Table B- 6, Value-added, hand-coded quality as covariate, all teachers 211

Table B- 7, Value-added, first-year teachers, hand-coded quality as covariate 212

Table B- 8, IV estimates, team indicator on growth, K-8 212

Table B- 9, IV estimates, team indicator on growth, high school 212

Table B- 10 IV estimates, team indicator on growth, K-8 213

Table B- 11, IV estimates, team indicator on growth, HS 213

Table B- 12. Association between indicators of team quality and retention, first-year teachers. 216

Table C- 1 Coding scheme 217

List of Figures

Figure 1. Diagram of Inquiry Team Process 6

Acknowledgements

I am deeply grateful to my advisor, Judith Scott-Clayton, whose continuous support, encouragement, and feedback have been critical to the completion of this project. Her keen ear for research questions that are big enough to be interesting and important, but focused enough to be answerable, has greatly shaped this work and my approach to policy research. I would also like to thank my second reader, Henry Levin, the most humanistic economist I know, for his support and very helpful comments throughout the dissertation process. Anand Marri has given generously of his time and provided me numerous opportunities to work on projects bridging research, policy, and practice, providing both intellectual and financial support for my research and keeping the work grounded in what would be most useful to practitioners. Thank you as well to Peter Bergman and Randy Reback, whose close reading and sharp comments and questions made this work considerably stronger.

I am also grateful to many colleagues at Teachers College and Columbia who have provided support in countless ways, most especially Maureen Grolnick and colleagues at the Understanding Fiscal Responsibility and Cowin Financial Literacy Projects; Brooks Bowden, Fiona Hollands, Clive Belfield, Yilin Pan, Barbara Hanisch-Cerda, Meridith Friedman and others at the Center for Benefit-Cost Studies of Education; and James Liebman, Alexa Shore, Charles Sabel, Elizabeth Chu, Jessica Wallenstein, Maren Hulden and others at the Center for Public Research and Leadership, who also collaborated on the qualitative project that was the genesis of this work.

Thank you as well to Jeff Henig and participants in the departmental dissertation seminar and the Economics and Education research workshop, for comments and suggestions; to Jonah Rockoff for very helpful discussions on early drafts of this work; and to the Zankel Urban

Fellowship, the Educational Policy and Social Analysis Department at Teachers College, the Institute for Education Sciences, the NoVo Foundation, TC Trustee Joyce Cowin, and the Peter G. Peterson Foundation for funding projects and grants that financed my graduate study and enabled me to complete this project. I am also thankful to the school district for sharing the data, and to the principals and teachers in the case study schools for welcoming us and candidly sharing their inquiry team experiences.

Last, but certainly not least, many, many thanks to my family and friends who supported and pushed me throughout this work, most especially my parents and my loving and patient wife, Michelle.

Dedication

To the teachers I have had and known, who endeavor to make the world better, especially my wife Michelle Leonor, the best teacher I know.

Chapter 1 INTRODUCTION

It has become a mantra among educational policymakers that the quality of individual teachers is the single most important in-school determinant of educational outcomes and further that teachers vary substantially in their abilities to increase student learning; significant research findings support this view (Rockoff, 2004). Nevertheless, very few observable characteristics of individuals entering the teaching profession have significant power to predict a teacher's future effectiveness (Kane, Rockoff & Staiger, 2008; Goldhaber & Anthony, 2004; Palardy & Rumberger, 2008). This information gap places significant burden on schools to be able to identify more and less effective teachers and find ways to help less effective teachers improve, or to replace them with more effective teachers.

Policies to increase teacher effectiveness fall broadly into three categories. The hypothesized effectiveness of policies under each category will depend, in part, on underlying beliefs about the nature of the education production function. Critically, the most effective policies will depend on whether the optimal mix of educational inputs and processes are fixed, in which case teachers must adapt to optimize student learning, or whether they are variable based on school, teacher, and even individual student-level factors. One school of thought makes relatively few assumptions about the nature of educational production and argues that teachers themselves are best poised to uncover the most productive processes and inputs for maximizing student learning, which may vary considerably based on context. Therefore, these policies aim to maximize teacher effectiveness by providing incentives for teachers to uncover the most productive educational techniques, adapted to their own areas of expertise and the unique learning needs of their students, on their own. This is achieved by better measuring their individual contributions to student learning and tying their compensation to measured performance, through some combination of incentive pay, heightened standards for achieving tenure protections, and/or increased risk of performance-based dismissal. A second school of thought holds that recruitment of quality teachers is not a problem, but that retention of high-quality teachers is problematic because of better labor market alternatives for the most effective teachers. Working conditions or compensating differentials play a critical role for this mechanism, as those who leave teaching often report poor working conditions as a more important reason than low salary. Further, working conditions tend to be worst in schools that serve the students with greatest needs, exacerbating inequities in access to quality teachers (Lankford, Loeb & Wyckoff, 2002). Policies in this category aim to improve retention of teachers, particularly of the highest-quality teachers in the highest-need schools, by improving working conditions and better compensating for poor working conditions. A final set of policies aims to directly increase the quantity of or improve the quality of inputs to the educational production function to increase teacher effectiveness, most commonly by raising human capital through in-service training.

Variations on incentive policies based on value-added measures are now being tested in several jurisdictions, in part in response to Race to the Top grants that encouraged such experimentation, but their long-term effects on student learning and teacher recruitment and retention are still unknown. Traditionally, professional development or in-service training was the most common policy to increase teacher performance. Nonetheless, very few of the professional development programs that have been subject to rigorous evaluation have shown evidence of effectiveness in improving student learning outcomes. A number of investigations of teacher attitudes on professional development reveal that teachers often view their training as irrelevant to daily practice and lacking appropriate coherence and follow-up (Jacob and Lefgren, 2004; Garet et al., 2008; Garet et al., 2010; Darling-Hammond and Richardson, 2009). Despite these limitations, educational practitioners and policymakers remain committed to professional development; estimates of the costs of professional development in the United States range from 3.3 to 5.7 percent of total educational expenditures[1] of $632 billion in 2011[2], or $20 to $36 billion annually.

Several school districts and teacher preparation programs have adopted models of ongoing teacher training based on structured collaboration. Notable among these examples is the nation of Finland, which is widely lauded for its performance in international assessments (Sabel et al., 2010). These models, which in many instances emphasize classroom-based action research, data analysis, and adaptation of instructional services to the unique needs of the students served, may address several of the inadequacies identified in traditional professional development. Teacher collaboration could lead to enhanced or more efficient curriculum development through joint production of instructional plans and materials, school improvement through better sharing of information among front-line workers and increased teacher leadership, and teacher professional development through knowledge sharing, learning from colleagues' experience, or peer pressure (Y. Goddard, R. Goddard, & Tschannen-Moran, 2007). Further, given the resources currently invested in in-service teacher training, increasing teacher collaboration may be a comparatively cost-effective method of increasing teacher effectiveness.

However, as previous literature on workplace collaboration reveals, not all collaboration is meaningful or fruitful. Productive collaboration must therefore be disentangled from activities that distract from or even actively impede instructional improvement. Research from organizational theory on team-based problem-solving, learning, and production, associated with

[1] https://www2.ed.gov/pubs/CPRE/t61/t61c.html
[2] http://nces.ed.gov/fastfacts/display.asp?id=66

the Japanese concept of *kaizen*, could help improve professional development (Dyer and Nobeoka, 1998), especially given that there is some evidence for positive teacher peer effects, meaning that having more effective colleagues tends to make teachers more effective (Jackson & Bruegmann, 2009). Evidence of persistent improvement as a result of qualitative evaluation and feedback suggests that some combination of context-specific information and peer pressure can lead to improvements in teacher effectiveness (Taylor and Tyler, 2012). In sum, the current research and policy consensus seems to be that teachers are extremely important, but we do not know with great confidence how to help teachers get better, and broadly speaking, current efforts are not working very well despite enormous expense. There are some indications that increasing the quantity of and improving the quality of collaboration could alleviate some of these concerns, and growing enthusiasm for teacher collaboration as a vehicle for school improvement, but relatively little causal research on the effects of any particular collaboration policy.

The Intervention

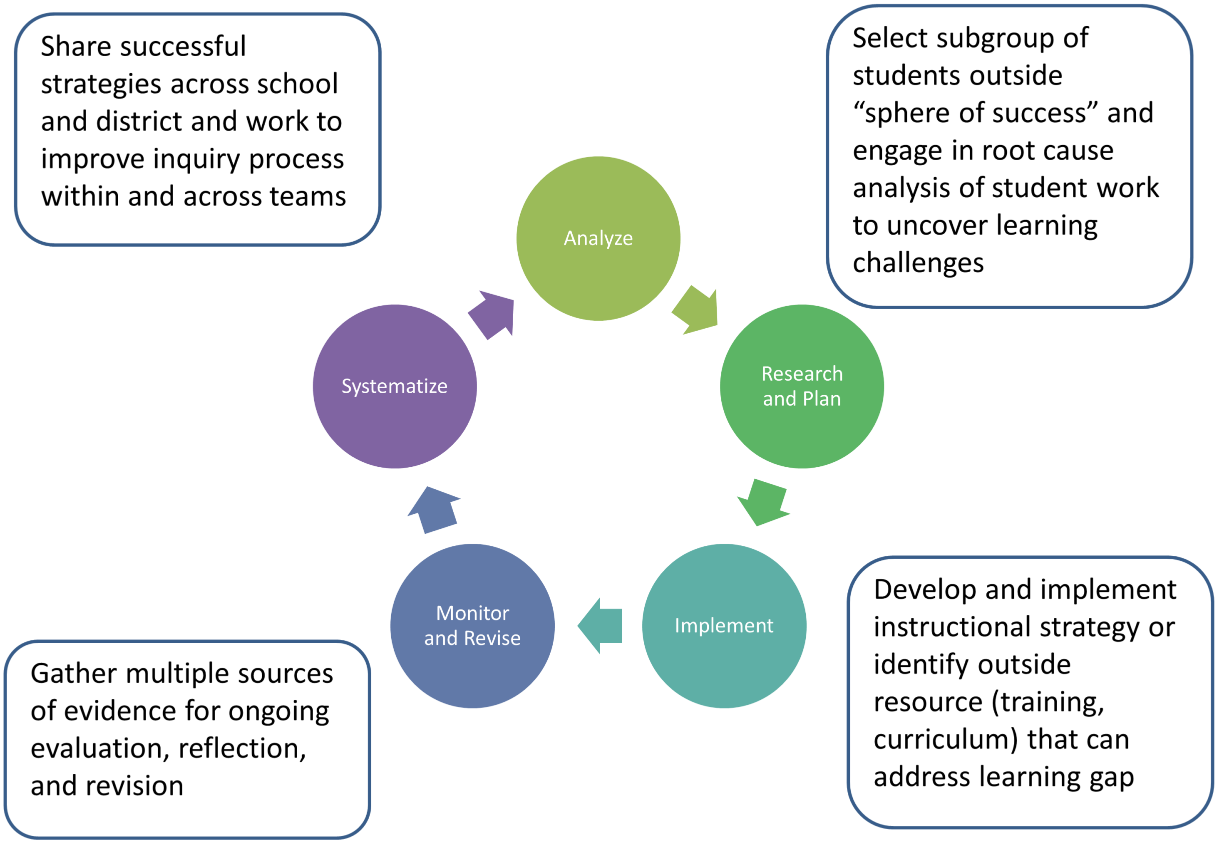
This study examines a policy intervention mandating teacher participation on inquiry teams, a particular form of teacher collaboration focused on action research, problem-solving, team learning, and organizational learning.[3] The intervention took place in a large, urban school district in the northeastern United States primarily between 2007 and 2012. Since then, although some schools still have inquiry teams and teachers still engage in many forms of collaboration, the emphasis at the district office has shifted to implementation of the Common Core learning standards and a new teacher evaluation system. Although there were tweaks to the process over the ensuing years, the basic notion of inquiry teams remained the same; as described by the school district, inquiry teams are groups of teachers engaged in structured work focused on

[3] The activities of inquiry teams have also been referred to as "collaborative inquiry" and "strategic inquiry."

analyzing the learning needs of small groups of students using a rigorous approach based in data. The inquiry team initiative was designed to identify and develop innovative, research-based instructional approaches with the aim of immediate, small-scale instructional improvement that would lead to wider organizational learning and change.

Inquiry teams were, by design, both structured and flexible. There were relatively few parameters surrounding who could be on a team or what a team could focus on, beyond the requirement in early years that teams selected a small subgroup of approximately ten to fifteen students and a narrowly defined skill to help them focus their work. Teams varied considerably in size, composition, and focus, but fell broadly into three categories: teams that focused on students in a particular grade level, teams that focused on a particular subject area, and teams that focused on a specific, high-needs subgroup such as English language learners or students with disabilities. Inquiry was defined as an iterative cycle, as shown in Figure 1, whereby a team used data and root cause analysis to identify and uncover underlying causes of learning breakdowns, sought instructional changes that could address this cause, developed precise assessment instruments to monitor progress toward measurable learning goals, and spread successful strategies to other teachers in the school (Panero and Talbert, 2013).

FIGURE 1. DIAGRAM OF INQUIRY TEAM PROCESS


Share successful strategies across school and district and work to improve inquiry process within and across teams
Analyze
Select subgroup of students outside "sphere of success" and engage in root cause analysis of student work to uncover learning challenges
Systematize
Research and Plan
Monitor and Revise
Implement
Develop and implement instructional strategy or identify outside resource (training, curriculum) that can address learning gap
Gather multiple sources of evidence for ongoing evaluation, reflection, and revision

Source: Adapted by the author from descriptions of inquiry team process by sponsoring school district.

At the district level, the inquiry team policy consisted of a phased mandate over three years that ultimately all teachers would participate on at least one team. During the first year, 2007-2008, each school was required to assemble one pilot team. Principals were expected to be members of the team, although they did not always participate in practice, and recruited teachers and other staff to join the team via a job posting. Teachers who applied and were selected often expressed an interest in using data and learning more about how data could inform instruction (Talbert, 2011). Therefore, teams in the first year exhibited two types of selection – teachers self-selected onto teams and principals encouraged teachers to apply and chose among those that did.

During subsequent years, schools were required to have multiple inquiry teams, with the goal of 90% of teachers participating by the 2009-2010 school year. One of the ultimate goals of the policy was to integrate inquiry into the "fabric of the school," fundamentally changing professional development, teacher meetings, and teacher leadership to shift the focus to the unique needs of the school and the students it serves, rooted in data (CPRE, 2008). The risks of selection bias therefore declined over time, as ultimately nearly all teachers were required to participate on teams.

The goals of the policy were essentially threefold: to improve learning outcomes of the teachers' current students, to improve instruction for all students by improving teachers' human capital in the classroom, and to increase organizational effectiveness by developing teacher leadership and providing structured avenues for knowledge-sharing and organizational learning. To help achieve these goals, the district provided substantial training, support, and resources in the first year, which declined as the initiative spread and in the face of budget constraints. Schools were required to designate a teacher or school leader as a Data Specialist, to receive additional training in the district's data and accountability systems, including the inquiry team initiative. Principals also received training from district leadership and were expected to share what they learned with their staffs. The district provided schools with additional funding for teacher overtime to support after- or before-school meetings and laptops for teachers to use to facilitate data analysis. Finally, senior district leaders, often experienced former principals, were designated Senior Achievement Facilitators (SAFs) and provided hands-on coaching to schools on various data and accountability systems, including inquiry teams. SAFs attended some inquiry team meetings at most schools, provided feedback on the process, answered questions from the principal and Data Specialist via telephone or email, and provided encouragement to teams to move along in the process. Several teams reported the support of SAFs as critical in the perceived success of the initiative (CPRE, 2008).

Textual analysis of the data teams reported on their activities from a sub-sample of teams, described in greater detail in Chapter 6, provides some descriptive trends on how teams organized themselves, what actions they undertook, potential issues of selection bias in team composition, and the obstacles even relatively strong teams faced that may have limited the overall success of the policy. For each of the three years under study, about half of all teams focused on a single grade level, with the other half focused on a subject area or demographic sub-group of students across multiple grades. The large majority – 59% of teams in 2007-2008, 72% in 2008-2009, and 62% in 2009-2010 – focused on English language arts (ELA) as a subject area.

In the first year, teams described team composition and the process by which teams were selected, which often entailed a combination of teachers volunteering and principals recruiting team members. Many teams included a number of non-teacher members, such as administrators, counselors, and other professionals, in addition to classroom teachers and specialists in special education and English as a Second Language (ESL). One possible mechanism by which team participation can enhance teacher and school effectiveness is by specialized professionals sharing expertise through the team. Teams mentioned experience with data analysis and the school district's data and accountability technology systems as criteria for team participation.

Several teams identified a subgroup of students within a subject and within or across grade levels that exhibited persistently low or declining performance on a state assessment in math or ELA. Teams then administered follow-up assessments to more precisely diagnose learning issues in students, uncovering gaps in areas such as spelling and decoding skills among first graders, making inferences among third graders, vocabulary among middle school students, and writing and algebra among high school students. Teams reported using several strategies, including creating portfolios of written work, administering supplemental instruction through small-group tutoring for targeted students, and testing new curricula and materials, to address these learning needs. With some exceptions, including a team that successfully addressed communication skills among students with autism, teams struggled with pacing and follow-up, as they spent much of the year diagnosing student learning needs, leaving little time to experiment with potential solutions. Some teams explicitly noted challenges in determining how to proceed from the diagnostic stage, either because the learning needs of their targeted subgroup of students were too broad and diverse, or because they lacked time and resources to do so. The program grew out of a school-leader training program developed by Baruch College and New Visions for Public Schools known as the Structured Apprenticeship Model (Talbert, 2011). The initiative shares some common features with two other well-known examples of structured teacher collaboration, professional learning communities (PLCs) and Japanese lesson study. In particular, the central idea of PLCs is that learning about teaching is fundamentally experiential and best transmitted through a structured process involving others with shared experience (Buysse, Sparkman & Wesley, 2003). Inquiry teams lie between PLCs and Japanese lesson study on a continuum of how structured and prescriptive the collaborative process is; like lesson study, inquiry teams are encouraged to follow protocols and keep their work tightly focused, but there is more room for experimentation and choice in terms of what that focus will be, along the lines of PLCs.

The intervention was part of a larger package of reforms, the central philosophy of which was an “autonomy for accountability” exchange. Schools, principals, and teachers were granted greater authority over budgets, hiring and staffing decisions, curriculum, and professional development, but were expected to strategically use their authority to achieve student learning targets as measured by growth on standardized tests. Schools that consistently failed to demonstrate growth in student learning were subject to sanction, up to and including closure, and schools that consistently showed outstanding growth received financial rewards for the principal and, in some cases, the teachers. District leaders saw inquiry as a critical tool for building teacher and school capacity to make use of their greater autonomy to close learning gaps. Operating under the theory that traditional professional development was too general and decontextualized to be effective at increasing teacher productivity or student learning, the district envisioned inquiry as a tool to help teachers make use of new data and accountability tools and shift their focus to individualized learning needs of students. The inquiry team initiative was one effort by the central office to promote capacity building and knowledge sharing from within to help schools accelerate student learning.

There is some extant literature on the inquiry team initiative. The Consortium for Policy Research in Education (CPRE) at Teachers College, Columbia University engaged in two implementation studies (CPRE 2008 & 2010) in conjunction with the school district. They found that inquiry teams in the first year generally implemented with fidelity, following the model laid out by the school district. Team members reported appreciation for the level of support provided, including funding for team meetings after school and training by SAFs. The variable seemingly most related to implementation quality was the role of the principal; teams where the principal played an active, but not overly prescriptive role were more effective overall than teams where the principal was either uninvolved or too directive. Despite early successes, teams did struggle with pacing, taking much of the year to analyze data to identify a target sub-group of students and very focused instructional skill, leaving little time to experiment with multiple cycles of instructional strategies and assessments. Teams also reported wanting more time and support for teamwork.

A follow-up analysis in 2010 reinforced the critical, yet difficult to balance, role of the principal and the need for protected time. As the initiative spread throughout the school, the report notes some shift in focus away from using inquiry as a tool to directly impact student achievement through instructional innovation and toward using inquiry as a teacher development tool to build capacity in analyzing data and differentiating instruction. The report also notes the integration of inquiry into other work, including the school's general improvement goals and pre-existing team structures such as grade-level and subject-area department meetings.

More recent inquiry work, in the same district but studying the intervention at a later time period with fewer prescriptive mandates from the central office, has uncovered more divergent findings. Talbert and Panero (2013) expanded upon Talbert's account of the history of inquiry teams, discussing several cases of successful teams that, through disciplined research on narrowly focused instructional skills, identified gaps in the writing curriculum as a root cause of student skill deficiencies. In contrast, Chu et al. (2012), in a study in which the qualitative data for the case studies in this dissertation were collected by a research team including me, found that teams no longer maintained focus on focused skill gaps among sub-groups of students. Instead, they aimed for general teacher capacity building, particularly in light of the implementation of the Common Core standards and a new teacher evaluation system.

RESEARCH QUESTIONS

This study seeks to expand what is known about teacher collaboration and teacher learning through an in-depth examination of the mechanisms by which teachers learn from one another and the factors that are associated with relatively more or less productive collaboration in one context using three empirical approaches. The data for the three approaches are drawn from the same district, intervention, and general population of teachers, but different years and different samples of teachers, so results across questions reflect in part evolution of the intervention over time. Although the three empirical analyses are separate, they logically connect in an explanatory sequential, mixed methods design (Creswell & Clark, 2011, p. 81-90), in which qualitative methods follow quantitative analyses to provide context and possible explanations for patterns of results.

This study utilizes the phased nature of the policy mandate to participate on inquiry teams estimate the effect of team participation on teacher retention, student learning, and teacher value-added in a series of difference-in-difference estimates to address concerns about selection of teachers onto teams. The quasi-experimental approach, combined with the use of administrative data in which teams reported their activities, as opposed to self-reported survey data on collaboration as has commonly been used in prior literature, represents a significant contribution. More importantly, however, this study provides a framework for considering the substantial measurement issues that arise when assessing teacher collaboration, which are difficult to capture and display significant heterogeneity across teachers, teams, and schools. While the small amount of signal relative to noise in the data may be a concern in evaluating the underlying value of teacher collaboration as a concept, in the case of a policy mandating collaboration, effects that are obscured by significant heterogeneity and weak implementation represents a significant finding. Whether or not to collaborate is a manipulable policy lever, whereas the quality, intensity, and authenticity of that collaboration is not and may require improved targeting, training, and support in order to have an effect.

This dissertation therefore contributes to the literature on teacher effectiveness, teacher labor markets, and teacher training and collaboration by answering three broad research questions:

- Does mandating collaboration through participation on an inquiry team improve teacher effectiveness as measured by value-added scores, teacher retention, and student achievement? I will also examine whether goal setting, leadership involvement, and use of particular types of data predict heterogeneity in the effectiveness of teams.
- Through what processes do teams of teachers engage in collaborative inquiry? What team and teacher-level conditions are associated with indicators of teacher learning, such as evidence of changing attitudes, dispositions, or practice?
- What are the costs inquiry teams, and how do the costs compare to the estimated benefits of teams, measured in monetary terms?

Conceptual Framework

Economic Model

The basic economic concept underlying this study is the notion of an educational production function, which formalizes the relationship between educational inputs, such as prior student achievement, school characteristics, and teacher characteristics, and various outputs, most commonly student achievement as measured by standardized test scores. This dissertation deviates from common formulations of educational production in an important respect – rather than including teacher education and experience as proxies for human capital, the teacher human capital function over two periods is explicitly modeled to take into account ongoing training and interaction with colleagues as important determinants of the returns to experience. Therefore, student *i* in the classroom of teacher *j* in school *s* in time *t* will have the following achievement production function:

(1) $A_{ijst} = f(A_{i-js,t-1}, \bar{A}_{-ijst}, X_{ijst}, X_{-ijst}, Z_{st}, \theta_{jst})$,

where A_{ijst} refers to a student learning outcome, which could include test scores as well as other important outcomes, $A_{i-js,t-1}$ is prior achievement for student *i* in the class of teacher j ≠ j, $\bar{A}_{-ijst}$ are peer effects, X_{ijst} and X_{-ijst} are demographic covariates for the students and his or her peers, respectively, Z_{st} are school-level covariates such as resources and leadership, and θ_{jst}can be conceived as teacher human capital. Gary Becker (1964), among other economists, formalized the concept of human capital as the knowledge and skills accumulated through education, training, and experience, which enhance worker productivity.

Teacher human capital is often estimated in practice as teacher fixed effects or teacher value added, $\hat{\theta}$. This is itself a function of education, experience, on-the-job learning, and teacher peer effects, as well as unobserved underlying teacher characteristics, and is assumed to be concave, or increasing at a decreasing rate in each of those dimensions:

(2) $\theta_{jst} = g(educ_{jst}, exp._{jst}\,, \theta_{-jst}, learning_{js,t-1})$

Assume: $\frac{\partial\theta}{\partial g} > 0$; $\frac{\partial^2\theta}{\partial g} < 0$

In any given period education is assumed to be given and is treated as a constant; this assumption may not hold, for example, for teachers who are in the process of obtaining their Master's degrees in the current period. The relevant decision for teachers as agents in this study is the extent to which they invest in their own learning, or ongoing training as an individual or with colleagues. Since in most cases wages do not vary with productivity, teachers will select how to allocate working time across various activities according to their own utility, which they may derive from intrinsic satisfaction, esteem of colleagues or parents, increased job security or reduced fear of sanctions by employers, subject to the constraints of time and the cost of effort, which may vary by activity.

In each period, teachers must divide time between non-productive activities ("shirking,"), individual effort such as planning lessons, collaborative work focused on building collegial relationships with colleagues, and collaborative work focused on team-based learning. The budget constraint incorporates time and effort by including time weighted by effort, whereby more mentally taxing or less pleasant activities "feel" like they take more time. Under similar assumptions as noted above for the human capital function – that utility increases at a decreasing rate for each of these options, and costs increase at an increasing rate - the optimal solution will occur when the ratios of the marginal returns to each activity to the marginal cost are all equal. There is an additional assumption required that the returns to time invested in collaborative learning will not be immediately felt, and therefore the time allocation will further depend on teachers' discount rates. Note that this simple model abstracts from several important practical realities and constraints; for instance, how teachers allocate time at work will be at least in part determined by their supervisors, and the choice to collaborate will also depend on colleagues' decisions, as individual teachers obviously cannot collaborate alone. Nonetheless, this model describes how teachers make time allocation choices on the margin, and how much effort they allocate to various tasks within their workdays.

This model has several important implications and raises questions that will be addressed in the quantitative and qualitative research designs. The extent to which teachers working on teams will focus their efforts on more productive activities, including experiential learning, joint production of curricula and assessments, and idiosyncratic learning about instruction in their particular school context, will depend on the extent to which teachers on teams believe these goals to be attainable through teamwork. Due to data limitations – namely, that I only observe the individual teachers on teams in the final year of the initiative and that data available on individual teacher characteristics that may predict team participation and its effectiveness are limited – this model motivates the issue of heterogeneity in the effects of team participation, explored in more detail in chapter 6. The main empirical models and results, in chapters 5 and 7, are therefore mainly estimating the effects of teams on teacher effectiveness and student learning at the team level, or on teachers nested in teams, as opposed to predicting at the teacher level whether and how they will participate.

Therefore, variation in the effectiveness of teamwork will occur along dimensions that affect the marginal returns to teamwork and the marginal costs, relative to other possible uses of teacher time. Some predictions from this model that can be empirically tested include, for example, that the marginal returns to teamwork will be higher for first-year teachers, teachers who are new to a school, and teachers switching to a new grade and subject, as those teachers will have the greatest incentives to learn new content and skills and the longest time horizon for future payouts to current investment. Similarly, changes to curricula, assessments, or the accountability context that teachers face could induce additional teamwork by requiring teachers to reinvest in their skills; unfortunately, many such changes, for example implementation of the Common Core standards, took place after the sample period for the quantitative data in this study, but can be observed descriptively in the qualitative data, which were collected later. Other factors that the literature suggests could impact the returns to teamwork or the costs of teamwork, including the size of the team, the homogeneity of the team in terms of teacher beliefs and learning needs, the extent to which leadership supports teamwork, and the complementarity of teacher skills across the team, can be tested, as well. One important distinction between teamwork among teachers and in other sectors is that in most settings under which teams have been studied in an economic framework, productivity of each individual member is at least partially directly observable by other members of the team, and in many cases, production processes are in fact joint. In education, while student learning undoubtedly depends upon the contributions of several teachers across grades and subjects, and team teaching scenarios do exist, for the most part the actual work of teaching is performed individually, so team members can observe one another's productivity only indirectly.

Further, the predictions of this model and the empirical literature on teacher collaboration, teamwork, and ongoing training may be surprising, given the predictions of the standard human capital model that returns to investment in human capital are maximized when such investments occur before work begins, and that firms will generally only invest in firm-specific human capital (Becker, 1964). These contradictory findings raise several possibilities that will be tested in this dissertation; specifically, the likelihood of participating on a team and the benefits of doing so may vary according to whether the team is focused on a group of students or a content area. The former could indicate that teaching is highly context-specific, whereas the latter may indicate either market failure in the quality of pre-service training or changes in standards and assessments. The empirical tests noted above on how the likelihood of team participation and the effects of team participation vary by years of experience overall and in a particular school can also help disentangle these mechanisms.

Finally, teachers' decision to participate in a team relies upon their perceptions of the costs and returns to teamwork, which may differ from the actual costs and returns in three cases. First, teachers may engage in hyperbolic discounting, in which teachers have a strong preference for returns in the present and aversion to costs in the present. Secondly, there may be asymmetric information, in which teachers do not know the returns to teamwork. Finally, teachers may be risk averse, and therefore unlikely to participate on teams or unwilling to change their teaching practice to something that is unfamiliar but potentially better, instead sticking with what is familiar and working but possibly sub-optimally. Without measures of teachers' discount rates and attitudes about risk, it is difficult to test these hypotheses with the given data, but I consider them in the qualitative analysis.

TEAM LEARNING MODEL

The process by which teams operate, the definitions of team processes and team learning, and some of the outcomes of team process may not be observable in the economic and framework described above. Therefore, I complement the economic model with a conceptual framework for team processes, particularly focused on team learning and problem-solving, to examine using a qualitative methodology. Examples of the types of more nuanced team characteristics and processes to be studied in this component of the work include Hoegl and Gemuenden's (2001) dimensions of effective teamwork: quantity and formality of communication, coordination of effort, balance of contributions across team members, mutual support, effort, and cohesion. The relative effectiveness of a teacher team may depend on nuanced aspects of how teachers interact. These interactions could promote varying degrees of attitudes toward conflict, including unproductive avoidance or acrimony or more productive discussion, as well as different levels of inclusion or exclusion across a community. Other research has suggested that group size, the role of school leaders, and the amount of time devoted to teamwork are important determinants of the quality of the team process (Scribner, 1999; Graham, 2007; Wayman, Midgley and Stringfield, 2006).

Therefore, based on part on the team learning model developed by Kasl, Marsick and Dechant (1997), I conceptualize teacher teamwork as a series of conditions, processes, and outcomes. While team process may evolve over time as conditions change and teams learn (including about the process of teamwork itself in a reflexive stance), teams may be characterized by overall modes of dynamics, problem solving, and learning at any given time, and do not necessarily proceed through various modes as discrete stages.

Conditions that determine team processes include:

- Structural features, such as team size, composition of team in terms of levels of experience and areas of expertise
- Team focus (e.g., whether the team's work is focused on a sub-set of students or all students in a particular subject or grade, or if the team has a grade-level or subject-area orientation)
- Time for collaboration
- Leadership support
- Individual and group process characteristics such as openness to new ideas, willingness to challenge norms and beliefs, and efficient processes for communication

The team's processes include:

- Identifying and defining instructional needs and issues that the group faces
- Analyzing root causes of instructional needs
- Identifying gaps in the group's expertise that may be inhibiting performance
- Locating and developing new instructional strategies or approaches
- Systematically testing and analyzing new approaches
- Reflecting on practice
- Challenging underlying beliefs that may be inhibiting change
- Spreading new learning and innovation to the larger community.

Positive outcomes may include increased effectiveness as measured by student learning, increased teacher satisfaction and retention, or evidence of professional learning, while negative outcomes could include frustration, resistance to change, excessive team conflict, or complete inaction.

Combining elements of conditions, processes, and outcomes leads me to hypothesize that there will be four major modes of teacher collaboration, synthesized by me; the specific conditions, processes, and outcomes come from existing literature on teamwork in education and other sectors, but the particular combinations and categories are new. Teams may work together in name only, or what I refer to as an *isolationist* mode. Teams may engage actively focus excessively on group harmony, with little substantive discussion, challenge to established norms, or evidence of any change in practice, in a *collegial* mode. Teams that promote thoughtful engagement with new ideas are engaged in the *problem-solving* mode, while in the *dynamic* mode, the team extends team problem-solving and learning to continuously improve its own team processes and effectively shares its discoveries with the broader school community (see Table 1-1 for examples).

TABLE 1-1 TEAM LEARNING FRAMEWORK

Mode	Conditions	Processes	Outcomes
Isolationist	Group is too small or too large, enabling shirking or lacking group cohesion and identity Group lacks leadership support or structures to support teamwork, such as dedicated time to meet	Group meetings are short, infrequent, perfunctory Group engages primarily in updates; does not address problems or group learning needs	Little to no change in practice, student outcomes likely to remain the same, or may slightly decline because teachers are substituting unproductive team work for more productive individual work Potential dissatisfaction, reduced retention
Compliance	Group has mandate to meet from leadership, but does not exhibit shared support for collaboration	Group follows a rigid protocol for engaging in collaborative inquiry, and may have some evidence	Little to no evidence of any change in practice; some evidence of resistance to change by

	or common beliefs about student learning; teachers may lack appropriate training in collaboration and/or research skills to engage in inquiry	of formal compliance, such as written agendas, but little evidence of meaningful engagement	teachers
Collegial	Group is moderately sized (literature suggests optimal size of 4-6 participants) Group has teachers of very similar experiences, backgrounds, and beliefs Group norms and processes emphasize efficiency, group harmony	Group addresses instructional needs without root cause analysis Suggested strategies are generally not experimental; do not challenge status quo or established norms	Teacher satisfaction may increase, but preliminary student outcomes will likely remain the same or decline; very little learning by group members, little to no discernible change in practice
Problem-solving	Group is of optimal size; group composition includes a mix of levels of experience, different areas of expertise, grounded in some common beliefs about learning while team members are open to new ideas Group has significant leadership support and time for meeting	Group engages in an experimental process that identifies potential issues and gaps in expertise, systematically analyzes root causes including challenging underlying beliefs that may be inhibiting performance, and seeks and tests out new approaches	Improvements in student outcomes, although may be slow to come as group experiments with new instructional approaches; effect on satisfaction and retention may be indeterminate, as some group members may be frustrated by process, at least at first
Dynamic	As above, with additional leadership support for organizational, not just team learning	Group reflects on its own process and systems are in place to share team learning across the school or larger system	Improvements in student outcomes, satisfaction, and learning that spills over to other teams; continuous improvement as group improves its own processes

In sum, the economic model generates predictions about which teachers will be most likely to participate in teamwork, and which teachers will receive the most benefit from teamwork, which I test in the quantitative analysis. There are further predictions about the conditions under which teachers will derive the greatest benefit from teamwork, which are descriptively analyzed using quantitative methods and explored more deeply in the qualitative analysis. Finally, in the qualitative analysis I examine aspects of the teams themselves, including the conditions and processes they use and how outcomes vary accordingly.

The dissertation proceeds as follows: Chapter 2 outlines existing literature on teacher quality, teacher development, and collaboration; Chapter 3 describes the quasi-experimental quantitative data and methods; Chapter 4 describes the qualitative and cost methods and data; Chapter 5 presents the quantitative results; Chapter 6 examines mechanisms and heterogeneity in these results; Chapter 7 presents the qualitative results; Chapter 8 presents the cost-benefit analysis; and Chapter 9 concludes.

Chapter 2 LITERATURE REVIEW

DEFINING AND MEASURING TEACHER EFFECTIVENESS

Among countless other researchers, Eide and Goldhaber (2004) have argued that teacher quality is the single most important school variable influencing student achievement, one of several desired outcomes of schooling, but find significant disagreement on how to define and measure teacher quality, as well as tensions between quality and quantity. Like many others, Eide and Goldhaber argue that quality is fundamentally the ability to produce growth in student achievement, although for a given teacher quality may be mutable and may vary by context. The most important characteristic in determining quality, according to the authors, is a teacher's own academic aptitude, which has been declining on average over the past several decades due to improved alternate opportunities for women. There is also a complex relationship between school quality and teacher quality; Loeb, Kalogrides and Béteille (2012) examined administrative data in Miami and found that more effective schools hire better teachers, retain better teachers, and help teachers improve more over time, although those findings raise important "chicken or egg" questions of causality, as the schools with more effective teachers are almost by definition the most effective schools. Empirically, there is evidence that a standard deviation increase in teacher effectiveness, measured using teacher fixed effects on test scores in a value-added approach, is associated with about a 0.1 standard deviation increase in student achievement on math and reading tests (Rockoff, 2004).

Many policymakers and researchers have attempted to uncover what makes some teachers more effective than others. While there are few clear answers at this point, one emerging finding is that what teachers do generally matters more than who they are. Palardy and Rumberger (2008) found that attitudes and practices are more important in predicting student achievement than observable characteristics. Using unusually detailed information on teacher practice tied to math and reading achievement data in Cincinnati, Kane and Taylor (2011) found that students of teachers who are relatively better at classroom management tended to do better in math, while students of teachers who are relatively better at discussion and questioning techniques tended to do better in reading. Consistent with the findings on the relative importance of teacher practice over teacher characteristics, Jacob and Walsh (2011) found that principals are fairly good at identifying effective teachers once they have begun teaching, as principal ratings are correlated with value-added measures, especially at the top and bottom of the distribution of value-added measures.

This difficulty in predicting teacher effectiveness *ex ante* highlights the need for human capital-enhancing policies; the evidence that actions matter more than characteristics and that teacher quality is closely related to school quality supports the need for knowledge sharing between teachers as a way to increase effectiveness over time.

PRE-SERVICE TRAINING PROGRAMS

One set of policies to increase teacher effectiveness aims to assess and increase the quality of teacher preparation programs, often based in universities. Although this approach has received much recent attention in policy debates and was featured as an important component of President Obama's "Race to the Top" education initiative, in general much more of the empirically observed variation in teacher effectiveness is within, rather than between, preparation programs. Accounting for clustering at the teacher level and including school fixed effects and measures of individual ability and institutional selectivity to account for non-random selection of teachers into programs and into schools, the difference in average teacher effectiveness between the highest and lowest-performing programs was 0.12 standard deviations in math and 0.19 in reading (Goldhaber, Liddle & Theobald, 2012). This may be an underestimate, as schools may tend to hire teachers of similar quality such that one school may have the best teacher from one program and the worst teacher from another program, washing out differences between them when looking at within-school variation. Using a similar approach in Missouri and clustering at the teacher level, Koedel and Ehlert (2012) found very little meaningful variation between programs; they speculated this is because teaching programs at highly selective universities tend to have more students from the lower tail of the ability distribution within that university as compared to teaching programs at less selective universities, suggesting that the average teacher across programs will be of about the same average intellectual ability. Neither of these studies, however, utilized an experimental or quasi-experimental design, so they are better viewed as descriptive, rather than causal findings.

In part to address concerns about the quality of many university-based teacher preparation programs, as well as to reduce barriers to entry to expand the pool of potential teachers, alternative certification programs such as Teach for America (TFA) have become popular policies across the United States and in other countries. These programs generally require limited or no prior coursework in education or student teaching experience and instead compress training into an intensive summer program, followed by ongoing training and support during a teaching commitment period, often of two years. While many scholars have expressed concern about alternatively certified teachers being less prepared and more likely to leave after two years than traditionally certified teachers, defenders argue that the programs attract applicants with higher average academic ability and from more selective universities than the average traditionally certified teacher. The empirical evidence thus far suggests that TFA teachers are not very different, in terms of effectiveness, from other beginning teachers. A randomized experiment found a 0.15 standard deviation effect of TFA on student achievement in math and no impact on reading (Glazerman & Decker, 2006), and a study using panel data found very small differences between groups of teachers with different types of certification, on the order of 0.01 standard deviations (Kane, Rockoff & Staiger, 2008).

Thus far, therefore, although concerns about the quality of pre-service training persist, there is limited empirical evidence to identify any specific policy or practice that would lead to substantial improvement. Further, evidence that alternatively certified teachers are similar to traditionally certified teachers, at least with regard to value-added, suggests that learning about teaching may be primarily experiential, and that improvements to in-service training through collaboration may be necessary to increase on-the-job learning. Further, team-based problem solving, as exists in the inquiry team initiative, may be a promising reform for pre-service training, as well.

Value-added Measures and Incentives

Some researchers and policymakers have suggested that, given our collective lack of knowledge about who will be an effective teacher and what effective teachers do, the best course of action would be to carefully measure and tie stronger incentives to increased student learning (Chetty, Friedman and Rockoff, 2013a and 2013b; Hanushek, 2007; Staiger and Rockoff, 2010). However, there are serious concerns about the validity and reliability of existing measures of teacher effectiveness, and empirical evidence on the effectiveness-enhancing impacts of incentive pay schemes is limited and mixed (Haertel, 2010; Rothstein, 2010; Darling-Hammond, et al. 2011). Further, there are substantial political objections to value-added measures of teacher performance and merit pay schemes, including concerns about over-reliance on standardized test scores, narrowing the curriculum and teaching to the test, creating incentives to game the system, and undermining teacher professionalism (Corcoran, 2010).

Several incentive programs for teachers have been tested in practice, often in randomized experimental settings, with drastically different results. This pattern of results suggests that the effects of incentive policies will be highly sensitive to context and the design of the incentive scheme. In the United States, two major incentive schemes have been recently evaluated: on a large scale, the District of Columbia unveiled its IMPACT evaluation system that featured the promise of large bonuses at the high end of performance, measured by principal and external evaluator observations and student test score gains, and the threat of dismissal at the low end. On a smaller scale, New York City tried a randomized experiment offering bonuses to schools, which could be distributed to teachers however schools wished, based on aggregate school performance. Dee and Wyckoff (2013) evaluated the DC IMPACT system using a regression discontinuity design, exploiting the cutoffs for rewards and sanctions to estimate the effect on otherwise similar teachers, and found that dismissal threats increased voluntary attrition of low-performing teachers by 11 percentage points and improved performance of those who remained by 0.27 standard deviations, while financial incentives improved performance of high-performing teachers by 0.24 standard deviations. In contrast, Fryer examined a randomized experiment that assigned bonuses of up to $3,000 to each teacher in randomly selected schools that met school performance targets and found no effect of the incentive on student achievement.

The National Center on Performance Incentives at Vanderbilt University has also performed a number of studies of performance incentive schemes in different contexts across the United States and with different design features. For the most part, the results have been quite modest, at best. The Project on Incentives in Teaching (POINT) experiment in Nashville, Tennessee offered bonuses to middle school math teachers in an RCT and, while there were effects at some grade levels, there was no overall statically significant effect (Springer et al., 2012a). Similarly, a two-year randomized study of team-based performance incentives in Texas did not yield any effects (Springer et al., 2012b).

Teacher incentive schemes in other countries have generally been more successful than those in the United States, with some exceptions. In Kenya, teachers ordinarily face particularly weak performance incentives given strong job protections, leading to high teacher absenteeism. A randomized experiment of an incentive valued at up to 43% of monthly salary resulted in significantly higher test scores, but no effect on teacher behavior except for intensive test preparation; the results also did not extend to another test, suggesting that they were highly specific to the testing instrument (Glewwe, Ilias & Kremer, 2003). In a quasi-experimental study of a tournament based incentive for relative performance in high school English and math in Israel, there were significant improvements in test-taking rates, pass rates, and mean test scores, driven by changes in teaching methods and increased after-school tutoring (Lavy, 2009).

Collectively, this literature suggests that stronger measures and incentives related to teacher performance will not be sufficient on their own to lead to widespread increases in teacher effectiveness. Value-added and incentive-based approaches on their own do not tell teachers how to improve, and most policies based on these approaches only target the very top and bottom of the distribution of teacher effectiveness, leading to little change for the majority of teachers (Hargreaves and Fullan, 2012). Nonetheless, value-added measures can be informative for helping to identify effective teacher practice, which can then be shared with colleagues via collaborative efforts such as inquiry teams.

Retention and Working Conditions

Linda Darling-Hammond and Gary Sykes (2003) argued that recruitment of quality teachers is less of a problem than retention of the most effective teachers, given that many teachers leave the profession after five or fewer years of teaching. The extent to which teacher turnover is a problem depends on two factors that are highly disputed in the literature: whether it is the most effective or least effective teachers who tend to leave and whether there are additional negative externalities of turnover due to high replacement costs or disruption to the school.

A key distinction in assessing differential attrition is how quality is defined and measured; based on a review of the literature on teacher recruitment and retention, the preponderance of evidence suggests that teachers with higher academic ability measured by their own test scores, math or science majors, and more selective undergraduate institutions are more likely to leave teaching. However, the evidence on differential attrition by measured effectiveness in increasing student learning is more limited and ambiguous (Guarino, Santibanez & Daley, 2006). There is relatively little causal research on the effects of teacher turnover on student achievement, but a recent longitudinal study of 850,000 New York City 4th and 5th graders examining school-by-grade turnover found substantial reduction in student achievement driven by turnover. Changes in the quality distribution due to replacement teachers being less experienced or effective than those who leave drive some of the results, but the authors also find spillover effects of turnover on students of teachers who remain, suggesting that there are disruptive effects of turnover on the entire school. Still, the effects of are substantively small, on the order of 0.01-0.02 standard deviations, even with 25% of teachers on a given grade level leaving a school (Ronfeldt, Loeb & Wyckoff, 2013).

In a descriptive study, Ladd (2009) found that working conditions, particularly the quality of school leadership, were highly predictive of teachers' stated intention to remain in schools in

North Carolina. Some working conditions, however, are by definition impossible to change. For example, there is evidence that teachers prefer to work with high-performing students, but a policy to promote equity may aim to assign the best teachers to the lowest-performing students. Compensating differentials could encourage teachers, particularly highly effective teachers, to work in less desirable conditions and with more challenging and higher-need students. However, unobserved school and teacher characteristics make it difficult to estimate how much extra would need to be paid to attract high-quality teachers to more difficult school environments, given the confounding of financial resources, working conditions, and teacher quality. Examining the relationship between teacher pay and working conditions, one study found that teachers actually tend to earn more in schools where they have more time to plan and where teachers report better student behavior, contrary to expectations (Goldhaber, Destler & Player, 2010). This is likely due to confounding working conditions and school district affluence, as well as difficulties in measuring teacher quality.

Of particular concern in raising teacher retention is the experience of teachers during the first few years, which can often involve a challenging and steep learning curve as they master basic teaching and classroom management skills, some of which may not be learned without hands-on experience. Several schools and districts have experimented with improved induction or mentoring to improve working conditions and increase effectiveness and retention of new teachers. In a descriptive study that uses a multinomial logistic model that measures the association between supportive working conditions such as mentoring and remaining in teaching, Smith and Ingersoll (2004) did find that having common planning time with other teachers reduced the risk of leaving teaching by about 43%. Although a fair amount of this effect could be driven by selection, it is substantively large and suggests that improved in-service training and teacher teamwork, discussed in the next two sections, could increase teacher retention. Overall, this line of research suggests that retaining teachers, particularly the most effective teachers, is an important outcome, that teachers value collaboration as a working condition that could contribute to retention, and that collaborative training could be particularly important during the first year of teaching.

In-service Training

Experimental and Quasi-Experimental Studies

One of the most common ways schools attempt to increase the effectiveness of their existing teachers, as well as provide incentives for existing teachers to stay to reduce the costs of turnover, is through in-service training or professional development programs. This category of interventions is so pervasive that the federal government spent $1.5 billion on professional development for teachers in 2004-2005 (Birman et al., 2007). In-service training programs have been studied extensively in the education literature, and somewhat less so in labor economics and the economics of education. The literature on in-service training programs provides a baseline measurement of effects for the most common policy alternative to teacher collaboration through inquiry teams, as well as insights on how collaborative efforts can enhance teacher learning by addressing perceived deficiencies in existing approaches.

The empirical evidence on the effectiveness of professional development programs is quite mixed, and such variability may be explained by differences in dosage or intensity, the quality of implementation, alignment with teacher work, the outcome measured, and the estimation methodology. Relatively few studies offer rigorous experimental or quasi-experimental estimates of the causal effect of particular programs on student achievement or teacher value-added, which are primary outcomes of interest. The small handful of recent causal studies in education and economics are discussed in detail below, followed by a brief discussion of a wider range of studies that are more descriptive, correlational, or qualitative, for additional context.

Yoon and colleagues (2007) surveyed 1300 studies of teacher professional development programs published between 1986 and 2003 for the What Works Clearinghouse (WWC) and found that only nine met evidence standards – six were published in peer-reviewed journals and three were doctoral dissertations, and six were randomized controlled trials (RCTs) while three were quasi-experimental estimates. Overall, they found the programs evaluated to have positive effects, with 18 out of 20 measured effects across the studies being positive and statistically significant, although this result may partly be due to publication bias.

Within the labor economics literature, five major quasi-experimental studies of teacher professional development programs have been published in recent years. The first examined the effects of an in-service training on pedagogical methods in elementary math and reading on test scores in Jerusalem schools (Angrist and Lavy, 2001). The authors noted the importance of teachers' on-the-job experience, and specifically in-service training, in determining their effectiveness and returns to experience, and the relative paucity of literature on this subject. This is surprising given the importance schools and teachers assign to such training – as Farrell and Oliviera (1993) noted, "pre-service training is essential to teach subject matter. In-service training is essential to teach teaching skills." Angrist and Lavy studied the 30 Towns intervention in Jersualem, a large infusion of additional resources for schools in a neighborhood with a high proportion of immigrant students and lower performance than the city average; much of the additional resources were used for teacher training. Since treatment was not randomly assigned, the authors used difference-in-differences, ordinary least squares regression controlling for observables, and non-parametric student-level matching to estimate the effects of the program, arguing that while none of the identification strategies was ideal, robust results across the multiple specifications would bolster confidence in program effectiveness, particularly because each strategy addressed different potential confounding factors. The estimates of program effectiveness for non-religious schools were positive, significant, and robust to specification, whereas estimates for religious schools were less robust. The authors argued that dosage and treatment intensity may explain this disparity. Based on a back-of-the-envelope cost analysis, the authors estimated that the program cost $12,000 per class in 2001 dollars and resulted in 0.25 of a standard deviation increase in math and reading test scores, comparable in cost-effectiveness to increasing instructional time and more cost-effective than reducing class size.

Similarly, Jacob and Lefgren (2004) took advantage of a reform targeting low-performing schools to estimate the effect of additional teacher training on student achievement in Chicago. In this case, the reform targeted schools with students achieving below a particular cutoff to receive both additional resources and the threat of sanctions, enabling a regression discontinuity design. Utilizing a regression discontinuity with student test scores as the running variable helped mitigate concern about teacher selection into training, as teachers in schools just above and just below the cutoff were arguably similar. Further, schools had little ability or incentive to precisely manipulate the running variable, as it was aggregated across many students, and the “treatment” was complex, consisting of both positive elements such as technical assistance and negative elements such as the threat of sanctions. The authors did find that teachers reported an increase in the quantity and quality of professional development they receive, but did not find any effect on student achievement.

Jacob and Lefgren explicitly noted the differences in their results from those of Angrist and Lavy, and hypothesized that this could be due to numerous factors related to the treatment, setting, or methodology. The complex nature of the treatment may have motivated schools just above the cutoff to seek alternate means to improve performance to avoid sanctions, the high-stakes setting may have reduced effectiveness of the training, or the training may have simply been too low-intensity or poorly implemented to be effective. They estimated that $108,000 per school was spent on additional training, in 2004 dollars; even with a conservative assumption of just 20 classrooms per school, the intensity of treatment as measured by resource use in the Jerusalem intervention was more than twice as high as the Chicago treatment.[4] Finally, Jacob and Lefgren did find that their results were more consistent with those of the literature than those of Angrist and Lavy; a meta-analysis by Kennedy in 1998 found only 12 studies out of 93 with positive effects of staff development, suggesting that much in-service training for teachers is low-intensity and low-quality, a finding discussed further below. Further, the quality of implementation and coaching, the level of follow-up and ongoing support, and integration with teachers' ongoing curricula and regular work lives all likely influence the effectiveness of any particular intervention.

Most recently, Harris and Sass (2011) examined administrative data in Florida that allowed them to link information on student achievement to teacher training and qualifications. The authors were able to take advantage of the panel structure of the data, along with unusually good measures of teacher training linked specifically to student achievement, to estimate essentially a value-added model with school and teacher fixed effects and a rich set of time-varying covariates. Although there were apparent effects of professional development when estimated using pooled ordinary least squares, only effects of content-specific training on achievement in middle and high school math remained when adding teacher fixed effects, thus

[4] Using the CPI, $12,000 in 2001 dollars is $15,908 in 2014 dollars, while $108,000 in 2004 dollars is $134,232 in 2014 dollars, meaning that an elementary school would need to have just 8 classrooms for the investments to be equivalent.

estimating the effect over time within individual teachers. This suggests that results in other subjects and for other types of training were driven primarily by positive selection of teachers into training.

Bridging the gap between pre-service and in-service training, Rockoff (2008) studied the effects of mentoring programs for beginning teachers in New York City on teacher retention and student achievement outcomes. Rockoff took advantage of the fact that teachers who transferred to New York City from other school districts did not receive mentoring, while those who were brand new to teaching did, to implement a difference-in-differences estimate of the effects of the program, comparing differences in effectiveness over time between the two groups. The effects were relatively limited, and included an increased likelihood to remain through the first year for those teachers who did receive mentoring; interestingly, having a mentor who taught at the same school was a predictor for retention, suggesting that school-specific knowledge, akin to firm-specific human capital, may play an important role in teacher development. Similarly, using a variety of quasi-experimental methods to analyze the effects of induction programs on teacher turnover in New York City and nationwide using the Schools and Staffing Survey, You (2012) found that, when taking into account endogeneity of participation in mentoring and induction programs, the effects of the programs were too widely variable to obtain a statistically significant point estimate.

Along similar lines, Bressoux, et al. (2009) examined the effects of training for novice teachers on student outcomes in France, taking advantage of an administrative forecasting mistake that led to some otherwise similar teachers receiving initial training and others delaying training until the following year. In the French system, all potential teachers are ranked and those with highest ranks are placed into strictly limited slots to receive training, while other students are placed on a waiting list and may teach without training if a vacancy arises. Ordinarily, this setup would not be ideal for research, as the trained and waiting list teachers are demonstrably different. In 1991, however, an unusually small number of teachers were selected for training, leading to a group of teachers who would have been trained any other year but instead were placed on a waiting list and filled vacancies if needed. The authors argue that other common concerns about selection into schools by teachers or students into teachers' classrooms are less salient in France, where novice teachers are typically assigned to schools wherever they are needed. Using a gain score model with class-level random effects, the authors found substantial effects of training in math, but not in reading, and not for the lowest-achieving students. It is unclear, however, whether this study is comparable to other in-service training studies, as in-service training is usually supplementary to whatever basic training teachers receive; it would be natural to expect training to have an effect when compared to teachers who receive no training at all, but the more policy-relevant question is whether training has an effect on the margin. Since the control condition in this study is unlike "business as usual" in most other contexts, the external validity of the study may be more limited.

The United States Department of Education has commissioned two recent randomized controlled trials (Garet, et al., 2008 and Garet, et al., 2010) of professional development programs of early reading and middle school math, respectively. As Garet and colleagues noted, No Child Left Behind has underscored the importance of teacher training by placing great emphasis on all schools having "highly qualified teachers," although the meaning of this term has been limited in practice to fully-licensed teachers, and providing $535 million in Title II aid for PD to states and districts in the 2002-03 school year. Nonetheless, concerns remain about the effectiveness of in-service training, particularly because, as noted by Jacob and Lefgren, it tends to be ad hoc and of low intensity. Eighty percent of elementary school teachers report 24 or fewer hours of professional development each year.

To test whether a more intensive approach, closely aligned to the curriculum, would improve outcomes, the Early Reading PD Interventions Study randomly assigned teachers to receive one of two treatments – an intensive summer institute with follow-up training over the school year or the same training plus intensive coaching of 60 hours per teacher on average – or a control condition receiving ordinary training. While the authors did find effects on teacher knowledge in the short-term, they saw no statistically significant effects on student outcomes, and the teacher knowledge effects faded after one year. These results are not attributable to low-intensity implementation, as treatment schools received 39-47 hours of PD, compared to 14 hours in control schools, and schools assigned to coaching received 62 hours per teacher, on average, compared to 4-6 in non-coaching schools, although implementation still may have been of poor quality, depending on the skill of the coaches.

In the Middle School Mathematics PD Study, 77 schools across 12 districts were randomly assigned to treatment or control conditions. There were no significant effects on teacher knowledge or student achievement, although researchers did observe treated teachers engaging students in critical thinking exercises more often, suggesting some change in teacher behavior that may have long-term benefit not captured by test scores.

Finally, an experimental study of a teacher-training program in the Netherlands examined the effects on math student achievement of a highly scripted math training program called Sigma (Van der Sijde, 1989). Thirty-three teachers were randomly assigned to one of four conditions, corresponding to different intensity levels of training. Those receiving treatment underwent training on preventative classroom management techniques, such as monitoring for signs of student distraction, effective pacing, and smooth transitions; outcomes included observations of teacher effectiveness by graduate students, student surveys, and math achievement. Observers did note changes in teacher behavior based on amount of training, but those did not translate into measurable differences in student achievement; nonetheless, the follow-up was only two and a half months after the program started and sample sizes were extremely small, so the improved techniques may have not had enough time to affect student achievement, or the sample may have been too small to detect any results.

Non-experimental Studies

Most other research on in-service training is primarily correlational, descriptive, or qualitative, and does not explicitly address selection of teachers into training as a potentially confounding variable. Further, relatively few studies feature student achievement as an outcome, and instead focus on teacher-reported satisfaction with training or changes in teacher behavior. Barrett, et al. (2012) explicitly addressed this issue when examining the effects of the Appalachian Math and Science Partnerships in Kentucky. While a value added model with teacher fixed effects showed no relationship between the training program and student outcomes, accounting for prior teacher effectiveness yielded positive effects of the program, implying that previously less effective teachers were more likely to participate. This fact may be idiosyncratic to the Kentucky program, however, as many teacher in-service training programs are voluntary and may be just as likely to feature positive selection.

Addressing similar concerns about the quality of research, Desimone (2009) made several suggestions for improving impact studies of professional development programs. Most importantly, Desimone suggested emphasizing student learning results, or the mechanisms by which programs will change teacher behavior to impact student learning, rather than teacher attitudes and satisfaction as important outcomes. Desimone also suggested several "critical features" of effective PD that stand out in the literature, including content focus, active learning, coherence, or the extent to which teacher learning is consistent with teachers' knowledge and beliefs, duration, and collective participation, and advocated for additional experimental and quasi-experimental studies. Effective follow-up and close alignment with the existing curriculum and the developmental needs of teachers and students can also contribute to the success of PD programs.

Several articles and reports in the education literature attempt to synthesize findings across these and many other studies of professional development programs to isolate key factors that determine their effectiveness. Corcoran (1995) found that the type of in-depth, ongoing PD that is suggested by the research is rare due to its high cost and time commitment; most PD instead takes the form of discrete workshops on "hot" topics taught by local experts. He suggested that PD that is integrated with teacher work, based on current research, and reliant on teachers as valuable experts and sources of information was most likely to be effective. Corcoran also proposed experimentation with models more commonly used in other countries, such as lesson studies common in Japan, in which teachers spend less time actively teaching and more time in planning, training, and collaboration. Employing such a shift would require fundamental restructuring of school, but Corcoran suggested this could be achieved by replacing some instructional time with computer-based or distance learning, community-service projects, and extracurricular activities led by volunteers.

Garet, et al. (2001) surveyed 1027 math and science teachers who had attended a federally-funded PD program and found that focus on content knowledge, opportunities for active learning, and coherence with other learning activities were the program features most correlated with effectiveness in terms of teacher knowledge, skills, and changes in classroom practice. Similarly, Darling-Hammond and Richardson (2009) found that teachers preferred hands-on training focused on their content areas, with an emphasis on student learning and active observation, reflection, and teaching. The most effective PD was sustained and intensive; the largest effects were among programs with 30-100 hours of training, and no effect was seen in programs of fewer than 14 hours. Increasingly, as well, professional learning communities as a form of teachers collaboratively learning from one another have been featured in the literature but, as noted below, simply bringing teachers together does not ensure effective collaboration. Finally, Darling-Hammond and Wei (2009) reported that a majority of teachers spent fewer than 16 hours per year in content-area training, while teachers said they needed about 50 hours per year. Compared with other countries, Darling-Hammond and Wei say that teachers in the United States spent more time actively teaching and less time training and collaborating with colleagues, limiting their ability to improve over time.

Day and Gu (2007) attempted to uncover the causes of variation in teachers' professional learning, using data on 300 teachers in 100 schools in the United Kingdom. While their mixed methods analysis was descriptive and not causal, they did find wide variation in the association between performance and experience, and evidence that teachers did not necessarily learn from experience. The key factors they identified in determining professional learning were commitment, resilience, and leadership, and they argued that recent "performativity," or emphasis on compliance with mandates and emphasis on accountability measures in the UK, would reduce intrinsic motivation. A similar study of the factors that influence teacher professional learning in Dutch schools (Sleegers, Stoel & Kru, 2009) examined the effects of teacher psychology, school organization, teacher collaboration, and leadership on teacher experimentation, innovation, reflection, and learning using structured equation modeling. They found that psychological factors, notably self-efficacy and internalization of goals, had stronger effects, while organizational and leadership factors had smaller and mostly indirect effects. Notably, however, collaboration was strongly related to experimentation and keeping up to date with the field.

A number of other teacher training and professional development programs have been evaluated in the literature on outcomes besides student achievement and with non-causal methods. Goldschmidt and Phelps (2010) assessed the effect of the California Professional Development Institutes on subject matter knowledge of teachers, following theories by Lee Shulman that pedagogical content knowledge, or knowledge within a specific subject area about the most useful ways to present a subject to make it understandable to others, is one of the most important aspects of teacher quality. The intervention consisted of 40 hours of summer training, 40 hours of follow-up training during the school year, and 40 hours of team meetings, and the outcome was teacher pedagogical content knowledge in reading measured by the Content Knowledge for Teaching Reading test. The authors used a multilevel growth model, assuming that teachers would not have shown growth in knowledge in the absence of the program, but did not have any quasi-experimental methods to support or test this assumption. They found positive effects of the program that faded over time.

A simple pre-post analysis of a PD for science teachers (Lee, et al., 2008) found significant gains in science achievement, but similarly did not employ any control or comparison group. Finally, Tournaki, Lyublinskaya and Carolan (2011) examined the effects of a professional development program on teacher effectiveness through classroom observations using Charlotte Danielson's *Enhancing Professional Practice: A Framework for Teaching* (2007). Measuring classroom environment, instruction, and planning and preparation, the authors only found effects under the domain of instruction. Overall, therefore, although evidence that traditional professional development programs enhances teacher human capital or student achievement on average is weak, there is suggestive descriptive evidence that programs with particular features may be more effective. These specific features, which may be more likely to be present in collaborative professional development programs such as inquiry teams, have generally not been subject to rigorous, experimental evaluation.

Returns to Experience and Diffusion of Expertise

A number of studies focused on professional development and training, but specifically emphasized the importance of collaboration and teamwork as an important ingredient to successful teacher learning. A study of the National Writing Project's school partnership on instructional practices used an randomized controlled trial of 39 schools, 20 of which were randomly assigned to a partnership condition to receive customized professional development (Sun, Penuel, Frank, Gallagher & Youngs, 2013). The authors used longitudinal and sociometric data to examine specifically how high-quality training could promote diffusion of effective teaching strategies through collaboration. The authors hypothesized that these spillover effects would increase productivity of colleagues, based in part on economic literature on human capital externalities, leading to a dual effect of training that was magnified when workers worked on teams. They did find statistically significant increases in the number of teachers helped by other teachers, although the coefficient of .012 additional teachers helped per teacher-hour of training is substantively small.

Along similar lines, Kraft and Papay (2013) used longitudinal administrative data from the Charlotte-Mecklenburg schools to examine how supportive professional environments affected differential returns to experience among teachers. They argued that average returns to experience masked large variation across individual teachers, and that working in more supportive professional environments could lead to greater increases in effectiveness over time. The authors were able to combine administrative data, including student achievement data on math, with the North Carolina Teacher Working Conditions Survey to analyze 280,000 student-year observations over 3,145 unique teachers. The measures of professional context included order and discipline, peer collaboration, principal leadership, professional development, school culture, and teacher evaluation. A factor analysis revealed these items loaded on a single factor, with internal consistency reliability estimates exceeding 0.90.

The identification strategy relied heavily on teacher fixed effects and teacher and school-level random effects; fixed effects control for time invariant teacher and school characteristics, whereas random effects allow slopes to vary to allow for variance in returns to experience. The authors finally interacted the measure of professional environment with experience to determine whether variable returns were systematic and related to professional context. They found that the average returns to ten years of experience are quite large, at 0.11 of a standard deviation in value-added, with significant heterogeneity at 0.025 of a standard deviation. School-specific random slopes explained about 30% of the variation in returns to experience, although they were substantively quite small at 0.007 of a standard deviation, and a one standard deviation increase in quality of professional environment was associated with an additional 0.0026 standard deviation increase in annual returns to teaching experience. The authors addressed the concern that teachers and students did not randomly sort into school environments by interacting the experience variable with student and teacher covariates and saw no change in coefficients, but they were only able to perform this test on observables.

In a companion paper, Papay and Kraft (2013) addressed some of the methodological issues that arose in these estimates. They further found evidence of returns to experience later in the career, contrary to many estimates that suggest that teacher effectiveness plateaus at around five years of experience. They identified the confounding of experience with year trends; the simplest approach to addressing this problem would be to omit year effects, assuming they are random shocks. Other options include a censored model, using only year effects for teachers with more than 10 years of experience on the assumption that they do not continue to improve beyond that point (Rockoff, 2004) or to bin experience across multiple years to still allow for year effects.

Papay and Kraft proposed a third option, using teachers who have non-traditional career trajectories to identify experience effects. The authors tested each model with simulated data with different "true" parameters and find that the censored model performed perfectly if the assumption of no improvement after 10 years held true. Even minor violations of that assumption, however, generated downward bias, while their proposed two stage model performed well if there was no general time trend, but had a downward bias otherwise. Based on their simulations and using a variety of specifications, the authors showed that generally there was a downward bias on the estimates to returns to experience in the literature, and teachers did tend to improve even after 10 years. Together, these findings emphasize the importance of ongoing teacher learning and the hypothesis that school culture and organization, as well as a teacher's peers, contribute substantially to that development. Nonetheless, the empirical support for any given policy or intervention on average is quite weak.

COLLABORATION AND TEAMWORK IN EDUCATION AND OTHER SECTORS

Evidence from economic theory and other sectors suggests that one way to improve teacher in-service training, which has mostly proven ineffective, and potentially to increase teacher effectiveness through other channels is to increase and enhance teacher teamwork or collaboration. Collaboration could enhance effectiveness or productivity through more relevant on-the-job learning and knowledge sharing between colleagues, through peer pressure or other social incentives, or through building intrinsic motivation to achieve shared goals, among other channels. Collaboration could also be subject to free-riding, encouragement of negative social norms, and other problems.

Even without obvious externalities from group-based production, socialization in the workplace can play an important role in determining productivity. Using within-worker fixed effects and examining variability in productivity based on whether an individual berry picker was working alongside self-identified friends, Bandiera, Barankay and Rasul (2010) found no average social effects on productivity in berry farming, but those averages masked considerable heterogeneity. Workers were more productive when working with higher-ability friends and less productive when working with less able friends; these effects appear to be driven primarily by conformism to adopted social norms. In other words, workers adapted their own practices to match the productivities of those around them, even when the production processes were entirely independent.

Similarly, Mas and Moretti (2006) utilized plausibly random shift changes in a grocery store to analyze how an individual cashier's productivity varied with that of his or her colleagues. They found that a 10% increase in the average productivity of those working with a cashier was associated with a 1.7% increase in that cashier's own productivity. Social pressure and peer monitoring likely drove the effect, as it was most pronounced when the cashier was visible to others. Peer monitoring extends beyond labor market productivity; one study found that peer monitoring and social ties reduced moral hazard in group lending in Eritrea (Hermes, Lensink & Teki, 2005).

In cases when the outcome is jointly determined by a group, unlike more individual efforts such as berry farming and grocery store cashiering, effects can be even more pronounced and work through other channels. In a simple experiment on the quality of decision-making, participants working with a group made fewer errors than those working alone, although the channel was unclear and likely driven by reduction in idiosyncratic error by pooling the decision (Chalos & Pickard, 1985). Other literature, however, emphasizes the potential for productivity losses and increases in error due to lack of individual accountability and time lost to group coordination, referred to by Steiner (1972) as “process loss.” The productivity of group processes and accuracy of group decisions depends a great deal on contextual factors, and the evidence overall on group versus individual decisions is mixed (Kerr & Tindale, 2004).

The adoption of group-based piece rates and team-based production at a garment factory increased worker productivity by 14%. Participation in teams was voluntary, and the researchers compared productivity within the same worker who was observed working individually and on a team (Hamilton, et al., 2003). There is some concern that some of the effect was driven by selection, as team participation was endogenous – those most likely to benefit from joining a team would be most likely to join. However, the increased productivity of teams compared to the aggregated productivity of the same individuals working alone does reduce concern about free-rider effects of teams. Further, the productivity of some teams exceeded the individual productivity of their most productive workers, suggesting some synergistic benefits of working as a group, and some individuals elected to join teams even if it reduced their pay, suggesting some non-pecuniary benefits to teamwork.

Kandel and Lazear (2012) created a theoretical model that could explain these empirical findings. Despite concern about free-riding, partnerships and profit-sharing mechanisms could enhance productivity through a sense of team spirit and peer pressure. Peer pressure could be the result of avoiding shame in cases where effort is observable or avoiding feelings of guilt for shirking when effort is not observable by colleagues – in other words, peer pressure can exist even without monitoring mechanisms. Kreps (1997) further suggested that concern for esteem of colleagues, particularly when work is ambiguous or creative as in teaching, may reduce the disutility of effort and increase worker productivity.

Empirical work suggests several factors that may contribute to or detract from the effectiveness of teams, teamwork, and the social pressure mechanisms described here. Team size is one important predictor, although the optimal team size likely depends on the context, and there is little consensus in the literature on that question. One experiment analyzed how teams of different sizes performed on a cognitive puzzle game and found that teams of four performed better than teams of one or two (Sutter, 2005). Using descriptive data on teamwork in the software industry, Hoegl and Gemuenden (2001) found that effective teamwork is divided into six dimensions: quantity and formality of communication, coordination of effort, balance of contributions across team members, mutual support, effort, and cohesion. They based their framework on empirical case study analysis and tested it with structural equation modeling; they found that all six dimensions of team quality were strongly associated with work satisfaction and learning. Using this framework to analyze the determinants of team quality, they found that proximity of team members was associated with almost all of the factors, but that team size was negatively correlated with them, implying that, at least above a certain level, teams could become counter-productively large.

There have been no causal studies of the effects of collaboration or teamwork on teacher effectiveness or retention, and very few empirical quantitative studies that have specifically focused on these associations, although some quantitative studies with a broader focus have included measures of collaboration as important covariates. There is, however, a rich qualitative literature on the factors that make teamwork in education relatively more or less effective in different settings.

The relative effectiveness of a teacher team may depend on nuanced aspects of how teachers interact. These interactions could promote varying degrees of attitudes toward conflict, including unproductive avoidance or acrimony or more productive discussion, as well as different levels of inclusion or exclusion across a community. A case study of the micropolitics of teacher collaboration in two schools in the San Francisco Bay Area found that one school effectively used collaboration to address conflict in a way that challenged institutional norms, sparked new ideas, and promoted institutional learning, while another school used collaboration to promote warm and collegial relationships among teachers, but saw little long-term change as a result (Achinstein, 2002).

A quantitative, but non-causal, study of the factors that affect professional community in Chicago schools found that, at the teacher level, experience was a predictor of professional community, and at the school level, strong leadership, trust, and higher prior achievement were strong predictors. The authors conceptualized professional community as comprising reflective dialogue, deprivatized practice, staff collegiality and collaboration, a focus on student learning, collective responsibility for school improvement, and new teacher socialization, and measured it using a survey of teachers. School size was the strongest predictor of professional community, with smaller schools having more community, but that effect seems to be a mediator, not a direct cause, as it disappeared when survey data on the school's social context were added to the model. Although the study analyzed rich survey data using multi-level modeling techniques, it could not control for selection of teachers into schools with particular features and levels of community, so only offers correlations for further study (Bryk, Camburn & Louis, 1999).

Using a similar quantitative but non-causal strategy to analyze unusually rich data about school characteristics in New York City, Dobbie and Fryer (2013) found that particular practices in charter schools were correlated with effectiveness, while teacher training and traditional factors such as class size and per pupil expenditures were not. These practices did not explicitly include teacher collaboration, but did include frequent feedback, tutoring, increased instructional time, high expectations, and the use of data to guide instruction. The last, in particular, could be related to effective team-based problem-solving by teachers.

In his 2006 review of the education literature on teacher collaboration as a workplace condition, Kelchtermans found that collaboration could increase effectiveness of teaching and enhance continuous school improvement, but simply increasing professional collegiality did not automatically confer these benefits. To the extent that the threat of interfering with collegial relationships may inhibit colleagues from discussing difficult issues, a culture of collegiality may become a culture of comfortable mediocrity dominated by unchallenged consensus or majority thinking. Kelchtermans found that collaboration was most effective when teachers engaged in collaborative problem solving that pushed them to deeply engage in underlying actions and beliefs. An early case study found dramatic range in the quality and quantity of conversations teachers had about teaching, supporting this hypothesis (Little, 1982). A case study of one form of teacher collaboration, Professional Learning Communities (PLCs), in a middle school similarly found variation in the extent to which PLCs contributed to depth of teacher learning and improvement in effectiveness; notably, one teacher on a less effective team observed that the collaboration was “About teaching… but not about student learning.” Researchers identified group size, leadership support, and time for common planning as elements that contributed to more effective collaboration (Graham, 2007).

Surveys and focus groups of teachers in Minnesota reported that over 90% of teachers found collaboration to be valuable and to improve the use of data to make decisions in schools (Huffman & Klanin, 2003). Case studies of collaborative professional culture in three elementary schools that dramatically improved in a short period of time revealed that conversations based on student data tended to promote purposeful improvement and self-efficacy (Strahan, 2003). Clearly, the effectiveness of any particular collaborative approach to teacher improvement and development would depend greatly on the underlying culture of the school (Hoy, 1990). In particular, since teacher collaboration will tend to promote the spread of dominant beliefs about teaching and learning to new teachers, schools with already positive cultures should see more positive effects from collaboration; it is unclear, however, whether collaboration on its own can help improve the culture of a dysfunctional school.

Increasingly popular research methodologies also offer promise for learning more about the effects of collaboration, as well as the factors that determine effective collaboration. In particular, social network analysis has been applied to teacher collaboration in Dutch schools, uncovering structures of teacher interactions that often differ from formal structures, serve multiple purposes and change over time. Notably, interactions did seem to be closely linked to teacher characteristics – teachers seemed to interact most frequently with other teachers who were similar to themselves with respect to gender, age, experience, ethnicity, beliefs about teaching, and grade and subject taught (Moolenaar, 2012).

A common refrain in the literature on teacher collaboration is the importance of strong principal leadership to facilitate more effective collaboration. There is no doubt that leadership matters, but there is a tension between leadership that is hands-on but that could become too prescriptive versus leadership that promotes teacher autonomy but could lead to a lack of quality control. Either leadership style appears to be associated with more effective collaboration unless it veers too far into the extreme in either case. School leaders also promote effective collaboration by providing protected time for common planning (Scribner, 1999). Based on work with four medium-sized districts, Wayman, Midgley and Stringfield (2006) emphasized the importance of "calibration," which they defined as developing a common understanding of teaching and learning and consensus on goals, as a critical contextual ingredient for successful collaborative data teams.

Overall, although the quantitative and in particular any experimental or quasi-experimental data is thin, there is reason to be optimistic about specific forms of collaboration that emphasize data use, teacher learning, and problem solving, as a way to enhance teacher learning and productivity and address gaps in teachers' pre-service training. An ongoing qualitative study of PLCs that emphasized collaborative inquiry presented initially promising results on teacher professional growth despite significant challenges in successfully forming such teams (Nelson, 2009). Further, a quantitative analysis of teacher satisfaction with professional development based on collaborative action research found promising results that hinged critically on specific design features (Burbank & Kauchak, 2003). Therefore, both in practice and in research there appears to be great potential, but much work to do in the area of identifying effective practices in teacher collaboration and teamwork.

Ronfeldt and colleagues recently published a study (2015) that quantitatively explores how within-school variation in the quality of collaboration relates to student achievement. Although the study is non-causal, it represents a significant contribution by incorporating both school-level and teacher-level variability in collaboration type and intensity in a multi-level model; further, the study examines the teacher-level factors that predict collaboration quality, measured by surveys of teachers on the intensiveness and helpfulness of various types of collaboration. The authors emphasized the importance of both conditions as indicators of quality, as collaboration that is helpful without being extensive is necessary but not sufficient, and collaboration that is extensive but unhelpful is merely a waste of time. They find positive and statistically significant results of more and better-quality collaboration, particularly at the school level, with much stronger effects in math than in reading. Once controlling for the school level in a random-effects, multi-level framework, only small teacher effects remain - less than 0.1 standard deviations in math, and no significant effects in reading.

In a five-year, quasi-experimental analysis of teacher inquiry among grade-level teams in nine Title 1 schools, effects were only seen in later years of the intervention, when the intervention was refined to provide additional training and implementation support (Saunders, Goldenberg & Gallimore, 2009; Gallimore et al., 2009). The authors note significant gaps in existing literature on teacher collaboration: few quantitative studies and even fewer experimental or quasi-experimental studies exist, few studies examine impacts on student achievement, and qualitative studies likely suffer from selection bias as teams are only selected for study on the condition of already being effective. While the quasi-experimental approach, matching 9 schools that selected the inquiry team improvement initiative to 6 comparison schools that received other initiatives and were similar at baseline, is an improvement upon descriptive analyses, it still may suffer from selection or management bias since schools elected to participate in the treatment. Further, while measured effects were substantially larger than other effects in the literature, at 0.8 standard deviations, those were only observed after intensive training and implementation support that was not randomly assigned.

Individual and Organizational Learning – Human and Social Capital Interactions

A number of authors have offered arguments for why collaboration may be productive in public sector settings in general, and in education in particular. While empirical evidence on this is somewhat limited, it does suggest specific potential benefits of collaboration and settings under which it is most likely to enhance individual and organizational productivity. Hargreaves and Fullan (2012) argued that professional capital – the product of human capital, social capital, and decisional capital – is the key lever by which to invest in better schools and teachers (p. 3). Any of three alone is insufficient, particularly human capital, in part because investing in social capital, defined as relationships and trust, can lead to improvements in human capital through knowledge sharing and peer effects, but not vice versa. Hargreaves and Fullan's argument is based in part on examination of the educational practices of high-performing nations such as Finland, arguing that school reforms must invest in raising the performance of all teachers, as opposed to narrowly focusing on eliminating a few at the lower tail of the performance distribution and rewarding a small handful at the top. They further base their argument on a McKinsey report based on a study of 20 national school systems that are consistently high performing and continuously improving. The authors conclude that a marker of school systems in transition from "Great" to "Excellent" is continuous improvement and innovation through peer-based learning, structured experimentation, and decentralized decision-making (Mourshed, Chijioke & Barber, 2010). One possible reason for the hypothesized interaction between human capital and social capital is that knowledge may be highly contextualized in school settings, and learning may be highly experiential – as Hargreaves and Fullan state, truth is “situational, not statistical” (p. 112).

Pil and Leana (2009) engaged in a study of 1,103 teachers in 239 grade teams to analyze the relative importance of human and social capital and their interactions. They used survey data to measure the number of social ties between teachers and their strength and incorporated those measures, along with human capital measures, in a multi-level analysis on their effects on student math scores. They conclude that social capital may have as much of an effect on student achievement as teacher human capital, in part because aggregated and accumulated human capital is itself a resource, deemed intellectual capital, that is shared through social capital. At the teacher level of analysis, human capital was an important predictor, but at the team level, social ties became important, as well.

Finally, collaboration may have implications broader than promoting individual learning and productivity. Ansell (2011) argues that there is a fundamental tension between democracy and governance - governance requires flexibility and discretion, especially in novel situations that street-level bureaucrats encounter for which there has not been time to develop a series of rules through the democratic process. Managing this tension, and allowing public servants sufficient latitude to solve problems while building their own and organizational capacity to do so while also maintaining public oversight and trust, is a central challenge of democracies. Two trends relevant to education make this tension even more challenging to manage: one is the growing complexity of systems and problems, due in part to technological change and the fact that new systems and institutions seldom fully entirely replace, but are rather layered on top of, old systems. Secondly, constituencies for public services – in this case, students – are highly differentiated, requiring unique and innovative responses on demand.

Democratic experimentalism - based on the philosophy of pragmatism developed by Peirce, James, Dewey, Mead and others, and applied to particular settings by Charles Sabel, Michael Dorf, and others – applies pragmatist philosophy to the relationship between democracy and governance, based on continuous improvement as seen in Japanese production methods. Public sector agencies engage in collaboration, monitoring, and problem-solving to continuously evolve and improve to meet public demand. This philosophy implies that effective collaboration can help teachers adapt to new conditions, manage changes in curriculum, innovate and solve new instructional problems, adapt to meet the unique learning needs of students, incorporate new research findings into their practice, and share successful strategies so that teachers and schools can learn from one another. Focusing collaboration on a particular problem, as is the case with the inquiry team intervention studied here, reflects the need to drill down to particulars, and an emphasis on what Ansell calls "analytical holism." The process is analytical because of the need to break complex problems into constituent parts, to address them in a focused, disciplined manner, while "holism" refers to need to fully consider the context in which problems occur.

Collectively, the literature suggests that there is relatively little consensus on the most effective policies to increase teacher productivity over time. Empirical evidence on pre-service training, in-service training, and incentive programs is generally inconsistent or weak, although elements of all three could ultimately be combined to enhance teacher effectiveness. There is relatively little quantitative literature on the effects of teacher collaboration, and most of the studies that do exist rely on descriptive methods and survey data. For the most part, the few existing studies that relate teacher collaboration to student achievement suggest that there are small but positive effects. There is a relatively rich qualitative literature on the practices of effective teacher teams, which generally suggests that leadership support and time are critical pre-conditions, and the most successful teams are willing to productively engage in conflict to promote learning; however, these studies generally are on teams that have already been determined to be effective, and thus may suffer from selection bias.

The recent study by Ronfeldt represents a significant contribution in that it was the first to look at within-school variation in collaboration type and quality using multi-level modeling. This study contributes to the literature and advances knowledge on the research questions by engaging in the first comprehensive, mixed methods examination of a single policy initiative over time using three empirical approaches: a quasi-experimental approach that incorporates heterogeneity, mechanisms, and within-school variation that is based on administrative, not survey data, a qualitative case study analysis that further examines heterogeneity in team quality, processes, and more proximal teacher learning outcomes, and a cost analysis. While there is significant literature on the practice of teacher collaboration, this study contributes to the literature by investigating a policy that intends to induce effective collaborative practices.

Chapter 3 DATA AND METHODS: QUANTITATIVE STRATEGY

DATA

Quantitative data for this study on schools, teams, teachers, and students come from administrative datasets from 2008-2010. The primary unit of analysis is teachers, nested in teams and schools. The quantitative data come from administrative datasets on school demographics and accountability; basic teacher biographical information including experience, education, and tenure within a particular school; information about team composition and focus that teams voluntarily entered into a central database in the 2007-2008, 2008-2009, and 2009-2010 school years; teacher value-added scores for teachers of math and English Language Arts in grades 4-8 that were released publicly from 2007-2010; and student-level information on attendance and graduation. Due to changes in the level of detail reported on team composition, the individual teachers serving on each team are only identifiable in the 2009-2010 school year. Data from other years will be used for alternative identification strategies as robustness checks and sensitivity analyses, and to measure changes over time on some outcome variables.

The number of teams grew from 1,455 in 2007-2008 to 2,605 in 2008-2009 to 9,176 in 2009-2010, and ranged an average of just over 1 per school in 2007-2008 to 15 per school in 2009-2010, although a small number of schools with an implausibly large number of teams (up to 85) indicates some possible data entry errors that may skew the mean number of teams per school upward. Note that a number of the identification strategies employed depend upon grade assignment or grade shifts, and therefore are limited to teams that focused on a single grade level. Approximately 55% of teams over the three years were grade-level teams; as one test of the generalizability of the results beyond grade-level teams, I run ordinary least squares (OLS) models separately on grade-specific and non-grade-specific teams. In most cases, results for grade-level teams are slightly smaller than those for the pooled sample, suggesting caution in generalizing the quasi-experimental results to other types of teams.

The data available on teams varies by year, as the district changed the questionnaire for the administrative database they used to collect information on team activities, and data entry was optional. Thus, the quantity and quality of data varies by school, a fact that can be exploited in descriptive analysis of heterogeneity in team quality and intensity. Data become generally more available over time, but include the school, number of teachers on the team, grade level and/or subject area focus of the team, characteristics of the student population on which the team focused (e.g., English language learners), and descriptive notes on the team's process, including assessments it used to measure student learning and any instructional changes the team made. Table 3-1 summarizes the data available each year and which outcomes are used in each empirical model, described in the next section, for each year.

TABLE 3-1 OUTCOMES BY MODEL AND YEAR

	2007-2008			2008-2009			2009-2010		
Model	Retention	VA	Test Scores	VA	Retention	Test Scores	VA	Retention	Test Scores
OLS with school fixed effects	X	X				X	X		X
Rationale	Estimate baseline effects	Estimate baseline effects	No data available	Only available for first-year teachers	Only available for first-year teachers	Estimate baseline effects	Estimate baseline effects	Only available for first-year teachers	Estimate baseline effects
OLS for first-year teachers (with school fixed effects)	X	X		X	X		X	X	
Rationale	No prior year data, must use OLS	No prior year data, must use OLS	No data available	D-in-D preferred	D-in-D preferred	Cannot tie teachers to test scores	D-in-D specification preferred	D-in-D preferred	Cannot tie teachers to test scores
Difference-in-differences for first-year teachers				X	X		X	X	
Rationale	No prior year data	No prior year data	No data available	Prior year available, D-in-D specification preferred	Prior year available, D-in-D specification preferred	Cannot tie teachers to test scores	Prior year available, D-in-D specification preferred	Prior year available, D-in-D specification preferred	Cannot tie teachers to test scores
Difference-in-differences for grade switchers							X		
Rationale	Cannot identify individual teachers or grades	Cannot identify individual teachers or grades	No data available	Cannot identify individual teachers or grades	Cannot identify individual teachers or grades	Cannot tie teachers to test scores	Only year for which individual teachers on teams are observable	First-year model preferred for retention	Cannot tie teachers to test scores
Instrumental Variables						X			X
Rationale	No class section data	No class section data	No class section data	Class sections linked to students, not teachers	Class sections linked to students, not teachers	Have test score and class section data	Class sections linked to students, not teachers	Class sections linked to students, not teachers	Have test score and class section data

Information about teachers comes from two sources – one is an administrative dataset with basic biographical information including education and experience on all teachers in the district from 2007-2011; although this dataset has information about all teachers, the data available are relatively thin and do not include, for example, the exact grade taught by each teacher. Therefore, some biographical information about teachers must come from the outcome dataset, which provides an estimate of value-added in math and English Language Arts for teachers in grades 4-8 from 2007-2010. Although this dataset only covers 14,651of the 96,680 teachers and administrators in the complete pedagogical information database, it will provide the primary sample for analysis because it includes grade level and value-added data, critical for the analysis under the primary identification strategy. Descriptive statistics on the teachers in this sample are summarized in Table 3-2.

Table 3-2. Descriptive Statistics on Teachers, 2007-2010

	2007-2008	2008-2009	2009-2010
n	82726	82700	80117
Share first year	0.09	0.07	0.03
Average years teaching	9.05	9.30	9.82

As noted above, one outcome of interest will be an estimate of teacher value-added, or the changes in student learning over time that can be attributed to an individual teacher by partialing out prior achievement and student characteristics. The school district under study contracted with researchers at the University of Wisconsin-Madison to obtain estimates of value-added for teachers in grades 4-8 in math and ELA. The analysis was limited to those grades because valid pre- and post- measures are required for value-added analysis, and teachers of younger grades lacked pre-scores while high school exams are not vertically aligned with exams from previous years to allow for a growth measurement. This value-added analysis was not for formal stakes, and was intended to provide information to teachers and principals about teacher performance that might be used to improve professional development; however, these scores were released to the public following a Freedom of Information Act request in September 2010 by several media outlets, following a similar release to the Los Angeles *Times*.

The value-added model is estimated in three stages, taking into account the standard error of measurement for pre- and posttests, prior scores, student characteristics, and peer characteristics within the classroom. The first stage regresses student test scores on prior scores, student characteristics, and classroom indicators; the second stage regresses the residuals from the first stage on classroom characteristics, and the third stage regresses the residuals from the second stage on teacher indicator variables, to take into account that some teachers teach multiple classes and some classrooms have multiple teachers (Value-added Research Center, 2010). Note that this analysis is therefore subject to limitations of analysis using value-added data, including limited scope of outcomes measured by test scores, potential bias due to systematic sorting of teachers and students, and imprecision and lack of reliability of measures of teacher effectiveness.

There are 14,651 unique teachers with value-added scores on math and/or ELA in the 2008-2010 time period, which includes the 2009-2010 year of nearly full inquiry team implementation and the prior school year as a baseline for the difference-in-differences specification. Of the 58,826 value added scores, which are substantially greater in number because teachers receive scores for multiple subjects, grades, and years, the majority are for elementary grades, as middle school teachers each teach more students, and many middle school teachers are not represented as they teach non-tested subjects. The scores are almost evenly divided between ELA (27,559) and math (29,267). Table 3-3 shows descriptive statistics on value-added measures; the main measures are standardized to have a mean of 0 and standard deviation of 0.2, which is very similar across subjects, grades, and years, although appears to be very slightly higher for math than ELA, for higher grades than lower grades, and for later years, on average.

Additional outcomes include teacher retention, obtained from the pedagogical information database, as well as student outcomes on attendance and graduation rates to analyze the effects on important outcomes besides test scores. The teacher retention outcomes allow for analysis with a wider sample, as data are available for all teachers.

TABLE 3-3. DESCRIPTIVE STATISTICS ON VALUE-ADDED, 2008-2010

	Obs	Mean	Std. Dev.	Min	Max
Value-added	56826	0.002	0.1720	-0.99	2.15
Subject					
ELA	27559	0.000	0.147	-0.78	2.15
Math	29267	0.003	0.192	-0.99	1.59
Grade					
4	17010	0.001	0.185	-0.87	1.51
5	16132	0.003	0.181	-0.84	2.15
6	6583	0.008	0.163	-0.68	1.33
7	4661	0.005	0.131	-0.84	0.97
8	5036	0.005	0.151	-0.73	0.87
Year					
2008	17697	0.001	0.130	-0.87	1.59
2009	16733	0.002	0.130	-0.69	1.14
2010	16640	0.006	0.237	-0.86	2.15

The average number of teams per school was high, at 15, with a maximum of 85 teams at one school. Since this initiative was supported by the school district's central administration, there may have been an incentive to inflate the number of teams or exaggerate the extent to which teamwork was actually happening at the school level when entering information about team activity into the database; for instance, the largest team had 127 members, and one teacher was listed as a member of 96 different teams. Therefore, to focus on teams of a more realistic scope, teams with more than 20 members and teachers on more than 10 teams were trimmed from the sample. Of the 9,176 teams originally in the database, 176 were trimmed in this way, leaving 9,000 teams in the analysis sample. Similarly, 383 of the 55,827 teachers represented in the team database were trimmed from the sample.[5] Of the remaining teams, 4,919 focused on a single grade level. There were slightly more teams at the elementary than at the middle school grade levels, potentially because middle school teams may have been more likely to be organized by subject area.

Official teacher IDs are not observable, so teachers were linked to the Inquiry Spaces data by name, creating potential for incorrect linkages due to spelling variations or two teachers sharing the same name. To mitigate this risk, all teacher names were converted to lower-case and trimmed of leading and trailing spaces using Stata's string functions. Teachers with the same first and last name but different schools within the same year were dropped from the sample to reduce the risk of connecting team data to the wrong teacher; this applied to 177 teachers during the sample period. Note that some of these teachers may genuinely be the same person who switched schools during the school year, and this procedure does not address the risk that teachers within the same school may have the same first and last names. Therefore, inferences about this sample cannot be extended to teachers who switch schools during a school year.

Outcome data on student test scores come from the State Education Department for the school district under study. Data are provided on Math and ELA test scores at the grade-demographic subgroup level for grades 3-8, on high school exit exams by cohort and subgroup, and on graduation rates for a smaller number of subgroups by cohort for grades 9-12. I matched these data to teams that focused on the same grade, subject, and subgroup, with subgroups including black or Hispanic students, English language learners, and students with disabilities. Teams that did not identify one of these subgroups or that identified multiple subgroups were matched with outcome data aggregated across all students for the respective grade and subject.

[5] The removal of these outlier observations did not qualitatively alter the results.

Note that because teams do not identify exactly which students were targeted by the team that these estimates are likely conservative, as the outcomes for the treatment group may include several students who were not direct recipients of the treatment; nonetheless, if the outcome of interest is general improvements to instruction and teacher productivity that affect all students then these estimates are appropriate. One limitation of these data are that the state does not report average scores for a grade-subject-subgroup cell if fewer than five students were tested in that group for that year. For smaller schools and for sub-groups for which many students are not tested, particularly students with disabilities and English language learners who may be exempt from certain tests, this results in a large amount of missing data.

To net out preexisting differences between students across schools, I used a gain score methodology in place of raw scores. The gain scores are similar in purpose to value-added measures, in that they attempt to isolate the effects of schooling on student learning from other factors, but differ in that they only adjust for prior scores and not other student characteristics, and that they are at the grade-subject-subgroup level, as opposed to the teacher level, so they are aggregated differently. Gain scores can be calculated in one of two ways – subtracting this year's score from last year's score at the previous grade to net out cohort effects or subtracting this year's score from last year's score at the same grade level to net out grade effects. The main results use the latter methodology, as results are reported at the school level and therefore using the cohort methodology does not allow for comparing students across years when they switch schools, such as from 5th grade in elementary school to 6th grade in middle school. The OLS results are qualitatively the same using either approach, but the instrumental variables results, discussed in the next section, differ, suggesting a potential problem with the instrument. At the high school level, the cohort approach is used for exit exams, as students take those exams at different grade levels and multiple times, so the final results at time of graduation are used. Graduation rates are used simply as raw rates, which could disadvantage schools serving students with greater needs who enter high school with lower expected graduation rates or average time to graduation; this may bias the results against inquiry teams if these are the students who are most likely to be served by the teams, but there is not a clear counterfactual against which to compare graduation outcomes. As a partial test of the limitation of this approach, I separately use four year, five year, and six year graduation rates and find qualitatively similar results.

Table 3-4 presents descriptive statistics on these outcomes. Elementary and middle school math and ELA test scores are generally centered around 670 and slightly decline through grades and over time. Students with disabilities and with limited English proficiency score substantially lower, on average, and students show about 5 points of growth on average in 2008-2009 and lose about 1 point on average in 2009-2010. At the high school level, exit exams are scored on a scale of 1-4, with scores of 3 or higher indicating college readiness and 2 being the minimum to obtain a state-certified diploma. Average scores are between 2 and 3 on this scale, are fairly consistent across years, and once again are substantially lower for students with disabilities and English language learners.

TABLE 3-4. SUMMARY STATISTICS ON TEST SCORE OUTCOMES, 2008-2010

Panel A. 2008-2009				*Panel B. 2009-2010*	
		Mean score	Growth	Mean score	Growth
Elementary/Middle Schools				*Elementary/Middle Schools*	
Grade					
	3	673.7	3.93	668.5	2.43
	4	670.87	3.15	664.94	1.05
	5	671.39	8.23	666.79	0.4
	6	662.03	4.56	658.25	-1.54
	7	661.93	9.03	657.6	-1.68
	8	655.12	3.28	651.79	-1.95
Subgroup					

All Students	668.95	5.27	671.96	-0.99
Black or African American	663.17	4.96	666.28	-1.07
Hispanic or Latino	664.52	5.69	668.09	-0.49
Limited English Proficiency	647.91	6.76	653.42	1.4
Students with Disabilities	645.05	7.78	651.12	2.29
High Schools			*High Schools*	
Subgroup				
All Students	2.54	0.04	2.56	0.03
Black or African American	2.53	0.05	2.55	0.03
Hispanic or Latino	2.49	0.03	2.51	0.02
Limited English Proficiency	1.89	0.02	1.9	-0.01
Students with Disabilities	1.69	0.06	1.75	0.06

SIMPLE MODEL

The most straightforward way to estimate the effect of team participation on teacher productivity, student learning, and other outcomes would be to regress an outcome measure, Y_{jst}, on an indicator for team participation, $TEAM_{jst}$, which would in effect calculate a difference in mean outcomes for teachers who do or do not participate on teams, as in equation 3:

(3) $Y_{jst} = \alpha + \beta TEAM_{jst} + \varepsilon_{jst}$

where, as above, *j* index teachers, *s* indexes schools, and *t* indexes time.

There are, however, potential problems with this approach. Two key challenges to isolating the effects of teacher-led, structured collaboration are that this type of collaboration is difficult to measure and is undoubtedly correlated with other important omitted variables. Nearly all principals and teachers would likely report that they are “collaborative,” but in order to have a measurable impact on students, that collaboration would likely have to be fundamentally different in subtle but important ways that are difficult to observe in quantitative data. Further, these subtle differences likely do not come about independently of other important factors that determine teacher effectiveness and overall student achievement, including the capacity of the school leader and overall school culture. Based on the assignment mechanism to teams, there are likely multiple potential sources of selection bias. These include voluntary participation on teams by teachers for various reasons, possibly based on important prior characteristics that may also determine outcomes, principal assignment of teachers to teams, and the quality of each individual team based on its focus and composition. In the first year of the initiative, descriptive data suggest that teachers primarily volunteered for teams, suggesting that the first source of selection bias may be the most problematic.

Specifically, the concern is that $cov(TEAM_{jst}, \varepsilon_{jst}) \neq 0$. In other words, there are likely omitted variables that determine both likelihood of team participation and outcomes of interest, such as individual motivation and effort and the choices of colleagues. It is highly likely that this represents a classic simultaneity problem in that it is difficult to determine whether good teachers collaborate more or whether more collaboration makes teachers better. The direction of the omitted variable bias is not certain, however; for instance, principals may assign weaker teachers to work with their colleagues to improve, which may mask positive effects of collaboration. Teachers themselves may select onto more or less selective teams, more effective colleagues may choose to work together, or principals may assign stronger teachers to work with colleagues who need support on teams.

A third problem, as noted in the review of the literature on teamwork and collaboration, is that not all teamwork is equal, and that while some types of team-based activities may contribute to productivity, others may have no effect or even be negative. This is particularly problematic in this setting, given that the policy under study technically mandates team participation, especially in the 2009-2010 school year. Both failure to capture variability in team quality, as well as inability to distinguish genuine team work from perfunctory compliance with external mandates, can be described as a case of measurement error, where $TEAM_{jst} =$

$TEAM *_{jst} + \mu$. $TEAM *_{jst}$ represents a team's true quality. If μ were uncorrelated with ε_{jst}, as in classical measurement error, this basic model would be biased toward zero. A more accurate measure of $TEAM *_{jst}$ could help mitigate this bias. However, there are very likely school and teacher-level unobservable factors that will influence both quality of teamwork and outcomes, implying that $cov(\varepsilon, TEAM *) \neq 0$. Therefore, attempts to reduce bias due to measurement error may exacerbate the aforementioned omitted variables bias. Further, some apparent gains to teamwork may represent a case of management bias, whereby productivity returns attributed to a particular input in a production process are in fact the result of the intangible skill and influence of the leader who chose that input (Mundlak, 1961).

Given these measurement, endogeneity, and heterogeneity challenges, I estimate the effect of collaboration under a series of natural experiments by which teachers were induced by plausibly random circumstances into collaboration. The study includes multiple quasi-experimental approaches, as opposed to just one, for two reasons: each approach may be subject to its own potential sources of bias, but if the sources of bias across approaches are not correlated with one another and the results are qualitatively similar, it increases confidence in any "true" effect of collaboration. There is, however, no way to empirically test whether or not any potential sources of bias cancel out, so while consistent results across models increase confidence, they do not guarantee valid causal estimates. Secondly, since the approaches isolate the effects of collaboration on peculiar groups of teachers, such as first-year teachers in particular grades and subjects and grade-switchers within a school, consistency across results increases confidence that results are generalizable and are not likely due to idiosyncrasies in the sub-samples used for identification of a local average treatment effect. Finally, given the complex and in some cases competing mechanisms by which collaboration can affect teacher productivity, differences across models can provide evidence on how teamwork affects teachers and which mechanisms are at play. Further analyses, incorporating measures of team quality as well as team characteristics as covariates and interaction effects, are more descriptive in nature and help inform the qualitative analysis.

Overall, I made methodological choices to obtain conservative estimates of the effects of inquiry. Where a decision could potentially induce bias, and when otherwise lacking in theoretical and empirical guidance, I chose the option that would more likely lead to downwardly biased estimates, both to err on the side of caution and to more clearly obtain a “lower-bound” estimate of the policy effects of inquiry for consistent interpretation. For example, all standard errors are clustered at the school level. It is highly likely that errors are correlated within schools, given the sorting of students and teachers into schools and the unobserved effects of leadership and school culture. Ignoring such patterns of correlation would lead to downwardly biased standard errors and incorrect inferences. Clustering at the appropriate level so as to model the error variance using observed correlational patterns within the data addresses both heteroskedasticity and serial correlation, leading to efficient and unbiased estimates; however, the correct level at which to cluster standard errors is not always clear, a priori. Clustering could occur at the teacher, grade, or subject levels within schools. As a general rule, clustering at higher levels within the data leads to larger standard errors, so school-level clustering is the conservative assumption (Cameron and Miller, 2014).

Similarly, the available outcome measures are likely to produce lower bound estimates of the direct effects of inquiry. In its original conception, inquiry teams were intended to focus their efforts on small sub-groups of students with similar instructional needs, identified as a skill gap common to a group of students who share the same grade, subject area, and/or demographic characteristics. The actual students targeted by inquiry teams are not identifiable in the data, so outcomes are more aggregated, at the teacher or grade-subject-subgroup level. This allows for the estimates to capture any spillover effects on students not directly targeted by inquiry teams, as well as more general effects on teacher productivity, but provides a conservative estimate of the direct effects of the inquiry process on the students it targets. As a sensitivity analysis in chapter 8, the cost-benefit analysis, I consider what assumptions or effects would be necessary in order for the policy to "break even," or for the benefits of the policy to exceed the costs.

MODEL #1: GRADE SWITCHERS

The first quasi-experiment takes advantage of the phase-in of the initiative, as well as teachers who switch grades within the same school from a grade without an inquiry team during the 2008-2009 school year to a grade with an inquiry team in the 2009-2010 school year, following a similar identification strategy used in Chetty, Friedman, and Rockoff (2013a and 2013b). Ordinarily, grade-switching might be problematic because it could indicate that, for example, principals are moving more effective teachers to high-stakes testing grades and less effective teachers away from those grades, but all teachers in this sample teach grades and subjects with high-stakes standardized tests. This model estimates the difference-in-differences in value-added outcomes for teachers who switch from a non-inquiry grade to a grade with an inquiry team in 2009-2010 to all other teachers, as in equation 4:

(4) $y_{jst} = \alpha + \beta_1 POST_{jst} + \beta_2 SWITCH_{jst} + \beta_3 POST_t SWITCH_{jst} + X_{jst}\beta_4 + Z_{st}\beta_5 + \varepsilon_{jst}$

where j indexes teachers, s indexes schools, t indexes time, y_{jst} is the outcome of interest, primarily teacher value-added, $POST_{jst}$ is an indicator for the 2009-2010 school year, $X_{jst}$is a vector of teacher-level controls, Z_{st} is a vector of school level controls, and β_3 is the coefficient of interest, with standard errors clustered at the school level. The critical assumption for this analysis is that the change over time in value-added measures for non-grade switchers is a valid counterfactual for changes over time in grade switchers. This assumption is plausible because even though many other policy changes occurring during a relatively tumultuous time for the school district might ordinarily be of concern as confounding factors, it is highly likely that switchers and non-switchers would be subject to the same alternative policies. Further, while switching grades may be endogenous, it is unlikely that a principal would switch the grade a teacher is teaching simply to place the teacher on an inquiry team, when creating a new team at the teacher's present grade level would be a far less disruptive way to achieve the same objective. One possible source of downward bias is that value-added may be expected to dip when teachers switch grades as they adjust to a new curriculum. As one test of this assumption, a robustness check for this analysis restricts the sample solely to grade-switchers to net out any effect of the switching itself. This model is restricted to 2009-2010 because that is the only year for which individual teachers on teams are identifiable, and is further restricted to teachers in grades 4-8 with students who took the state math and/or English Language Arts (ELA) exams, as those are the only teachers for whom the grade levels they taught are observable.

MODEL #2: FIRST-YEAR TEACHERS

The grade-switcher model only allows for estimation of the effect of the policy for a small sub-set of teachers, and only in the year 2009-2010, when the objective was for 90% of teachers to be participating in a team. The second model allows for focusing specifically on teachers in their first year, which may be a particularly critical year for the development of teacher human capital. In particular if the mechanism through which teams operate is to address market failure in teacher preparation programs because learning to teach is highly experiential, or if teaching is idiosyncratic to the specific context and knowledge and skills are highly school-specific, then the effects would be expected to be particularly pronounced among new teachers. Further, because new teachers are hired to fill vacancies that may arise as late as the summer, after planning for the next school year is already underway, it is less likely that first-year teachers are strategically placed on inquiry teams and more likely that such placement is idiosyncratic and uncorrelated with inquiry teams at a school. One possible threat to validity would occur if teachers were especially likely to leave the previous year if they were on poor-quality inquiry teams, implying that grade-subject combinations with vacancies and inquiry teams represent an unusually weak sub-sample of teams. This would bias results against teams. However, heterogeneity analysis in Chapter 6 suggests that most teams implement with low-intensity, suggesting that poor inquiry work is unlikely to be a sufficient reason for a teacher to leave a school. Finally, the first-year teacher model allows for exploration of prior years of data that may indicate enthusiastic, early adopter effects of the first wave of inquiry teams, or improving effects over time due to gaining experience with teamwork. One limitation of this model is that individual teachers are not identifiable in the data in the 2007-2008 and 2008-2009 school years, so the assumption must be made that if a team exists at a particular grade level with a first-year teacher, the new teacher is a member of that team. There are two potential concerns created by this assumption – one is that while principals may not strategically place first-year teachers on grades with teams, they may strategically elect to have teams on grades with first-year teachers, and the second is this once again limits the sample to those teachers for whom the grade they teach is observable, which is the set of teachers in the value-added data.

Of particular concern among new teachers is the expected payoff to investment in their human capital, given high turnover rates, especially in high-poverty urban schools. Therefore, the primary outcome of interest will be whether or not a first-year teacher remains for a second year, as well as how many years the teacher remains in teaching, up through the 2010-2011 school year, the last period for which retention data are observable. Note that, although these outcomes may suggest the appropriateness of non-linear models – probit or logit for the binary outcome of returning the following year and a survival model for the duration analysis – the properties of these models in conjunction with quasi-experimental methods such as the difference-in-differences specification used here are not well established. Therefore, the primary results will assume linearity. The key issue this raises relates to the distribution of the error term; for a number of reasons, when analyzing duration as an outcome the error is unlikely to be distributed normally. Most obviously, there are no negative durations. There is also right censoring, as I only observe teacher retention through 2011, which is not nearly enough time for all teachers who began teaching in the 2007-2010 time period to exit teaching. Finally, teacher retention is likely bimodal, with some teachers for whom the profession is not a good match self-selecting out relatively quickly and others remaining for their full careers (Cleves, Gould and Gutierrez, 2002). In practice, however, none of these issues are likely to significantly alter the results of the analysis, and assuming normality and linearity simplifies interpretation of results and allows for use of quasi-experimental methods, as in equation 5, specified similarly to equation 4:

$$(5)\ y_{jsgt} = \beta_1 POST_t + \beta_2 TEAM_{jsgt} + \beta_3 POST_t TEAM_{jsgt} + X_s\beta_4 + \varepsilon_{jsgt}$$

where y_{jst}represents a range of outcomes for teacher j in school *s* at grade g and time *t.* These outcomes include whether or not a teacher returns for a second year, total years teaching through the 2010-2011 school year, and value-added measures.

MODEL #3: INSTRUMENTAL VARIABLES

The third model uses two-stage least squares (2SLS) to estimate the effect of inquiry teams on various student-level outcomes, based on the predicted incidence of a team at the grade or subject level. The prediction is based on the number of class sections offered at that grade or in that subject within a school, as shown in equations 6 and 7, under the theory that as the policy phases in across a school, teams are most likely to form where there is a "critical mass" of teachers. This instrument is based on the literature on optimal team size and on homogeneity as a predictor of team effectiveness. Although the outcome of the first stage regression of team existence on the instrument of class sections is binary, this regression is estimated using OLS. As Angrist (2001) notes, estimating a binary first-stage using probit or logit yields an inconsistent second stage unless the first-stage is correctly specified, whereas the second stage results are consistent even with a linear approximation in the first-stage. Data on class sections come from the Class Size Reports published by the school district each year; these reports provide grade-specific enrollments, number of class sections, and average class size for each grade in elementary and middle schools and for each subject in high schools. Unfortunately, class size data are not disaggregated at the grade level in 2007-2008, so I ran these models for the 2008-2009 and 2009-2010 inquiry data only.

(6) $x_{gst} = \alpha + \beta_1 Sections_{gst} + \beta_2 Sections_{gst}^2 + \gamma_s + \varepsilon_{gst}$

(7) $Y_{igst} = \gamma_1 + \gamma_2 \widehat{x_{gst}} + \varepsilon_{igst}$

$x_{gst}$is an indicator for having a team at grade g in school *s* at time *t*, $Sections_{gst}$ is the number of class sections at that grade level, γ_s is a school fixed effect, $\widehat{x_{gst}}$is the predicted probability of having a team based on the number of sections, and $Y_{igst}$is a vector of outcomes for student *i* in the grade, which includes attendance and graduation.

Several assumptions are required for an instrumental variables estimate to be valid. First, the instrument cannot be weak - the covariance between the instrument and the endogenous variable, in this case, the existence of a team, must be greater than zero. Secondly, the instrument must satisfy the exclusion restriction, implying that the instrument only acts on the outcome through the channel of the endogenous variable and is uncorrelated with the error in the second-stage equation. Note that one limitation of this analysis is that, because of the way school planning occurs, there is likely to be relatively little variation between grades in enrollment and number of sections, so the prediction for the existence of teams at a particular grade level or subject area is based on a small amount of variance. This could potentially lead to a weak instrument problem, which exacerbates the bias of 2SLS estimates in small samples. The first assumption is testable using the F-statistic for the first-stage regression, among other tests, but the exclusion restriction cannot be tested directly.

In general, instrumental variables estimates provide a causal effect of the treatment on compliers – in this case, those grades and subjects which have a team because of the number of class sections at that grade or subject, excluding those which would always have a team and those which would never have a team. Compliers are not directly observable, as those grades and subjects with teams also include always-takers, or those that would have a team regardless of the number of sections. An important assumption for this interpretation, however, is that there are no defiers – no grades or subjects that would elect *not* to have a team because of more class sections and would have a team with fewer class sections. This monotonicity assumption is potentially problematic in this case, as the likelihood of having a team does not increase indefinitely with the number of teachers at a grade or in a subject area. There is likely a point beyond which there are too many teachers for efficient teamwork, and they may have two or more teams or may elect to have a team at another grade level or subject area where the numbers are more manageable. Therefore, to test this assumption the first stage is specified in multiple ways, including linearly, with a quadratic term to allow the likelihood of having a team to decrease with class sections above a critical point, and for sub-samples of schools with smaller numbers of class sections, to determine robustness of results to the specification of the first-stage and to this assumption, as suggested by Dieterle and Snell (2014).

DESCRIPTIVE ANALYSIS OF HETEROGENEITY

As noted above, all of these models are subject to potentially severe attenuation bias due to measurement error in the indicator variable for having a team, particularly given that the quality of team collaboration and the actions teachers take as a result of teamwork is likely to vary tremendously between schools, is not able to be adequately captured in data, and is itself likely to be endogenous. Although the instrumental variables approach can mitigate measurement error, model 3 instruments for team participation, not for team quality. I explore additional measures of quality of team participation, as well as possible instrumental variables to use to predict quality, as a way to address measurement error. Measures of quality include the number of "inquiry cycles" the team completed – in 2009-2010, for instance, it was possible to input up to 5 cycles in the database recording inquiry activity, but only four out of over 9,000 teams entered data for all five cycles – and descriptive coding of the team process, focus, and activities based on reading a sub-sample of entries into the inquiry team database. Differences in accountability pressure due to staggering of the qualitative accountability system are considered as an instrumental variable for team quality in Appendix B.

Chapter 4 METHODS AND DATA: QUALITATIVE AND COST ANALYSES

QUALITATIVE ANALYSIS

There is extensive literature on the heterogeneity of quality of collaboration, as well as the difficulties in measuring quality with validity, reliability, and precision. Further, the quantitative analysis estimates the effects of a policy mandating teamwork, but due to heterogeneity and measurement issues, does not estimate the effect of teamwork itself, particularly if that teamwork is of high quality. Accordingly, I complement the quantitative analysis on the effects of teamwork, as well as descriptive analysis on the heterogeneity of effects and the conditions that contribute to success, with a qualitative case study analysis of four teams to further explore the conditions and processes that constitute effective teamwork. The qualitative data were collected prior to the quantitative analysis, although they represent a later time period in the evolution of the same intervention in the school district. The qualitative analysis was informed by, and occurred subsequent to, the quantitative analysis in a mixed methods, sequential explanatory design (Creswell and Clark, 2011). In essence, given the low intensity of implementation seen in the quantitative analysis and high degree of heterogeneity, the qualitative analysis was designed to attempt to explain and contextualize quantitative findings, to uncover practices and attitudes that may contribute to more successful inquiry, and to generate hypotheses for further quantitative analysis.

The topic, research questions, and setting lend themselves to a case study analysis for a number of reasons. Teams within schools constitute what qualitative methodologists, starting with Louis Smith, refer to as a "bounded system," whereby each team has distinct characteristics and recognizable edges. Case studies provide an in-depth description and analysis of each particular context, as well as cross-cutting analysis comparing and contrasting cases, based on multiple data sources including observations and interviews. The case study methodology emphasizes exploration and discovery over testing specific hypotheses, and therefore is well suited to uncovering team processes and the conditions that lead to team success (Merriam, 2009).

As noted above, the data available for analysis include transcripts from between 8-10 observations of team meetings and 1-2 semi-structured interviews with individual or groups of team members at each of four schools. The schools were selected to represent a range of student demographics among schools that showed great promise in team practice based on analysis of the team database and recommendations by central office staff. All data were collected and transcribed by me and a team of students and professors at Teachers College and Columbia Law School; transcriptions have been entered into the Dedoose qualitative analysis software program to facilitate analysis. All observation and interview transcripts were carefully read and coded according to a topical and analytical coding scheme developed in accordance with the conceptual framework and the literature on teamwork (see Appendix C for coding scheme). I then wrote brief summary descriptions of the processes teams underwent in each case. Using the analytic codes and case descriptions, I have summarized the findings for each case using data tables based on analytical categories from my conceptual framework, including instances of particular conditions, processes, and outcomes, and compared and contrasted findings within and across cases (Yin, 2013).

The case study method has several potential limitations. In particular, the sample for case studies is necessarily small and nowhere near as large and representative as the 13,000 teams in the quantitative sample. Therefore, findings may be less generalizable. Further, the results are primarily descriptive, not causal, as important, unobservable elements of the school and team context may be influencing conditions, processes, and outcomes. The case study sample was deliberately selected to reflect high-quality implementation of the inquiry team initiative, and thus results may reflect selection bias – schools with strong leadership and teachers may be more likely to collaborate well and see positive results, as opposed to collaboration contributing to teacher quality. Relatedly, qualitative analysis relies heavily on methodological choices and interpretation by the researcher. A strength and limitation is its inherent subjectivity, whereby it is possible to more deeply explore nuance, context, and mechanisms more deeply than in quantitative methodology, but the results may be sensitive to limitations of sample, context, and the researcher's own judgment. The primary guard against potential biases is the systematic development and testing of alternative hypotheses that may also explain observed patterns in the data.

The research team collected qualitative data on four teams in one elementary and three middle schools during the 2011-2012 school year. The primary unit of analysis for the qualitative study is the team. A purposive sample of four teams has been selected, each comprising between 5 and 10 teachers; two administrators (a principal and an assistant principal) are a permanent part of one team and school administrators are occasionally members of the other teams.

Schools were selected based on quantitative measures of school performance, a qualitative evaluation of collaboration by outside experts using a school quality rubric, and recommendations by non-profit organizations that support the schools. Two of the teams represent grade-level teams, with teachers of different subjects sharing interdisciplinary practice and evaluating the instructional needs of individual students. One team represents a group of teachers who joined together to focus on similar pedagogical skills, and a final team represents a school-wide team meant to coordinate the activities of other teams and align curriculum and instruction with research-based best practices across the school. While the sample is not random, it includes a range of school demographics, one school with a very high population of English language learners, and one school with an above-average population of students with disabilities (see Table 4-1). Each team was observed between eight and ten times, and at least one interview was conducted with each team. Observations and semi-structured interviews following an established protocol with iterative follow-up questions were audio-recorded and transcribed. The observations intended to document team practice, focusing on whether teams followed established protocols and agendas, leadership on the team, group dynamics, how teams made decisions and the types of evidence consulted, and whether and how teams followed up on plans discussed at each meeting. Interview questions addressed similar topics, but gave teachers and principals the opportunity to reflect upon and discuss the extent to which choices made were deliberate, the process for choosing how to implement inquiry teams, and any proximal outcomes, including changes in teacher attitudes or instructional practice.

TABLE 4-1. DESCRIPTION OF CASE STUDY SAMPLE

	School A	School B	School C	School D
School Characteristics				
Enrollment	190	500	250	400
Grades Served	6-8	Pre-k-5	6-8	6-8
Race/Ethnicity				
Black	40%	5%	20%	5%
Hispanic	50%	20%	70%	90%
Asian	<5%	30%	5%	--
White	10%	45%	5%	--
American Indian or Alaska Native	<5%	--	<5%	--
English Language Learners (ELLs)	10%	30%	5%	50%
Individualized Education Programs (IEPs)	40%	15%	20%	20%
Free or Reduced-Price Lunch	75%	80%	75%	95%

Note: Rounded to nearest 5% to protect identity of schools.

COST ANALYSIS

To address the questions of the costs of inquiry teams and how those costs compare to potential benefits, I followed the ingredients method, developed by Henry Levin (Levin and McEwan, 2001). As noted in the theoretical model, a critical element in the time allocation model by which teachers choose how to perform their work is the marginal cost of each activity, relative to its marginal benefit and compared to alternatives, expressed in the model in terms of time and effort.

While the theoretical model focuses on teachers, school leaders play an important role - their policies and choices strongly influence teachers' choices, and the costs and benefits of how teachers spend their time clearly relate to school leaders' objective functions and budget constraints. Even when the direct costs of additional teacher collaboration are minimal, there are clearly opportunity costs when teachers engage in collaborative activities; they could be engaging in individual work that may be more productive, they could be providing additional professional services to the school such as tutoring struggling students or helping with administrative tasks, and there may even be direct financial outlays if collaborative work requires, for example, overtime wages, as it did in some cases with the inquiry team initiative. Therefore, a full economic evaluation of the initiative requires evaluation of not just the effects, but also the costs. Given continued investment in teacher professional development, and renewed interest in collaboration as a professional development tool – the contract with the teachers' union in the school district in this study now mandates weekly peer collaboration – information about the costs and benefits of collaboration is particularly important.

The ingredients method, in contrast to analysis of expenditures or budgets, attempts to account for all resources, or ingredients, used in an intervention to fully account for the economic or opportunity cost. Simply looking at budgets or other sources of data on financial outlays may miss important ingredients that are donated or provided in-kind, such as time teachers reallocate to inquiry teams from other work, or uncompensated time they may spend after school on inquiry and related work. Budgets may also not account for the reallocation of existing resources from one use to another, and may not take into account important costs that are not typically reported in annual budgets, such as the costs of fringe benefits or the depreciation of capital goods. Further, the careful tabulation of ingredients based on document analysis and interviews with stakeholders provides a more thorough picture of exactly what an intervention entails, acting as a rudimentary implementation analysis, as well.

To gather ingredients data, I drew upon the rich implementation data available in the team databases over three years, as well as the implementation studies performed by CPRE (2008, 2010) and Talbert (2010). Data on ingredients were then combined with information about national average prices for educational resources from the Bureau of Labor Statistics, National Center for Education Statistics, and other sources, gathered to form the Educational Resource Price Database by the Center for Benefit-Cost Studies of Education at Teachers College, Columbia University, to obtain a range of per-school costs as well as a pooled average per student cost for the program.

Cost data were then combined with effectiveness data from the quantitative analysis for cost-benefit analysis. A cost-benefit analysis assesses the ratio of the benefits to the costs of a program, both measured in monetary terms. The necessary condition for implementing a program is that the benefits exceed the costs – in other words, the net benefits are positive or the benefit-cost ratio is greater than one. A sufficient condition for implementing the program is that the benefits exceed the costs by a margin greater than all alternatives, but that analysis requires the existence of meaningful alternative programs that are evaluated using similar methodology and assumptions.

To perform the within-program cost-benefit analysis to determine whether the program is worthwhile on its own, I applied the measured effects to shadow prices for various outcomes that have been reported in the literature; for instance, Chetty, Friedman, and Rockoff (2013b) have estimates of the value of increases in teacher value-added based on differences in students' future earnings, and Ronfeldt, Loeb, and Wyckoff (2013) have estimates of the detrimental effects of turnover on student achievement, which have been translated into monetary terms based on associations between achievement, probability of high school graduation, and labor market outcomes.

Chapter 5 EFFECTS OF INQUIRY TEAM PARTICIPATION ON MEASURES OF TEACHER PRODUCTIVITY

This chapter presents findings on the association between team participation and various outcomes that serve as indicators of teacher human capital or productivity, including measures of value-added, gains in student test scores, student graduation rates, and teacher retention rates. This chapter also includes estimates of the causal effect of inquiry on these outcomes using the three quasi-experimental models described above. Consistent results across the models will provide strong evidence of an overall effect of inquiry on the development of teacher human capital, whereas differences in results across models are more difficult to interpret and may be due to differences in outcomes, in local average treatment effects (LATEs) for each model, or potential bias, wherein one estimate is valid and another is invalid due to violations of the necessary assumptions for internal validity. Whenever possible, I ran robustness checks across models; for example, the value-added outcomes are the primary outcomes of interest for the grade switcher model, but are also used in the instrumental variables and first-year teacher models to determine whether differences in estimated effects are due to different outcome measures or different analytic approaches.

SIMPLE MODEL

Missing Data

Given that data for the value-added models, including the simple model described here and the grade switcher model in the next section, come from a variety of different administrative datasets over multiple years, bias or inefficiency in estimates due to missing data is a potential concern. Most significantly, as noted above, the outcome variable is restricted to teachers for whom there is a value-added score and therefore who taught math and/or ELA in grades 4-8 in

2009-2010. For the difference-in-differences model, there is the further sample restriction to those teachers for whom value-added scores and grades taught are observable in 2008-2009 and 2009-2010. This represents 33,354 observations across the two years, out of 98,852 total different teachers observed over those two years in the team and pedagogical information databases. Within these sample restrictions, data is occasionally missing on some covariates, particularly pre-intervention value-added and the standard deviation of prior value-added at the team level, which requires additional prior years of data. Further, school-level demographic covariates are not observed in the database provided by the district for a very small number of schools, possibly because the schools were phasing out or because they were too small to report summary statistics without risking privacy violations.

The available data on covariates, including number of observations, are presented in Table 5-1. The main results presented in this section and the next section exclude observations for which there is incomplete information. As a robustness check, I also ran models that imputed zero or the mean across other observations for missing variables, with dummy variables to indicate missingness, and found results that were qualitatively similar. In some cases, point estimates were very slightly higher and estimates that were marginally significant in the main specification became significant with the imputed data, possibly because of greater power due to larger samples. However, the preferred results are those that do not impute zero or the mean for missing data, as they are generally more conservative and avert the risk of overfitting the model due to imputation.

Table 5-1 Summary Statistics on Covariates in Team-Switcher Model

Variable	Obs	Mean	Std. Dev.
Value-added	33,354	0.004	0.191
On a Team	98,852	0.737	0.440
School ID	98,852	N/A	N/A

% Black (school-level demographic)	91,387	0.302	0.284
Prior VA	21,998	0.007	0.130
Team-level Prior VA SD	42,061	0.116	0.044

To examine the basic association between being a member of a team and teacher value-added, I ran a series of models using Ordinary Least Squares (OLS); Table 5-2 presents the results. Given that many teachers were on multiple teams, the data were collapsed to the teacher-subject-grade-year level for ease of interpretation and consistency with the value-added outcomes, which were also at that level. The key independent variable is a dummy variable indicator for whether a teacher is on any teams. Column 1 shows the basic correlation between team participation and value-added measures, which is small and statistically insignificant. Columns 2-4 show the results with various covariates added to the model, including measures of value-added prior to the intervention, a measure of the spread of value-added on the team to address heterogeneity of teacher effectiveness on teams, an indicator for middle schools, and school-level student demographic measures. Although the coefficients become larger and are estimated with somewhat greater precision, they are still substantively small and statistically insignificant. Given that standard errors are also substantively small, at less than 0.01 standard deviations, it is likely that the raw average effect of inquiry team participation on teacher value-added in the same year is zero. Further, the R^2 statistic on models without prior value-added is exceptionally low, at less than 0.01, suggesting that inquiry team participation explains very little of the variation in value-added. Since several of the following identification strategies rely upon inquiry teams at a single grade level, I also test whether grade teams and other teams are systematically different by excluding teams that focus on only one grade in column 5. The coefficient of 0.018 is marginally greater than the coefficient of 0.013 in the similar model for all teams, suggesting that non-grade level teams are slightly more effective than grade-level teams. The coefficient excluding grade teams is similarly small and not statistically significant, and the two coefficients are not statistically different from one another using a Chi square test.

TABLE 5-2 ORDINARY LEAST SQUARES ESTIMATES OF THE ASSOCIATION BETWEEN TEAM PARTICIPATION AND VALUE-ADDED

	(1)	(2)	(3)	(4)	(5)
	VA	VA	VA	VA	VA
On a team	0.0000476	0.00995	0.00424	0.0129	0.0181
	(0.00537)	(0.00687)	(0.00531)	(0.00700)	(0.0127)
Prior VA		0.713***		0.711***	0.706***
		(0.0238)		(0.0239)	(0.0252)
SD of Prior VA on Team		0.177*		0.177	0.251
		(0.0896)		(0.0911)	(0.143)
Constant	0.00629	-0.0140	0.0341*	0.0125	0.005
	(0.00357)	(0.0127)	(0.0168)	(0.0213)	(0.0256)
Observations	16628	11803	16223	11644	9687
Demographic Covariates			X	X	X
Excluding Grade Teams					X

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As noted above, there are a number of concerns with this basic approach to estimating the causal effects of teamwork on teacher value-added. Team participation is not the only margin on which there could be selection bias. There are likely complex interaction effects related to team composition and the quality of teamwork, as well as issues related to the reflection problem, whereby it is difficult to disentangle the effects of the group on an individual and the group effects as the aggregated individual effects (Manski, 1993). Column 1 in Table 5-3 adds school fixed effects to estimate within-school variation in teacher value-added based on whether or not teachers participate on teams. This model captures unobserved elements of leadership and school quality that might influence team participation and team quality. The remaining effect is small and remains statistically insignificant, which could be because any effects of teams are actually due to differences in leadership and culture across schools, or could be because very little variation in team participation and value added exists within schools.

The school fixed effects specification may address some of those sources of bias to the extent that they are consistent across the school, but does not address underlying teacher characteristics that could simultaneously determine team participation and outcomes. A teacher-subject fixed effect estimate nets out time invariant teacher characteristics within each subject area; the results of this model are presented in Column 2 of Table 5-3. The fixed effects are identified based on the introduction of the program over time, and measure changes in value-added within the same teacher and subject. The effect is positive and statistically significant, although is still substantively quite small. The fact that this coefficient is higher than the OLS estimate indicates that selection into teams, at least by the 2009-2010 school year, could be negatively associated with value-added, as teachers who need the most help may be most likely to be assigned to teams by principals. Note that there are several major limitations to this model that caution against interpreting the results causally: one is that unobserved teacher characteristics that change over time are not captured, and secondly is that since the model captures changes in value-added between 2009 and 2010, some portion of the effect attributed to team work may partly be a year shock, as the mean value-added is 0.004 units higher in 2010 than in 2009. Further, because the amount of variation is limited to changes in teamwork over time, the issues of measurement error and any remaining omitted variables bias that is not captured by the teacher fixed effects are exacerbated, as the total variation is significantly reduced. Note that these models were similarly run imputing missing data and splitting the sample according to grade-level teams, and results were qualitatively similar.

TABLE 5-3 FIXED EFFECTS ESTIMATES OF THE EFFECTS OF TEAMWORK ON VALUE-ADDED

	(1) VA	(2) VA
On a team	0.0018	0.0113*
	(0.00645)	(0.00506)
Prior VA	0.674***	-0.300***
	(0.0163)	(0.0159)
SD of Prior VA on Team	0.0618	
	(0.0819)	
School FE	X	
Teacher-Subject FE		X
Demographic Covariates		X
Observations	11644	21998

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

One limitation of the fixed effects estimate is that the only source of variation is changes over time within the same exact teacher, limiting the sample to teachers with observations for both years and excluding all between-teacher variation.

MODEL #1: GRADE SWITCHERS

The first quasi-experimental model takes advantage of the gradual phase-in of the inquiry team initiative. This analysis focuses on teachers who switch grades within a school from a grade without a team in 2008-2009 to one with a team in 2009-2010.

Of the 14,651 teachers who appear in the value-added data, and for whom there is data for both the 2008-2009 and 2009-2010 school years, 1,975 switched grades. Of those, 777 switched to a grade with a team, and of these 191 switched from a grade without a team to a grade with a team, which is the treatment identified in the difference-in-differences specification. Note that this analysis is restricted to teams that focused on a single grade level; some teams focused on multiple grade levels, on subject areas, or on particular subgroups of students, and are excluded from the analysis. Therefore, to the extent that those teams may have contributed to teacher value-added, this exclusion leads to a downward bias in the estimated effects of teamwork. Note that Rothstein (2014) critiqued the switching quasi-experiment used by Chetty, Friedman, and Rockoff (2013a) to estimate the bias of value-added estimates using data from North Carolina and found that teacher switching is correlated with changes in student prior test scores. In general, Rothstein finds that teachers leaving a grade are replaced by others with higher prior VA scores when student test scores are increasing and lower prior VA scores when student test scores are decreasing. The model that restricts the sample to switchers nets out this effect, and may be the most credible due to this concern.

Table 5-4 presents the results from the main difference-in-differences specification, described in equation 4. The key coefficient of interest is the interaction between an indicator for switching from a grade without a team in 2009 to a grade with a team in 2010 with a fixed effect for the year 2010. The indicator for switching to a grade with a team is applied to all teachers who switch to that grade within a school in 2009, so the relevant coefficient estimates the differences over time between those who switched to a grade with a team and those who did not. The counterfactual therefore includes teachers who did not switch grades and switchers who were always on a team or never on a team. Column 1 presents the results of the basic model, which are positive but small and not statistically significant.

Columns 2 and 3 of Table 5-4 add covariates, both as a check on the validity of the difference-in-difference assumptions and to increase the precision of the estimates. If the difference-in-difference estimates are valid and the assumptions hold, the coefficients should not change dramatically between the models; the coefficient on the interaction term remains insignificant and becomes trivially negative when including prior VA scores, and the standard error actually increases, but these differences are likely due to sample size restrictions based on the availability of value-added scores. Model 4 replaces school-level demographic covariates with school fixed effects, and the results become slightly more negative, but are still small and not statistically different from zero. Models were also estimated with zero imputed for missing values on prior value-added and the standard deviation of prior value-added for the team, along with dummy variables to indicate missing values, and were qualitatively similar. Although they were estimated with slightly greater precision, the coefficients of interest were still not statistically significant.

TABLE 5-4 DIFFERENCE-IN-DIFFERENCES ESTIMATE OF THE EFFECT OF TEAMWORK ON VALUE-ADDED

	(1) Value Added	(2) Value Added	(3) Value Added	(4) Value Added
Switch to team in 2010	0.00230	0.00318	-0.000485	-0.0023
	(0.0140)	(0.0144)	(0.0183)	(0.0230)
2010	0.00464	0.00461	0.00615	0.0042
	(0.00312)	(0.00315)	(0.00364)	(0.0187)
Switch to team grade	-0.00777	-0.00641	0.00125	0.0043
	(0.00824)	(0.00832)	(0.00941)	(0.0187)
2006-2007 VA Score			0.512^{***}	0.4610***
			(0.0162)	(0.0101)
Team SD - Prior VA			0.204^{***}	0.1378**
			(0.0477)	(0.0449)
Constant	0.00172	0.0242^{*}	0.000867	-0.0103
	(0.00168)	(0.0104)	(0.0124)	(0.0059)
Demographic Covariates		X	X	
School FEs				X
Observations	33354	32484	21338	21675

Standard errors, clustered at school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Two potential concerns with this analysis are that switching grades is not exogenous, therefore violating the assumption that differences over time between switchers and non-switchers would be expected to be similar in the absence of inquiry teams, and that switching itself may have an effect on value-added. Therefore, Table 5-5 presents the results of this analysis restricted just to the sample of 1,975 teachers who switched grades. The results are qualitatively similar to the results for the full sample, but are generally much smaller, in part due to smaller samples; none of the coefficients under any specification of the model are statistically significant.

TABLE 5-5 DIFFERENCE-IN-DIFFERENCES ESTIMATE OF EFFECT OF TEAMWORK ON VALUE-ADDED, GRADE SWITCHERS

	(1) Value Added	(2) Value Added	(3) Value Added	(4) Value Added
Switch to team in 2010	-0.00148	0.000141	0.00069	0.0005
	(0.0152)	(0.0156)	(0.0189)	(0.024)
2010	0.00842	0.00736	0.00652	-0.0016
	(0.00698)	(0.00716)	(0.00791)	(0.0068)
Switch to team grade	-0.00738	-0.00536	0.00501	0.0168
	(0.00882)	(0.00886)	(0.0104)	(0.0214)
2006-2007 VA Score			0.334^{***}	0.1610***
			(0.0402)	(0.0279)
Team SD - Prior VA			0.167	0.1722
			(0.100)	(0.1297)
Constant	0.00133	0.0249	0.00498	-0.0159
	(0.00345)	(0.0198)	(0.0248)	(0.0171)
Demographic Covariates		X	X	
School FEs				X
Observations	4535	4364	3550	3680

Standard errors, clustered at school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The teacher data reports estimated value-added in multiple ways and for various sub-samples of students. One key metric is value-added using multiple years of student data, which is estimated with greater precision than value-added using a single year of data. Single-year value-added measures, in particular, have been criticized for imprecision and lack of stability over time, suggesting that they are more subject to differences in idiosyncratic student populations and measurement error. Table 5-6 reports estimates of the OLS and difference-in-difference models, including demographic covariates, using various outcomes. Column 1 reports OLS results using the multi-year value-added outcome measure, and column 4 reports the difference-in-differences estimate; in neither case is the result statistically significant. The value-added data also include estimated effects on particular subgroups of students; based on the intended purpose of the inquiry team initiative and anecdotal evidence on teams, several teams targeted English Language Learners and students in the lowest 3rd of the school by performance. Therefore, effects of team participation on a teacher's percentile ranking on these measures is potentially an outcome of interest that may capture more directly the effects of inquiry on the target population of students; columns 2 and 3 report the OLS estimates on these respective outcomes, and columns 5 and 6 report the difference in differences estimates. The only statistically significant coefficient is the difference-in-differences estimate of the effect of team participation on value-added for students in the lowest 3rd. This may be because inquiry targets this group of students, but this coefficient should be interpreted with caution, as the multiple inferences across models and outcomes increase the risk of Type I errors, or false positives (Benjamini & Yekutieli, 2001).

TABLE 5-6 OLS AND DIFFERENCE-IN-DIFFERENCE ESTIMATES OF THE EFFECTS OF TEAMWORK ON OTHER VALUE-ADDED MEASURES

	(1) Multi-year VA	(2) VA Percentile - ELLs	(3) VA Percentile Lowest 3rd	(4) Multi-year VA	(5) VA Percentile Lowest 3rd	(6) VA Percentile - ELLs
Switch to team in 2010				0.00964	8.221*	-2.035
				(0.0173)	(3.878)	(8.834)
On a team	-0.00587	-1.306	-0.858			
	(0.00523)	(1.396)	(0.832)			
2010				0.0118***	1.028*	1.487*
				(0.00238)	(0.402)	(0.703)
Switch to team grade				-0.00146	-2.043	0.433
				(0.0108)	(2.937)	(6.683)
Constant	0.0343*	33.26***	38.65***	0.0218	38.11***	35.37***
	(0.0157)	(5.724)	(3.154)	(0.0112)	(2.865)	(5.203)
Demographic Covariates	X	X	X	X	X	X
Observations	10444	3228	10259	20295	19575	5837

Standard errors, clustered at school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Discussion

Overall, the results of the grade-switcher models and the related OLS and fixed effects estimates of the effects of teamwork on teacher value-added in 2009-2010 are quite modest. Almost all estimated coefficients are positive, but they are all substantively quite small and only two – the coefficient on teamwork in the individual teacher fixed effects model and the coefficient on the difference-in-differences interaction of the effect of teamwork on value added for students in the lowest third – are statistically significant. The coefficient on the main value-added effects of approximately 0.01 represents the equivalent of a 0.04 effect size on teacher value-added, and the standard error is roughly twice the size of the coefficient. Although the effects of collaboration measured in other quantitative literature are quite small, in the range of 0.1 standard deviation increases in test scores, these effects are small enough to suggest that there are no notable effects of inquiry on value-added, with the possible exception of value-added on students in the lowest third, academically. Given the number of different models tested, the two significant results may be the result of multiple hypothesis testing and therefore may not hold in the population. There are a number of possible reasons for the lack of effects some of which are empirically tested in subsequent sections.

One reason for the modest results in Model 1 could be that inquiry teams simply do not affect teacher productivity, on average – the intervention could be ineffective, could be working on another outcome, or could be implemented with such heterogeneity that positive and negative effects net to approximately zero on average. Given the various theoretical mechanisms through which collaboration can affect teacher productivity, some of which are positive and some negative, it is quite possible that zero net effects are masking considerable heterogeneity, and some teams are quite effective, some teams are ineffective, and a number of teams may be doing very little substantive inquiry work in reality, adding a great deal of noise to the data. The next chapter on team quality, mechanisms, and heterogeneity will further explore and test this hypothesis.

Another possible reason for modest effects relates to the choice of outcome measure and the time horizon. Note that the theoretical model makes predictions about the effects of inquiry based on time spent collaborating as an investment in teacher human capital, which may involve a tradeoff in current productivity for greater future productivity. The value-added measures are based on tests taken during the 2009-2010 school year, at the same time as the inquiry work; if the primary effect of inquiry work is through the channel of teacher human capital, it may take longer for that investment to pay off and gains may not be seen until future school years. Unfortunately, this hypothesis cannot be directly tested, as value-added measures are only available through the 2009-2010 school year; however, the other models will include some longer-term outcomes. It is also possible that value-added outcomes are not the best measure of teacher productivity or human capital, as they are too imprecisely measured and subject to potential bias related to the sorting of students. It should be noted, further, that very few interventions have a proven effect on teacher value added, which may be a particularly difficult outcome to change. Finally, the difference-in-differences estimates focus specifically on the local average treatment effect on teachers who switch from a grade without an inquiry team to a grade with a team; this may be a peculiar sub-sample of teachers who is not representative of the larger group, although we do not observe statistically significant coefficients in the OLS model, either.

MODEL #2: FIRST-YEAR TEACHER MODEL

Due to differences in the availability of data, the models for first-year teachers are estimated separately for the 2007-2008, 2008-2009, and 2009-2010 school years. This also allows for comparison between years, as well as comparisons in the estimated effect of teamwork on value-added in 2009-2010 as a robustness check on the grade switcher model, but does reduce statistical power relative to pooling and precludes the inclusion of year fixed effects to control for year-specific shocks. The key difference in modeling between years is that the grade level taught is not observable prior to 2007-2008; therefore, there is no prior year to use in a difference-in-differences specification, and effects in that year can only be estimated using OLS.

Missing Data

Since the covariates and outcome variables for this model come from pedagogical databases with information on all teachers, there is very little missing data, with the exception of the sample restriction on teachers with value-added scores for that outcome. Table 5-7 summarizes descriptive statistics for this model.

Table 5-7 Descriptive statistics on first-year teacher model, 2007-2008

Variable	Obs	Mean	Std. Dev.
Value-added	15634	0.00	0.13
Still teaching next year	88535	0.92	0.27
Years teaching	88535	4.49	0.97
On a team	88535	0.10	0.30
First Year Teaching	88535	0.09	0.29
On a grade team	88535	0.04	0.19
% Black	88534	0.32	0.28

These results are presented in Table 5-8. Participating on a team in the first year of the inquiry team initiative appears to increase the probability of returning the following year by about 4-5 percentage points, depending upon specification and on a baseline of approximately

89% of non-team first-year teachers returning, but has no statistically significant effect on total years teaching through 2011, which is 3.4 years on average. Importantly, however, the first year of the initiative was subject to the greatest potential selection bias, since principals recruited team members and members volunteered to participate. Therefore, those who would have been less likely to return anyway may have been less likely to participate on teams, whereas first-year teachers in subsequent years are more likely to be placed on preexisting teams, reducing endogeneity.

TABLE 5-8 OLS ESTIMATES OF ASSOCIATION BETWEEN TEAM PARTICIPATION AND RETENTION, FIRST-YEAR TEACHERS, 2007-2008

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Still teaching next year	Still teaching next year	Still teaching next year	Still teaching next year	Years teaching	Years teaching	Years teaching	Years teaching
On a team	0.0485***	0.0482***	0.0375*	0.0468***	0.009	0.0105	-0.033	-0.001
	(0.0104)	(0.0104)	(0.0153)	(0.0137)	(0.0472)	(0.0468)	(0.0503)	(0.064)
Constant	0.887***	0.932***	0.888***	0.932***	3.431***	3.751***	3.434***	3.7***
	(0.0149)	(0.0150)	(0.004)	(0.015)	(0.0147)	(0.053)	(0.0121)	(0.05)
Demographic Covariates		X				X		
School FEs			X				X	
Excluding grade teams				X				X
Observations	7543	7542	7543	7292	7543	7542	7543	7542

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5-9 shows the association between being on a grade with an inquiry team and measures of value-added, specified using school-level covariates, school fixed effects, and only teams that did not focus on a single grade. The results are small and not statistically significant at the 0.05 level; the association with the fixed effects specification is marginally significant at the 0.10 level. Overall, there appears to be some evidence of a small but limited correlation between team participation and both short-term retention and value-added for first-year teachers in the first year of the inquiry team initiative, suggesting that the effects of inquiry may be more pronounced for early career teachers and that, on average, the initial implementation of the initiative was stronger than in later years. Nonetheless, these results cannot be interpreted causally without the fairly strong assumption that principals did not strategically elect to have inquiry teams on grades with new teachers to aid in their development or that teachers who were most likely to return were most likely to join teams; this assumption is plausible, but not empirically testable. Further, the R^2 measure across these models is quite small, at approximately 0.005, or half of one percent of variance in retention explained by the models.

TABLE 5-9 OLS ESTIMATES OF ASSOCIATION BETWEEN TEAM PARTICIPATION AND VALUE-ADDED, FIRST-YEAR TEACHERS, 2007-2008

	(1) Value-added	(2) Value-added	(3) Value-added
Grade with team	0.002	0.015	-0.001
	(0.006)	(0.0088)	(0.0069)
Constant	-0.0022	-0.0124	-0.0101
	(0.0156)	(0.0055)	(0.0183)
Demographic Covariates	X		
School FEs		X	
Excluding grade-level teams			X
Observations	1102	1102	852

Standard errors, clustered at school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Additional prior years of data allow for estimation using a difference-in-differences methodology, which compares changes over time for first year teachers in grades that have an inquiry team in 2008-2009 to grades that do not, therefore netting out any grade-specific effects. The results of this analysis on teacher retention are shown in Table 5-10; columns 1-6 show OLS results for 2008-2009 on whether a teacher is still teaching the following year and total years of teaching, without and with demographic controls and with school fixed effects. The results are similar across specifications, small, and not statistically significant. Columns 7-8 show similar results using the difference-in-differences approach on the two outcomes with school fixed effects. The coefficient of interest is on the interaction of being on a grade with a team and the indicator for the treatment year, 2009; the effects are negative and not statistically significant across specifications. Notably, the effects are qualitatively similar across OLS and difference-in-difference specifications, indicating that differences between this model and the OLS estimates for 2007-2008 are likely due to year-specific shocks or better implementation in 2007-2008, rather than methodological differences due to the estimation strategy. One clear difference between years is that the baseline next-year retention rate of 95.3% is significantly higher in 2008-2009 than it was in 2007-2008.

TABLE 5-10 TEAM PARTICIPATION AND RETENTION OF FIRST-YEAR TEACHERS, 2008-2009

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Still teaching next year	Still teaching next year	Still teaching next year	Years teaching	Years teaching	Years teaching	Still teaching next year	Years teaching
On a team	-0.0172	-0.0172	-0.0219	0.00407	0.001	0.0024	-0.0006	0.0416
	(0.0123)	(0.0124)	(0.0235)	(0.0419)	(0.0424)	(0.0496)	(0.0149)	(0.0449)
On a team in 2009							-0.0133	-0.0294
							(0.0175)	(0.0525)
2009							0.0325^{*}	-0.597^{***}
							(0.0140)	(0.0422)
Constant	0.960^{***}	0.968^{***}	0.963^{***}	2.745^{***}	2.847^{***}	2.746^{***}	0.931^{***}	3.369^{***}
	(0.0086)	(0.0272)	(0.0156)	(0.0342)	(0.0907)	(0.0330)	(0.0111)	(0.0335)
Demographic Covariates		X			X			
School FEs			X			X	X	X

Observations	1381	1343	1381	1381	1343	1381	4443	4443

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Effects on teacher value-added are shown in Table 5-11. The effects are insignificant across specifications. The results for the difference-in-differences specifications, presented in columns 3-4, are quite similar, indicating that there are limited grade-specific effects.

TABLE 5-11 EFFECTS OF TEAM PARTICIPATION ON VALUE-ADDED, FIRST-YEAR TEACHERS, 2008-2009

	(1)	(2)	(3)	(4)
	Value added	Value added	Value added	Value added
On a team	0.0127	0.0184	-0.00213	0.0120
	(0.00856)	(0.0119)	(0.00740)	(0.00735)
Team in 2009			0.0150	0.00228
			(0.0109)	(0.00859)
2009			-0.0308***	-0.0120
			(0.00896)	(0.00690)
Constant	-0.0273	-0.0377***	0.00227	-0.0189***
	(0.0205)	(0.00791)	(0.0151)	(0.00547)
Demographic covariates	X		X	
School FEs		X		X
Observations	1343	1381	4287	4443

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5-12 shows the association between team participation and measures of retention for the 2009-2010 school year. Note that the years teaching outcome is less meaningful for this year, as data are only available through 2010-2011, meaning that first-year teachers in 2009-2010 can only be observed for up to two years. Columns 1-6 show OLS outcomes on still teaching next year and total years teaching, without and with controls and with school fixed effects, respectively; no results are statistically significant. Columns 7 and 8 show the difference-in-differences results for these two outcomes, including school fixed effects. Once again, the baseline one-year retention rate is significantly higher, at 96.7%, indicating a general upward trend in retention of first-year teachers for a second year during this time period. These results are negative and, in the case the years teaching outcome, statistically significant, implying that teachers who were placed on a grade with a team in 2010 were less likely to remain in teaching compared with teachers placed in the same grades in prior years, as compared with the difference between years for first-year teachers in non-team grades. This could be indicative of year-specific shocks in 2009-2010 or a decline in the overall quality of teams with the spread of the initiative.

TABLE 5-12 2009-2010 RETENTION

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Still teaching next year	Still teaching next year	Still teaching next year	Years teaching	Years teaching	Years teaching	Still teaching next year	Years teaching
On a team	0.0182	0.0176	-0.052	0.0024	-0.006	-0.138	0.112^{***}	0.340^{***}
	(0.017)	(0.017)	(0.035)	(0.036)	(0.037)	(0.072)	(0.027)	(0.068)
On a team in 2010							-0.0589	-0.261^{*}
							(0.042)	(0.104)

2010							0.0486^{*}	-0.64^{***}
							(0.021)	(0.052)
Constant	0.967^{***}	0.990^{***}	0.988^{***}	2.018^{***}	2.220^{***}	2.061^{***}	0.923^{***}	2.675^{***}
	(0.012)	(0.043)	(0.012)	(0.026)	(0.106)	(0.021)	(0.008)	(0.020)
Demographic covariates		X			X			
School FEs			X			X	X	X
Observations	324	320	324	324	320	324	1339	1339

Standard errors, clustered at the school level, in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Finally, Table 5-13 shows the effects of team participation on value-added measures. Columns 1 and 2 show OLS results and columns 3 and 4 show the difference-in-differences results; results are quite similar across specifications, and statistically significant and positive for the difference-in-differences model using school fixed effects. These value-added results are measured in the same year as the initiative, so they do not necessarily represent positive selection by the teachers who do not attrit; nonetheless, if teachers are more likely to leave as a result of participating on a team, and also have marginally higher-value added scores, the initiative may have led more effective teachers to exit the profession.

TABLE 5-13EFFECTS OF TEAM PARTICIPATION ON VALUE-ADDED, FIRST-YEAR TEACHERS, 2009-2010

	(1) Value-added	(2) Value-added	(3) Value-added	(4) Value-added
On a team	0.0388	0.0690	-0.00183	-0.0236
	(0.0260)	(0.0540)	(0.00846)	(0.0181)
On a team in 2010			0.0373	0.0552*
			(0.0277)	(0.0278)
2010			-0.0102	-0.000470
			(0.0153)	(0.0138)
Constant	-0.0795	-0.0450*	-0.0199	-0.0242***
	(0.0599)	(0.0189)	(0.0226)	(0.00542)
Demographic Covariates	X		X	
School FEs		X		X
Observations	320	324	1299	1339

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Discussion

Overall, the first-year teacher models show limited effects of inquiry teams on beginning teachers, with some marginally significant effects on value-added and some effects on retention in the first year of the initiative. These results could be due to measurement error and attenuation bias as a result of differences in implementation intensity and quality; the fact that results were strongest in the first year is consistent with, but does not irrefutably prove, that hypothesis, as implementation may have been strongest with enthusiastic early adopters within each school.

The consistency of results across OLS and difference-in-differences models, as well as across models estimating the effect of inquiry teams on value-added for grade-switchers and for first-year teachers in 2009-2010, provides some evidence in support of the validity of the identification strategies employed here.

MODEL #3: INSTRUMENTAL VARIABLES

A final set of models allows for exploration of a wider range of outcomes by examining the effects of inquiry at the team, rather than the teacher level of analysis. These models also rely upon the gradual phase-in of the initiative and make the assumption that, as principals expand inquiry teams across the school and select new sub-groups of teachers to serve on teams that teams are more likely to arise when a "critical mass" of teachers exists at a particular grade level or in a particular subject area. Variation within schools in terms of where teams are more or less likely to occur therefore is based on grade or subject-specific enrollment shocks due to plausibly random, year-to-year variation in cohort size. This assumption is based in part on literature on team process and optimal team size, which suggests that most types of teams operate best with about 4-6 members (Sutter, 2005).

Missing Data

Once again, the primary concern regarding missing data in this set of models is missing outcome data, as test outcomes are not reported on subgroups with fewer than five students. Of the 35,985 school-grade-subject-subgroup combinations in 2008-2009 that match up to subgroups identified as targets by inquiry teams, 28,289 have reported scores, 25,310 have reported scores from a previous year, and 23,649 have a growth score, which requires two years of reported scores. Patterns of missing data vary by subgroup – for teams that focused on multiple subgroups or did not identify any subgroup and were therefore matched with scores for "All Students," only 7 out of 7,450 cells are missing. For other subgroups, missing outcomes ranged from 10.6% for "Students with Disabilities," which is a common target for inquiry teams, to 38.9% for "Black or African American."

Table 5-14 presents basic OLS results of the association between a grade-subject-subgroup cell being the target of an inquiry team and average differences in scores for that group in 2009 compared to the same group in that grade in 2008. There is a small but statistically significant association that is robust to the inclusion of demographic covariates of about 1-1.5 points in growth, depending upon the specification, which is about 12% of the average level of growth for this time period across all grade-subject-subgroup combinations. Note that restricting the sample to teams that focused on a single grade, as in column 3, results in a point estimate that is roughly half that of the other specifications and not significant, suggesting that grade-specific teams are, on average, less effective than teams that focused on multiple grades within a subject area or student subgroup.

TABLE 5-14 OLS ESTIMATES OF EFFECTS OF INQUIRY TEAM PARTICIPATION ON MATH AND ELA GAIN SCORES, K-8 SCHOOLS, 2008-2009

	(1) Growth	(2) Growth	(3) Growth	(4) Growth
Team at grade-subject-subgroup cell	1.578***	1.542***	0.652	1.305***
	(0.317)	(0.315)	(0.452)	(0.256)
Demographic covariates		X	X	
Single grade teams only			X	
School FEs				X
Constant	5.983***	2.215***	0.919	6.010***
	(0.144)	(0.570)	(0.976)	(0.0759)
Observations	23649	23649	6065	23649

The first stage, reduced form, and two-stage least squares results for K-8 schools in 2008-2009 are shown in Table 5-15. Columns 1-4 present first stage and reduced form results including demographic covariates and school fixed effects, respectively, while columns 5 and 6 present two-stage least squares results using demographic covariates and school fixed effects.

While there is a statistically significant relationship between the number of classes and having a team in the first stage, it is minuscule at 0.003, increasing the odds of having a team by less than a percentage point, and even smaller and not statistically significant in the school fixed effects specification. The two-stage least squares results are highly unstable, ranging from 7 points in growth associated with having a team using the demographic covariate model to an implausible 198 points of growth in the fixed effects model. These results are likely due to a weak instrument at the first stage, particularly in the school fixed effects model, suggesting that the within-school variation in number of classes at each grade level and subject is too small to identify variation in team formation. Table 5-16 shows the results including covariates and using a quadratic specification for the instrument. The two-stage least squares results are very similar across specifications of the instrument , suggesting that the monotonicity assumption may hold. Neither the first stage nor reduced form coefficients are significant in this specification, however.

In all cases, however, the F-statistic from the first-stage regression is small. Using the preferred school fixed effects specification, the F-statistic is 0.25, indicating strong potential for a weak instrument; although the instrument is a significant predictor of having a team with the linear specification, it is not significant in the quadratic specification. Using multiple instruments, as is the case in the quadratic specification, allows for an overidentification test, which tests for the validity of all instruments under the assumption that at least one of the instruments is valid. Ordinarily, this would be tested using a Sargan test, which determines whether any exogenous variables are correlated with the residuals from the two-stage least squares estimate. In the case of clustered standard errors, Hansen’s J statistic, which follows a Chi-square distribution, can be used for this test. In this case, the J-statistic is 0.019, so we do not reject the null hypothesis and conclude that at least one instrument is exogenous under the assumption that the other is; given that one instrument is a transformation of the other instrument, this assumption is plausible.

TABLE 5-15 INSTRUMENTAL VARIABLES ESTIMATES OF EFFECTS OF INQUIRY TEAM PARTICIPATION ON MATH AND ELA GAIN SCORES, K-8 SCHOOLS, 2008-2009

	(1) Is a team (First stage)	(2) Growth (Reduced form)	(3) Is a team (First stage)	(4) Growth (Reduced form)	(5) 2SLS	(6) 2SLS
Number of classes	0.00288**	0.0232	0.000703	-1.148***		
	(0.00108)	(0.0341)	(0.00141)	(0.124)		
Team at subject-grade-subgroup cell					7.037	198.1
					(10.42)	(113.2)
Constant	0.0448***	2.088**	0.0884***	11.62***	1.633	-14.11
	(0.0136)	(0.660)	(0.00657)	(0.588)	(1.156)	(11.61)
Demographic Covariates	X	X			X	
School FEs			X	X		X
Observations	66664	22601	66664	22601	22601	22601

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 5-16 QUADRATIC INSTRUMENTAL VARIABLES ESTIMATES OF EFFECTS OF INQUIRY TEAM PARTICIPATION ON MATH AND ELA GAIN SCORES, K-8 SCHOOLS, 2008-2009

	(1) Is a team (first stage)	(2) Growth (reduced form)	(3) Is a team (first stage)	(4) Growth (reduced form)	(5) 2SLS	(6) 2SLS
Number of classes	-0.000521	-0.141	-0.00390	-1.504***		
	(0.00278)	(0.106)	(0.00279)	(0.228)		
Number of classes squared	0.000202	0.00950	0.000285	0.0282		
	(0.000165)	(0.00485)	(0.000149)	(0.0152)		
Team at grade-subject-subgroup cell					13.96	130.1*
					(7.630)	(65.37)
Constant	0.0546***	2.585***	0.101***	12.41***	1.031	-7.132
	(0.0148)	(0.750)	(0.00924)	(0.725)	(0.964)	(6.707)
Demographic Covariates	X	X			X	
School FEs			X	X		X
Observations	66664	22601	66664	22601	22601	22601

Standard errors, clustered at the school level, in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

While they cannot be tested formally, a number of other diagnostics can be run with instrumental variables models to qualitatively assess the validity of the required assumptions. One simple test for exogeneity is to regress seemingly unrelated variables that may be correlated with omitted variables on the instrument; these regressions should not reveal a statistically significant relationship. Columns 1 and 2 of Table 5-17 report this test, regressing prior test scores and free and reduced price lunch rates on the number of class sections at a grade level. The instrument does not pass this test, calling the exclusion restriction into question, although controlling for these pre-existing characteristics does help capture some of these unobserved characteristics to the extent that they are correlated with unobservable school characteristics. A test of the monotonicity assumption is to restrict the sample to areas where it is more likely to hold and determine if results are robust to this restriction. Columns 3 and 4 show these tests, with column 3 restricting the sample to grades with fewer than 8 class sections, and column 4 restricting the sample to elementary schools, which tend to be smaller and have fewer sections than middle schools. In both cases, the coefficient varies wildly and is estimated with great imprecision, indicating that the behavior of the instrument is unpredictable; further, since the instrument is weak, sample restrictions may exacerbate any bias. Columns 5 and 6 report similar tests using OLS and the results are more consistent, suggesting that the problem lies with the instrument itself.

TABLE 5-17 INSTRUMENTAL VARIABLES DIAGNOSTICS, 2008-2009, K-8 SCHOOLS

	(1) Prior score	(2) % Free/Reduced Lunch	(3) Gain scores, small schools	(4) Gain scores, Elementary schools	(5) Gain scores, small schools, OLS	(6) Gain scores, Elementary schools, OLS
Number of class sections	-0.578***	-0.00398***				
	(0.0476)	(0.000514)				

On team			-30.57	261.6	0.585*	1.536***
			(27.78)	(473.1)	(0.269)	(0.319)
Constant	657.0***	0.838***	5.905*	-18.49	4.063***	3.475***
	(0.275)	(0.00297)	(2.432)	(41.33)	(0.417)	(0.473)
Demographic Covariates			X	X	X	X
Observations	22903	22903	20534	15178	16909	8107

Standard errors, clustered at school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

At the high school level, the outcomes of interest are increases relative to prior cohorts on exit exam scores and graduation rates. OLS results of the association between teamwork and these outcomes are presented in Table 5-18. Columns 1-4 show the association between teamwork and gains on exit exam scores under various specifications, including with and without demographic covariates, for only teams that focused on a single grade, and with school fixed effects. Under no specification are these results statistically significant. Similarly, the effects on graduation rates are shown in columns 5-8. Although these effects are larger, they are still not significant under either specification. This could be in part because of the highly aggregated way that graduation rates are measured, as they do not take into account any prior probability of graduation, and they are reported for fewer subgroups and thus include many students who are not targeted by teams.

TABLE 5-18 OLS ESTIMATES OF ASSOCIATION BETWEEN TEAMWORK AND HIGH SCHOOL OUTCOMES, 2008-2009.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Growth	Growth	Growth	Growth	Graduation Rate	Graduation Rate	Graduation Rate	Graduation Rate
Team at grade subject subgroup cell	0.00532	0.00450	0.00733	-0.00392	0.0316	0.0272	0.0275	-0.0267
	(0.0164)	(0.0166)	(0.0291)	(0.0164)	(0.0243)	(0.0235)	(0.0446)	(0.0154)
Demographic Covariates		X	X			X	X	
School FEs				X				X
Single grade teams only			X				X	

Constant	0.0434***	0.0464	0.0449	0.0439***	0.577***	0.895***	1.006***	0.579***
	(0.00576)	(0.0245)	(0.0274)	(0.00356)	(0.0114)	(0.0510)	(0.0337)	(0.00260)
Observations	10356	10318	1932	10356	3845	3838	596	3845

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5-19 provides results of the instrumental variables estimate of the effects of having a team on high school outcomes. Similarly to the K-8 results, the instrument may be weak, as the first-stage F-statistic is 7.38 when the instrument is specified with demographic covariates and 25.0 when specified with school fixed effects. The relationship between class sections and team formation is stronger for high schools than for K-8 schools, as the bivariate relationships are significant and the F statistics are somewhat larger; however, neither the reduced form results nor the two-stage least square results are significant. Panel A shows the effects on student achievement gains, while Panel B shows the effects on graduation rates. Columns 1-4 of Panel A show the first stage and reduced form results of the quadratic specification, using demographic covariates and school fixed effects, respectively. The number of classes is associated with the probability of having a team, and the square is negatively associated, as would be expected if there is an optimal team size above which collaboration becomes counterproductive, although the relationship is substantively quite small. The reduced form and two-stage least squares results are not significant for test scores or graduation rates.

TABLE 5-19 INSTRUMENTAL VARIABLES (2SLS) ESTIMATES OF EFFECTS OF INQUIRY TEAM PARTICIPATION ON HIGH SCHOOL OUTCOMES, 2008-2009

Panel A. Student achievement growth

	(1)	(2)	(3)	(4)	(5)	(6)
	First stage	Reduced	First stage	Reduced Form	2SLS	2SLS
	(Team)	Form	(Team)	(Growth)		
		(Growth)				
Number of classes	0.0131^{***}	0.00130	0.00710^{***}	-0.000433		
	(0.00279)	(0.00111)	(0.00122)	(0.00169)		
Number of classes squared	-0.000126^{*}	-0.0000275	-0.0000738^{***}	-0.00000293		
	(0.0000556)	(0.0000190)	(0.0000215)	(0.0000299)		
Team at grade-subject-Subgroup cell					0.0350	-0.125
					(0.0608)	(0.193)
Constant	0.000682	0.0802^{*}	0.0774^{***}	0.0260^{**}	0.0819^{*}	0.0388
	(0.0631)	(0.0319)	(0.00673)	(0.00990)	(0.0322)	(0.0258)
Demographic Covariates	X	X			X	
School FEs			X	X		X
Observations	8504	6415	8524	6435	6415	6435

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Panel B. Graduation rates

	(1)	(2)	(3)	(4)
	Reduced form	Reduced form	2SLS	2SLS
	(Graduation rates)	(Graduation rates)		
Number of classes	-0.00729^{*}	-0.000134		
	(0.00339)	(0.00259)		
Number of classes squared	0.000152^{**}	0.00000218		
	(0.0000546)	(0.0000612)		
Team at grade-subject-subgroup cell			-0.583	-0.0179
			(0.365)	(0.343)
Constant	0.918^{***}	0.637^{***}	0.915^{***}	0.638^{***}
	(0.0806)	(0.0116)	(0.0816)	(0.0272)
Demographic covariates	X		X	
School FEs		X		X
Observations	1582	1585	1582	1585

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As with the K-8 instrumental variables results, some more informal diagnostic tests can provide insight as to whether the exclusion restriction and monotonicity assumptions are justified, even if those cannot be formally tested. The results of these diagnostics are reported in Appendix A. While the diagnostics for high school suggest that the exclusion restriction may be more likely to apply in this case than in K-8 schools, the monotonicity assumption may once again be violated. Given the concerns raised regarding the instrumental variables estimates – the weak instrument and potential violations of monotonicity and exclusion restriction assumptions – results for 2009-2010 are presented in Appendix A. They are qualitatively similar to the 2008-2009 results, with the exception that the two-stage least squares results for K-8 schools in 2009-2010 become negative and statistically significant.

Discussion

Overall, the instrumental variables results appear to be significantly less valid and reliable than the results from the other models. The most worrisome aspects of the instrumental variables results are the significant evidence that the proposed instrument is quite weak, exacerbating the bias of the two-stage least squares estimate, and the fact that IV results vary considerably by year, outcome, and specification, whereas OLS results are generally more stable across these dimensions. Evidence on the validity of the other required assumptions for instrumental variables is more mixed. Results are similar whether the instrument is specified linearly or in a quadratic term, providing evidence for monotonicity, but they are not robust to taking sub-samples of schools with fewer class sections or elementary vs. middle schools, while OLS results are. Similarly, the Hansen’s J statistic test of overidentifying restrictions provides some evidence that this instrument meets the exclusion restriction, but the instrument’s correlation with prior variables brings that assumption into question. Finally, while OLS results are robust to how gain scores are calculated, and very similar whether they compare scores this year to scores for the same cohort in the prior grade last year, or scores for the same grade last year, IV results are quite sensitive to this choice. Therefore, of all the models it appears that the IV results hold up least well to scrutiny.

INTERPRETATION OF RESULTS

Across the models, the general evidence seems to be the effects of inquiry team participation on general teacher productivity and student learning are quite small, on average, if any causal effects exist at all. While most point estimates are positive, few are statistically significant and those that are in most cases do not rise to the level of substantive policy significance, being the equivalent of less than 0.1 standard deviations. These estimates are somewhat smaller than, though still roughly in line with, the most rigorous existing estimates of the effects of teacher collaboration on student achievement, and are not entirely surprising given the generally weak empirical evidence for teacher effectiveness-enhancing programs overall, suggesting that teacher productivity is a particularly difficult outcome to measure and to change through policy.

This is a difficult question to answer causally, given the many complex and interrelated issues of selection bias, management bias, measurement error, and confounding with school leadership and culture. Further, given that the students who are the direct targets of inquiry are not directly observable, any effects measured will be more distal on general learning outcomes and teacher effectiveness at a particular grade level and subject or for a particular subgroup of students. Given these concerns, even the relatively modest results that appear, particularly for struggling students and in the first year of the inquiry team initiative, are still promising and merit further study. The next chapter will further examine the mechanisms by which inquiry could enhance teacher effectiveness and study heterogeneity, which may be masked by very small average results. Given the limitations of the grade-switcher and instrumental variables models, namely that the grade switcher model is limited to the last, lowest-intensity year of implementation and focused on potentially problematic value-added measures as an outcome and that the instrumental variables estimates suffer from a weak instrument and other potential issues, the first-year teacher model is preferred going forward. Based on that model, as well as descriptive results from OLS models, there is evidence for some retention outcomes and possibly some test score outcomes in the first year of the initiative, but most other evidence suggests that the inquiry team initiative had little measurable effect on teacher productivity, retention, and student learning.

Chapter 6 ANALYSIS OF MECHANISMS AND HETEROGENEITY

The literature on collaboration and teamwork in education and other sectors suggests that not all teamwork is positive, which could help explain the modest and mostly null results in the previous chapter. Teamwork can enhance teacher productivity and effectiveness, as well as student learning, through a number of channels but can also have no effect or even detract from productivity through other channels. Further, teams vary considerably in the intensity with which they implemented the inquiry initiative, resulting in a great deal of noise in the data that could obscure effects of more meaningful teamwork. This chapter examines possible causal mechanisms and heterogeneity due to variation in team processes and intensity, primarily in a descriptive and exploratory way, as these differences are likely to be strongly correlated with other unobserved factors that determine outcomes, including underlying teacher quality, school culture, and leadership.

POTENTIAL MECHANISMS

Suggestive evidence for several possible mechanisms can be observed in the quasi-experimental models through patterns in outcomes and sample affected, as well as by including covariates and interaction terms and examining results for sub-samples. The most straightforward mechanism by which teamwork could impact teacher productivity and student learning is through knowledge-sharing, whereby individuals on teams benefit from the team's collective knowledge and experience, which supplements and complements their own. Without a deeper dive into the narrative responses to scan for explicit examples of this, the closest way this mechanism can be assessed is by testing whether the effect of teamwork is greater for teachers whose prior value-added is lower than the team average, indicating these teachers might be gaining from the knowledge of their colleagues.

A second mechanism by which inquiry can operate is through instructional innovation that benefits all teachers and students, regardless of their starting point. Since inquiry is an iterative research process, rooted in theories of democratic experimentalism and abduction, this would involve adopting or developing new instructional strategies or making changes to the curriculum to adapt to the learning needs of sub-groups of students (Talbert, 2011). This mechanism cannot be tested directly, but will be explored in the case studies in the next chapter. Related mechanisms, including the extent to which teamwork facilitates challenging teacher mindsets, particularly if they hold negative views about students' abilities, will also be explored in case studies.

A third possible mechanism through which inquiry can enhance teacher productivity specifically is by addressing gaps in teacher preparation programs. This could indicate market failure of teacher preparation programs to adequately prepare teachers for the classroom and may be especially prevalent if teaching skills are primarily experiential and must be learned hands-on. One way to test this mechanism would be to assess whether the initiative has a particularly strong effect on first-year teachers, as it appears to.

Teams could also enhance productivity through peer monitoring, peer pressure, and enhancing intrinsic motivation due to desire to achieve shared goals and reinforcing a common mission. This might be observed through stronger effects in smaller teams or teams that are more homogeneous with regard to grade-level, subject-area, level of experience, and prior value-added (Kandel and Lazear, 2012; Kreps 1997).

Conversely, teams that implement the inquiry initiative at very low intensity would not be likely to have any effect on teacher productivity at all, and in some cases, teamwork may even have a negative effect on teachers. Teamwork can create problems with free-riding or shirking if teachers believe that others on the team will do work for them, could substitute for more productive but less pleasant individual work, or could reinforce negative social norms around student learning. This may be exacerbated in this scenario, when participation on inquiry teams was mandated by 2009-2010; therefore, any measured effects of inquiry teams are effects of the policy, rather than of substantive engagement in the inquiry process itself. Further examination of these mechanisms, as well as identifying teams that are operating at very low intensity and likely to do little to affect outcomes, requires deeper examination of data on the teams themselves.

Defining and Measuring Quality Teamwork

The data teams entered about their processes and outcomes in an administrative database operated by the school district to gather data on inquiry implementation that could be used to assess heterogeneity varied each year. Data in the 2007-2008 school year, with the fewest number of teams and at the start of the initiative, was generally the richest in terms of narrative detail and team reflection. Representatives of teams were asked to enter information about the grade level, content area, skill, and demographic subgroup focus of the team, the goal the team set for its target population, the assessments used to measure progress toward that goal, the team's findings, ways the team impacted overall school culture, reflections on how the school would change and expand the inquiry team initiative the following year, and a series of smaller sub-goals and instructional strategies designed to achieve those goals while working toward the larger goal. Note, however, that although the data in the first year was richest in these narrative and reflective details, teams did not report the number of teachers on the team, the number of students in the target group, nor the identities of teachers or students.

The data teams entered in 2008-2009 were less rich overall. Since there were multiple teams in each school, schools entered information about how inquiry was organized at the school (e.g., by grade level vs. by subject area; whether or not the school employed a hub-and-spoke style system with a central coordinating team including representatives from each team), and how many teachers across the school were involved in inquiry work. Each team entered data on the number of students in the target population, grade level and demographic subgroups targeted, content and skill areas targeted, goals for the inquiry work, and assessments used to identify students, set goals, and measure progress. All other information was organized by "cycles," a series of mini-inquiries in which teams narrowed their focus to small subskills, test strategies, assess, and make adjustments as necessary. The number of cycles for which teams enter data is itself an indicator of the sustainability of the inquiry process, as it indicates flexibility in the process, openness to making changes, and ongoing learning. The vast majority of teams, however, only entered information for the first cycle. For each cycle, teams were asked about goals, plans to achieve goals, measured effects of the strategies implemented, and reflections on what they learned from the cycle and how it could be extended to a larger group of students and the school community.

The data for 2009-2010, as mentioned above, were the only set for which the individual teachers on teams are identified, although students are not identifiable. The inquiry data for this year was folded into a larger, centralized database system, to more easily link with teacher and student data. Teams entered information about team composition, subject area and skill focus, the question that guided the team's work for the year, goal and assessments used to measure progress toward the goal, narrative on why the team chose its focus group of students and skill, instructional strategies implemented for up to five inquiry cycles, and reflective questions on the effects of inquiry.

In order to more closely evaluate heterogeneity in team processes and how that correlates with outcomes, some analyses will focus on a textual analysis of a random sub-sample of 100 teams from each year. For this sub-sample, the team's responses to all questions was carefully read and the intensity and fidelity of the team's implementation of the inquiry initiative was assessed on four dimensions according to a rubric developed by the author based on the literature on teamwork. Note that this exercise was not intended to evaluate the inquiry teams, as that is not one of the research questions for this study and is beyond the scope of this dissertation; it would also likely be impossible to do so in a valid and reliable manner given the limitations in available data, particularly on the outcomes of teamwork and whether any individual teachers implemented practices developed by the teams in their classrooms. Rather, the purpose is to summarize the data on the wide range of team processes embedded in detailed textual responses into a smaller number of numerical values to analyze how those processes covary with outcomes. The four dimensions, based in part on Hoegl and Gemuenden's analysis of the dimensions of high-quality collaboration, as well as the central school district's evaluation of inquiry work in the Quality Review rubric and survey scales used to assess teamwork developed by Joan Talbert and colleagues at the Center for Research on the Context of Teaching at Stanford University, are Focus, Diagnosticity, Sustainability, and Process. Focus refers primarily to the team's ability to set targeted, measurable, ambitious yet attainable goals that are primarily instructional in nature. It also includes the extent to which there is evidence that the team's efforts aligned with its stated goals. Diagnosticity and use of evidence refers to the team's problem-solving orientation, including its comfort and familiarity with engaging in a range of sources of evidence, and willingness to experiment and test new ideas. Sustainability and follow-up refers to the extent to which teams engaged in longer-term pursuit of inquiry work, including whether or not they engaged in multiple inquiry cycles. Process and balance of contributions refers to evidence of team dynamics, including whether there is evidence of structure or protocol at team meetings, whether team members rotate roles and responsibilities, and whether there is discussion of sharing effort among team members. Overall, Focus and Sustainability were the dimensions on which there is the most concrete evidence, such as the presence or absence of a goal and the presence or absence of additional cycles of teamwork, whereas evidence on the other dimensions is more limited and required subjective judgment. Table 6-1 summarizes the rubric used to assess these 300 teams.

TABLE 6-1 RUBRIC USED TO ASSESS INTENSITY OF IMPLEMENTATION OF INQUIRY TEAMS

Category	**1**	**2**	**3**
Focus	No clearly-defined goals or apparent area of focus, whether it is a sub-group of students, particular instructional skill, or particular area of teacher professional development; team does not use time together to discuss teaching and learning	Some evidence of goals and narrowing of team's focus, but goals are not well-defined, not measurable, or are overly-broad	Clearly-defined, measurable, time-bound goals focused on the instructional needs of particular students and/or specific skills and instructional areas
Diagnosticity and use of evidence	Little to no evidence of a problem-solving orientation or any diagnostic process; little to no evidence of systematic use of data to inform decisions or test results; no rationale provided for decision-making	Some evidence of a problem-solving orientation and decisions informed by use of data; mentions sources of evidence; goal and instructional strategies are related to assessment of student learning needs	Teams have a clear diagnostic process and a problem-solving approach that involves analysis of multiple types of evidence and student data, including analysis of student work; evidence of root cause analysis; team engages in

			multiple cycles, refining and testing instructional approaches
Sustainability and follow-up	Little to no evidence of structures to promote sustainability, e.g., regular meetings, sufficient time to meet, clearly defined next steps	Teams meet regularly and have sufficient time to meet; some limited evidence of instructional change as a result of inquiry process	Frequent, regular meetings; clear evidence of changes to curriculum, instruction, professional development, etc., as a result of inquiry; effective leveraging of outside resources to address identified instructional needs; ongoing reflection, monitoring of progress, and adjustments
Team process and balance of contributions	No evidence of structured processes to promote effective collaboration (e.g., agendas, protocols, facilitators); evidence of domination by one or a few members, or free-riding by one or more members	Team has some processes in place, such as agendas and protocols, to ensure effective meetings, and may have one particular individual who primarily leads the work	Team has clear, agreed-upon structures to promote effective use of time, including protocols and agendas, and team members rotate roles and all play an active role in decision-making

Sources: Quality Review Rubric, indicator 4.2; Inquiry Capacity Continuum; SAM Evaluation scale from the Center for Research on the Context of Teaching, Stanford University; Hoegl and Gemuenden, 2001

Given the inherent subjectivity of this rating process, as well as power limitations inherent in using a sample of 100 teams in each year, most analyses of heterogeneity will proceed using proxies for team quality that can be readily calculated for the full sample. These include the number of inquiry cycles for which teams entered information, whether or not teams set a goal, and whether or not teams include student work in the evidence they use to identify target students and assess student progress. These quality indicators are cross-checked with the subjective ratings for the sub-sample to determine correlation between the two measures of quality, and whenever possible heterogeneity analyses are run using both the sub-sample with more detailed quality ratings and the larger sample, with rougher quality indicators for all teams.

Critically, since the team's self-reports of their activities in the inquiry database are the only source of data on team process for the entire sample of 13,425 teams over the three years, the unbiasedness of all these analyses rests on the assumption that what teams actually did is correlated with what they say they did. It does not need to be exactly the same, as small differences will wash out on average and the analysis only requires that teams who said they did more on average did more; in other words, the relationship between data entry and real team activity must be at least weakly monotonic. There are a few plausible scenarios in which this assumption could be violated due to social desirability bias, given that the data were being collected by the school district that implemented the policy – one is the case of teams dominated by a single individual who enthusiastically entered a great deal of data on team activity, even when the team itself did little. In that case, these teams may in fact be less effective than average, because they do not represent a balance of contributions and effort and instead represent the efforts of a single individual, but would appear to be better in the database. Similarly, teams that were highly concerned about complying with central mandates on teamwork may have exaggerated the extent of their teamwork to appear better to central district officials. It could be the case that those teams most concerned with compliance would enter the most information but also engage in relatively less authentic inquiry work because they are more risk-averse and less willing to engage in experimentation and challenge preconceptions. Due to these possible violations, in addition to the fact that intensity of team activity is itself endogenous, all analyses of quality and heterogeneity are descriptive, rather than causal.

DESCRIPTIVE ANALYSIS OF HETEROGENEITY AND QUALITY

Descriptive statistics on the intensity of inquiry team activity by year are presented in Tables 6-2 and 6-3. For each year, the average scores out of 3 on Focus, Diagnosticity, Sustainability, and Process for the sub-sample of 100 hand-coded teams are shown, as well as the unweighted average across the three scores. In addition, for all teams, the share of teams that have no goal, that analyze student work, and that engage in multiple inquiry cycles, as well as the average number of cycles, are shown. Finally, the table shows correlations between Focus and having no goal, Diagnosticity and analyzing student work, and Sustainability and having multiple inquiry cycles, to assess the suitability of these proxy measures of team quality.

TABLE 6-2 DESCRIPTIVE STATISTICS ON TEAM QUALITY AND INTENSITY OF IMPLEMENTATION

Variable	Mean	Standard Deviation	Min	Max
Panel A: 2007-2008				
Focus	1.91	0.4734	1	3
Diagnosticity	1.74	0.5794	1	3
Sustainability	1.67	0.4935	1	3
Process	1.98	0.449	1	3
Overall	1.825	0.3901	1	3
No goal	0.0069	0.0826	0	1
Analyzes student work	0.1821	0.3861	0	1
Engages in multiple cycles	0.422	0.494	0	1
Number of cycles	1.904	2.158	0	23
Panel B: 2008-2009				
Focus	1.49	0.5024	1	2
Diagnosticity	1.25	0.4578	1	3
Sustainability	1.24	0.474	1	3
Process	1.33	0.5136	1	3
Overall	1.327	0.3688	1	2.75
No Goal	0.0837	0.277	0	1
Analyzes Student Work	0.1163	0.3207	0	1
Multiple Inquiry Cycles	0.3217	0.4672	0	1
Number of Cycles	0.8714	1.139	0	10
Panel C: 2009-2010				
Focus	1.61	0.4902	1	2
Diagnosticity	1.2	0.402	1	2
Sustainability	1.12	0.3562	1	3
Process	1.18	0.4115	1	3
Overall	1.278	0.3216	1	2.5

No Goal	0.2832	0.4506	0	1
Analyzes Student Work	0.7614	0.4262	0	1
Multiple Inquiry Cycles	0.0551	0.2283	0	1
Number of Cycles	1.088	0.4138	1	5

TABLE 6-3 CORRELATIONS (2009-2010)

	Focus	Diagnosticity	Sustainability	Overall	Process	No Goal	Student work
Focus	1.0000						
Diagnosticity	0.2973	1.0000					
Sustainability	0.2707	0.6067	1.0000				
Overall	0.6454	0.8361	0.7867	1.0000			
Process	0.3015	0.7572	0.6782	0.8593	1.0000		
No Goal	-0.7261	-0.2358	-0.2323	-0.4777	-0.1969	1.0000	
Student Work	0.1215	0.0992	-0.0323	0.0549	-0.0420	-0.1285	1.0000
Number of cycles	0.0977	0.1833	0.2344	0.2186	0.1850	-0.0838	0.0781

Overall, the intensity of inquiry team implementation declined over time. As more and more teachers participated on teams to reach the goal of 90% participation in 2009-2010, teams became less likely to set goals and engage in multiple cycles of inquiry work. This could be for a number of reasons which are not possible to directly test with the given data, but which can be observed anecdotally in a sub-sample of teams: principals are less likely to directly participate on all teams when there are many teams per school, teams receive less coaching and support from networks and the central office over time, and teams in the first year are composed of enthusiastic early adopters who signal their willingness and ability to implement the inquiry team initiative through volunteering to participate at the outset. In the first year, nearly all teams (over 99%) set goals, nearly half of teams engage in multiple cycles of work, and the hand-coded ratings of implementation intensity are by far the highest, at an average of nearly 2 out of 3. In fact, the first year is the only year for which there are any teams that receive the highest score on all dimensions of implementation. Nearly all measures of intensity of implementation drop every year, with the exception of the percentage of teams analyzing student work, which increases sharply in the final year. This could be an artifact of the data entry system, as teams were allowed to select "Student Work" from a checklist of sources of evidence they used, as opposed to prior years in which teams manually entered information about evidence. Notably, by 2009-2010 nearly 30% of teams did not even set a goal, indicating that these were teams in name only that schools entered to demonstrate compliance with the 90% participation goal, but in practice the teams may not have even ever met.

Responses that the sub-sample of teams selected for quality ratings entered into the inquiry database reveal some general trends in how teams implemented inquiry teams. Overall, even when teams do show evidence of higher-intensity implementation, such as goal-setting, looking at multiple sources of evidence, and engaging in multiple cycles of inquiry work, the number of teams that fully adhere to the spirit of inquiry teams by engaging in root-cause analysis that precipitates instructional changes is low. Many teams set vague goals with broad target areas, such as general increases in student reading levels or state test scores. The instructional strategies implemented were also often quite broad and general, such as "differentiate instruction," "use data to inform instruction," and "provide professional development to teachers." One team listed its instructional strategies as "Small group instruction. Differentiating instruction. Flexible grouping. Technology integration."

In the first year in particular, although teams adhered to the inquiry team model more closely, the instructional strategies they tried were often not scalable beyond the targeted subgroup of students without substantial increases in resources. For example, many strategies involved small group or one-on-one tutoring during lunch or after school by the inquiry team members, or moving the targeted students to smaller classes. Therefore, any results of this work may be the result of students being aware they are the target of an initiative, as well as reshuffling resources to target those particular students' needs away from the general student body, as opposed to any deeper changes in teacher work or the culture of the school. Talbert (2010) noted this trend as well, which was initially troubling to central district leaders, as it went against the intent of the policy of using inquiry for instructional innovation, teacher capacity-building and deeper systemic change. Ultimately, however, district leaders tolerated this interpretation of inquiry as a step toward using the process for more authentic change.

Possibly as a result of a conscious effort by the central office to move away from these sorts of non-scalable interventions and refocus on teacher capacity, operationalized through the elimination of the requirement to target a specific subgroup of students with inquiry work, there is a notable shift in 2009-2010 toward teacher-focused strategies. While this shift did help to avoid directing inquiry activities toward resource-shifting strategies such as smaller classes and extended day programs for small sub-groups of students, it also is generally associated with broader questions and vaguer goals. For instance, one team's guiding question in 2009-2010 was "How well are students mastering specific Performance Indicators within specific classes and across the entire grade?"

Some of the most promising strategies teams identify revolve around improving communication and knowledge-sharing among various stakeholders in student learning, including other teachers and parents. Therefore, although the evidence from the textual analysis that inquiry teams succeeded in engaging in root cause analysis or deep innovation as a result of action research is limited, the capacity of teams as a vehicle for improving communication and sharing existing information across the school seems stronger, indicating that as one possible mechanism for any positive effects of teamwork.

Incorporating Heterogeneity in Main Models

To test some of these mechanisms, as well as to reduce noise in the data created by extremely low-intensity teams such as the 30% that did not set a goal in 2009-2010, indicators of implementation fidelity were added to selected models analyzed in the previous chapter. These indicators were added as covariates and in place of the team indicator variable. Ideally, these indicators would also be added as interactions with team indicators to assess heterogeneity; however, because there is only data on implementation fidelity for teachers and/or grade-subject cells that actually have a team (zero is imputed for all others), the interaction effect is perfectly collinear with the main effects and thus adds no information to the model. Overall, results are fairly consistent across specification of quality, including whether it is used as a replacement for the team indicator or added as a covariate and which measure of quality is used, and similar to the main results, although somewhat more positive and more precisely measured. Representative results are summarized here, and estimates from all specifications are reported in Appendix B.

The preferred specification uses the indicator for whether or not the team sets a goal as a basic measure of quality that indicates whether or not a team actually engages in inquiry work, even at a minimal level. Results from specifications using the hand-coded sub-sample are highly erratic and measured with very little precision, given very low power from the sample of just 100 teams per year. Other measures of quality and intensity of implementation are only weakly associated with outcomes, are statistically insignificant in almost all specifications, and are in some cases negative. This is likely due in part to the collinearity of these measures. Surprisingly, the indicator for whether a team analyzes student work is often negative, albeit insignificant. This indicates that, on average, outcomes are better for teams that do not analyze student work, contrary to expectations given prior literature. One possible explanation is measurement error; teams could report analyzing student work at much higher rates than what they actually do.

The association between being on a team that has a goal and retention outcomes in 2007-2008 for all teachers and first-year teachers, respectively, are reported in Tables 6-4 and 6-5. These results are very similar to those reported for the first-year model using the team indicator, in part because nearly all teams have a goal in 2007-2008. Notably, the correlation for all teachers is particularly strong, indicating a 15 percentage point increase in one-year retention rates, but this result is strongly subject to selection bias related to likelihood of volunteering or being chosen to participate on a team.

TABLE 6-4 ASSOCIATION BETWEEN TEAM QUALITY AND RETENTION, ALL TEACHERS, 2007-2008

	(1) Still teaching next year	(2) Years teaching	(3) Years teaching	(4) Still teaching next year
Has goal	0.157^{***}	0.0654^{***}	0.0545^{***}	0.149^{***}
	(0.00320)	(0.0154)	(0.0149)	(0.00316)
Constant	0.800^{***}	4.487^{***}	4.680^{***}	0.877^{***}
	(0.00238)	(0.00623)	(0.0240)	(0.00894)
Demographic Covariates			X	X
Observations	108937	88535	88534	108935

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 6-5 ASSOCIATION BETWEEN TEAM QUALITY AND RETENTION, FIRST-YEAR TEACHERS, 2007-2008

	(1) Still teaching next year	(2) Years teaching	(3) Still teaching next year	(4) Years teaching
Has goal	0.0558^{***}	0.0172	0.0560^{***}	0.0232
	(0.00859)	(0.0456)	(0.00862)	(0.0450)
Constant	0.889^{***}	3.433^{***}	0.939^{***}	3.797^{***}
	(0.00409)	(0.0149)	(0.0167)	(0.0647)
Demographic Covariates			X	X
Observations	8130	8130	8129	8129

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Similarly, the association between having a goal and estimates of teacher value-added, reported in Tables 6-6 and 6-7 for the full sample of teachers and first-year teachers, respectively, is quite similar to the association between the team indicator and value-added. The association is consistently positive but small and only statistically significant for value-added for students in the lowest third.

TABLE 6-6 ASSOCIATION BETWEEN TEAM QUALITY AND VALUE-ADDED, ALL TEACHERS, 2007-2008

	(1) VA	(2) VA Percentile Multi-year	(3) VA Percentile Lowest 3rd	(4) VA Percentile ELL
Has goal	0.000807	0.857	1.329^{*}	0.550
	(0.00332)	(0.558)	(0.675)	(1.387)
Constant	0.0124	49.18^{***}	41.11^{***}	47.40^{***}
	(0.0100)	(1.451)	(1.961)	(4.551)
Demographic Covariates	X	X	X	X
Observations	17697	16684	10214	2271

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 6-7 ASSOCIATION BETWEEN TEAM QUALITY AND VALUE-ADDED, FIRST-YEAR TEACHERS, 2007-2008

	(1) VA	(2) VA Percentile Multi-year	(3) VA Percentile Lowest 3rd	(4) VA Percentile ELL
Has goal	0.00805	2.098	3.232	2.754
	(0.00634)	(1.409)	(2.149)	(4.343)
Constant	-0.00796	51.30***	42.65***	62.59**
	(0.0173)	(3.488)	(6.585)	(18.81)
Demographic Covariates	X	X	X	X
Observations	1689	1552	714	171

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2008-2009

For 2008-2009, the heterogeneity analysis focuses on the effect of having a team at a particular grade-subject cell on test score growth. Although the instrumental variables specification in the previous chapter addresses concerns with selection bias related to the placement of teams at particular grades and subject areas, it is not the preferred specification in this case because of the previously discussed weak instrument and monotonicity problems that lead to potential bias and difficulties in interpreting the IV estimates. Therefore, the analysis focuses on the OLS results, which should be interpreted as correlational, not causal. Column 1 of Table 6-8 presents the baseline association between the existence of a team at a particular grade and subject combination and test score growth, which is positive and statistically significant. Grade-subject-subgroup combinations with teams are associated with about 1 more point of growth from a baseline of about 8 points on average, or about 0.1 of a standard deviation in growth. Column 2 adds quality measures on whether a team has a goal, whether the team engages in multiple cycles of inquiry work, and whether a team analyzes student work as covariates. Due to collinearity, none of these estimates are statistically significant, but it appears that much of the variation in the association between teamwork and score growth is related to having a goal and engaging in multiple cycles, whereas the coefficient on student work is negative. Column 3 replaces the team indicator with an indicator for whether there is a team that sets a goal, and the coefficient is slightly higher than the basic team indicator, suggesting some small variation in team quality captured by goal-setting. Columns 4 and 5 show results incorporating the hand-coded quality measures, which are very imprecisely estimated due to low power.

TABLE 6-8 OLS ESTIMATES OF ASSOCIATION BETWEEN TEST SCORE GROWTH AND TEAM QUALITY MEASURES, K-8

	(1) Growth	(2) Growth	(3) Growth	(4) Growth	(5) Growth
Team indicator	1.117^{***}	0.175			-2.185
	(0.273)	(0.935)			(3.292)
Has goal		0.786	1.168^{***}		
		(0.947)	(0.281)		
Multiple cycles		0.732			
		(0.570)			
Analyzes student work		-0.614			
		(0.747)			
Overall				-0.285	1.162
				(0.520)	(2.185)
Constant	3.651^{***}	3.652^{***}	3.657^{***}	3.677^{***}	3.679^{***}
	(0.550)	(0.551)	(0.550)	(0.719)	(0.719)
Demographic Covariates	X	X	X	X	X
Observations	25016	25016	25016	9012	9012

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6-9 shows the same results for high schools. The baseline estimates in column 1 are small and not statistically significant. Somewhat surprisingly, the results on the team indicator become positive and marginally significant when adding quality indicators as covariates, shown in column 2, suggesting that teams with higher scores on quality indicators are associated with *lower* test score gains. This could be because these indicators are not reasonable proxies for team quality, due for instance to violations of the assumptions listed above, could indicate teams operating differently at the high school level, or could be a statistical fluke given the number of models, specifications, and outcomes being tested. Columns 3 and 5 incorporate the hand-coded quality measure and column 4 replaces the team indicator with an indicator for whether the team has set a goal; in no case are estimates statistically significant.

TABLE 6-9 OLS ESTIMATES OF ASSOCIATION BETWEEN TEST SCORE GROWTH AND TEAM QUALITY MEASURES, HS

	(1) Growth	(2) Growth	(3) Growth	(4) Growth	(5) Growth
Team indicator	0.0105	0.155^{*}	0.115		
	(0.0169)	(0.0741)	(0.294)		
Has goal		-0.108		0.00523	
		(0.0731)		(0.0174)	
Analyzes student work		-0.0685			
		(0.0429)			
Multiple cycles		-0.0780^{*}			
		(0.0312)			
Overall			-0.137		-0.0567
			(0.242)		(0.0316)
Constant	0.0541^{*}	0.0531^{*}	0.0440^{***}	0.0544^{*}	0.0400
	(0.0254)	(0.0254)	(0.00738)	(0.0254)	(0.0208)
Demographic Covariates	X	X		X	X
Observations	9542	9542	2946	9542	2946

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2009-2010

For comparison, results for 2009-2010 are shown for both retention and value-added for the full sample and first-year teachers, as well as for test score gains at the grade-subject-subgroup level. Tables 6-10 and 6-11 present the association between measures of the quality and intensity of team participation and retention outcomes for all teachers and first-year teachers, respectively. The results for first-year teachers are more subject to causal inference, as it is less likely that brand new teachers would be strategically placed on grades and in subjects with teams. In the models with all teachers, there is a small but positive and statistically significant relationship between being on a team and remaining in school the following year; the relationship becomes insignificant when adding quality measures as covariates, in Column 3 of Table 5.9, but is very slightly larger when using an indicator for a team having a goal in place of an indicator for whether a team exists (Column 5). Once again, the relationship with the hand-coded quality measure is small and insignificant, with the exception of a modest relationship with the Years of Teaching outcome. For the much smaller sample of first-year teachers, no relationships are statistically significant. Note that the results for the hand-coded sample are omitted, as no first-year teachers were on teams selected for the hand-coding sample.

TABLE 6-10 ASSOCIATION BETWEEN INDICATORS OF TEAM QUALITY AND RETENTION, ALL TEACHERS.

	(1) Still teaching next year	(2) Years teaching	(3) Still teaching next year	(4) Years teaching	(5) Still teaching next year	(6) Years teaching	(7) Still teaching next year	(8) Years teaching
Team indicator	0.0123***	0.105***	0.00828	0.126**				
	(0.00345)	(0.0205)	(0.00670)	(0.0387)				
Has goal			0.00599	- 0.00477	0.0129***	0.0890***		
			(0.00556)	(0.0325)	(0.00350)	(0.0215)		
Analyzes student work			- 0.000512	-0.0219				
			(0.00558)	(0.0310)				
Multiple cycles			0.00142	- 0.00344				
			(0.00799)	(0.0482)				
Overall							0.00762	0.275***
							(0.0221)	(0.0388)
Constant	0.968***	4.650***	0.968***	4.650***	0.969***	4.661***	0.971***	4.606***
	(0.00836)	(0.0526)	(0.00837)	(0.0525)	(0.00836)	(0.0532)	(0.0107)	(0.0610)
Demographic Covariates	X	X	X	X	X	X	X	X
Observations	12447	12429	12447	12429	12447	12429	7992	7975

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 6-11 ASSOCIATION BETWEEN INDICATORS OF TEAM QUALITY AND RETENTION, FIRST-YEAR TEACHERS.

	(1) Still teaching next year	(2) Years teaching	(3) Still teaching next year	(4) Years teaching	(5) Still teaching next year	(6) Years teaching
Team indicator	0.0173	-0.00441	-0.00211	-0.0702		
	(0.0174)	(0.0382)	(0.0240)	(0.0435)		
Has goal			0.0270	0.0599	0.0238	0.0166
			(0.0344)	(0.0439)	(0.0129)	(0.0392)
Analyzes student work			0.0128	0.0424		
			(0.0170)	(0.0342)		
Multiple cycles			-0.205	-0.259		
			(0.184)	(0.177)		
Overall						
Constant	0.946***	2.175***	0.946***	2.169***	0.941***	2.170***
	(0.0367)	(0.0974)	(0.0371)	(0.0976)	(0.0367)	(0.0975)
Demographic Covariates	X	X	X	X	X	X
Observations	324	324	324	324	324	324

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Similarly, Tables 6-12 and 6-13 show results for all teachers and for first-year teachers on value-added outcomes. None of the coefficients on team and team quality indicator variables are statistically significant in either case, although the point estimates are somewhat larger and more positive for the quality indicator variables, in particular the indicator for having a goal, than for the team indicator variable. This provides some supportive evidence of the hypothesis that low-intensity teams bring down the average impact estimates, but even with quality indicators the effect is still small and not very precisely measured.

TABLE 6-12 RELATIONSHIP BETWEEN TEAM QUALITY INDICATORS AND VALUE-ADDED, ALL TEACHERS

	(1) Value added	(2) Value added	(3) Value added	(4) Value added	(5) Value added	(6) Value added	(7) Value added
Team indicator	-0.00130	-0.0133	-0.00173	0.00546	-0.00657		
	(0.00593)	(0.00905)	(0.00612)	(0.00966)	(0.0115)		
Has goal		0.0166			0.0165	0.00496	
		(0.00975)			(0.00985)	(0.00635)	
Multiple cycles			0.00556		0.00215		
			(0.0142)		(0.0143)		
Analyzes student work				-0.00830	-0.00842		
				(0.0101)	(0.0101)		
Overall							0.0134
							(0.0409)
Constant	0.0219	0.0215	0.0220	0.0220	0.0217	0.0202	0.0240
	(0.0162)	(0.0162)	(0.0162)	(0.0163)	(0.0162)	(0.0162)	(0.0187)
Demographic Covariates	X	X	X	X	X	X	X
Observations	13585	13585	13585	13585	13585	13585	8854

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 6-13 RELATIONSHIP BETWEEN TEAM QUALITY INDICATORS AND VALUE-ADDED, FIRST-YEAR TEACHERS

	(1) Value added	(2) Value added	(3) Value added	(4) Value added	(5) Value added	(6) Value added
Team indicator	0.0394	0.0155	0.0377	0.164	0.140	
	(0.0258)	(0.0371)	(0.0263)	(0.103)	(0.0975)	
Has goal		0.0336			0.0355	0.0473
		(0.0438)			(0.0437)	(0.0300)
Multiple cycles			0.0322		0.0482	
			(0.0876)		(0.0884)	
Analyzes student work				-0.143	-0.147	
				(0.103)	(0.101)	
Overall						
Constant	-0.113	-0.119^{*}	-0.113	-0.103	-0.111	-0.120^{*}
	(0.0581)	(0.0589)	(0.0580)	(0.0562)	(0.0567)	(0.0591)

Demographic Covariates	X	X	X	X	X	X
Observations	324	324	324	324	324	324

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Finally, Tables 6-14 and 6-15 show the association between the existence of a team at a grade-subject-subgroup cell and test score gains for K-8 schools and high schools, respectively. For the most part, the estimates do not change substantially using various indicators for quality, although notably the estimates are somewhat smaller on the coefficient for a team having a goal than for the simple team indicator, the estimate on analyzing student work is larger, positive, and approaching statistical significance, unlike in other models, and the coefficient on the hand-coded quality measure is positive and significant. This is somewhat surprising, given the substantial proportion of teams that do not even set a goal, as well as the very large number of teams that report analyzing student work. It could indicate that whether or not a team sets a goal is too rough an indicator of quality, and coding a larger sample of teams could be worthwhile for further study. Once again, there are very few statistically significant relationships at the high school level with the exception of the hand-coded quality measure.

TABLE 6-14 OLS ESTIMATES OF ASSOCIATION BETWEEN TEST SCORE GROWTH AND TEAM QUALITY MEASURES, K-8

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Growth	Growth	Growth	Growth	Growth	Growth	Growth
Team indicator	1.708***	1.621***	1.693***	0.823	0.714		
	(0.232)	(0.341)	(0.247)	(0.592)	(0.649)		
Has goal		0.158			0.163	1.659***	
		(0.443)			(0.452)	(0.302)	
Multiple cycles			0.209		0.175		
			(0.837)		(0.873)		
Student work				1.115	1.125		
				(0.692)	(0.692)		
Overall							1.443*
							(0.684)
Constant	-1.758**	-1.758**	-1.756**	-1.756**	-1.754**	-1.687**	-1.396
	(0.622)	(0.622)	(0.622)	(0.621)	(0.622)	(0.627)	(0.918)
Demographic Covariates	X	X	X	X	X	X	X
Observations	23671	23671	23671	23671	23671	23671	9765

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 6-15 OLS ESTIMATES OF ASSOCIATION BETWEEN TEST SCORE GROWTH AND TEAM QUALITY MEASURES, HS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Growth	Growth	Growth	Growth	Growth	Growth	Growth
Team indicator	0.0198	0.0221	0.0205	0.0280	0.0302		
	(0.0122)	(0.0129)	(0.0124)	(0.0270)	(0.0285)		
Has goal		-0.00448			-0.00438	0.0167	
		(0.0227)			(0.0228)	(0.0194)	
Multiple cycles			-0.0358		-0.0324		
			(0.0716)		(0.0730)		
Student work				-0.0104	-0.00955		
				(0.0308)	(0.0308)		
Overall							0.0886**
							(0.0318)
Constant	0.0712**	0.0709**	0.0713**	0.0711**	0.0709**	0.0742**	0.0830*
	(0.0246)	(0.0245)	(0.0246)	(0.0247)	(0.0245)	(0.0246)	(0.0326)
Demographic Covariates	X	X	X	X	X	X	X
Observations	11476	11476	11476	11476	11476	11476	2774

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Discussion

With some exceptions, the results were generally consistent across models, years and specifications, and despite a small number of positive and statistically significant relationships, still consistently weak. As a general rule, the key variable of interest on which teams varied was whether or not they set a goal, which serves a basic indicator as to whether or not a team was "real." The results are reasonably consistent with those presented in the main models, although overall somewhat larger and more positive, providing some limited support for the hypothesis that results are modest in part due to noise and heterogeneity. Overall, implementation appears to be strongest in the first year and the effects of teamwork are largest in that year. Still, adding measures of quality does not dramatically alter the results, suggesting that effects may still be small in reality, or that significant measurement error remains and the proxy measurements used here are weak. The general conclusion that remains fairly robust across models is that teamwork has a small effect, primarily on first-year teachers in terms of retention, possibly a small effect on student test scores in elementary and middle school grades, and some possible small effects on teacher value-added, especially with regard to teaching the lowest-performing students.

Given some surprising findings, particularly that whether or not a team looks at student work does not seem to significantly alter its effects, continued examination of heterogeneity and team quality is warranted. In addition to poor measures not adequately capturing quality, there is the possibility that quality measures do not significantly alter the results because teams were not, on average, implementing inquiry teams particularly well or with very high intensity. Even in the first year, when most quality measures are highest, the average score for the hand-coded sub-sample of teams was low and only one team in the sample of 100 achieved the highest possible score. Therefore, although there is some promising evidence for productivity-enhancing effects of inquiry, it appears that additional training and support would be needed in order to implement teams effectively at scale.

Finally, as noted above, all of these findings are subject to bias due not only to selection of teachers onto teams but also due to confounding of team quality with other factors, such as the preexisting, unobserved quality of the teachers on the team and unobserved elements of school culture and leadership. An attempt was made to address this concern by isolating possibly exogenous variation in team quality and intensity of implementation by using variability in accountability pressure due to the staggered nature of the accountability system as an instrumental variable. However, the instrument in this case suffered from several potential validity threats, including weak instrument problems and possible violations of the exclusion restriction; for this reason, the results are presented in Appendix B for illustration purposes, but are not presented as main findings.

Chapter 7 QUALITATIVE CASE STUDIES OF TEAMS IN ACTION

The purpose of the qualitative case study is to understand the process of teacher collaboration on inquiry teams and in particular how teachers interact on teams. While the unit of analysis for the quantitative analyses is the individual teacher, the unit of analysis for the qualitative analysis is the team. The major qualitative research question focuses descriptively on the processes by which teams work together to improve their practice and develop innovative solutions to instructional problems. Although the research design for the qualitative analysis does not have the goal of making causal inferences, descriptive analysis provides some evidence for proximal outcomes of successful teamwork, such as deeper questioning and abductive reasoning by teachers, as well as the conditions and processes associated with those outcomes, as noted in the conceptual framework. Overall, these findings can help contextualize the fairly modest results from the quantitative analyses by uncovering processes and conditions that lead teams to be more successful, identifying challenges and obstacles to the success of the inquiry team initiative, and suggesting appropriate proximal outcome measures that may show positive results of teamwork before value-added measures, which may require a longer time period to change.

Answers to these research questions will help address some gaps in the literature identified above and inform practice by honing in on specific practices when teachers are collaborating and learning on teams that could improve professional development and teamwork and collaboration in schools. To that end, teachers were asked reflective questions about how the team organized itself, why the team followed the processes it did, and any initial impacts in terms of teacher learning and changes in practice. Teams can further be a vehicle for identifying the practices that more effective teachers use and disseminating those practices to other teachers within a school, as well as other schools, thereby helping to address the information gap created by the difficulty in predicting teacher effectiveness based on observable characteristics.

CODING SCHEME

An initial coding scheme, based upon the literature and conceptual framework described above, the description of the intervention by the school district and Joan Talbert, the author's own experience working on a team of teachers, results from a prior smaller pilot study of a smaller group of teams, and initial impressions from the data collection phase, was created. From a list of approximately 100 characteristics that constitute effective teams, including aspects from the conceptual framework of conditions, processes, and outcomes, conceptually similar categories were grouped together to create 53 codes. For each code, Appendix Table C.1 documents the title, a detailed description, a quotation that serves as an example, and the source, whether it be from the literature, experience, or the data.

Major codes include examples of the formality of communication, the balance of contributions, organization and structure of meetings, leadership support, and openness to change among participants, generally reflecting "Conditions" in the conceptual framework. Other codes include experimentation, a focus on an individual student, investigation of student data or work, instructional strategies, and peer monitoring or peer pressure, as examples of "Processes." Finally, some codes represent proximal outcomes of the inquiry team process, including evidence of organizational learning or improvement upon the inquiry process itself, statements of the benefits of teamwork by participants, evidence of improvements in student learning, and evidence of changes in attitudes, thinking or instructional practices by participants.

FINDINGS

PROTOTYPICAL CASE

While the four teams in the sample differ in purpose and composition, there are clear consistent patterns across teams that provide important findings for further exploration and analysis. First, each team has a clear leader, although cases differ as to whether that role is explicit or implicit. On one team with more informal leadership, the apparent leader was an English teacher, whereas on another team, the leader did have positional authority as Assistant Principal, but seemed to supersede the Principal, also present at the meetings, in terms of setting agendas, facilitating the meeting, moving toward decisions and action, and ensuring follow-up. Two other teams did have clearly designated leaders who represented their teams of teachers – one grouped based on commonly-identified areas for instructional improvement and another comprising third grade teachers – to school leadership at a core inquiry team meeting.

A second similarity is that the teams have clearly gotten a signal that basing decisions on evidence, meaning specifically student work and student achievement data, is an expectation, as nearly all discussions were framed in terms of student evidence, and teams spent some time in nearly every meeting jointly analyzing student work or student data. Nonetheless, it is also clear that teachers have not yet received adequate training on meaningfully engaging with student data to make decisions, as in some cases teachers appear to be discussing student data in name only, or in a way that is not purposeful or strategic, and even tangential to the purpose of the meeting. While teams spend time discussing student data and examining student needs, as well as discussing instructional interventions to address student needs, there is often an apparent gap between these two discussions. In spite of clear differences between the four teams in terms of the degree of strategic follow-up based on data analysis, in all four cases teams struggled with connecting analysis to action and strategically taking advantage of resources to help address identified needs. This may help explain why examining student work was not a useful predictor of team quality in the quantitative analysis.

Finally, although some portion of every meeting is devoted to inquiry, defined as broadly as possible to capture any discussion of student data or student work, examination of the learning needs of an individual student or sub-group of students, and experimentation with or reflection on instructional strategies, teams also spend meeting time addressing other issues. These include student behavior issues, logistical issues such as planning field trips and parent-teacher conferences, as well as other off-topic discussions. Nonetheless, many of these teams have multiple purposes by design, so this observation is not intended as a critique of their inquiry process.

School A

Inquiry teams, known as Collaborative Learning Communities (CLCs) at School A, had been an integral part of teacher professional development at the school for at least three years prior to the 2011-2012 school year. At that time, the school made a conscious shift from inquiry teams focused on specific sub-groups of students or specific academic skills and toward teams focused on teacher learning goals, although the ultimate aim remained to increase student learning and teachers were encouraged to measure their own success with inquiry through their students' learning. In part as a result of a Race to the Top grant that required all school districts in the state to develop more rigorous teacher evaluation systems, the district was piloting a new evaluation and feedback system based on Charlotte Danielson's *Enhancing Professional Practice: A Framework for Teaching* (2007). The *Framework* breaks down teaching into research-based strands and domains, such as "Designing Student Assessments" and "Using Questioning and Discussion Techniques." Teachers reflected upon the strengths and weaknesses in their own practice and joined teams based upon common areas for improvement under the

Danielson rubric. These teams were led by teachers experienced with inquiry who had undergone training to serve as team facilitators and who represented the teams in a core team led by the principal.

This case study focuses primarily on one such team – the Engagement team, which encompasses goals related to questioning and discussion strategies and classroom management – and its relationship to the core team. The Engagement team was composed primarily of first-year teachers, and therefore the case serves as an example of inquiry work as professional development specifically for beginning teachers. Working on a team as a new teacher could address failure of teacher pre-service training to fully equip students with some of the skills they need to successfully teach; some of these, particularly related to engagement, questioning, and classroom management, may be experiential and best learned on-the-job. This case also examines unique challenges faced by a team composed mainly of inexperienced teachers, although it seems likely that an experienced facilitator with a clearly defined leadership role is particularly important under these circumstances. Given the quantitative findings that teamwork may be especially beneficial for first-year teachers, understanding the processes by which a team of beginning teachers works is particularly important.

While the teachers on the Engagement team had all selected similar professional development goals for inquiry, they each developed their own mini-research projects, with a question, a measurable goal, an instructional strategy they would try, and a strategy for gathering and analyzing data to assess their success. Therefore, much of the work that would be done collectively in the original inquiry team model was done by individuals who provided one another with support and feedback. One limitation of this approach is that, at least for the first-year teacher group, much of the communication observed by the research team was between individual members and the facilitator, as opposed to among the members themselves.

At team meetings, teachers shared their goals and how they planned to assess them. The team facilitator, as well as occasionally other teachers, gave individuals feedback and team members then revised their goals. The primary thrust of the feedback was for goals to be more narrowly focused and measurable. One teacher went through an illustrative process by thinking aloud:

> But here's the thing, like, my question has to do with students being invested in the learning. And students who are, like, often disengaged and unmotivated to even care what they're doing. So I don't think my question is as much about, like, them doing great on a… quiz. It's almost more about getting them more invested in what they're learning, why they're learning it, so maybe class work is my best gauge because these are the same kids who don't complete any work either. And they're not gonna get to the stage of success if they don't get to hear this first.
>
> And I was like, "Well, how do I, like, measure that with student data? And how can I provide something-- concrete where I can -- that's actually measurable?" So I was thinking… there's this little-- group of students that seemed very unmotivated to, like, do anything… And we, like, have a little, you know, writing in journals at the beginning of class every day. And they're never writing while the rest of the class is. When we get to do group work, they're the ones in the group that's not contributing.

In addition to helping new teachers refine their goals, the team facilitator served as a focal point for sharing instructional resources and suggesting tools to assist with data analysis. In addition to the direct support that sharing examples provided, it also served as a subtle form of peer pressure that led teachers to work on improving their own data collection and analysis practice.

The team facilitator represented the team at a core inquiry team facilitated by the principal. The principal served in an active management role in this team, directly impacting all of the teams throughout the school, soliciting updates on how the teams were progressing, providing feedback on their work, and giving directives on how to proceed with the next meetings. In this way, although the teams were semi-autonomous and individual teachers had a great deal of leeway in directing their work with regard to subject matter, goals, subgroups of students, instructional strategies, and data to collect and analyze, the process itself was prescriptive. The principal further reinforced the push toward more narrow, focused goals and concrete sources of evidence. The core team further worked to institutionalize learning from teams across the school by establishing a shared Dropbox folder and a newsletter for teachers.

There are some possible leading indicators of success of this approach. The Engagement team facilitator noted that she adapted her expectations when working with first-year teachers, such that much of the initial learning would be about the research skills needed for the process itself, as opposed to more direct learning about instructional skills. Even so, school leadership noticed some general improvement in instruction as a spillover effect, as seen in teacher lesson plans and classroom observations, perhaps due to the sustained focus and structured reflection inherent in the inquiry process.

> What ended up happening though was we kinda realized that by improving that questioning and… by making better lesson plans and having better management it kinda then led to the better teaching anyway which was kind of what the core team was kinda hoping to get around to in the end anyway.

Further, teachers themselves made note of the value of the process in uncovering new findings. One teacher commented on his surprise at the results of inquiry that, in retrospect, he should not have found surprising, which ultimately suggests that he is open to challenging some preconceptions about teaching and about his students:

> I feel like the things that I found out are all things that I almost, like, could have assumed would have been the case…
>
> I'm basing this on, like, three activities-- specific activities I've done with them. And even if everything's tight and ready to go, and like I feel like I've covered all my grounds, like, I'm still having some issues with some of them. But others… have actually been

> doing really well with me, and they're normally pretty defiant… because the three activities, especially the last one we did, we made these critters to talk about traits, and like-- it was something very doable for them.
>
> And I think… it would seem obvious that… having just something that they could do would help them to reduce behavior issues. But like, I don't know, that was, like, kind of like a big revelation for me, actually.

Relatedly, when another teacher was discussing a particularly disengaged student whom he had struggled to reach, other teachers on the team were willing to push back on the teacher's assumption and suggested that if he gave the student challenging, independent work to do that he would master the concepts for the class. This suggests some capacity for teachers working on teams to share their unique experiences and insights about individual students, as well as the benefits of a willingness to challenge one another's assumptions. Teachers are acutely aware of the challenges of inquiry and the additional investment required in terms of time and effort, but at this school have ultimately decided that the costs are worthwhile:

> I mean-- it's-- it's been a helpful process that-- where in the beginning I think it was kinda like, "Oh, this is something extra to do--" like, three years ago it's totally not viewed like that anymore here even by teachers that have been here the whole time-- because people do see results. And even if they don't see a result always, you know, the gains definitely are outweighing any disadvantage or, you know, failed inquiry by not seeing the results that we're hoping for. So it's been helpful.

Overall, this case provides a useful framework for a team primarily composed of inexperienced teachers with an experienced facilitator, which may be a useful model given the quantitative analyses. It also provides some context for explaining the general lack of quantitative results – the outcomes of interest were focused on social and emotional learning outcomes not captured by student test scores or value-added results, and even with strong leadership and sustained focus, teachers struggled with analyzing data effectively and mastering the inquiry process.

SCHOOL B

Inquiry teams at School B were organized around development of new math assessments that were aligned to the Common Core standards at each grade level. Therefore, this case is a particularly good example of how teacher collaboration may be particularly important when the curriculum, standards, or assessments change. Upon examining the school's prior student achievement data, the administration and core inquiry team concluded that the school needed to focus on math and in particular on the "Number sense and operations" skill, which comprised 47% of the new state test and the area in which students struggled the most. The skill focused on a broad range of arithmetic concepts, including counting, patterns, and skip counting at the younger grades, fractions and decimals at older grades, and addition, subtraction, multiplication, and division of different types of numbers across the elementary grades.

In the first part of the year, each grade-level team focused on constructing new assessments to measure student progress in this area, consulting a wide variety of published math materials. The assessment-creation process itself served as a tool for professional development, as teachers needed to familiarize themselves with the new standards and expectations and consider what areas were most important to assess student mastery, as well as anticipate possible areas of confusion for students. Further, this process was an opportunity for collaborative learning along two dimensions – teachers gave one another feedback on their assessment instruments within grades and across grade levels in the core inquiry team, facilitating individual learning, and the group made discoveries about the new math curriculum and student learning needs through the entire process. The structured, focused process also provided an opportunity for stronger teachers to assist teachers who needed help with assessment and inquiry skills, along with their own mathematics knowledge, as the lead third grade teacher selected by the principal provided a great deal of support to the relatively new Kindergarten teacher in the process. Additionally, the development of these assessments and then the subsequent implementation of the tests with students and analyzing the data for patterns of results gave a great deal of structure to the inquiry work, taking advantage of the Common Core implementation as an opportunity for teachers to deeply revisit their approach at a time when they may be unusually open to change:

> Well, it was a good jumping off point to do a lot of collaboration. You know, because certain topics, you know, when you're teaching for a while you teach things a certain way and you kinda don't really think about doing it another way. You know, like, for example I was having a hard time with equivalent fractions, I tried a couple different things.
>
> And you know, when we look at the data together and we say, "My kids are having trouble with equivalent fractions. Like, how do you, you know, approach that topic? How do you?" And then it's good for us to kinda talk it out and share things. And, I mean, not that we don't collaborate to begin with, but it gives us, you know, a more clear focus of, you know, something I'm doing is not working. And, like, let's talk about how we can help each other to come up with new ways to teach things.

After implementing the assessments, the core team discussed trends in student performance and brainstormed instructional strategies to address key gaps in student understanding. Interestingly, this process uncovered some gaps in teachers' own conceptual understanding of mathematics. As teachers debated the wording of questions about place value, some teachers noted that a question that asked about how many tens are in 900 could be answered with "90," when the teacher was looking for zero, the number that is in the tens place. Another teacher on the team noted that students would be correct in saying that there are, in fact, 90 tens in 900, launching a discussion about deepening teachers' and students' conceptual understanding about mathematics and ability to communicate these ideas clearly and effectively. While the difference in understanding may have been purely semantic, it did lead to some discussion about a professional development that a few teachers had attended about teaching math more conceptually; unfortunately, due to time and resource constraints, there was not evidence follow-up from this discussion in the subsequent meetings that were observed.

Following up from the analysis of the assessment results, the core team decided that it would be most helpful to gather more evidence about breakdowns in student understanding. Grade teams therefore selected a number of "case study" students for whom to closely examine student work with individual, multi-step problems to diagnose issues that may be representative of a larger group of students and test instructional strategies. This was in some ways a scaling-back from inquiry work in previous years, which involved more peer observation, low-inference note-taking, and closely observing students with serious learning or behavioral challenges in a number of settings. Teachers had generally reported that such processes required a great deal of time and work that did not seem to pay off in terms of improved instruction, suggesting that generalizing from highly idiosyncratic learning needs of students may be a challenge in inquiry teams leading to more systemic change. Therefore, inquiry work focused instead on carefully examining student work together using the "What Comes Up" protocol from *The Power of Protocols: An Educator's Guide to Better Practice* (McDonald et al, 2003). According to the protocol, teachers first observe low-inference "noticings" about the work before engaging in more free-form conversation about potential causes of the things they notice. Once teachers have collectively developed a hypothesis as to the cause of any gap in student understanding, they brainstorm possible remedies.

This process did contribute to deeper thinking and root cause analysis among the teachers, as they tried to uncover the source of gaps in student understanding. When examining how students respond to the problems, teachers noticed that students were often fixated on showing their work, using graphic organizers and drawings, in a way that was not clearly connected to the problem or to mathematical reasoning. Even in cases when answers to the problems were correct, student reasoning was opaque. In one example, students were asked to derive the optimal number of students and balloons at each table for a party. Several students spent a great deal of time drawing individual tables, students, and balloons and ultimately showed that they arrived at the correct answer through multiplication, even though the most efficient process would have used division and the drawings did not add anything to the analysis. Teachers therefore devised a strategy to better get at student thinking by asking them to prove their responses, as opposed to just showing their work, and providing less guidance in terms of graphic organizers and sentence starters to avoid explanations that added little to their understanding. In another case, a student responding to a problem related to recognizing a pattern successfully filled in a chart to help him see the pattern, but then wrote a number sentence that was seemingly unrelated.

> If he could've followed the chart, he would've been correct. That's why I said there's no-- I don't know how he got from the chart to this number sentence, to get--
>
> TEACHER 1:
>
> What's interesting thing about the number sentence is, even adding-- the numbers have no relation to the chart.
>
> But the answer for 10 and six and four should be 20.
>
> TEACHER 2:
>
> Right.
>
> TEACHER 1:
>
> But he wrote 14, which is close to the real answer.
>
> TEACHER 2:
>
> It's still not that, either.

In a follow-up discussion, teachers tried to understand what led the student to write a number sentence that was unrelated to other parts of the problem and also incorrect. They hypothesized that it could have been related to the balloon drawings, in that students were following a mechanical set of problem-solving steps without really understanding what they were doing, and therefore suggested a teaching strategy that encouraged them to step back and think more deeply about how they approached word problems in math:

> See, so I think doing the number sentence… I think he goes-- he just knows he has to do a number sentence, so he's just put it in there. He didn't really know why. So I think maybe-- I mean, I think the next step is if, like, everyone says, "Okay. We're gonna teach this like we're teaching reading a story, but that-- that will be what we do."

In spite of the apparent successes of deepening teachers' own understanding of the content through developing and analyzing assessments, and uncovering gaps in student understanding through close examination of student work, major challenges remained with regard to what teachers could do with that information. One challenge was how to generalize findings from a single student to a group of students, or from a single skill to the larger curriculum, given the need to address the learning needs of all students and the need to cover the full curriculum. Ideally, root cause analysis through inquiry will uncover instructional gaps that can be addressed with strategies that benefit all students or through general improvements in teacher human capital. A next-best scenario may be that teachers become more comfortable with flexible grouping and pacing strategies so that, while student learning gaps cannot necessarily be addressed in a holistic fashion that benefits all students, strategic re-teaching and small group instruction and tutoring can help address specific deficiencies. There is substantially more evidence for the latter response than the former in this case, suggesting that applying the *kaizen* model of team learning and continuous improvement to education may be particularly challenging and providing some explanation as to why results were small even when focusing on high-quality teams. Teachers did, however, adapt their pacing and instructional calendars in response to the findings:

> And also, they were thoughtful when they looked at the results. When they looked at the results from the first administration, they rearranged their curriculum so that they could meet the needs of, you know, for the test. Not only for the test, but what they need to go to the next grade, meet the needs of the standards.
>
> So they became more cyclical in their teaching. They knew they had to come back and review other, you know, topics that they introduced earlier on in the year. They moved ahead and moved topics closer, you know, closer to the test. They knew that some topics could wait. I think they had a better understanding of the expectations for the grade and what the common core standards ask. So I thought it was a very thoughtful process.

The lead third grade teacher and the assistant principal both cited success of the inquiry teams in pushing teachers to think in new ways and providing some structured opportunities to challenge teachers who conventionally were not open to change. The assessment development opened conversations about what truly constituted "mastery" – for instance, if a student understood a concept in a December administration but not in March, had they ever really understood it, or whether teachers should award partial credit when the standard calls for complete understanding. Grounding the inquiry work in concrete assessment and data analysis tasks provided a launching point for challenging reflections and conversations:

> And I think that going off of it we've really pushed ourselves to, like, go outside of what we knew, and you know, do more, look for more stuff, you know, teach things in different ways… And I think that as a majority of teachers, we did that. And it was a really successful project.

Finally, participants noted some costs and constraints they faced in the inquiry work. The assistant principal noted that, although grade-level representatives to the core team helped to facilitate the grade-level inquiry meetings, more direct leadership participation on all inquiry teams would be helpful if it were feasible:

> No, the only thing is, like, you know, you meet with a core member and they have to go back to their teams. It would just be, you know, a perfect world if we could, you know, have the teams meeting together and us going in and, you know, facilitating or, you know, it's just you meet with a team and they go back, you meet with a team and they go back. And it kinda holds you back a little bit.

Teachers also noted that inquiry competed with other demands on their scarce time, both individually and as a grade team. Therefore, teachers and school leaders needed to be strategic in how much time they allocate to inquiry teamwork and how to use their resources most effectively. This constraint has implications for the cost analysis, as well, and may call for enhancing the inquiry team initiative by providing funding for after-school meetings or by relieving teachers and administrators who participate on inquiry teams of other responsibilities.

School C

Data from School C comes from the 7th grade team, which regularly comprised the English Language Arts (ELA), math, and social studies teachers, with occasional representation by the science, English as a Second Language (ESL), and special education teachers. The team met quite regularly, as often as three times per week, and focused on other topics as well as inquiry. The frequency of meetings may have actually been an impediment to meaningful engagement in team learning, as team members appeared to take meetings for granted, meetings were often canceled or rescheduled without notice, and much meeting time was devoted to non-inquiry tasks. Although this is to be expected to some degree, given that inquiry was not the only stated purpose of the grade-level meetings and teachers had other important demands on their time, collectively and individually, it is notable that more time is not a sufficient condition for consistent engagement in inquiry work on its own.

At meetings when the team did follow the inquiry process, they followed a regular protocol by which one teacher brought a set of student work for the group to jointly examine, diagnose particular learning needs, and brainstorm instructional strategies that could be applied in that particular course or across the grade. At the next meeting, the team reflected upon the success of the strategy and discusses further refinements before moving on to another skill and strategy, repeating the cycle. Although she was not the formal leader, the ELA teacher appeared to facilitate the meetings, implicitly setting an agenda even when there was not a written agenda. There was a notable difference in tone and focus of meetings based on whether this one individual was present or absent from meetings. While the team seemed very comfortable with this protocol, it is unclear whether it was particularly effective given a few limitations. First, I was unable to observe in the data any follow-up or long-term reflection that occurred after trying an instructional strategy; however, given the frequency of the meetings, it was not possible to observe all of them, so some of this work may have happened when researchers were not present. Further, a large portion of meeting time was spent on non-inquiry tasks (e.g., a discussion of pencil sharpening routines that took up about a third of one short meeting). Finally, as discussed in more detail in the analysis section below, the team appeared to be somewhat limited in its ability to connect learning needs to instructional strategies in sophisticated and nuanced ways, instead gravitating toward somewhat simplistic and overly broad solutions, such as teaching essay-writing in identical formats across subject areas to limit student confusion.

At one meeting, teachers each carefully read essays students had written in ELA class to diagnose strengths and weaknesses in writing skills that could have more general implications for instruction in other classes, particularly social studies. They followed a simple protocol that involved first sharing low-inference observations about strengths and weaknesses and then engaging in more open conversation about possible causes of the weaknesses and strategies to try to alleviate them. The meeting and the protocol provided a structured opportunity for teachers to engage outside resources, including the *Write to Learn* assessment tools and teaching strategies from *Teaching Basic Writing Skills.* Teachers identified a number of strengths and weaknesses, praising students' ability to use supporting facts and details while noting that students struggled to organize and prioritize those details into a more coherent organizational structure.

Nonetheless, the discussion became focused on whether teachers should be requiring a consistent format for introductory and concluding paragraphs for all writing assignments across the grade, based on a format some teachers had learned at a workshop on teaching writing skills. One teacher suggested that some genres of writing did not call for formal introductions and conclusions, and another teacher wanted longer, more detailed introductions than the format suggested by the workshop. After some conversation, all teachers agree to try the 3-sentence strategy in their classes.

While there are some positive indicators in this interaction, including willingness of the team to engage in productive conflict and express disagreement and the examination of underlying beliefs about what constitutes developmentally appropriate writing, the resolution is ultimately formulaic, based on the assumption that clear and consistent expectations will improve student writing skills. The instructional strategy selected does not seem to be based on a clear connection between the student learning needs identified from the analysis of work and the learning standards. Although it was said in a joking manner, the following exchange seems revelatory in suggesting that the teachers themselves may have felt the proposed solution left issues unresolved:

> TEACHER A:
> All right, so that's our strategy.
> TEACHER B:
> Thank you.
> TEACHER A:
> And everyone have a lovely day.

In an interview, the team indicated that the regularity of meetings, support from school leadership for the work, and the student work protocol were all contributing to the team’s success. They appreciated the consistency, and if anything, wanted more consistent participation by the more peripheral members, who could only sporadically participate in team meetings due to scheduling conflicts:

> But if we gonna start off with a team, I understand that we're the core because we have been here together since September. But I wish that they would not, like, take… one person out.

In spite of competing demands for the team's time, there is evidence that they devoted a great deal of time to analyzing student work, an important part of the inquiry and the team learning process. Team members also exhibited a positive and open attitude regarding conflict and instructional change and a willingness to experiment:

> Thanks for-- sorry. I know it's hard to change your teaching practice.
> TEACHER B:
> No, it's okay.
> TEACHER A:
> I mean, you've done it for so long.
> TEACHER B:
> No, no. I mean, like-- listen, it's-- it's all about applying things and learning, you know?

Overall, there is some initial evidence for the team's success, but also some limitation. Notably, teachers placed strong value on collaboration and reflected that their collaborative processes had improved over time, mainly through communication channels becoming more frequent and informal, a condition predicted by Hoegl and Gemuenden. They also noted some anecdotal evidence that student work and course grades had improved since the previous year, when collaboration was less structured. In part due to the inquiry work, teachers became more comfortable observing one another, sharing strategies, and asking for help, and data collection became more systematic. Finally, teachers noted that some innovations developed in the 7th grade team had spread around the school, although many of these were not directly instructional and concerned primarily with school culture, including the implementation of reading logs and reward field trips to incentivize student reading.

The frequency of meetings and the level of freedom, informality and comfort that the teachers had with one another, along with some structural issues including shuffling of personnel at meetings, frequent rescheduling, and supplanting inquiry work with urgent issues, may have ultimately contributed to inquiry work that was successful in small cycles but ultimately ad hoc, disconnected from a larger strategy for teacher learning or instructional change. While the principal was very supportive of inquiry, structuring all of the school's collaborative planning time around inquiry work based on a charter school model and sometimes attending meetings herself, the lack of any prescriptive directive from leadership as well as the elimination of the requirement to focus on a specific sub-group of students for a sustained period of time may have led to inquiry work that was too diffuse to achieve any long-lasting change, even if there were some short-term benefits.

Finally, while a significant portion of the team's collaborative work was not oriented around inquiry or, more broadly, team-based problem-solving, action research, or team learning, that is not to say that the work did not have value for teachers or students. In addition to time the team spent on planning events to celebrate and enrich student learning, they also spent a great deal of time discussing behavioral challenges and needs of individual students and appropriate strategies for follow-up, including communications with parents and referring to other school staff to evaluate student emotional and behavioral needs and provide additional services, if needed. Therefore, further quantitative examination of non-academic outcomes, including social and emotional learning measures, student behavior, and attendance, is important for further research. While it is outside the scope of this dissertation, teacher collaboration can take many forms beyond those outlined in the conceptual framework in Chapter 1 and the policy under investigation here, and teachers and teams make choices about how to allocate their time based on the perceived relative costs and benefits of those different activities. It should not be assumed, therefore, that teams are necessarily making an incorrect decision when they substitute other types of collaboration, which may ultimately be more or less productive, for inquiry work.

School D

The team at School D comprised three administrators and a variable number of teachers who served as leaders of their own subject and grade-level teams. This team, known as the School-wide Data Team, came together to discuss school-wide trends in student learning, develop instructional strategies for the entire school, and focus on the professional development needs of teachers. Therefore, although the team followed the inquiry approach, the focus was less on direct instructional interventions and the needs of a particular group of students, and more on the general needs of teachers across the school to help them better meet the needs of students.

The team reviewed high-level student achievement data and identified academic vocabulary as a skill limitation that was impeding learning for many of the students across grades and subjects; many students were English Language Learners, so academic vocabulary acquisition in English was a particularly important skill within the school. The team quickly identified polysemous words with distinct yet related meanings across disciplines, such as *rate*, as a high-leverage area for skill development. Similarly to the School C team, however, the team transitioned from an evidence-based approach focused on identifying the learning needs of each individual student and each individual teacher to a one-size-fits-all solution that emphasized consistency above adaptation: the team spent much of the remainder of meetings discussing implementation and assessment of a school-wide Word of the Day initiative, whereby all students would be exposed to a multi-meaning academic vocabulary word in every subject each day and then tested on using the words from the previous week in multiple contexts each Friday.

It is unclear exactly why the team chose to value consistency over differentiation to meet the identified needs of individual students and teachers, although the data point to some possible reasons. One is that administrators on the team were acutely aware that teachers were feeling a great deal of pressure from a number of initiatives, including the implementation of the Common Core and a new teacher evaluation system; there was therefore a desire to keep inquiry work minimally intrusive so as to avoid further damage to teacher morale, which overall seemed to be lower at this school than in the others. Secondly, the team expressed a desire to gather data on a sub-sample of students across classes and grades so as to monitor and celebrate progress and identify any possible trends in performance that might help locate the most effective instructional strategies to share throughout the school. Therefore, although the proposed strategy is quite broad and somewhat disconnected from the original purpose of using inquiry to tailor instruction to student needs, it seems that the intent was for the Word of the Day initiative to be but a first step in the inquiry process. As appears to be a common obstacle, given the descriptive statistics on inquiry cycles from 2007-2010 and the findings of the CPRE inquiry studies, the team struggled to find the time to get the initiative off the ground quickly enough to be able to follow up with next steps.

As with the team in School C, an apparent leader played a role in pushing conversations to challenge underlying assumptions, as well as to move toward actionable decisions. This person had some positional authority, as one of the two Assistant Principals on the team, but was not the most senior member of the team, which also contained the Principal. This leader also repeatedly raised concerns about teacher buy-in, teacher capacity to complete this work, and the availability of scarce time, resources, and tools in order to help teachers be successful in implementing inquiry and data-related initiatives, suggesting some underlying concerns with school culture, shared values around collaboration, and capacity constraints:

> I mean, if we keep looking at all these different initiatives as separate pieces, it-- it-- we have to start seein' how they fit together. Otherwise it's like that's why people freak out and get stressed 'cause, "Oh, here's another initiative."
> "And another one and another one and another." We don't see how they lock together to make a whole picture.

There was much discussion about how to assess vocabulary acquisition, including whether such assessment would take place in just one subject or across subjects and for all students or a representative sample of students, and how to make sure the assessment was authentic and represented higher-order thinking skills. There is less tangible evidence, however, that any data from the vocabulary assessments was used to engage in root cause analysis to discover patterns in how particular instructional strategies relate to student learning outcomes. After collecting data on two rounds of pre-tests and post-tests for a representative sample of students from across the school, teachers and administrators met to analyze trends and decide upon next steps. A number of logistical challenges, including finding time to administer the assessments to enough students and finding meaningful patterns in the data given the very small sample, precluded significant follow-up.

Therefore, it appears that School D was putting into place several of the important ingredients for successful inquiry, including sophisticated data collection and difficult conversations challenging teacher assumptions about student learning, but due to an overly-broad focus, lack of teacher buy-in, and capacity constraints in teachers' time and ability to effectively use student data to inform instruction, did not see many indicators of positive results from inquiry in the time period under study. Nonetheless, there were some indicators of positive results of inquiry work, including the development and sharing of engaging materials that used audio, video, and images to help teach vocabulary, as well as increased capacity to engage in instructional research, including assessment, data collection, and analysis, among some teachers who led the initiative.

CROSS-CASE ANALYSIS

One clear conclusion that emerges from these four cases is that inquiry teamwork, as envisioned by the school district and stylized in Figure 1, is difficult to do well. Even with a sampling strategy designed to locate examples of the best inquiry work in the district, at none of the four schools were inquiry teams implemented with perfect fidelity to the original model or without substantial challenges and constraints. Notably, at the two schools that arguably achieved greater success with inquiry (Schools A and B), the model departed most radically from the original vision focused on targeted sub-skills and sub-groups of students, with an explicit focus on teacher development, as found by Chu et al. (2012). Three hypotheses can explain these findings, in conjunction with the quantitative findings: (1) inquiry team work is too challenging to work well under any circumstances, on average, given resource constraints and competing demands for teacher and administrator time; (2) in order to work well, inquiry teams require more sustained investment of time, resources, effort, and leadership support than is currently provided, on average; or (3) the products of inquiry team work are not well-measured by the existing outcome measures in the quantitative and qualitative data. The patterns of findings across these cases, as well as in the quantitative work, provide some suggestive evidence for the second and third hypotheses.

Table 7-1 shows counts for each code at each school, as well as the codes that most commonly co-occur with that code, as text excerpts can receive more than one code. Note that for brevity, co-occurrences are only listed once; for instance “Balance of contributions” and “Leadership support” tend to be coded together, so “Leadership support” is listed as a commonly co-occurring code under “Balance of contributions,” but the reverse is not true to avoid repetition. Additionally, if no codes co-occurred with a particular code in at least 3 text excerpts, nothing is listed in the code co-occurrence column.

TABLE 7-1 CODE APPLICATION BY SCHOOL AND CO-OCCURRENCE

Code	Commonly co-occurring codes	A	B	C	D
Conditions					
Balance of contributions	Leadership support	3	1		11
Coordination	Leadership support, Organization/structure	4	7		1
Free-riding					
Specialization					
Courage/Openness to change	Willingness to engage in conflict, Experimentation	2	5	5	2
Formal vs. informal communication		1	2	1	1
Formal communication			1		
Informal communication		2	1		
Leadership support	Organization/structure, Gathering useful data	29	14	1	25
Organization/Structure	Gathering useful data	16	14	4	29
Duration		4	1		
Frequency		2	2		4
Research skills	Leadership support	5	2	1	7
Asking good questions	Gathering useful data, Focus	16	13		24
Analyzing and interpreting results		3	4	2	2
Gathering useful data	Focus	14	18	5	43
Shared values and norms	Leadership support		2	1	4
Willingness to engage in conflict	Asking good questions	2	11	2	6
Processes					
Analyzing data	Goal-setting, Instructional strategies	1	17	14	6
Reviewing student work	Root cause analysis		10	10	5
Experimentation		7	3	2	5
Giving feedback		1	1	2	1
Receiving Feedback				1	1
Focus - depth vs. breadth	Framing, Asking good questions	15	4	1	4
Breadth		2			
Depth		2	1		
Framing	Goal-setting	17	13	3	4
Goal-setting	Leadership support	9	7	1	21
Individual student	Analyzing data	2	5	9	
Instructional strategies	Taking advantage of outside resources	4	10	21	13
Content		1	2	4	5
Skills	Goal-setting	1	6	8	11
Logistics/other non-inquiry				7	3
Peer monitoring	Leadership support, Peer pressure	5	5	4	2
Peer pressure	Peer monitoring	7			

Reflecting and Adjusting	Analyzing data	9	9	14	6
Root cause analysis		2	16	2	2
Taking advantage of outside resources	Skills	5	16	11	2
Outcomes					
Abductive Reasoning		2	2		
Consistency vs. Adaptation	Instructional strategies			14	18
Adaptation					1
Consistency			2		
Deeper questioning and thinking		1	4		
Individual vs. team learning			2	1	
Individual	Student learning, reflection and adjusting	6	5		
Team	Reflecting and adjusting		9	1	
Institutionalization	Organization/structure	7			2
Non-pecuniary benefits				1	
Organizational learning	Institutionalization	6	2	1	1
Student learning		1	9	8	1

One emerging theme suggested by patterns of code application across schools and code co-occurrence is that any one condition for the success of inquiry is insufficient without other conditions being in place, as well. For instance, school leadership was most directly involved in inquiry work at Schools A and D. Nonetheless, the role of leadership varied across those schools, as it seems leadership at School A was much more involved in framing problems and encouraging focus. Some processes also appear to be associated with one another across teams, and may lead to better outcomes – for instance, closely examining student work appears to be a catalyst for engaging in difficult conversations that challenge teacher preconceptions and contribute to deeper root cause analysis and strategic leveraging of outside resources.

Relatedly, these patterns of findings help explain why the quantitative results only improved slightly when taking into account quality and heterogeneity – the quality differences between teams are often quite nuanced. Setting a goal is not a sufficient indicator of team commitment, as all four teams set goals. Rather, the specificity of the goal and the extent to which the team effectively follows up on assessing progress toward the goal and making necessary adjustments matters more. Similarly, all teams analyzed student work, but it proved quite challenging to effectively make use of the findings of their analyses.

Notably, discussion at school C revolved substantially around processes, whereas discussion at school D revolved more around conditions, suggesting that they may, in fact, be at different stages of development in terms of building capacity for effective teamwork. Given the more active, direct role of school leadership in the school-wide data team at school D, it is unsurprising that there was much more evidence of leadership involvement and discussion of organization and structure of teamwork, whereas the protocols and structures seem to be more solidly in place at school C and therefore less explicitly addressed. At Schools A and B, on the other hand, there seems to be more balance in code application between conditions and processes, suggesting some complementarity between the two. Notable as well is some possible reverse causality in the conceptual framework. For instance, according to the pattern of code application and the rich description, School D may be best described as in the “Compliance” mode, rigidly following aspects of the inquiry process but facing some substantial resistance to change by teachers. It may indeed be that the outcomes drive the conditions and processes, rather than the other way around.

One unanticipated finding across schools is the balance between consistency and adaptation. The stated purpose of inquiry teams is to explore evidence-based instructional strategies to adapt to individual student learning needs. Teams at Schools C and D seemed to gravitate in the opposite direction, with school D representing almost the opposite extreme, whereby the strategy being tested was to teach every single student in the school the exact same content every day. In both cases, albeit in different ways, the connection between data analysis and strategic follow-up is tenuous. It appears that a careful balance between flexibility and rigidity, with enough structure to give teams focus but not so much so as to limit their capacity to experiment, is needed and contributes to better outcomes at Schools A and B.

Other themes include discussion of capacity constraints, including limitations on teacher time, competing initiatives, limited resources, tools, and technology to successfully complete this work, and potentially even skill gaps that teachers have in analyzing data and collaborating effectively. Relatedly, there seems to be some concern, particularly in school D, with teacher buy-in, and some fear of overwhelming teachers with new initiatives. There is even, at one point, explicit discussion of how additional burdens on teachers created by assessments and data analysis could run afoul of union contracts. Despite concerns about time, capacity, and resource constraints, there interestingly is no evidence of free-riding or specialization, two possible and divergent outcomes of teamwork predicted by economic theory. This could be because teachers are still ultimately responsible for their own classrooms and students, so there are limits to how much of these phenomena can occur.

Overall, the findings of these case studies are mostly consistent with the prior literature on teacher collaboration, as well as the previous investigations of inquiry. As with Chu et al. (2012) and the CPRE reports, these cases emphasize the importance of school leadership, the shift in focus from student to teacher learning, and in spite of some very promising practices and clear differences between schools, great challenges in implementing inquiry work even at the strongest schools. They help explain the limited quantitative findings, given challenges in implementation, and suggest student and teacher outcomes for further analysis. Further, they suggest that inquiry work that engages teachers in rigorous, research and evidence-based problem-solving is difficult and requires sustained effort, training, and culture change that may take some time to take hold, and require substantial investment of resources.

Chapter 8 COST-BENEFIT ANALYSIS

A full analysis of inquiry teams requires examination of not just the effects of teacher collaboration, but also the costs. Although regardless of the cost it makes little sense to implement a program that has no effects or measurable benefits, since inquiry teams seem to have some small effect when implemented well, and since many of the traditional alternatives to inquiry teams have little measured effects in the literature, a cost analysis can still provide useful information to policymakers.

A full accounting of all resources or ingredients required to achieve a particular measured result, and their associated properties such as experience, education, and special training or qualifications for personnel, taking into account all resources with alternative uses or opportunity costs including in-kind contributions, volunteer time, and reallocation of already-purchased resources from another use, gives the fullest and most accurate picture of the economic costs of an intervention (Levin and McEwan, 2001).

To obtain a conservatively high estimate of costs, this analysis will proceed by taking account of the costs of implementing for all teachers (as in 2009-2010). This is because, ex ante, decision-makers cannot likely anticipate the quality of implementation; therefore, it may be necessary to implement teams for all teachers but only to expect positive results from a small sub-sample of teachers who cannot be identified beforehand.

COST ANALYSIS

Ideally, ingredients data would be gathered from informants who directly implemented the policy in interviews. Such informants were not available at the time of this analysis, although they may be for further research; however, given the unusually rich implementation data available from the team databases over three years, as well as the history of the inquiry team initiative written by Talbert (2010) and the two implementation studies by CPRE (2008 and 2010), the ingredients necessary for replication can be inferred. Among all teams, a small sample within three levels of intensity of implementation were randomly selected and their responses in the inquiry database read to estimate the ingredients required and their associated quantities and qualities. Information from Talbert and the CPRE reports was used to verify and supplement this information with additional details, for instance on the coaching, training, and support provided by central district personnel. The three levels of intensity were low, defined as a team not setting a goal and therefore likely existing in name only; medium, defined as setting a goal and engaging in one inquiry cycle but no more; and high, defined as engaging in multiple inquiry cycles. The percentage of teams meeting each criterion among all teams in 2009-2010 was then applied to these estimates to obtain a weighted, pooled average of the cost of implementing inquiry teams to all teachers. These weights were 28% for low-intensity, 67% for medium-intensity, and 4% for high-intensity. These weights are varied in sensitivity analyses.

Relevant economic metrics can be calculated from these estimates. The weighted average cost per team and total cost for all teams are only meaningful when expressed relative to a unit of outcome. Therefore, the metrics will be cost per student per standard deviation gain in test scores, and cost per teacher per additional year teaching. There was a 3.8% increase in probability of returning for 7543 first-year teachers in 2007-2008, and about a 1.3 point (0.04 standard deviation) increase in student test scores in 2008-2009 and 2009-2010.

In addition to the cost per unit of output, in order to incorporate multiple outcomes and provide a metric that can be evaluated without a comparison, these outcomes can be converted into monetary values to directly ascertain whether the program is worthwhile. In other words, once the outcomes are monetarily valued, the net present value, or discounted benefits minus discounted costs, can be calculated to determine if the program meets the basic criterion of its benefits exceeding its costs. There is unlikely to be a market price for teacher retention or student achievement, but the values of these outcomes can be estimated using shadow pricing methods which assess willingness to pay. Calculating these shadow prices directly can be done via several methods, including stated preferences via contingent valuation, or revealed preference methods such as defensive expenditures and hedonic modeling, but doing so is beyond the scope of this dissertation. Shadow prices for these outcomes have been estimated in the literature and will be used here.

Note also that the perspective taken here is social and incremental. Social costs encompass all resources used in the intervention, regardless of who pays; in almost all cases with inquiry teams, the entity responsible for the costs will be the public school district. Social benefits similarly encompass all benefits, regardless of who receives them; the implicit assumption is that taxpayers have chosen to invest in public education because they value improved educational outcomes, even if they do not directly receive them, in part because of the positive externalities associated with education.

Incremental costs refer to costs over and above business-as-usual. Since in many cases inquiry work is happening during the school day, when teachers and principals are already being paid to work, it is difficult to disentangle exactly what is incremental. In many cases, particularly early in the phase-in of the initiative, teachers were paid overtime for inquiry team meetings after school, in which case included those additional hours as a cost of the program is clearly appropriate. When inquiry meetings are happening during the school day, whether or not the costs are incremental depends on what they are substituting; for instance, if teachers are taken out of class, necessitating substitute teachers, the cost is clearly incremental (although counting both teacher and substitute time would be double-counting). In other cases, such as when other types of teacher professional development or meetings are replaced by inquiry, whether or not the cost is incremental is less clear. To be conservative, I assume here that all costs of inquiry are incremental.

Cost estimates are presented in Table 8-1. Low intensity teams, which comprise 28% of the sample in 2009-2010, essentially meet once to organize but since they do not set a goal or report any follow-up, I assume that they then disband. The costs are therefore very low, essentially comprising about one hour of time for team members and for use of facilities and a computer to input the data on the team for that time. Total costs reported here are the product of unit costs and national average prices as reported by the National Center for Education Statistics and the Census Bureau; costs are in 2013 dollars and personnel costs reflect 29.5% average fringe benefit rates for K-12 educators.[6]

National prices were selected for teachers with Master's degrees and 10-14 years of experience, which is close to the average teacher in the school district, and for Assistant Principals and Principals with any level of experience and education. The average salary for Assistant Superintendents was used as the price for SAFs, as these personnel were generally senior school administrators. Average annual salaries were divided by 1440 hours in an academic year (8 hours per day for 180 school days) for teachers and counselors and the BLS definition of a working year, 2080 hours (8 hours per day for 260 weekdays per year) for administrators to obtain estimates of hourly wages. In the absence of a market for rental of educational facilities, average new construction prices as reported by School Planning and Management Magazine[7] were amortized at 3.5% interest over 30 years to annualize costs per square foot. These annual

[6] http://www.bls.gov/news.release/archives/ecec_12082010.pdf

[7] http://www.peterli.com/spm/pdfs/SchoolConstructionReport2011.pdf

costs were divided into the number of usable hours (1440 per year, to be conservative) to estimate an hourly cost per square foot of facilities. It is assumed that teams meet in a small classroom or conference room, about half the size of the average 900 square foot classroom.[8]

The CPRE implementation reports note that average team size was approximately 6, that most teams met about once or twice a month, and that the Senior Achievement Facilitator who trained and supported Inquiry Teams met with most teams about 2-3 times during the school year. These reports provide the basis for the Medium-Intensity team estimate. Based on data reported in the CPRE reports and inquiry databases, it is assumed that the Assistant Principal and Guidance Counselor are each members of the team, and the principal attends about half of the team meetings. High intensity teams are similar, but they are assumed to meet once per week and the SAF attends four times per year, which is the higher end reported by CPRE. Based on these estimates, the average cost for inquiry is approximately $4,360 per team, or $40,027,440 for all 9,176 teams in 2009-2010. These costs are broken down by ingredient and by team intensity in Table 8-1. In each case, teacher time constitutes the majority of the costs of inquiry teams, although for higher intensity teams, in which leadership participation is an important ingredient, administrator and counselor time is also a significant cost.

One additional consideration is the possibility of induced costs through the inquiry process itself. These costs can include purchasing new curriculum materials to test instructional strategies, engaging in small-group or individual tutoring with targeted students, and other costs of the interventions devised through the inquiry process and the external resources teachers sought for support. The CPRE implementation reports asked principals and teachers about these external costs and reported that they were very low.

[8] http://www.dpi.state.nd.us/finance/construct/sqfoot.pdf

TABLE 8-1 COSTS OF TEAMS BY INTENSITY

	Low Intensity		Medium Intensity		High Intensity		Pooled	
Ingredient	**Total Cost**	**%**	**Total Cost**	**%**	**Total Cost**	**%**	**Total Cost**	**%**
Personnel								
Teachers	$ 260.00	88%	$ 3,120.00	56%	$ 8,310.00	58%	$ 2,500.00	57%
AP	$ -	0%	$ 790.00	14%	$ 2,110.00	15%	$ 610.00	14%
Principal	$ 30.00	10%	$ 430.00	8%	$ 1,150.00	8%	$ 340.00	8%
SAF	$ -	0%	$ 250.00	4%	$ 330.00	2%	$ 180.00	4%
Other school staff - Guidance counselors	$ -	0%	$ 870.00	16%	$ 2,310.00	16%	$ 670.00	15%
Facilities								
Classroom/conference room for meetings	$ 10.00	2%	$ 70.00	1%	$ 190.00	1%	$ 60.00	1%
Materials								
Computers	$ 1.00	0%	$ 1.00	0%	$ 1.00	0%	$ 1.00	0%
Total cost per team	$ 300.00		$ 5,530.00		$ 14,410.00		$ 4,360.00	
Share of Teams	28%		67%		4%			

Notes: All dollar figures rounded to nearest $10, figures under $5 rounded to $1. Sources: Inquiry Spaces, 2007-2010; CPRE (2008, 2010); Talbert (2010)

BENEFIT-COST ANALYSIS

Given the multiple outcomes of inquiry teams and the need for a single economic metric that can evaluate whether it pays off as a social investment, a benefit-cost analysis that applies shadow prices to estimate the monetary value of the outcomes discussed above is appropriate. Several studies have attempted to estimate the costs of teacher turnover; a report commissioned by the National Commission on Teaching & America's Future attempts to do so using the "cost of illness" method, gathering data from five school districts on direct and indirect expenditures caused by teacher turnover. These could be low estimates of the costs due to negative spillover effects on achievement due to general disruption to school culture caused by turnover (Ronfeldt, Loeb, and Wyckoff, 2013). These spillover effects are omitted to avoid double-counting of benefits due to increased student achievement, which are themselves reduced form results that could be the net result of direct and indirect policy impacts. The costs of turnover collected include recruitment and advertising, special incentives, administrative processing, training for new hires, a reduction in achievement due to relatively less experienced teachers, and administrative costs such as substitute teachers and paperwork associated with teacher transfers. The average cost of turnover across four districts studied is $13,360, adjusted to 2013 dollars (Barnes, Crowe, and Schaefer, 2007). However, inducing a teacher to stay an additional year does not guarantee the teacher won't leave the following year; to be conservative, therefore, these benefits are divided by the average teacher tenure in the dataset, which is about 9 years, yielding $1,480 in benefits per teacher-year.

The economic benefits from increased achievement are estimated based on the total fiscal and social benefits of math achievement, estimated by Levin and Belfield (2009, Table 5), converted from 2006 to 2013 dollars and scaled from 0.25 to 0.04 standard deviations. Levin and

Belfield calculate these benefits using the cost of illness method, estimating the association between changes in math achievement and increases in the probability of high school graduation, and the concomitant economic benefits related to labor market, health, crime, and welfare outcomes. Because these estimates are based solely on math achievement, the number of students receiving the benefits are scaled by the percentage of inquiry teams focusing on math, which is 24%. Assuming each inquiry team targeted approximately 15 students, 33,034 students are expected to receive total social benefits of approximately $955 each, or $31,966,500 in benefits due to achievement. This is likely a lower-bound estimate, as it omits any benefits from ELA achievement, any spillover effects on math scores from teams that do not directly focus on math, and any benefits not captured by the outcomes measured in this study.

These benefits are combined and the total costs of implementing inquiry teams across the district in 2009-2010 are subtracted to calculate net benefits; benefits are also divided by costs to calculate a benefit-cost ratio. These results are reported in Table 8-2. Under these fairly stringent assumptions, inquiry teams do not pass a cost-benefit test, as the net benefits are negative, at approximately a $9.5 million loss, and the benefit-cost ratio is 0.76. Since it is less than one, the recommendation would be not to implement the policy if these are the outcomes of interest.

TABLE 8-2 BENEFIT-COST ANALYSIS OF INQUIRY TEAMS

Benefits per teacher-year	$	1,480
Total retention benefits	$	418,640
Benefits per 0.04 SD	$	960.00
Students		33033.6
Total achievement benefits	$	31,547,860
Total benefits	$	31,966,500
Total cost	$	40,007,360
Net benefits	$	(8,040,860)
B-C ratio		0.80

Sources: Barnes, Crowe, and Schaefer, 2007; Levin and Belfield, 2009

Sensitivity Analysis

Nonetheless, a number of assumptions are required in order to calculate these economic metrics. Some of these assumptions may be invalid, so a standard approach in cost analyses is to test whether the results are robust to assumptions. This can be done in a number of ways. The two most common are one-way sensitivity analysis, which consists of direct variation of the parameters in the model likely to have the greatest uncertainty to calculate, for example, best-case and worst-case scenarios, and break-even analysis, which estimates what the parameter values would need to be to change the recommendation from the analysis, in order to subjectively ascertain whether such values would be likely to be seen in practice (Levin and McEwan, 2001, p. 141-144).

The critical assumptions in this model are the weights applied to the various levels of intensity of team participation, the exact benefits to include, and the weights applied to derive the value of a teacher-year from the general costs of turnover and math scores from general achievement. Since assumptions were generally selected to be conservative, meaning to err on the side of estimating high costs and low benefits, sensitivity analyses will generally select assumptions that may yield results more favorable to the intervention. I therefore performed three sensitivity analyses: first, a best-case scenario assumes that ELA achievement results are worth half as much as math achievement results. The second analysis is a break-even analysis to determine what combination of weights between low and medium intensity teams will lead to net benefits being zero. Finally, an additional break-even analysis determines what share of students would need to receive the benefits from additional math scores in order for the intervention to break even; this assumption relates to both the value of ELA scores and the spillover effect on math scores from being the target of a team that does not explicitly target math.

The first sensitivity analysis estimates a benefit-cost ratio of 2.06 and net benefits of $42,254,260 from investing in the intervention. For the second sensitivity analysis, weights would have to be set such that the total costs of the intervention were $31,966,500, or equal to the benefits from the main estimate. This is achieved if we assume that roughly 46% of teams are low-intensity, 50% are medium-intensity, and 4% are high-intensity. Given that merely setting a goal and reporting one instructional strategy is a low bar to be considered "medium-intensity," for which it is assumed that teachers met about 15 times throughout the school year, this assumption seems plausible. Finally, if the math benefits are applied to 30% of students, as opposed to the 24% who were actually the target of a team focused on math, the net benefits equal zero. Once again, this assumption seems plausible; it implies that there are spillover effects on math from at least 6% of teams that are not focused on math, or that ELA achievement is worth at least 10% as much as math achievement in economic value. Given that the two break-even analyses result in plausible parameter values, the conclusion from the sensitivity analyses is there is a great deal of uncertainty around the benefit-cost estimates that merits further study. Specifically, uncertainty about the costs of medium-intensity teams, which represent a large share of the teams and therefore may include a wide range of implementation variability within that category, and about the benefits of ELA test scores merit further examination.

DISCUSSION

Overall, although the main benefit-cost results and the cost-effectiveness ratios are not especially promising for inquiry teams, there is reason to believe that inquiry teams may still be a worthwhile investment under certain conditions and depending upon the outcome of interest. Notably, the sensitivity analyses point to significant uncertainty around the benefit-cost estimates, even if the main estimates are negative. It is important to emphasize that the negative results are under extremely stringent assumptions, including the omission of several possible benefits of inquiry and only focusing on math achievement results.

One point worth emphasizing from the cost analyses, analysis of heterogeneity, and qualitative analysis is that, although low-intensity teams bring down the average cost of inquiry, they do add to the costs while likely contributing nothing to the results. A possible recommendation that emerges from this study is that future policies around inquiry team focus on deepening, rather than broadening teams. It may be the case that even though they require greater up-front investment, teams which have the resources, time, and support to do inquiry well achieve the greatest results, whereas the large share of low-intensity teams cost significantly less, but ultimately achieve nothing.

Chapter 9 CONCLUSION

SUMMARY OF FINDINGS

Overall, this dissertation contributes to the literature on teacher effectiveness, quality, training, and collaboration by providing quasi-experimental estimates of the effects of inquiry teams on various measures of teacher and student learning, and exploring in-depth heterogeneity of results and mechanisms through descriptive, qualitative, and cost analyses. One striking finding is that the overall effects of the inquiry team policy mandate are small and in many cases not statistically significant. This may be surprising given recent enthusiasm for teacher collaboration, and the widespread view that collaboration itself is unquestionably good. The small effects of the policy indicate some challenges in measuring and studying collaboration, and the risks associated with using a mandate as a policy lever. While it is likely that there is considerable noise in the team participation data, masking some much more positive effects of true collaboration, as a policy recommendation it is clear that mandating that teachers collaborate more is insufficient to achieve desired results. Nonetheless, there are indicators that under the right conditions the policy could be an effective tool for teacher development. Further, while the net benefits of the policy are negative under a stringent set of assumptions, there is evidence that the policy would pass a benefit-cost test under plausible assumptions regarding parameter values. Given several measurement and data challenges, including the inability to identify the exact students targeted by inquiry and relatively weak proxies for quality in the team data, even very small effects are promising for inquiry as a practice, if not as a policy.

The findings of this dissertation are largely consistent with the literature on team learning, the economics of teamwork and collaboration, and teacher collaboration. The effects of the initiative seem most promising in the first year and gradually decline, suggesting that support and attention from school and district leadership, including training and coaching that teams received in earlier years, is critical for its success. Nonetheless, the case studies reveal that these ingredients alone are not sufficient for successful inquiry; even among four teams sampled because of their promising practice, all of which implement inquiry with much higher-than-average intensity and with significant leadership support, there is substantial variation in conditions, processes, and outcomes.

Relative to economic theory on human capital and on the economics of teamwork, both the case studies and the quantitative results suggest that firm-specific training, experiential or idiosyncratic learning that is difficult to obtain in general education and pre-service training, and knowledge-sharing are the primary mechanisms by which inquiry teams work, as opposed to the kinds of innovative instructional solutions developed through team-based problem-solving and abductive reasoning that were originally envisioned by the policy.

POLICY IMPLICATIONS

Given the small, often not statistically significant, albeit usually positive results, some tempering of unbridled enthusiasm for teacher collaboration is in order. Collaboration by itself is clearly not a panacea. Structures and context matter a great deal, and collaborating effectively is a challenge that requires substantial resources and support. Reorganizing all school activities and professional learning around teamwork is unlikely to be effective on average, and the challenges teams face, such as generalizing from the learning needs of individual students or adopting new instructional strategies based on previously taught skills when teachers need to move on in the curriculum, suggest that adopting the *kaizen* model of team-based problem-solving in manufacturing to education is especially challenging.

Nonetheless, given some promising results, especially in the first year and for first-year teachers, combined with the weak evidence on alternatives, suggests that inquiry teams do have a place in enhancing teacher productivity. Rather than rapidly scaling up in a way that may undermine the authenticity of the process, it seems that concentrating efforts and resources in doing inquiry particularly well in targeted areas where it is most likely to have an effect, especially for beginning teachers paired with volunteers who may find inquiry work most beneficial, would be more promising. Findings from this and further study on what makes inquiry effective could also inform pre-service training programs, alternative certification and teacher residency programs, and other avenues for increasing teacher quality by emphasizing experiential and idiosyncratic learning.

Finally, although there was the least evidence for this effect in the dissertation, there is still promise for inquiry as an avenue for broader organizational learning and discovery, including sharing of knowledge, innovation around intractable instructional problems, such as persistently struggling or disruptive students, and adapting instruction to meet the learning needs of students. Testing these effects of inquiry is beyond the scope of this dissertation, but at the strongest case study schools, there was some evidence of using inquiry to share effective practices, if not direct evidence of organizational learning.

LIMITATIONS AND SUGGESTIONS FOR FURTHER RESEARCH

Teacher collaboration is a particularly difficult subject to study using the emerging econometric tools of causal inference. Myriad selection issues, related to teachers volunteering on teams, principals assigning teachers to teams, the reflection problem created by team composition itself, and the endogeneity of team processes and implementation fidelity, make causal identification particularly difficult in this context. Further, team processes and the intensity and quality of team work are not adequately captured in existing datasets. The literature on collaboration suggests, and the case study analysis finds empirical support, that small, nuanced details – not just whether the principal participates on a team, but how, or the team's willingness to engage in conflict – make the critical difference. This dissertation has made a contribution to frameworks for investigating workplace collaboration and team-based learning, particularly among teachers, but more work is to be done in terms of developing stronger identification strategies and improving data collection.

Notably, by using administrative data on what teams actually did, as opposed to survey data on how teachers collaborate more generally, this dissertation examines data more grounded in empirical reality. The data still, however, were limited by potential social desirability bias and other inaccuracies due to self-reporting by teachers and teams. Team members may have felt pressure to exaggerate the extent of their actual teamwork in administrative databases, particularly since the data were collected by the school district sponsoring the initiative. Therefore, further work examining inquiry teams as a practice, as opposed to a policy, can uncover more nuanced and more realistic aspects of teamwork through data gathered by external observers, reporting on how often teams met, leadership participation, team roles, and follow-up in terms of changes in classroom practice. Further, while administrative data offer some advantages over surveys, the voluntary nature of data entry on inquiry may reflect solely the views of those who found inquiry most helpful or were most concerned about how they appeared to district leaders. Therefore, an anonymous survey that did not identify individual teachers but could be linked to administrative data on teams could gather more nuanced information about team interaction, including from those who did not have a positive experience with teams.

As an additional candidate for further study, the instrumental variable selected in the third natural experiment, based on the literature on optimal team size, turned out to have a number of problems. Additional sources of exogenous variation in collaboration, including policy phase-ins across schools in addition to within schools and natural facilitators or impediments to collaboration, such as the physical layout of schools or demographic similarities among teachers, should be explored for further causal study. Social network theory and analysis could explore other sources of variation in whether and how teachers work together. One limitation of several of the identification strategies employed in this dissertation is that they emphasized grade-level teams as a source of exogenous variation in teamwork; however, some descriptive evidence suggests that, on some outcomes in some years, grade-level teams were somewhat less effective than the teams that focused on a subject area or a subgroup of students across grades. Alternative instruments can help address this issue.

Additionally, although this dissertation explored mechanisms and heterogeneity through analysis of detailed records of team activities, these activities proved insufficient to successfully differentiate between teams without significant, laborious, and highly subjective judgment. Proxies for quality, such as goal-setting and investigation of student work, were not strong predictors of outcomes. This could be because they simply are not related to outcomes, but it seems more likely that they fail to capture important nuance. Future analysis therefore might consider data mining techniques to uncover further patterns in the data, although such analysis risks spurious and ex post facto findings and must be done with caution. Follow-up work to this dissertation may extend the hand-coding of teams for markers of quality, in part to gain statistical power for further analysis with this sub-sample and in part to uncover patterns in the data that could indicate more about team processes.

Further research will also more precisely estimate costs by adding data from principals who implemented the initiative. Further cost analyses can consider variation in costs and benefits as the implementation scales up. On a per team basis, the initiative was most costly but also most effective during first year – further research is needed to determine the implications of this for policy.

Finally, this dissertation examined a fairly narrow range of academic outcomes, defining teacher productivity primarily around test score gains, whether as measured by student growth or teacher value-added. This was largely a limitation due to the availability of data, as data on non-academic outcomes was not disaggregated to an extent that would make it possible to link to specific teams, but given that the emphasis of inquiry work is on tailoring teacher development to student needs, it may be reasonable to expect that the bulk of the effects would be on critical non-academic outcomes, such as social and emotional learning, attendance, and suspensions. Future work should examine that link and broaden the definition of teacher productivity.

Contribution

This dissertation represents a significant contribution to our knowledge about teacher collaboration, and collaborative inquiry teams in particular. It is the first comprehensive examination of a single collaboration initiative that incorporates quasi-experimental, quantitative analyses utilizing administrative rather than survey data, analysis of heterogeneity and mechanisms, qualitative case studies of the practices and proximal outcomes of effective teams, and a cost-benefit analysis using the ingredients method. It includes one of the few quantitative analyses of teacher collaboration, and only the second quasi-experimental analysis.

While other quantitative analyses of teacher collaboration have tended to find substantively small effects, my analysis found no effects on teacher productivity or student learning in almost all cases. The only small effects which may exist are on teacher retention among first-year teachers and possibly on student learning, both only in the first year or two of the intervention. These findings, in conjunction with results from the heterogeneity and qualitative analyses, suggest reevaluation of teacher collaboration policies are in order. While there is little doubt that some teams are implementing inquiry teams especially well, based on examination of the inquiry database and the case study analysis, and some evidence from the case studies for more proximal indicators of teacher learning which suggest that longer-term analyses would be beneficial, it is clear that on average teams are not implementing inquiry intensively and are not seeing many positive results. In fact, the qualitative analysis suggests that even four especially strong teams could benefit from additional time, support, and resources, and struggle with particular parts of the process, especially appropriate follow-up. The baseline recommendation from the cost-benefit analysis would be to not implement inquiry teams, at least at the present scale, when a large investment in fairly weak implementation across the entire district only appears to be paying off in small results for a subset of teams and teachers. Based on this analysis, a more targeted initiative, focused on beginning teachers and those particularly interested in collaboration, with more time and support, would likely be far more effective and a more efficient investment of scarce school resources.

REFERENCES

Achinstein, B. (2002). Conflict Amid Community: The micropolitics of teacher collaboration. *Teachers College Record*, *104*(3), 421–455. doi:10.1111/1467-9620.00168

Ai & Norton. (2003). Interaction terms in logit and probit models. Economics Letters 80(1):123−129.

Angrist, J.D. (2001). Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business and Economic Statistics, 19*(1): 2-28.

Angrist, J. D., & Lavy, V. (2001). Does teacher training affect pupil learning? Evidence from Matched Comparisons in Jerusalem Public Schools. *Journal of Labour Economics*, *19*(2), 343–369.

Ansell, C. (2011). *Pragmatist Democracy: Evolutionary Learning as Public Philosophy.* New York: Oxford University Press.

Bandiera, O., Barankay, I., & Rasul, I. (2010). Social incentives in the workplace. *Review of Economic Studies*, *77*(2), 417–458. doi:10.1111/j.1467-937X.2009.00574.x

Barnes, G., Crowe, E., & Schaefer, B. (2007). *The Cost of Teacher Turnover in Five School Districts: A Pilot Study.* National Commission on Teaching and America's Future. Retrieved from http://nctaf.org/wp-content/uploads/CTTFullReportfinal.pdf.

Barrett, N., Butler, J. S., & Toma, E. F. (2012). Do less effective teachers choose professional development does it matter? *Evaluation Review*, *36*(5), 346–74. doi:10.1177/0193841X12473304

Becker, G.S. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education.* Chicago, IL: The University of Chicago Press.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*(4), 1165-1188.

Birman B, Le Floch KC, Klekotka A, Ludwig M, Taylor J, Walters K, et al. (2007). State and local implementation of the No Child Left Behind Act: Vol 2 Teacher quality under NCLB: Interim report (U.S. Department of Education, Washington, DC). Office of Planning, Evaluation and Policy Development; Policy and Program Studies Service.

Bressoux, P., Kramarz, F., & Prost, C. (2009). Teachers' training, class size and students' outcomes: Learning from administrative forecasting mistakes. *The Economic Journal*, *119*(March), 540–561.

Bryk, A., Camburn, E., & Louis, K. S. (1999). Professional community in Chicago elementary schools: Facilitating factors and organizational consequences. *Educational Administration Quarterly*, *35*(5), 751–781. doi:10.1177/0013161X99355004

Burbank, M. D., & Kauchak, D. (2003). An alternative model for professional development: investigations into effective collaboration. *Teaching and Teacher Education*, *19*(5), 499–514. doi:10.1016/S0742-051X(03)00048-9

Buysse, V., Sparkman, K. L., & Wesley, P. W. (2003). Communities of practice: Connecting what we know with what we do. *Exceptional Children, 69*(3), 263–277.

Cameron, A.C. and Miller, D.L. (2014). *A Practitioner's Guide to Cluster-Robust Inference.* Retrieved from: http://www.econ.ucdavis.edu/faculty/cameron/research/Cameron_Miller_JHR_2014_July_09.pdf.

Chalos, P., & Pickard, S. (1985). Information choice and cue use: An experiment in group information processing. *Journal of Applied Psychology*, *70*(4), 634–641. doi:10.1037//0021-9010.70.4.634

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2013a). *Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates* (No. 19423) (pp. 1–43). Retrieved from http://www.nber.org/papers/w19423

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2013b). *Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood* (No. 19424).

Chong, W.H. & Kong, C.A. (2012). Teacher collaborative learning and teacher self-efficacy: the case of lesson study. *The Journal of Experimental Education, 80*(3), 263-283.

Chu, E., Hulden, M., Sabel, C., Shand, R., & Wallenstein, J. (2012). *Getting Big to Go Small: Case Studies of Collaborative Inquiry Teams in New York City.* Working paper: Columbia University.

Cleves, M.A., Gould, W.M., and Guitierrez, R.G. (2002). *An Introduction to Survival Analysis Using Stata.* College Station, Texas: Stata Corporation.

Consortium for Policy Research in Education (2008). *A Formative Study of the Implementation of the Inquiry Team Process in New York City Public Schools: 2007-08 Findings.* New York, NY: Robinson, M.A., Kannapel, P., Gujarati, J., Williams, H., & Oettinger, A.

Consortium for Policy Research in Education (2010). *School Perspectives on Collaborative Inquiry: Lessons Learned From New York City: 2009-2010.* New York, NY: Robinson, M.A.

Corcoran, S. (2010). *Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy & Practice* (Providence, RI: Annenberg Institute for School Reform of Brown University)

Corcoran, T. B. (1995). *CPRE Policy Brief: Professional Development* (pp. 1–11).

Creswell, J.W. & Plano Clark, V.L. (2011). *Designing and Conducting Mixed Methods Research, 2nd Ed.* Thousand Oaks, CA: Sage Publications.

Danielson, C. (2007). *Enhancing Professional Practice: A Framework for Teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. H., & Rothstein, J. (2011). *Capitol Hill Research Briefing: Getting Teacher Evaluation Right: A Background Paper for Policy Makers*.

Darling-Hammond, L., & Richardson, N. (2009). Research Review / Teacher Learning : What Matters ? *Educational Leadership*, *66*(5), 46–53.

Darling-Hammond, L., & Sykes, G. (2003). Wanted: A National Teacher Supply Policy for Education: The Right Way to Meet the 'Highly Qualified' Teacher Challenge. *Educational Policy Analysis Archives, 11*(33), http://epaa.asu.edu/epaa/v11n33/.

Darling-Hammond, L., & Wei, R. C. (2009). Professional Learning in the Learning Profession : A Status Report on Teacher Development in the United States and Abroad A Status Report on Teacher Development in the United States and Abroad.

Day, C., & Gu, Q. (2007). Variations in the conditions for teachers' professional learning and development: sustaining commitment and effectiveness over a career. *Oxford Review of Education*, *33*(4), 423–443. doi:10.1080/03054980701450746

Dee, T., & Wyckoff, J. (2013). *Incentives, Selection, and Teacher Performance: Evidence from IMPACT* (No. 19529).

Desimone, L. M. (2009). Improving Impact Studies of Teachers' Professional Development : Toward Better Conceptualizations and Measures. *Educational Researcher*, *38*(3), 181–199.

Dieterle, S. and Snell, A. (2014). *It's Hip to Be Square: Using Quadratic First Stages to Investigate Instrument Validity and Heterogeneous Effects.* Working paper retrieved from: http://homepages.econ.ed.ac.uk/~sdieterl/researchpapers/DieterleSnellIV.pdf.

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. doi:10.1257/app.5.4.28

Dyer, J. & Nobeoka, K. (1998). Creating and managing a high-performance knowledge-sharing network: the Toyota case. Working paper W-0147b. Retrieved from http://hdl.handle.net/1721.1/1441.

Eide, E., Goldhaber, D., & Brewer, D. (2004). The teacher labour market and teacher quality. *Oxford Review of Economic Policy*, *20*(2), 230–244.

Farrell, J.P. & Oliveira, J. (1993). Teacher costs and teacher effectiveness in developing countries. In *Teachers in Developing Countries: Improving Effectiveness and Managing Costs,* edited by J.P. Farrell and J.B. Oliveira, pp. 175-86. Economic Development Institute Seminar Series. Washington, DC: World Bank.

Fryer, R. G. (2013). Teacher Incentives and Student Achievement : Evidence from New York City Public Schools. *Journal of Labor Economics*, *31*(2), 373–407.

Gallimore, R., Ermeling, B., Saunders, W., & Goldenberg, C. (2009). Moving the learning of teaching closer to practice: Teacher education implications of school-based inquiry teams. *The Elementary School Journal, 109*(5), 537-553.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., … Silverberg, M. (2008). *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement.*

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What Makes Professional Development Effective? Results From a National Sample of Teachers. *American Educational Research Journal*, *38*(4), 915–945. doi:10.3102/00028312038004915

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Walters, K., Song, M., … Warner, E. (2010). *Middle School Mathematics Professional Development Impact Study Findings After the First Year of Implementation*. Washington, DC.

Glazerman, S., & Decker, P. (2006). Alternative Routes to Teaching : The Impacts of Teach for America on Student Achievement and Other Abstract. *Journal of Policy Analysis and Management*, *25*(1), 75–96. doi:10.1002/pam

Glewwe, P., Ilias, N., & Kremer, M. (2003). *Teacher Incentives* (No. 9671).

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National Board certification as a signal of effective teaching. *The Review of Economics and Statistics*, *89*(1), 134–150.

Goldhaber, D., Destler, K., & Player, D. (2010). Teacher labor markets and the perils of using hedonics to estimate compensating differentials in the public sector. *Economics of Education Review*, *29*(1), 1–17. doi:10.1016/j.econedurev.2009.07.010

Goldhaber, D., Liddle, S., & Theobald, R. (2012). *The Gateway to the Profession: Assessing Teacher Preparation Programs Based on Student Achievement* (No. 4).

Goldschmidt, P., & Phelps, G. (2010). Does teacher professional development affect content and pedagogical knowledge: How much and for how long? *Economics of Education Review*, *29*(3), 432–439. doi:10.1016/j.econedurev.2009.10.002

Graham, P. (2007). Improving Teacher Effectiveness through Structured Collaboration : A Case Study of a Professional Learning Community. *Research in Middle Level Education*, *31*(1).

Guarino, C. M., Santibanez, L., & Daley, G. A. (2006). Teacher Recruitment and Retention: A Review of the Recent Empirical Literature. *Review of Educational Research*, *76*(2), 173–208. doi:10.3102/00346543076002173

Haertel, E. H. (2013). *Reliability and Validity of Inferences About Teachers Based on Student Test Scores* (pp. 1–28). Princeton, New Jersey.

Hamilton, B. H., Nickerson, J. A., & Owan, H. (2003). Team Incentives and Worker Heterogeneity : An Empirical Analysis of the Impact of Teams on Productivity and Participation. *Journal of Political Economy*, *111*(3), 465–497.

Hanushek, E. (2007). “The single salary schedule and other issues of teacher pay.” *Peabody Journal of Education, 82*(4), 574-586.

Hanushek, E. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*(3): 466-479.

Hargreaves, A. & Fullan, M. (2012). *Professional Capital: Transforming Teaching in Every School*. New York: Teachers College Press.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7-8), 798–812. doi:10.1016/j.jpubeco.2010.11.009

Hermes, N., Lensink, R., & Mehrteab, H.T. (2005). Peer monitoring, social ties, and moral hazard in group lending programs: Evidence from Eritrea. *World Development, 33*(1), 149-169.

Hoegl, H., & Gemuenden, G. (2001). Teamwork and the Success of Quality A Theoretical Concept Projects : Innovative and Empirical Evidence. *Organizational Science*, *12*(4), 435–449.

Hoy, W. (1990). Organizational climate and culture: a conceptual analysis of the school workplace. *Journal of Educational and Psychological Consultation, 1*(2), 149-168.

Huffman, D., & Kalnin, J. (2003). Collaborative inquiry to make data-based decisions in schools. *Teaching and Teacher Education*, *19*(6), 569–580

Jackson, C.K. & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics, 1*(4), 85-108.

Jacob, B. A., & Lefgren, L. (2004). The Impact of Teacher Training on Reform Efforts in Chicago. *The Journal of Human Resources*, *39*(1), 50–78.

Jacob, B. A., & Walsh, E. (2011). What’s in a rating? *Economics of Education Review*, *30*(3), 434–448. doi:10.1016/j.econedurev.2010.12.009

Kasl, E., Marsick, V.J., and Dechant, K. (1997). Teams as learners: A research-based model of team learning. *Journal of Applied Behavioral Science, 33*(2), 227-246.

Kandel, E., & Lazear, E. P. (2012). Peer Pressure and Partnerships. *Journal of Political Economy*, *100*(4), 801–817.

Kane, T., Rockoff, J., & Staiger, D. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0272775707000775

Kane, T., & Taylor, E. (2011). Identifying Effective Classroom Practices Using Student Achievement Data. *Journal of Human Resources, 46*(3), 587–613. Retrieved from http://jhr.uwpress.org/content/46/3/587.short

Kasl, E., Marsick, V. & Dechant, K. (1997). Teams as learners: A research-based model of team learning. *Journal of Applied Behavioral Science, 33*(2), 227-246.

Kelchtermans, G. (2006). Teacher collaboration and collegiality as workplace conditions: A review. *Zeitschrift Fur Padagogik*, *52*(2), 220–237.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*, 623–55. doi:10.1146/annurev.psych.55.090902.142009

Koedel, C., & Ehlert, M. (2012). *Teacher Preparation Programs and Teacher Quality : Are There Real Differences Across Programs ?*

Kraft, M. A., & Papay, J. P. (2013). *Can Professional Environments in Schools Promote Teacher Development? Explaining Heterogeneity in Returns to Teaching Experience*.

Kreps, D. (1997). Intrinsic Motivation and Extrinsic Incentives. *American Economic Review*, *87*(2), 359–364.

Ladd, H. (2009). *Teachers' Perceptions of Their Working Conditions: How Predictive of Policy-Relevant Outcomes?* (No. 33).

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis. *Educational Evaluation and Policy Analysis 24*(1), 37-62.

Lavy, V. (2009). Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review*, *99*(5), 1979–2011.

Lee, O., Deaktor, R., Enders, C., & Lambert, J. (2008). Impact of a multiyear professional development intervention on science achievement of culturally and linguistically diverse elementary students. *Journal of Research in Science Teaching*, *45*(6), 726–747. doi:10.1002/tea.20231

Levin, H.M. & Belfield, C.R. (2009). *Some Economic Consequences of Improving Mathematics Performance.* Menlo Park, CA: SRI International. Retrieved from http://cbcse.org/wordpress/wp-content/uploads/2012/10/CliveCTL_economic_consquences_math_report_092009.pdf.

Levin, H.M. & McEwan, P.J. (2001). *Cost-effectiveness Analysis: Methods and Applications* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Little, J. W. (1982). Norms of Collegiality and Experimentation: Workplace Conditions of School Success. *American Educational Research Journal*, *19*(3), 325–340. doi:10.3102/00028312019003325

Loeb, S., Kalogrides, D., & Béteille, T. (2012). Effective Schools: Teacher Hiring, Assignment, Development, and Retention. *Education Finance and Policy*, *7*(3), 269–304. doi:10.1162/EDFP_a_00068

Manski, C.F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies, 60*(3): 531-542.

Mas, A., & Moretti, E. (2006). *Peers at Work* (No. 2292).

McDonald, J.P., Mohr, N., Dichter, A., & McDonald, E.C. (2003). *The Power of Protocols: An Educator's Guide to Better Practice.* New York, NY: Teachers College Press.

Merriam, S.B. (2009). *Qualitative Research: A Guide to Design and Implementation.* San Francisco, CA: Jossey-Bass.

Moolenaar, N. M. (2012). A Social Network Perspective on Teacher Collaboration in Schools: Theory, Methodology, and Applications. *American Journal of Education*, *119*(1), 7–39.

Mourshed, M., Chijioke, C., & Barber, M. (2010). *How the World's Most Improved School Systems Keep Getting Better.* McKinsey & Co. Retrieved from: http://www.mckinsey.com/client_service/social_sector/latest_thinking/worlds_most_improved_schools

Mundlak, Y. (1961). Empirical production function free of management bias. *Journal of Farm Economics, 43*(1), 44-56.

Nelson, T. H. (2009). Teachers' collaborative inquiry and professional growth: Should we be optimistic? *Science Teacher Education*, *93*(3), 548–580. doi:10.1002/sce.20302

Palardy, G. J., & Rumberger, R. W. (2008). *Teacher Effectiveness in First Grade: The Importance of Background Qualifications, Attitudes, and Instructional Practices for Student Learning*. *Educational Evaluation and Policy Analysis* (Vol. 30, pp. 111–140). doi:10.3102/0162373708317680

Papay, J. P., & Kraft, M. A. (2013). *Productivity Returns to Experience in the Teacher Labor Market: Methodological Challenges and New Evidence on Long-Term Career Growth*.

Panero, N.S. & Talbert, J.E. (2013). *Strategic Inquiry: Starting Small for Big Results in Education.* Cambridge, Massachusetts: Harvard Education Press.

Pil, F.K. & Leana, C. (2009). Applying organizational research to public school reform: The effects of teacher human and social capital on student performance. *Academy of Management Journal, 52*(6), 1101-1124.

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *AEA Papers and Proceedings*, *94*(2), 247–252.

Rockoff, J. E. (2008). *Does Mentoring Reduce Turnover and Improve Skills of New Employees? Evidence from Teachers in New York City* (No. 13868). Retrieved from http://www.nber.org/papers/w13868

Ronfeldt, M., Farmer, S.O., McQueen, K., & Grissom, J.A. (2015). Teacher collaboration in instructional teams and student achievement. *American Educational Research Journal, 52*(3), 475-514.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How Teacher Turnover Harms Student Achievement. *American Educational Research Journal*, *50*(1), 4–36. doi:10.3102/0002831212463813

Rothstein, J. (2010). Teacher quality in educational production: tracking, decay, and student achievement. *Quarterly Journal of Economics*, *25*(1), 175–214.

Rothstein, J. (2014). *Revisiting the Impacts of Teachers.* Retrieved from: http://eml.berkeley.edu/~jrothst/workingpapers/rothstein_cfr.pdf.

Sabel, C., Sazenian, A., Miettinen, R., Kristensen, P. & Hautamaki, J. (2010). Individualized service provision in the new welfare state: Lessons from special education in Finland. Report prepared for SITRA.

Saunders, W.M., Goldenberg, C.N. & Gallimore, R. (2009). Increasing achievement by focusing grade-level teams on improving classroom learning: A prospective, quasi-experimental study of Title I schools. *American Educational Research Journal, 46*(4), 1006-1033.

Savelsbergh, C.M.J.H., van der Heijden, B.I.J.M. & Poell, R.F. (2009). The development and empirical validation of a multidimensional measurement instrument for team learning behaviors. *Small Group Research, 40*(5), 578-607.

Scribner, J. P. (1999). Professional Development: Untangling the Influence of Work Context on Teacher Learning. *Educational Administration Quarterly*, *35*(2), 238–266. doi:10.1177/0013161X99352004

Sleegers, P. J. C., Stoel, R. D., & Kru, M. L. (2009). The effect of teacher and leadership factors on teachers ' professional learning in Dutch schools. *The Elementary School Journal*, *109*(4), 406–427.

Smith, T. M., Ingersoll, R. M., Smith, T. M., & Ingersoll, R. M. (2004). What Are the Effects of Induction and Mentoring on Beginning Teacher Turnover? *American Educational Research Journal*, *41*(3), 681–714. doi:10.3102/00028312041003681

Springer, M.G., Ballou, D., Hamilton, L., Le, V., Lockwood, J.R., McCaffrey, D., Pepper, M., and Stecher, B. (2012). Final Report: Experimental Evidence from the Project on Incentives in Teaching. Nashville, TN: National Center on Performance Incentives at Vanderbilt University

Springer, M.G., Pane, J.F., Le, V., McCaffrey, D.F., Burns, S.F., Hamilton, L.S. & Stecher, B. (2012). Team pay for performance: Experimental evidence from the Round Rock Pilot Project on team incentives. *Educational Evaluation and Policy Analysis, 34*(4): 367-390.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, *24*(3), 97–118. doi:10.1257/jep.24.3.97

Steiner, I.D. (1972). *Group Process and Productivity.* New York: Academic Press, Inc.

Strahan, D. (2003). Promoting a Collaborative Professional Culture in Three Elementary Schools That Have Beaten the Odds. *The Elementary School Journal*, *104*(2), 127–146.

Sun, M., Penuel, W. R., Frank, K. a., Gallagher, H. A., & Youngs, P. (2013). Shaping Professional Development to Promote the Diffusion of Instructional Expertise Among Teachers. *Educational Evaluation and Policy Analysis*, *35*(3), 344–369. doi:10.3102/0162373713482763

Sutter, M. (2005). Are four heads better than two? An experimental beauty-contest game with teams of different size. *Economics Letters*, *88*(1), 41–46. doi:10.1016/j.econlet.2004.12.024

Talbert, J.E. (2011) Collaborative inquiry to expand student success in New York City schools. In J.A. O'Day, C.S. Bitter, L.M. Gomez (Eds.), *Education Reform in New York City: Ambitious Change in the Nation's Most Complex School System.* (pp. 131-155). Cambridge, MA: Harvard Education Press.

Taylor, E.S. & Tyler, J.H. (2012). The effect of evaluation on teacher performance. *American Economic Review,* 102(7), 3268-3651.

Tournaki, E., Lyublinskaya, I., & Carolan, B. (2011). An Ongoing Professional Development Program and Its Impact on Teacher Effectiveness. *The Teacher Educator*, *46*(4), 299–315. doi:10.1080/08878730.2011.604711

Value-added Research Center (2010). *Technical Report on the NYC Value-Added Model.* Wisconsin Center for Education Research: University of Wisconsin-Madison.

Van der Sijde, P. C. (1989). The effect of a brief teacher training on student achievement. *Teaching and Teacher Education*, *5*(4), 303–314.

Wayman, J. C., Midgley, S., & Stringfield, S. (2006). Leadership for Data-Based Decision-Making: Collaborative Educator Teams. In *American Educational Research Association Annual Meeting* (pp. 1–16).

Yin, R. (2014). *Case Study Research: Design and Methods.* Thousand Oaks, CA: Sage Publications.

Yoon, K. S., Duncan, T., Lee, S. W., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (pp. 1–55).

You, Y. (2012). *Evaluating the Effect of New-Teacher Induction Programs on Teacher Turnover.* (Doctoral dissertation). Retrieved from Columbia University Academic Commons.

APPENDIX A. ADDITIONAL INSTRUMENTAL VARIABLE DIAGNOSTICS AND RESULTS

Columns 1 and 2 of Table A-1 report regressions of preexisting variables, including prior test scores and the percentage of students in the school receiving free or reduced price lunch, on the number of class sections. There should not be a relationship between these variables if the instrument is valid. There is no relationship between the number of class sections and prior test scores, and while the relationship with free and reduced price lunch is statistically significant, it is substantively very small, indicating that unlike with elementary and middle schools, the exclusion restriction assumption may be valid for high schools. On the other hand, column 3 reports a test of the monotonicity assumption by restricting the sample for the two-stage least squares estimate of the effect of teamwork on test scores to those subjects with fewer than 8 class sections. The results are not statistically significant, but they do change sign from the results for the full sample, indicating that the monotonicity assumption may be violated in this case.

TABLE A- 1 INSTRUMENTAL VARIABLES DIAGNOSTICS, HIGH SCHOOLS, 2008-2009

	(1) Previous score	(2) % Free/Reduced Lunch	(3) Gain scores
Number of class sections	0.000222	-0.00619***	
	(0.000958)	(0.000243)	
Team			-0.0396
			(0.175)
Constant	2.230***	0.790***	0.00708
	(0.0123)	(0.00293)	(0.0367)
Demographic Covariates			X
Observations	5321	6672	3470

Standard errors, clustered at school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A-2 presents the results of a variety of diagnostic tests and robustness checks for the instrumental variables specification. Columns 1 and 2 provide correlations between the instrument and preexisting variables, which are statistically significant in both cases, although once again, the association with free and reduced price lunch is substantively small. This calls the exclusion restriction into question for this instrument. Columns 3 and 4 provide instrumental variables estimates for sub-samples with fewer class sections as a test of the monotonicity assumption. Restricting the sample to grades with fewer than 8 class sections, in column 3, does not appreciably alter the results, but restricting the sample to elementary schools, as shown in column 4, produces results that are large, positive, and not statistically significant. On the other hand, columns 5 and 6 show the OLS results restricting the sample to elementary school and middle school, respectively; the elementary school results are very similar to the pooled OLS results, although the middle school results become very close to 0. Overall, it seems that the OLS results are more robust across specifications and sub-samples, indicating potential issues with the instrument.

TABLE A- 2 INSTRUMENTAL VARIABLES DIAGNOSTICS, K-8 SCHOOLS, 2008-2009

	(1) Prior score	(2) % Free/Reduced Lunch	(3) Gain scores	(4) Gain scores	(5) Gain scores	(6) Gain scores
Number of class sections	-0.350***	-0.00119***				
	(0.0438)	(0.000273)				
Team			-3.633	11.19	1.780***	0.0310
			(8.574)	(24.81)	(0.181)	(0.157)
Constant	664.0***	0.865***	-0.909	-2.606	-1.431**	-2.280***
	(0.241)	(0.00147)	(1.066)	(3.406)	(0.503)	(0.510)
Demographic Covariates			X	X	X	X
Observations	24845	67801	20681	14885	15450	8221

Standard errors, clustered at school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A-3 presents instrumental variables diagnostics for the 2009-2010 high school outcomes. This year, the association with both prior variables is statistically significant, although it is once again substantively small. When restricting the sample to subjects and schools with fewer than 8 class sections to test the monotonicity assumption, the point estimate remains statistically insignificant, although the sign changes, indicating possible violations of the monotonicity assumption.

TABLE A- 3 INSTRUMENTAL VARIABLES DIAGNOSTICS, HIGH SCHOOL, 2009-2010

	(1) Prior score	(2) % Free/Reduced Lunch	(3) Gain scores
Number of class sections	0.00207^{*}	-0.00385^{***}	
	(0.000925)	(0.000197)	
Team			-0.0546
			(0.101)
Constant	2.232^{***}	0.833^{***}	0.0791^{*}
	(0.0105)	(0.00213)	(0.0375)
Demographic Covariates			X
Observations	6880	8504	4764

Standard errors, clustered at school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A-4 reports K-8 OLS results for the 2009-2010 school year. Columns 1 and 2 report the association between teams and gain scores and find very similar results to 2008-2009. Columns 3-5 report various specifications of the association between teamwork and value-added as a check on the value-added estimates in Model 1. The coefficient of interest is on whether there is a team at the grade level; results are qualitatively similar to those for 2008-2009.

TABLE A- 4 OLS ESTIMATES OF THE ASSOCIATION BETWEEN TEAMWORK AND TEST SCORE OUTCOMES, K-8 SCHOOLS, 2009-2010

	(1)	(2)	(3)	(4)	(5)
	Growth	Growth	Multi-year VA	VA	VA
Team at grade level	1.697***	1.708***	1.732	0.00927	0.00878
	(0.234)	(0.232)	(0.971)	(0.00629)	(0.00626)
2006-2007 VA Score				0.706***	0.682***
				(0.136)	(0.126)
Team SD - Prior VA				0.0951	0.0935
				(0.111)	(0.112)
Team*Prior VA					0.0839
					(0.140)
Constant	0.0178	-1.758**	50.42***	0.0251	0.0257
	(0.131)	(0.622)	(4.406)	(0.0349)	(0.0348)
Demographic Covariates		X	X	X	X
Observations	23671	23671	24570	18152	18152

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The instrumental variables results for K-8 schools in 2009-2010 are presented in Table A-5. Although the OLS results are quite consistent from 2008-2009 to 2009-2010, the IV results shift from large and positive but statistically insignificant to negative and statistically significant. Columns 1, 2, 5, and 6 present the first stage, reduced form, and 2SLS results with and without controls, respectively, with the linear specification and columns 3, 4, and 7 present the results for the quadratic specification. Once again, there is likely a weak instrument problem, as the F-statistics for the first stages are 4.28 and 5.02, respectively. The Hansen J-statistic, however, is 0.80, below the critical value for rejecting the null hypothesis that one of the instruments is valid assuming that the other is valid.

TABLE A- 5 INSTRUMENTAL VARIABLES RESULTS, EFFECTS OF TEAMWORK ON GAIN SCORES, K-8 SCHOOLS, 2009-2010

	(1) First stage (Team)	(2) Reduced form (Growth)	(3) First stage (Team)	(4) Reduced form (Growth)	(5) 2SLS (Linear)	(6) 2SLS (Linear)	(7) 2SLS (Quadratic)
Classes at grade or subject	0.00871^{***}	-0.0843^{**}	-0.00269	-0.0343			
	(0.00242)	(0.0318)	(0.00630)	(0.105)			
Number of classes squared			0.000691	-0.00297			
			(0.000434)	(0.00505)			
Team					-10.86^{*}	-8.662^{*}	-6.653^{*}
					(4.655)	(4.085)	(3.092)
Constant	-0.00529	-0.920	0.0261	-1.071	2.017^{**}	-0.751	-0.927
	(0.0242)	(0.702)	(0.0262)	(0.773)	(0.753)	(0.802)	(0.754)
Demographic Covariates	X	X	X	X		X	X
Observations	67801	22712	67801	22712	22712	22712	22712

Standard errors, clustered at school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Finally, high school outcomes for 2009-2010 are presented in Tables A-6 and A-7. There are small positive associations between teams and gain scores in this year, as well as teams and graduation rates, although all are substantively small and none is significant at the 0.05 level. The instrument is once again weak, with first-stage F-statistics of 4.76 and 3.94 when using the linear and quadratic specifications, respectively. Effects on test scores appear to be quite small and not statistically significant in the instrumental variables model. The Hansen J-statistic for this model is larger, at 1.773, but still smaller than the critical value to reject the null hypothesis, again supporting the exclusion restriction assuming that at least one of the instruments is valid.

TABLE A- 6 OLS ESTIMATES OF ASSOCIATION BETWEEN TEAMWORK AND HIGH SCHOOL OUTCOMES, 2009-2010

	(1) Growth	(2) Growth	(3) Graduation rate	(4) Graduation rate
Team at grade level	0.0205	0.0198	0.0341	0.0276
	(0.0123)	(0.0122)	(0.0205)	(0.0199)
Constant	0.0255***	0.0712**	0.572***	0.873***
	(0.00537)	(0.0246)	(0.0121)	(0.0708)
Demographic Covariates		X		X
Observations	11516	11476	3519	3513

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE A- 7 INSTRUMENTAL VARIABLES ESTIMATES OF ASSOCIATION BETWEEN TEAMWORK AND HIGH SCHOOL OUTCOMES, 2009-2010

Panel A. Gain scores

	(1) First stage (Team)	(2) Reduced form (Growth)	(3) First stage (Team)	(4) Reduced form (Growth)	(5) 2SLS (Linear)	(6) 2SLS (Quadratic)
Classes at grade or subject	0.00774***	0.000107	0.0131***	0.00130		
	(0.00121)	(0.000457)	(0.00279)	(0.00111)		
Number of classes squared			-0.000126*	-0.0000275		
			(0.0000556)	(0.0000190)		
Team					0.0350	0.0135
					(0.0608)	(0.0577)
Constant	0.0284	0.0868**	0.000682	0.0802*	0.0819*	0.0862**
	(0.0605)	(0.0308)	(0.0631)	(0.0319)	(0.0322)	(0.0317)
Demographic Covariates	X	X	X	X	X	X
Observations	8504	6415	8504	6415	6415	6415

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Panel B. Graduation rates.

	(7) First stage (Team)	(8) Reduced form (Graduation rate)	(9) First stage (Team)	(10) Reduced form (Graduation rate)	(11) 2SLS (Linear)	(12) 2SLS (Quadratic)
Classes at grade or subject	0.00774***	-0.00254	0.0131***	-0.00729*		
	(0.00121)	(0.00241)	(0.00279)	(0.00339)		
Number of classes squared			-0.000126*	0.000152**		
			(0.0000556)	(0.0000546)		

Team					-0.583	-0.649
					(0.365)	(0.701)
Constant	0.0284	0.906***	0.000682	0.918***	0.915***	0.920***
	(0.0605)	(0.0804)	(0.0631)	(0.0806)	(0.0816)	(0.106)
Demographic Covariates	X	X	X	X	X	X
	8504	1582	8504	1582	1582	1582

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

APPENDIX B. ADDITIONAL HETEROGENEITY RESULTS

2007-2008

TABLE B- 1 RETENTION – 100 TEAMS CODED FOR QUALITY, ALL TEACHERS

	(1) Still teaching next year	(2) Years teaching	(3) Years teaching	(4) Still teaching next year	(5) Years teaching
Overall	0.00248	0.0184	0.0179	0.00280	0.0179
	(0.00457)	(0.0276)	(0.0256)	(0.00441)	(0.0256)
Constant	0.954^{***}	4.558^{***}	4.808^{***}	0.980^{***}	4.808^{***}
	(0.00288)	(0.0167)	(0.0679)	(0.0114)	(0.0679)
Observations	7374	7355	7153	7172	7153

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B- 2 RETENTION 100 TEAMS, FIRST YEAR TEACHERS

	(1) Still teaching next year	(2) Years teaching	(3) Years teaching	(4) Still teaching next year	(5) Years teaching
Overall	0.0251^{**}	0.150^{*}	0.179^{**}	0.0317^{***}	0.179^{**}
	(0.00870)	(0.0740)	(0.0674)	(0.00809)	(0.0674)
Constant	0.936^{***}	3.434^{***}	3.804^{***}	0.958^{***}	3.804^{***}
	(0.00942)	(0.0508)	(0.220)	(0.0439)	(0.220)
Observations	768	768	728	728	728

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B- 3, VALUE-ADDED, 100 HAND-CODED TEAMS, ALL TEACHERS

	(1) Value added	(2) VA Percentile Multi-year	(3) VA Percentile Lowest 3rd	(4) VA Percentile ELL
Overall	-0.000536	0.787	0.600	-0.0454
	(0.00456)	(0.797)	(1.012)	(2.123)
Constant	0.00973	48.66^{***}	38.96^{***}	45.31^{***}
	(0.0119)	(1.999)	(2.671)	(5.737)
Observations	8064	7633	4723	1065

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B- 4, VALUE-ADDED, 100 HAND-CODED TEAMS, FIRST-YEAR TEACHERS

	(1)	(2)	(3)	(4)
	Value added	VA Percentile	VA Percentile	VA Percentile ELL
		Multi-year	Lowest 3rd	
Overall	0.00879	1.953	-3.005	10.18
	(0.0105)	(2.011)	(4.446)	(13.16)
Constant	-0.0212	52.23***	41.66***	57.52*
	(0.0256)	(4.981)	(9.699)	(26.43)
Observations	728	686	312	70

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B- 5, RETENTION, FIRST-YEAR TEACHERS, HAND-CODED QUALITY AS COVARIATE

	(1)	(2)	(3)	(4)	(5)
	Still teaching	Years teaching	Years teaching	Still teaching	Years teaching
	next year			next year	
Team indicator	-0.109	-1.374	-1.304	-0.0951	-1.304
	(0.166)	(1.391)	(1.283)	(0.152)	(1.283)
Overall	0.0817	0.865	0.857	0.0811	0.857
	(0.0799)	(0.680)	(0.634)	(0.0743)	(0.634)
Constant	0.936***	3.436***	3.797***	0.958***	3.797***
	(0.00940)	(0.0507)	(0.217)	(0.0437)	(0.217)
Observations	768	768	728	728	728

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B- 6, VALUE-ADDED, HAND-CODED QUALITY AS COVARIATE, ALL TEACHERS

	(1)	(2)	(3)	(4)
	VA	VA Percentile	VA Percentile	VA Percentile ELL
		Multi-year	Lowest 3rd	
Team indicator	-0.0189	0.976	6.211	2.553
	(0.0507)	(8.788)	(10.46)	(25.11)
Overall	0.00915	0.287	-2.644	-1.363
	(0.0265)	(4.583)	(5.579)	(12.76)
Constant	0.00959	48.67***	39.04***	45.35***
	(0.0120)	(1.999)	(2.666)	(5.729)
Observations	8064	7633	4723	1065

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B- 7, VALUE-ADDED, FIRST-YEAR TEACHERS, HAND-CODED QUALITY AS COVARIATE

	(1)	(2)	(3)	(4)
	VA	VA Percentile	VA Percentile	VA Percentile ELL
		Multi-year	Lowest 3rd	
Team indicator	0.0755	12.99	15.51	36.70
	(0.155)	(30.26)	(47.70)	(127.3)
Overall	-0.0304	-4.821	-11.85	-11.20
	(0.0785)	(15.10)	(26.52)	(78.42)
Constant	-0.0208	52.32***	42.03***	57.71*
	(0.0256)	(4.990)	(9.867)	(26.69)
Observations	728	686	312	70

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2008-2009

TABLE B- 8, IV ESTIMATES, TEAM INDICATOR ON GROWTH, K-8

	(1)	(2)	(3)
	growth	growth	growth
Team indicator	18.81		
	(10.12)		
Has a goal		18.34	
		(9.956)	
Overall			-35.44
			(40.70)
Constant	2.226*	2.448*	4.648**
	(1.103)	(1.008)	(1.534)
Observations	22903	22903	8751

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B- 9, IV ESTIMATES, TEAM INDICATOR ON GROWTH, HIGH SCHOOL

	(1)	(2)	(3)
	growth	growth	growth
Team indicator	0.302		
	(0.163)		
Has a goal		0.395	
		(0.231)	
Overall			1.637
			(2.620)
Constant	0.00987	0.00343	0.0361
	(0.0286)	(0.0322)	(0.0350)
Observations	4953	4953	2676

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2009-2010

TABLE B- 10 IV ESTIMATES, TEAM INDICATOR ON GROWTH, K-8

	(1) Growth	(2) Growth	(3) Growth
Team indicator	-6.653*		
	(3.092)		
Has a goal		91.47	
		(139.2)	
Overall			-46.02
			(31.75)
Constant	-0.927	-5.626	1.789
	(0.754)	(6.663)	(2.644)
Observations	22712	22712	9752

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B- 11, IV ESTIMATES, TEAM INDICATOR ON GROWTH, HS

	(1) Growth	(2) Growth	(3) Growth
Team indicator	0.0350		
	(0.0608)		
Has a goal		0.0636	
		(0.129)	
Overall			0.932
			(1.436)
Constant	0.0819*	0.0872**	-0.00757
	(0.0322)	(0.0298)	(0.114)
Observations	6415	6415	2175

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

INSTRUMENTAL VARIABLES QUALITY ESTIMATES

One possibly exogenous source of variation in treatment intensity is partially random variation in accountability pressure schools faced due to the staggered nature of the school accountability system. In 2007-2008, to complement quantitative score report cards that were primarily based on student test score growth at the elementary and middle school levels and test scores, credit accumulation, and graduation rates at the high school level with a more holistic, qualitative assessment of the school learning environment to give schools earlier and more actionable feedback on leading indicators of student learning, the school district implemented a qualitative review system. Under this system, district officials and external reviewers would engage in a 2-day site visit to observe classroom practice, meet with administrators, teachers, students, and parents, and observe the inputs and processes that were contributing to or impeding student learning. All schools received a review in 2007-2008, but due to the resource-intensive nature of this system, reviews were staggered in subsequent years. Schools that were deemed to be low-performing for various reasons, including poor performance on quantitative accountability measures and low scores on prior qualitative accountability measures, were scheduled to receive reviews more often. For schools that were average-performing or better, a random sub-set of approximately one-third of schools were selected to be reviewed each year starting in the 2008-2009 school year.

Therefore, there is between-school variation with a random component in the degree of accountability pressure schools faced. Even though this pressure was only directly felt during a 2-day review, it could potentially have reverberating effects throughout the school year. Schools may have implemented inquiry teams with greater fidelity in anticipation of the review, of which evaluation of the school's inquiry work was a part. Further, even if schools received reviews earlier in the school year, there may be cascading effects by which efforts to organize teams early in the year in preparation for the review could pay dividends in terms of stronger teamwork throughout the year.

This association between accountability pressure induced by being selected to receive a qualitative evaluation and teamwork sets up an instrumental variables approach, using a 2SLS framework:

(8) $Quality_{jst} = \beta_0 + \beta_1 QR_{st} + X_{st}\beta_2 + \beta_3 Year_t + \varepsilon_{jst}$(first-stage)

(9) $Y_{jst} = \gamma_0 + \gamma_1 \widehat{Quality}_{jst} + X_{st}\gamma_2 + \gamma_3 Year_t + \mu_{jst}$(second-stage)

where j indexes teams at a grade-subject-subgroup level, s indexes schools, and t indexes time. $QR_{st}$is a dichotomous variable indicating whether school s received a qualitative evaluation at time t, $X_{st}$is a vector of school-level controls, and $Year_t$ represents year fixed effects. The effects of team quality on a vector of outcomes, Y_{jst}, will therefore be assessed based on predicted quality based on accountability pressure. For γ_1, the coefficient of interest, to be a valid causal estimate of the effects on team quality on outcomes, several assumptions must be met. Most notably in this case, $cov(QR_{st}, Quality_{jst}) \neq 0$, meaning that the instrument is not weak, and the instrument must only affect the outcome through the channel of enhancing the quality of teamwork, or the exclusion restriction. Given the broad-based nature of the qualitative evaluations, it is quite possible that the exclusion restriction is violated in this case, depending upon the outcome in question, but given the centrality of inquiry teams to the school district's strategy at this time, this assumption may be valid for at least some outcomes.

Results of this model are presented in Table B-12. Somewhat surprisingly, greater accountability pressure is negatively associated with the two selected quality measures – whether a team has a goal and the hand-coded measure – although in neither case is the relationship statistically significant and in both cases the F-statistic is quite small, indicating a weak instrument. In the reduced form, being subject to qualitative accountability pressure is associated with lower growth overall, making the 2SLS results positive but statistically insignificant.

TABLE B- 12. ASSOCIATION BETWEEN INDICATORS OF TEAM QUALITY AND RETENTION, FIRST-YEAR TEACHERS.

	(1)	(2)	(3)	(4)	(5)
	First stage – Has a goal	First stage - Quality	Reduced form	2SLS – Growth	2SLS – Growth
Qual. Account.	-0.0082	-0.0022	-0.0161		
	(0.0049)	(0.0075)	(0.1723)		
Has goal				1.9544	
				(20.856)	
Year	-0.0119**	0.0057	-5.7768***	-5.754***	-5.808***
	(.0044)	(0.0063)	(.1701)		
				(.3209)	(1.653)
Overall					70.581
					(270.982)
Constant	24.0478***	-11.446	11607.48***	11560.48***	11667.98***
	(8.8766)	(12.7309)	(341.8101)	(645.58)	(3314.13)
Observations	66,706	24,373	66,706	66,706	24,373

Standard errors, clustered at the school level, in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

APPENDIX C. QUALITATIVE CODING SCHEME

TABLE C- 1 CODING SCHEME

Code	Description/Meaning	Example (if applicable)	Source
Formal communication	Planned, structured communication through specified channels, e.g., memos	"So typically when we have the data in pre-meeting, it's the grade representative's job to kind of bring it back to our team meeting. And then we discuss how we want to, you know, we analyze our grades data, you know, amongst ourselves and then decide what we need to do to improve the scores."	Hoegl and Gemuenden (2001)
Informal communication	Spontaneous (e.g., chats in the hallway) as opposed to planned (e.g., status memos) communication.	"The other thing I was thinking about, from a standpoint of collaboration, it could force teachers. For instance, I have this--my big pet peeve, as I said it last time when (name redacted) was here, that what we need to do is get past--knowing where to find the data and using the data, inform where we are with each kid and inform instruction, but we need to get together in groups, common prep groups, the ELA guy, the math guy, the social studies guy and the science guy, getting together about those kids in 701. And making sure that we don't just talk the talk, but we're really looking at an integrated curriculum"	Hoegl and Gemuenden (2001)
Balance of contributions	Respect for different ideas,	"And then we decide,	Hoegl and Gemuenden

	perspectives, areas of expertise; shared effort; specialization as opposed to diffusion of responsibility	you know, which is the best and We give them an incentive. Something. But then we're the ones that are making the-- those decisions instead of burdening the teachers. They will have them write in their class. And then they will have to select the best five. But then we will select-- every month, we have a winner in social studies, ELA…”	(2001)
Connecting student actions to teacher actions	Examples of “root cause” analysis – observing behaviors and trends and tracing backward, causally	“I think that a strength is these statements were lovely. Because I did a good job modeling them.”	From literature on *kaizen*, Toyota production model, and continuous improvement organizations; Sabel et al. on Finland
Consistency	Focus on interdisciplinary connections and consistency in instructional strategies, terminology, etc., across classes,	“I think that between all of us, we can do this, you know? 'Cause they're not only gonna see it your class. They'll see it in mine—“	Inductively derived from data and experience
Adaptation	Focus on adapting to needs of particular subject areas and students	“How 'bout if we turn those into subjects? So it would be the word and then subject. Social studies. (Name) got an 80. But in ELA, got a 70. Or blah, blah, blah, got a 90. So those-- instead of those being just sort of like so ELA driven because all of that is ELA, change it into a subject. So the student understands the particular word, the meaning of the word in this subject.”	Inductively derived from data and experience
Experimenting, feedback	Group is oriented toward trying out new ideas, open to learning, receiving	“See? Yeah, we have to have something that we can collect so that we	Kasl, et al. (1997)

	feedback	can then look at it and say, 'Okay, why are these students getting it and these students over here aren't?' And then we can go to the teacher and see what practices are being done, and learn from each other."	
Focus on individual student	Sustained discussion of one particular student, or a group of students, especially if focused on learning needs	"He's been-- yeah. And his paragraph is up on the bulletin board. And he's like, "Did you notice my home chemistry lab?" I saw-- you know, he's very-- he really want-- and you know what? What I think we need to work with Oscar is motivating him. That, like, he second-- he second guesses himself. Maybe that might be the alpha in him. But he does, like, he thinks-- like, "Can you double check this? 'Cause-- are you sure it's okay?" I mean, it's like—"	Inductively derived from data and experience
Instruction/Instructional strategies	Topical codes focused on what team is discussing – content area, skills, particular strategy	"702, if you're working with 702, the L's, they're doin' the quick outline, just to let you know. On-- the Middle Ages, which we started yesterday. And Adage and I are gonna continue working with them. So that way if you're meeting with Manny or any of them, you can work with the quick outlines with them. And they're doin' it on the Black Death: The Bubonic Plague."	From basic model of intervention; description of the content of team discussions
Organization/Structure –	Regular meetings over a	"Let's be mindful of the	Darling-Hammond

frequency, duration, sustained	long enough period of time	time. It's… we need to get ready for next week, and how we make use of next week…"	(2009)
Shared values, norms, collaborative culture	A common, basic set of beliefs, in particular about student learning, the mission of the school, and the purpose of the team	"So this was his baby and I thought it made sense overall. I don't think everybody really bought into it. All the teachers didn't buy into it."	Little (1982)
Courage – willingness to change/adapt, willingness to make teaching public	General openness, willingness to try new things, adjust practice. There is much literature on teaching being private, and an essential element of teacher collaboration being making it "public"	'Thanks for-- sorry. I know it's hard to change your teaching practice. FEMALE VOICE: No, it's okay. FEMALE VOICE: I mean, you've done it for so long. FEMALE VOICE: No, no. I mean, like-- listen, it's-- it's all about applying things and learning, you know?"	Kelchtermanns (2006); Little (1982)
Framing, exploration	Problem definition and redefinition; willingness and ability to see problems in new light	"I also wanna go back to what Russo was saying earlier. We also have the-- it's-- I don't think this team-- maybe I'm wrong, but part of this team's job is to analyze data, but we also have to support teachers"	Savelsbergh, et al. (2009)
Peer monitoring, peer pressure	Encouragement by peers to successfully complete work tasks with high quality; implicit or explicit social pressure to perform	"FEMALE VOICE: Just say I'll do it. I meant to (but) caught up with something else. FEMALE VOICE: By the-- complete by-- by the weekend? FEMALE VOICE: Yes."	Mas and Moretti (2006); Kandel and Lazear (2012)
Examining student learning	Rooting work in student	"Even, like, when--	Chong and Kong (2012)

needs	learning; making inferences about student learning based upon data and analysis of work	when they were drafting off of the outlines. Like, a lot of them were, like, "Okay, what do I do?" Like, they didn't know where to put the thesis.”	
Analyzing data	Rooting judgments in empirical analysis; seeking evidence to support theories and assertions; grounding work in evidence on student achievement	“We continued the Right to Learn (sic) assessment today in 701. So the scores are all different now. Like, they went up, pretty much all of them.”	Collaborative Inquiry Process; Sabel et al.
Reflection and revision	Making changes to team process, or instructional strategies adopted by team, based upon ongoing feedback loops	“But ne-- now it seems that we're at a point where the processes and systems are in place and we really have to-- we-- we really do have to work as a coordinated team. I'll just use an example. We have a kid like—(name redacted), who's a very bright kid. But he has strengths and weaknesses in everyone would call content areas before. And I don't think, from my view, that I am doing a very good job of articulating to the rest of my group and-- nor are they with me, what that kid's plan, his education plan, should be in each one of those classes. I don't-- I think we just kind of understand in a nebulous way what the school wants, but we're not really, you know, getting down and making it really happen day-to-day. It takes a lot of planning. Probably gets back to my original point that we need to get	From basic model of intervention; description of the content of team discussions

		together and really hash it out.”	
Focus	Tradeoff between focus on “big” things in a superficial way or smaller things in a deep way	“What about organization for the body? FEMALE VOICE: We just focused on the introduction and conclusion. (LAUGHTER) But that will be-- FEMALE VOICE: Another time.”	Inductively derived from data; partially based on description of intervention by Talbert (2010)
Taking advantage of resources	Utilizing outside resources, such as curricula, professional development, technology, etc., to help address learning needs identified by team, of students and/or of teachers	“It just made so much sense when they presented it in the workshop. It was kinda like, "Duh, that's an easy, like, formulaic way." I don't know if that's a word.”	CPRE (2008, 2010)
Leadership/support	Support by school leadership, either in direct participation, providing resources, providing time for team to meet, not interfering with inquiry process	“--staff to work together, you-- you know? That as a team, you bring it to the and we listen to-- because if we want buy-in, they have to be a part of it. It's-- again, we're gonna throw this at them and say, "Look, this is what you do." … But if we say, "This is what we're brainstorming, and we need some input, you know? How can we do it together?" I think that would be more beneficial.”	Severa – Lee, Zhang, Yin (2011)
Individual learning	Through team process an individual learns – e.g., knowledge-sharing, inquiry teams as PD	“And you know, when we look at the data together and we say, "My kids are having trouble with equivalent fractions. Like, how do you, you know, approach that topic?	Inductively derived from data and experience

		How do you?" And then it's good for us to kinda talk it out and share things. And, I mean, not that we don't collaborate to begin with, but it gives us, you know, a more clear focus of, you know, something I'm doing is not working. And, like, let's talk about how we can help each other to come up with new ways to teach things."	
Team learning	Team itself learns – e.g., abduction	"It's kind of-- it's kind of a good way to reflect and see, like-- like, a good-- you know, what are our do's and don'ts."	Buysse, Sparkman, Wesley (2003)
Reproduction cycle – passing on teamwork; institutionalizing teamwork	Sharing learning about *team process* itself with the school, with other teachers/teams, with new staff, etc.	"Right. It has to be something that those teachers also bring. I don't want it to just be-- the data team looking at this. I would like to see other teachers use it as well."	Literature on Professional Learning Communities
Organizational learning – systematizing learning from teamwork	Sharing learning *about student and/or teacher learning* ***from*** the team process with other teachers/teams, other schools	"Just everything that we talk about in here, and then it translates to the whole grade. And a lot of our things have been taken to the whole school."	Crossan, Lane, and White (1999)
Attitudes about conflict – openness	Willingness of team members to productively engage difficult topics, broach areas of disagreement, vs. focus on maintaining cordiality	"You guys are drivin' me nuts here."	Kelchtermans (2006); Achinstein (2002)
Non-pecuniary benefits – compensating differentials of teamwork	Evidence that team members value teamwork, see it as a workplace amenity that may enhance their jobs, skills, entice them to remain teaching	"We're doing something good—"	Hamilton et al. (2003)
Research skills – asking good	Evidence that teachers	"All right, my negative,	Fernandez (2002)

questions, gathering useful data, analyzing/interpreting results	have the necessary skills to successfully engage in the inquiry process	I see that they need organization. Their thoughts are all over the place. They're there, but now how do we organize them to say, "This is what goes with--" I mean, you know, they're there. Again, they just need organization now.”	