View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by IssueLab

An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) After Four Years

Final Report

March 7, 2012

Steven Glazerman
Allison Seifullah

MATHEMATICA
Policy Research

This page has been left blank for double-sided copying.

CAPE Reference Number:
2010-001-01

Mathematica Reference Number:
06736-520

Submitted to:
The Joyce Foundation
70 West Madison Street, Suite 2750
Chicago, IL 60602-4317
Project Officer: John Luczak

Submitted by:
Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Allison Seifullah

An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) After Four Years

Final Report

March 7, 2012

Steven Glazerman
Allison Seifullah

MATHEMATICA
Policy Research

This page has been left blank for double-sided copying.

ACKNOWLEDGMENTS

This study is the product of many people's efforts. At Chicago Public Schools, Ann Chan, Laurel Crown, Sylvia Flowers, Jessica Foster, Maria Rizzetto, Victoria Van-Nguyen, and other Chicago TAP staff provided data, offered useful feedback on earlier presentations and drafts, and cooperated with the study in many ways. John Luczak at the Joyce Foundation offered useful guidance.

At Mathematica Policy Research, Nancy Carey led the administration of the teacher survey, with Ellen Siegel and Katherine Burnett helping to manage the data collection activities. Jeremy Page, Sarah Prenovitz, and Dominic Harris provided expert assistance cleaning and analyzing data, with Ji-Hyeun Kwon-Min also supporting the data analysis. Philip Gleason, Eric Isenberg, and Allen Schirm provided helpful comments on the study design and carefully reviewed and commented on drafts of the report. The report was prepared for publication by Donna Dorsey and edited by Anne Himmelfarb.

This page has been left blank for double-sided copying.

CONTENTS

III *(continued)*

TABLES

FIGURES

This page has been left blank for double-sided copying.

EXECUTIVE SUMMARY

In 2007, using funds from the federal Teacher Incentive Fund (TIF) and private foundations, the Chicago Public Schools (CPS) began piloting its version of a schoolwide reform model called the Teacher Advancement Program (TAP). Under the TAP model, teachers can earn extra pay and take on increased responsibilities through promotion (to mentor teacher or master teacher), and they become eligible for annual performance bonuses based on a combination of their contribution to student achievement (known as "value added") and observed performance in the classroom. The model calls for weekly meetings of teachers and mentors ("cluster groups"), and regular classroom observations by a school leadership team to help teachers meet their performance goals. The idea behind TAP is that giving teachers performance incentives, along with tools to track their performance and improve instruction, will help schools attract and retain talented teachers and help all teachers raise student achievement.

This report is the last in a series of reports providing evidence on the impacts of CPS' version of TAP, called "Chicago TAP." It presents findings from the four-year implementation period, with special emphasis on the 2009–2010 and 2010–2011 school years, the third and fourth years of the program's rollout in Chicago. Earlier reports (Glazerman et al. 2009; Glazerman and Seifullah 2010) provide detailed data on each of the first two years of the program, respectively. CPS implemented Chicago TAP as a pilot program intended for 40 high-need schools. The program began in 10 schools in the first year (cohort 1) with a rollout plan to add 10 more Chicago TAP schools (cohorts 2, 3, and 4) in each year of the TIF grant's four-year implementation period.

We address three research questions regarding Chicago TAP:

1. How was the program implemented?
2. What impact did the program have on student achievement?
3. What impact did the program have on teacher retention within schools?

The first question, about implementation, is needed to understand the answers to the next two, about impacts. We go on to address subquestions within each of these areas. For example, we explore whether impacts grow larger over time or as schools gain more experience with Chicago TAP.

Chicago TAP was originally based on a national TAP model developed by the Milken Family Foundation in the late 1990s, but it includes some local adaptations. For instance, Chicago TAP determines the compensation given to teachers, and it offers performance pay for principals and other school staff in addition to teachers. Teacher performance was measured using a Skills, Knowledge and Responsibility (SKR) observation rubric, but unlike with the national model, value added performance was not measured at the individual teacher level in Chicago. It was initially measured at the school level and in its final two years, at the school-grade team level. On the basis of these performance measures, teachers in our data received an average bonus of approximately $1,100 in the first three years of district rollout (2007–2008, 2008–2009, and 2009–2010), rising to $1,400 for new Chicago TAP implementers in 2010–2011, with a maximum payout in these schools of less than $2,700 in any year. For a school's second and third years of implementation, the average payout was approximately $2,500, with a maximum payout of $6,400, although in the fourth year of implementation the average payout was lower, about $1,900 with a maximum of less than $4,600.

Teachers who were selected to serve as mentor teachers received an additional salary augmentation of $7,000, and lead (master) teachers received $15,000 with a possibility for $20,000 in the final year for teachers designated as "lead plus."

Study Design and Data

Our approach to estimating the impacts of Chicago TAP is based on a hybrid study design that relies on both the random assignment of schools to year of implementation ("experimental design") and the careful matching of Chicago TAP schools to non-TAP schools in the district ("quasi-experimental design"). The assignment of Chicago TAP schools to implementation year occurred in two rounds. In spring 2007, 16 noncharter elementary (K–8) schools, 2 charter elementary schools, and 2 high schools voluntarily applied for Chicago TAP and successfully completed the selection process. We used a lottery to determine the implementation year for the noncharter elementary schools, randomly assigning 8 to a treatment group that began Chicago TAP in fall 2007 (cohort 1) and 8 to a control group that began Chicago TAP in fall 2008 (cohort 2); CPS purposively assigned the charter and high schools. In spring 2009, Chicago TAP staff selected a second batch of 20 schools from a new set of applicants and also replaced 2 schools that had exited the program. We again used a lottery to determine the program start date for 16 noncharter elementary schools and the 2 charter schools, randomly assigning 9 to begin Chicago TAP in fall 2009 (cohort 3) and the other 9 to begin Chicago TAP in fall 2010 (cohort 4).

The experimental strategy has the advantage that any systematic differences in outcomes between cohorts 1 and 2 and between cohorts 3 and 4 can be attributed to the opportunity to implement one additional year of Chicago TAP. In the initial year, the design allows us to compare outcomes in schools that had been implementing Chicago TAP to a control group of schools that had not yet begun implementation, yielding an unbiased estimate of program impact. The drawbacks with this approach, however, are that we are limited to 34 schools and that our strategy for estimating impacts beyond the first year of implementation is indirect. In later years we can only estimate the impact of two versus one year of implementation, three versus two, and four versus three. This is because the control group was allowed to begin implementing Chicago TAP starting in the second year after randomization.

To complement the experimental analysis, we used propensity score matching procedures to form for each year of rollout a non-TAP comparison group that was not implementing the program. The term "propensity score" refers to the summary measure, a combination of several different variables, that captures the probability, or "propensity," that a school will be selected to implement Chicago TAP. By selecting non-TAP schools whose propensity scores are similar to those of the Chicago TAP schools, we hope to mimic an experiment; hence we call this design "quasi-experimental."

To conduct the quasi-experimental analysis we gathered administrative data on over 300 CPS schools that were not participating in Chicago TAP and identified those that were most closely matched to the Chicago TAP schools on preintervention characteristics such as size, school demographics, student achievement, and teacher retention measured prior to the Chicago TAP school selection. This quasi-experimental strategy does not offer the same protection against bias due to unobservable differences that the experimental strategy does. In comparing Chicago TAP schools to matched comparison schools, we can infer program impacts only if we assume that the observable characteristics used to match schools are similar and comprehensive enough that the remaining differences in outcomes can be attributed to Chicago TAP itself and not to unobserved

factors, such as a dynamic principal or an especially motivated teaching staff. Nevertheless, the matched comparison group can be much larger than the experimental control group.

Another advantage of the quasi-experimental strategy is that it enables us to compare Chicago TAP schools to a group of matched comparison schools that remain non-TAP schools throughout the study. For outcomes such as teacher retention that can be affected by knowledge of future implementation, the randomized control group is never a pure standard of comparison because staff in the control schools know that their schools are slated to begin Chicago TAP the following school year; thus, the control group is also affected by Chicago TAP, although less directly than the treatment group. Consequently, our analysis of teacher retention is based on the matched comparison group rather than the randomized control group. For the period of overlap—the first year of implementation—we were able to compare outcomes for the quasi-experimental comparison group to those of the experimental control group to reconcile the two methods and validate the more inclusive quasi-experimental methodology.

Data used for the study include teacher surveys and principal interviews as well as administrative data obtained from CPS. We administered a teacher questionnaire in spring 2008 and spring 2010 and interviewed principals in fall 2008. We obtained CPS student test score files covering all years beginning in 2006–2007 and teacher administrative records covering all years beginning in 2005–2006. CPS also provided us with the following data on Chicago TAP: teacher scores on the SKR classroom observation rubric performance, payouts by teacher, and scores on a program review that tells each Chicago TAP school how well they have been implementing the program over the current school year.

Findings

We present findings addressing each of the three research questions.

Program implementation. To assess the first year under Chicago TAP for schools that began the program in fall 2009 (cohort 3), we looked at how teacher development and compensation practices in Chicago TAP schools differ from practices normally implemented in CPS schools. We found that teachers in Chicago TAP schools reported receiving significantly more mentoring support than teachers in similar non-TAP (control) schools. This finding reflects the fact that under the Chicago TAP model, teachers are guided by mentor teachers, and cluster groups meet weekly. We also found that veteran teachers in Chicago TAP schools were more likely than their control group counterparts to provide mentoring support to their colleagues; this finding is consistent with the fact that under Chicago TAP, teachers have the opportunity to assume leadership roles and responsibilities as Chicago TAP mentor or lead teachers. Teachers in Chicago TAP schools (veteran and novice) were aware of their eligibility for performance-based compensation. We found that the amount of compensation they expected approached the amount that was eventually paid out; that is, the average expectation was about $900, and the actual amount paid out in bonuses to this group was an average of about $1,100 per teacher. We generally did not find evidence of an impact of Chicago TAP on teacher attitudes or school climate.

Program review scores provide evidence on the extent to which Chicago TAP schools implemented the program with fidelity to the national TAP model. Average scores were around 3 out of 5 for each cohort of schools in each of the first three years of district rollout. The National Institute for Excellence in Teaching (NIET), which oversees the TAP system nationally, concluded that elements of TAP had been introduced, but TAP implementation had not been "rigorous." After

the third year of rollout, NIET informed CPS that the district had not implemented the TAP system. In the final year of rollout, when CPS conducted the program reviews instead of NIET, the average scores were higher than 4 out of 5.

We found two salient areas in which program implementation did not occur as initially planned, both of which were related to performance-based compensation. Average performance-based payouts were smaller than the originally stated targets. In addition, the teacher-level value-added component of performance-based compensation that was originally scheduled to begin in each school's second year of program implementation was not implemented because the data that were needed to reliably link students and teachers were not available.

Impacts on student achievement. While the introduction of Chicago TAP led to real changes inside the schools, the program did not consistently raise student achievement as measured by growth in Illinois Standards Achievement Test (ISAT) scores. We found evidence of both positive and negative test score impacts in selected subjects, years, and cohorts of schools, but overall there was no detectable impact on math, reading, or science achievement that was robust to different methods of estimation. For example, impacts on science scores overall (across years and cohorts) were positive, but not statistically significant unless we used one particular matching method that excluded some Chicago TAP schools from the analysis.

Impacts on teacher retention. We did find evidence suggesting that Chicago TAP increased schools' retention of teachers, although the impacts were not uniform or universal across years, cohorts, and subgroups of teachers. We found that teachers who were working in Chicago TAP schools in 2007 returned in each of the following three years at higher rates than teachers in comparable non-TAP schools. For example, we found that 67 percent of classroom teachers in cohort 1 schools in fall 2007 returned to their same school in fall 2010 compared to about 56 percent of teachers in non-TAP schools, an impact of nearly 12 percentage points. In other words, teachers in Chicago TAP schools in fall 2007 were about 20% more likely than teachers in comparison schools to be in those same schools three years later. When we looked at teachers who were working in schools that started Chicago TAP in later years, some of the impact estimates were not statistically significant. We also found some evidence of impacts on retention for subgroups of teachers, such as those with less experience, but the pattern of findings was not consistent. When we considered retention of teachers in the district, we did not find consistent evidence of a measurable impact. Given that Chicago TAP is a school-specific program, our main focus was on school-level retention, as opposed to retention in the district.

Conclusions

Chicago TAP was only partially successfully in achieving its goals. Implementation of Chicago TAP increased the amount of mentoring, promotion opportunity, and compensation relative to non-TAP schools, and these increases alone may have translated into making Chicago TAP schools a more desirable place to continue working, as evidenced by the positive impacts on retention. However, these changes did not, in turn, pay off in terms of higher student achievement within the four-year rollout period in Chicago. This result provides a caution to funders investing in future programs in terms of what to expect over a four-year period. However, designers of new policies might consider how to change selected program elements to produce more favorable outcomes in the future.

I. INTRODUCTION

The practice of paying and promoting teachers based on classroom performance is gaining momentum in the United States. One program in particular, the Teacher Advancement Program (TAP), has become a model for schools around the country. TAP links teacher performance measures to pay and aligns mentoring and professional development with the performance measures so that teachers have the resources to improve their practice. This report provides evidence on the implementation and impacts of Chicago Public Schools' version of TAP, known as Chicago TAP.

A. Improving Schools Through TAP

TAP was developed in the late 1990s by the Milken Family Foundation (MFF) as a schoolwide program to improve schools by raising teacher quality. Under the TAP model, teachers can earn extra pay and be given increasing responsibilities through promotion (to mentor teacher or master teacher), and they are eligible for annual performance bonuses based on a combination of their contribution to student achievement (known as "value added") and observed performance in the classroom. The model also calls for weekly meetings of teachers and mentors ("cluster groups"), and regular classroom observations by a school leadership team to help teachers meet their performance goals. The idea behind the program is that giving teachers performance incentives, along with tools to track their performance and improve instruction, will help schools attract and retain talented teachers and help all teachers raise student achievement.

B. More Evidence Needed

TAP has been implemented in more than 200 schools in 13 states around the country and is overseen by the National Institute for Excellence in Teaching (NIET), an organization started by MFF. The most recent expansion of TAP came via the U.S. Department of Education's Teacher Incentive Fund (TIF), which makes grants to states and localities implementing performance-based compensation systems in high-need schools. These and related efforts to reform teacher pay and promotion (by the Bill and Melinda Gates Foundation, among others) have raised a great deal of interest and controversy. The question for researchers is whether there is any evidence that TAP or other teacher pay reforms improve the teaching workforce and raise student achievement.

Apart from the current study, much of the existing evidence about the effects of TAP comes from six reports. The program developers have conducted four studies of their own program (Schacter et al. 2002, 2004; Solmon et al. 2007; Daley and Kim 2010); one independent research team conducted a study using schools in two unnamed states (Springer et al. 2008), and an undergraduate student wrote a thesis on TAP (Hudson 2010) using data from the same 10 states studied by Daley and Kim. All of these studies were quasi-experimental, meaning that the researchers relied on self-selected samples of TAP schools and non-TAP schools for their comparisons. The fundamental challenge for program impact evaluations is that unmeasured factors can determine both a school's decision to participate in TAP (or be selected by NIET to participate) and the outcomes being studied. When researchers compare TAP and non-TAP schools, there is a danger that these unmeasured factors will be confounded with the true impact, a problem known as selection bias. The studies each used different methods—for example, statistical matching—to address this problem.

The current study is the first to use random assignment of schools to program implementation status as a way to control selection bias and is also the first to validate quasi-experimental methods (such as matching) using experimental methods (selection via random assignment). This validation is discussed further below.

The first two NIET studies (Schacter et al. 2002, 2004) relied on comparison groups that were small, self-selected samples. The more recent NIET report, by Solmon et al. (2007), includes larger numbers of comparison schools and teachers, a total of 61 TAP and 285 non-TAP schools across six states. As in the two earlier reports, the comparison schools were chosen as a convenience sample of schools from other districts that agreed to supply data and may not be representative of the outcomes that TAP schools would have realized if they had not adopted the program. Using comparison schools from different districts confounds district differences with TAP effects. Because TAP schools are carefully selected and typically volunteer to go through the many steps required to adopt the program, comparisons with nonselected schools could lead to biased program impact estimates. For example, schools that are able to attract the funding required to mount TAP may already be more effective than schools that are not able to attract such funding. Similarly, a school that NIET determines is "ready" to implement TAP may have faculty and leaders more willing to improve than a school that is not ready to implement, and this disparity would appear as a positive difference regardless of the true impact of TAP. Each of the NIET-sponsored reports found evidence of high percentages of TAP teachers and schools outperforming their comparison counterparts. The studies did not compare the mean outcomes from the student-level data so it is unclear if the results obtained by aggregating to the school level would also hold if the results had been generated by counting every student or teacher equally.

Springer et al. (2008) used a panel data set of math scores from TAP and non-TAP schools in two states and found positive impacts for elementary grades but undetectable or negative impacts for middle and high school grades. Importantly, the Springer et al. report presents evidence of selection effects by explicitly modeling the process of adopting TAP with an ordered probit model. This result is not surprising given the screening and self-selection that must take place for a school to adopt TAP. To become a TAP school, the faculty must vote to adopt the program, usually must raise substantial funds to finance the bonus pool, and often must be found worthy of the investment by NIET or a state or local sponsor.

Hudson relied on a synthetic control group method, a method similar to propensity score matching, in which each TAP school was matched to a weighted average of non-TAP schools in the same state. The author also used a difference-in-difference model, which controlled for the lower average test scores in TAP schools to isolate a TAP effect from the effect of other variables that do not change over time. The study found positive effects of TAP on math scores, with effect sizes ranging from 0.13 to 0.24 depending on the model. Reading score effect sizes ranged from 0.05 (not statistically significant) to 0.13 (significant).

The current study, about which two earlier reports have been released (Glazerman et al. 2009; Glazerman and Seifullah 2010), aims to expand this body of knowledge by using random assignment methods and focusing on a single district's experience implementing the TAP model. Importantly, the non-TAP schools all come from the same district as the TAP schools.

C. Chicago TAP and the Evaluation

1. Research Questions

This report focuses on one TIF grantee, the Chicago Public Schools (CPS). We address three research questions regarding Chicago TAP:

1. How was the program implemented?

2. What impact did the program have on student achievement?

3. What impact did the program have on teacher retention within schools?

The first question, about implementation, is needed to understand the answers to the next two, about impacts. We go on to address subquestions within each of these areas. For example, we explore whether impacts grow larger over time or as schools gain more experience with Chicago TAP. We also address methodological questions needed to interpret the impact estimates. Specifically, as explained below, we exploit the timing of the rollout of Chicago TAP to validate the statistical methods used throughout the impact analysis.

2. The Program Under Study

The school system implemented Chicago TAP as a pilot program intended for 40 high-need schools. The program began in 10 schools in fall 2007 (cohort 1) with a rollout plan to add 10 new Chicago TAP schools in each year of the TIF grant's four-year implementation period. Chicago TAP was originally based on the national TAP model, but it makes some local adaptations. For instance, Chicago TAP determines the compensation given to teachers, and it offers performance pay for principals and other school staff in addition to teachers.

For performing their extra duties, mentor teachers receive an additional $7,000 per year and lead teachers an additional $15,000.[1] Program documents indicate that in the first year of implementing Chicago TAP, the pool for teacher performance bonuses was supposed to support an average bonus of $2,000 per teacher based on value added to student achievement and observed classroom performance (Chicago Board of Education and Chicago Teachers Union 2007; Chicago Public Education Fund n.d.). In subsequent years, the target average payout was supposed to rise to $4,000 per teacher. We show in Chapter III that payouts were lower than these initially planned levels, with an average for most cohorts of about $1,100 in the first year of a school's implementation, $2,500 in the second and third years, and $1,900 in the fourth year. Principals can earn up to $5,000 each year based on the quality of program implementation and schoolwide value added. Other school staff can receive up to $500 in the first year and $1,000 in subsequent years based on schoolwide value added.

[1] For the final year of program rollout, Chicago TAP added a "lead plus" position in which a lead teacher supports two other Chicago TAP schools in addition to his or her home Chicago TAP school. Lead plus teachers receive an additional $5,000 per year, for a total stipend of $20,000.

3. Study Design and the Current Report

Chicago TAP provides a unique opportunity to learn about the impacts of the popular TAP model as adapted and implemented in a particular school district. To address issues of selection bias raised above, we designed a randomized experiment to estimate the impacts of Chicago TAP on student and teacher outcomes. School officials who wanted their school to implement Chicago TAP had to submit an initial application, and the selection process also involved site visits by Chicago TAP and CPS staff, a faculty vote (with at least 75 percent approval), and a successful final application with responses to essay questions. Schools had two opportunities to apply for Chicago TAP, one during the 2006–2007 school year and one during the 2008–2009 school year.

The experimental design, discussed in more detail in the next chapter, was implemented in two rounds, corresponding to the two opportunities for schools to apply to Chicago TAP. Of the 16 noncharter elementary schools that went through this process and were selected by district officials as finalists in spring 2007, we randomly assigned eight to a treatment group that began implementing Chicago TAP in 2007–2008 (cohort 1) and the other eight to a control group that delayed implementation until 2008–2009 (cohort 2). In spring 2009, we randomly assigned a new group of successful Chicago TAP applicant schools, including the charter schools: eight CPS elementary schools and one charter school were assigned to implement Chicago TAP in 2009–2010 (cohort 3) and another eight CPS elementary schools and the other charter school were assigned to implement in the final year of the grant program, 2010–2011 (cohort 4).[2] In the first year after random assignment, the difference between the outcomes for schools assigned to implement early and schools assigned to delay implementation represents the impact of Chicago TAP during that time period, because the only systematic difference between the two sets of schools is the ability to begin implementing Chicago TAP. In later years, the difference between the groups can be used to generate a similarly unbiased impact estimate, but that impact represents the effect of having a one-year head start implementing the program, rather than the impact of implementing versus not implementing it.

The quasi-experimental design, which is also discussed in the next chapter, relies on matching of Chicago TAP schools to non-TAP schools. This approach is especially promising in Chicago because the limited funding available to implement the program meant that there was a large pool of nonparticipating schools, many of which could be promising candidates for the program and hence make "good" matches. It is also promising in this setting because the overlap of experimental and quasi-experimental samples means that we can validate the matching methods to gauge the extent to which the matched comparison group replicates the experience of having a randomized control group. Finally, we have baseline test score and teacher turnover data that allow us to calculate preintervention measures of the key outcomes of interest and use them to estimate quasi-experimental impacts. These are all factors that have been associated with more successful (lower-bias) quasi-experimental evaluation designs (Glazerman et al. 2003). The difference between the

[2] All 34 schools had grades K–8 and are considered by CPS to be elementary schools. In addition to these randomly assigned elementary schools, the district selected four high schools and six other K–8 schools as Chicago TAP schools. In spring 2007, the district purposively assigned two high schools and two charter schools to implement Chicago TAP—one of each beginning in 2007 and the others in 2008. In spring 2009, CPS purposively assigned two new high school applicants to begin Chicago TAP in either 2009 or 2010. The district also replaced two elementary schools that had exited the program with two new schools selected by Chicago TAP staff to start Chicago TAP in fall 2009 and two additional schools to begin in 2010, replacing two cohort 4 schools that had dropped out.

Chicago TAP and matched comparison schools represents the estimated impact of the program for a given time period and cohort. Nevertheless, we caution that quasi-experimental results are potentially subject to selection bias and should be interpreted with caution.

This report presents experimental findings on the impacts of Chicago TAP on all 34 CPS elementary (K–8) schools and charter schools that had been randomly assigned to implement the program in each of the four rollout years (cohorts 1, 2, 3, and 4, assigned to begin implementing in 2007, 2008, 2009, and 2010, respectively). We also present quasi-experimental evidence on Chicago TAP effects for the same group expanded to include charter schools and "replacement" schools that were added to the roster of Chicago TAP schools when other schools closed or stopped implementing the program.

The current report summarizes the findings from all four years of Chicago TAP rollout. Additional detail on the first two years' experiences can be found in earlier reports (Glazerman et al. 2009; Glazerman and Seifullah 2010).

This page has been left blank for double-sided copying.

II. METHODS AND DATA

Our approach to estimating the impacts of Chicago TAP is based on a hybrid study design that relies on both the random assignment of schools to year of implementation and the careful matching of Chicago TAP schools to non-TAP CPS schools. Below, we discuss these methods and the data on which the analysis is based, and we present some simple descriptive statistics on the study sample.

A. Hybrid Study Design

The hybrid study design is shown in Figure II.1. Each box in the figure represents a group of schools. To estimate the impacts of the program in a given year, we compare the schools across the row: Chicago TAP schools to control schools, Chicago TAP schools to comparison schools, or Chicago TAP schools with more experience implementing the program to Chicago TAP schools with less experience. The baseline year for the first three columns of schools is 2007. We follow these schools for up to four years. For the other schools, the baseline is 2009 and we follow them for two years, because the study ended in 2011.

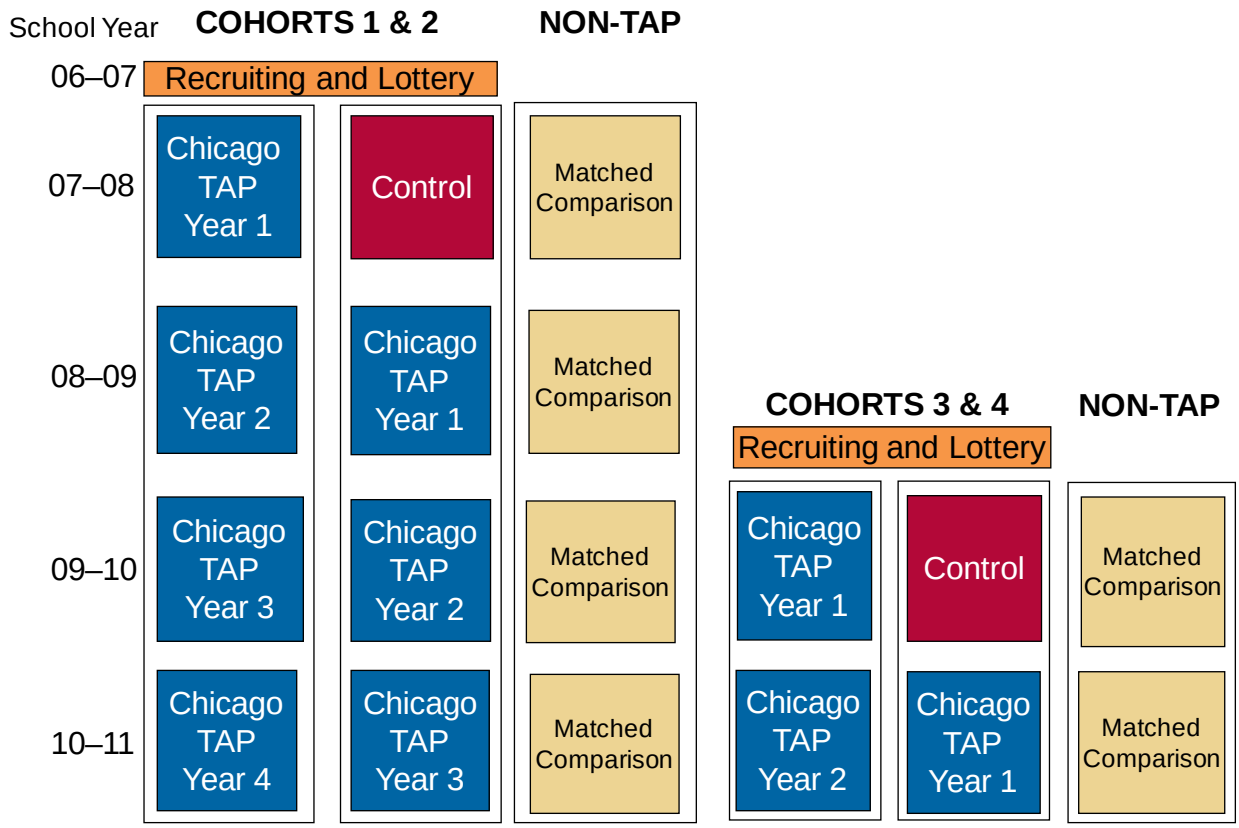
Specifically, the random assignment component, which we call the "experimental" design, proceeded as follows. We randomly assigned 16 preselected K–8 schools to begin implementing Chicago TAP in either fall 2007 or fall 2008.[3] In spring 2009 we repeated this process and assigned another 18 preselected K–8 schools to begin implementing in either fall 2009 or fall 2010.[4] We refer to these groups of schools as cohorts 1 through 4, respectively. Comparisons for 2007–2008 could be carried out for cohorts 1 ("treatment," represented by the blue box) and 2 ("control," represented by the red box), since cohort 2 had not yet begun implementing Chicago TAP during this period. Similarly, comparisons for 2009–2010 could be carried out for cohorts 3 ("treatment") and 4 ("control"), since cohort 4 had not yet begun implementing Chicago TAP during this period. Results from the first comparison were presented in the year 1 report (Glazerman et al. 2009); the current report includes results from the second experimental comparison and also combines the two, to represent the impact of Chicago TAP in its first year of implementation.

We continued with the experimental design by comparing cohort 1 to cohort 2 in spring 2009, spring 2010, and spring 2011. We also followed up on the cohort 3–cohort 4 comparison in spring 2011. All of these comparisons are made between adjacent pairs of blue boxes in Figure II.1. In each of these cases, both groups had been implementing Chicago TAP, but the experimental comparison represents the effect of having one additional year of implementation experience. At the outset of the study, we hypothesized that having more time to implement Chicago TAP might result in better outcomes.

[3] The schools had been preselected by CPS using a series of school information sessions, initial interest applications, and site visits to assess staff's ability and readiness to benefit from the program.

[4] In spring 2007, 16 noncharter K–8 elementary schools were randomly assigned and two K–8 charter schools were purposively assigned. In spring 2009, all elementary schools—16 noncharter and 2 charter—were randomly assigned. We also collected data on four high schools—one of each was purposively assigned to start each fall. Because the high schools were purposively assigned by CPS and comparable data were not available on test scores, this report focuses on elementary (K–8) schools. Charter schools were excluded from the teacher retention analysis because comparable data were not available for teaching assignments.

Figure II.1. Hybrid Study Design



The other component of the study design, the matched comparison, is "quasi-experimental." In this approach, we gathered administrative data on over 300 Chicago schools that were not participating in Chicago TAP, and then used statistical methods described below to identify the schools that were most closely matched to each group of Chicago TAP schools. These matching non-TAP schools formed the comparison group, with weights representing the degree to which each comparison school mostly closely matched a Chicago TAP school.

Each design strategy has advantages that complements the other by offsetting its disadvantages. The experimental strategy has the advantage of methodological rigor—any systematic differences in outcomes between cohorts 1 and 2 or between cohorts 3 and 4 can be causally attributed to the early implementation of the program in cohorts 1 and 3. In the first and third years of the study's observation period, the 2007–2008 and 2009–2010 school years, this provides us with a comparison between Chicago TAP and non-TAP schools. In subsequent years, we must interpret the differences in outcomes between each pair of cohorts (1 versus 2 and 3 versus 4) as the effect of having one extra year of experience implementing the program.

A disadvantage of the experimental design is the reliance on just 34 schools. There may be chance differences between the 17 treatment and 17 control schools that make it difficult to detect any true impacts.

Another drawback of random assignment to a delayed implementation status is that once the delay period is over, the control group is no longer untreated. This means that the experimental design is only helpful in generating a "clean" impact estimate for the first year of implementation.

Moreover, for outcomes that can be affected by knowledge of future implementation, such as teacher retention, the randomized control group is never a pure standard of comparison. That is, the control group is also affected by Chicago TAP, although less directly than the treatment group, because the school's staff know that they will be implementing Chicago TAP in the future, and that knowledge could change their behavior. On the other hand, the matched comparison group can be much larger than the randomized control group, even if the treatment group is limited in size; moreover, its members continue to be non-TAP schools during the course of the study, allowing more years for a comparison between Chicago TAP and non-TAP schools. However, in order to use the matched comparison group to infer program impacts, we must assume that the observable characteristics used to match schools are sufficiently similar and comprehensive that the remaining differences in outcomes can be attributed to Chicago TAP itself and not to other factors that were not observed, such as a dynamic principal or a motivated teaching staff.

1. Random Assignment Procedures

Whenever random assignment is conducted with a small number of experimental units, chance differences between the assigned groups—differences that have nothing to do with actual treatment status—are more likely to arise. This likelihood is reflected in the statistical hypothesis tests performed and standard errors calculated along with the estimates of program impact. Even with a given sample, it is possible to achieve through careful random assignment procedures a sample whose treatment and control groups are better matched on observable characteristics and which therefore supports more precise estimation of program impacts.

Specifically, we used the many observable characteristics of schools to improve the chances that schools assigned to the treatment versus control group would have similar characteristics. This method is based on constrained randomization, sometimes referred to as dynamic minimization (see Glazerman et al. [2006] for an application to an experiment on teacher induction and McEntegart [2003] for a more general discussion). In assigning cohorts 1 and 2 we imposed constraints on the randomization so that the largest and smallest school (in terms of student enrollment) were in the same Chicago TAP cohort (treatment or control), the three schools with a predominantly non–African American student body were not in the same cohort, and neither cohort had more than one pair of schools from the same geographic area of the city. In assigning cohorts 3 and 4 we imposed constraints so that the difference in average enrollment between cohorts was less than 100 students, the difference in average total score on the Illinois Standards Achievement Test (ISAT) was less than six scale points, the three schools where fewer than 95 percent of students were African-American were not all in the same cohort, and both cohorts had representation of each geographic area that contained more than one school.[5]

After the lottery assigning schools into cohorts was conducted, the research team sent the school assignments to CPS, and the district informed principals in the spring to plan early implementation procedures such as orientation and hiring of mentor and master teachers.[6] Teachers in Chicago TAP schools were provided an orientation over the summer.

[5] To achieve balance on geographic location in assigning cohorts 3 and 4, we imposed constraints so that the two area 3 schools were not in the same cohort, the three area 12 schools were not in the same cohort, and fewer than four of the five area 17 schools were in the same cohort. Each of the remaining schools came from different areas.

[6] The lotteries were conducted in May 2007 and March 2009.

2. Propensity Score Matching Procedures

Propensity score matching methods were used to identify non-TAP schools whose students were as nearly similar to those in Chicago TAP schools as possible (see Appendix A for a more technical discussion of the matches). We first eliminated schools where fewer than 50 percent of students were eligible for free or reduced-price lunch. In order to be selected for Chicago TAP, a school had to have at least 75 percent of its students qualifying for free or reduced-price lunch. We also eliminated from consideration for the teacher retention analysis any schools that had been selected to implement Chicago TAP beginning in either fall 2009 or fall 2010.[7] Teachers' behavior in these schools might have been influenced by the knowledge that the school was slated to begin the program in the near future. Then we matched the remaining schools in each year along dimensions we believed were related to student achievement and teacher retention, such as prior average achievement and retention as well as other factors listed below. All variables were measured before the schools implemented Chicago TAP; in particular, all variables except teacher retention were measured in 2006–2007.[8] Because knowledge of future implementation might have affected the decision to return to a school, we measured teacher retention before schools in each cohort learned of their Chicago TAP implementation year. Teacher retention was measured as the percentage of teachers in fall 2005 who returned to the school in fall 2006. We matched along the following dimensions:

- School size, measured using student enrollment and student enrollment squared (to capture nonlinearities in the relationship between size and outcomes)
- Teacher retention, measured as the percentage of novice teachers (having less than 5 years of experience) who returned to the school and the percentage of midcareer teachers (having 5–24 years of experience) who returned
- School accountability status, measured as indicators of the number of years since the school last met goals for adequate yearly progress
- Student achievement, measured using average math and reading scores on the ISAT, standardized within grade across the district
- Student race/ethnicity, measured as the percentage of students who were non-Hispanic African American and the percentage of students who were Hispanic, with the two variables collapsed into three categories: more than two-thirds, between one-third and two-thirds, and less than one-third African American or Hispanic, respectively
- Student poverty, measured as the percentage of students qualifying for free or reduced-price lunch
- Student special education status, measured as the percentage of students who had an Individualized Education Program

[7] An exception occurred for analysis of fall 2007 to fall 2008 retention rates for cohort 1 schools. Because teachers in cohort 3 and 4 schools did not learn of their future Chicago TAP status until spring 2009, fall 2007 to fall 2008 retention decisions could not have been affected by knowledge of their future Chicago TAP status.

[8] For a small number of schools that expanded by more than two grades between 2006–2007 and 2008–2009, we used a later baseline. All variables except teacher retention for these schools were measured in 2008–2009; teacher retention was measured as the percentage of teachers in fall 2007 who returned to the school in fall 2008.

- Student language proficiency, measured as the percentage of students who were limited English proficient

- Charter school status, measured as an indicator for whether the school was a charter school[9]

The procedure itself is a propensity score match using the nearest five neighbors with replacement. That means that we listed the observable factors that predict selection into the Chicago TAP finalist pool and used them in a logistic regression model to predict the probability of being in that pool. The predicted probability ("propensity score") from this model was used to rank all the schools sequentially along a number line, and each Chicago TAP school in each year was matched to the five non-TAP schools whose propensity scores were closest to that of the Chicago TAP school, regardless of whether those schools were also among the nearest five neighbors of any other Chicago TAP schools.[10] If non-TAP schools matched with more than one Chicago TAP school, they received proportionally more weight in the analysis.

The result of this matching was a group of non-TAP schools that was observationally similar to the Chicago TAP schools once we applied the appropriate weights. The degree of similarity is illustrated in Section D of this chapter; Appendix A provides more technical details on the matching. The particular matching method we used was chosen based solely on the quality of the matches produced—that is, before seeing the outcome data.

B. Impact Estimation

The impacts of Chicago TAP can be estimated by comparing the outcomes observed in Chicago TAP schools to those observed in similar non-TAP schools. The non-TAP schools are used to approximate the counterfactual condition, that is, the outcomes that would have been observed had the Chicago TAP schools not implemented the program. After four years of Chicago TAP, we have several comparisons to choose from, depending on the year and the outcome. Each time period and subgroup has a slightly different interpretation in light of the hypothesized learning effects with Chicago TAP. Within each impact chapter we discuss in more detail the specific contrasts we examine, but here we discuss the experimental contrasts.

For the first and third years after rollout, there was a randomized control group of schools that were not implementing Chicago TAP at all. In those years, the test score outcomes can be compared because they would not be influenced by the prospect of implementing Chicago TAP in the future. For test score outcomes in other years, or teacher retention outcomes in all years, there are several ways to form a (nonexperimental) comparison group of non-TAP schools, depending on which matching method we follow.

[9] Charter school status was used only for the student achievement analysis. As noted earlier, charter schools were excluded from the teacher retention analysis because administrative data on teaching assignments were not available for charter schools.

[10] We determined that nearest five neighbors produced the most efficient matches, but we repeated all the analyses using the single nearest neighbor as well as other matching algorithms, including ones that selected all neighbors within a certain distance or "caliper" size, which we varied, and all neighbors with weights related to distance (known as kernel density matching).

For the student achievement analysis, we first present experimental results comparing two groups: the first is cohort 1 and cohort 3 Chicago TAP schools in their first year of implementation, and the second is cohort 2 and cohort 4 schools that had not yet begun Chicago TAP. We then present quasi-experimental results comparing Chicago TAP schools in each year to matched comparison schools. We perform robustness checks and reconcile the findings for the two sets of analyses.

Knowledge that cohort 2 schools would implement Chicago TAP in 2008–2009 may have affected the decision of teachers to return in fall 2008 (and similarly for cohort 4 schools in 2010–2011), so we designed the study to use quasi-experimental methods for estimating the impacts of teacher retention even for the first year of implementation. We compared retention of teachers in Chicago TAP schools to retention of teachers in carefully constructed matched comparison groups through fall 2010. We did not analyze retention through fall 2011 because teachers knew during the 2010–2011 school year that Chicago TAP would not be continued in 2011–2012; therefore there is no reason to expect that Chicago TAP would influence teachers' decisions to return to former Chicago TAP schools in fall 2011.

1. Dropouts, Consolidations, and School Closures

Several Chicago TAP schools have experienced substantial changes that pose a complication for the estimation of program impacts. Effective for the 2009–2010 school year (the third year of the rollout), two cohort 1 schools discontinued Chicago TAP; in addition, a cohort 2 school was closed. Effective for the 2010–2011 school year, a cohort 1 school was closed and its students and faculty sent to another school that began implementing Chicago TAP that year. Also, a cohort 3 school was reconstituted with all new staff, and two cohort 4 schools randomly assigned to begin Chicago TAP did not implement the program.

For the experimental analysis of test scores we estimated impacts including all schools that had been randomly assigned to Chicago TAP, even if they had exited the program. These estimates based on the full sample of randomly assigned schools (called "intention-to-treat" estimates) exploit the statistical equivalence between treatment and control groups created by random assignment. That is, randomly assigning schools ensures that there are no systematic differences between the entire treatment group and the entire control group prior to starting Chicago TAP. The difference in average outcomes between the full treatment group and the full control group yields an unbiased estimate of the impact of being *offered* Chicago TAP, not necessarily the impact of implementing Chicago TAP. This distinction mattered only for the last two years of the analysis.

For the quasi-experimental analysis of test scores by cohort, we dropped schools that were closed or that discontinued the program—along with their matched comparison schools—beginning in the school year the changes went into effect. For example, if the change went into effect for the 2010–2011 school year, the staff knew of the change late in the prior (2009–2010) school year. Nevertheless, we assume the March 2010 test score results were unaffected.

For the teacher retention analysis, however, we dropped schools beginning in the school year prior to the change going into effect. Consider, for example, changes effective for the 2010–2011 school year. The retention analysis measures the outcome (teachers returning to the same school or to the district) as of fall 2010, after the school transitions had gone into effect. Consequently, we dropped all of the schools affected by transitions in 2009–2010 or 2010–2011 and their comparison group counterparts.

2. Regression-Adjusted Means

When presenting outcomes in this report (in Chapters IV and V), we typically show "regression-adjusted" means. A regression-adjusted mean for a particular group (for example, Chicago TAP schools) represents a predicted average outcome for the entire analysis sample (both Chicago TAP and non-TAP schools) if each one had been assigned to that group. The prediction is based on a regression model—a linear model for continuous outcomes and a logistic model for dichotomous outcomes—that controls for a range of teacher or student characteristics. Regression-adjusted means have the useful property that their difference equals the impact estimate, although they do typically differ slightly from unadjusted means. Thus nearly all means reported in the impact chapters of this report should be interpreted as the mean outcome for the whole sample as if all sample members had been implementing Chicago TAP or all sample members had not been implementing Chicago TAP.

For teacher retention analyses, the regression model controlled for the following variables:

- Teacher education (having a master's degree or higher)
- Teaching assignment (teaching a tested or nontested academic subject in grades four through eight)
- Years of service in CPS
- Prior teacher retention at the school (percentage of novice teachers and percentage of midcareer teachers in fall 2005, or fall 2007 for cohort 3, who returned to the school in fall 2006, or fall 2008 for cohort 3)
- Prior student achievement (average math and reading ISAT scores)
- Student race/ethnicity (percentage of students who were African American or Hispanic)
- Student language (percentage of students who had limited English proficiency)
- Student poverty (percentage of students who qualified for free/reduced-price lunch)
- School size

For student test score analyses, the model controlled for the following variables:

- Prior student achievement (math and reading ISAT scores)
- Family poverty (eligibility for free/reduced-price lunch)
- Special needs (whether an Individualized Education Program was in place)
- Language (whether limited English proficient)
- Race/ethnicity
- Grade level
- Age (whether older than the normal age for a grade)

We accounted for the clustering of students or teachers within schools by estimating robust standard errors. In addition, we conducted numerous sensitivity tests to determine whether the results were robust to the choice of regression model or other decisions. Those tests are described in more detail in the discussion of findings.

C. Data

The study's data fall into two broad categories: (1) data collected directly from teachers and principals in the sampled schools by Mathematica Policy Research, and (2) administrative data provided by CPS.

In the first category, Mathematica administered a questionnaire in spring 2008, towards the end of the first year of the program, to a group of teachers in Chicago TAP cohort 1 schools, cohort 2 schools, and a group of schools matched to cohorts 1 and 2. The survey was repeated in 2010 with a new sample consisting of teachers from cohort 3 schools, cohort 4 schools, and an additional comparison group of schools matched to cohorts 3 and 4. The results presented here focus on the 2010 survey, whose findings were similar to those of the 2008 survey, presented in an earlier report (Glazerman et al. 2009). The school principal interviews were conducted in 2008 as well and are also summarized in that same report. Those interviews were mainly used to validate the teacher retention analysis and to detect the possible presence of a harmful effect of Chicago TAP on school climate. We did not find any such evidence, nor did we find any reason to doubt the administrative data, so we did not repeat the principal interviews and they are not discussed in this report.

In the second category, we used CPS human resources records for information on teachers, and CPS student testing records for information on student assessments. In addition, we used data provided by CPS on program review scores (which measure fidelity of implementation to the TAP model), teacher observation scores, and teacher bonus payouts under Chicago TAP.

1. Teacher Survey

In spring 2010 we administered a questionnaire to all eligible teachers in cohort 3, cohort 4, and sampled matched comparison schools. We included Chicago TAP lead teachers and all sampled schools' teachers who had a program code identifying them as a regular classroom teacher of an academic subject. The questionnaire gathered data that were not available in the CPS administrative records. It included six sections focusing on the following areas: (1) teachers' educational background and professional experience, their certification status, and their current teaching assignment; (2) the types of professional development and support that teachers receive at their schools; (3) the leadership roles and responsibilities teachers have assumed in addition to their regular classroom teaching duties; (4) the compensation, or potential for compensation, associated with teachers' performance and that of their students; (5) teachers' attitudes about and satisfaction with various aspects of their school and the opportunities provided to them; and (6) teachers' basic demographic characteristics.

We mailed the questionnaire to teachers at their schools in mid-March, and continued collecting responses through mid-July, either as self-administered paper questionnaires returned by the teachers or through telephone interviews. Of the 826 eligible teachers, 617 completed the survey, for a final response rate of 75 percent (79 percent for treatment, 78 percent for control, and 69 percent

for matched comparison teachers).[11] Nonresponse adjustment weights were used in all analyses to account for any observable differences between respondents and nonrespondents. We computed the nonresponse adjustment weights using teacher experience level, whether the teacher has a master's degree, and school characteristics (such as years since the school last made adequate yearly progress, attendance rate, and percentages of students who were limited English proficient, low-income, African American, and Hispanic).

2. Administrative Data on Students, All Teachers, and Chicago TAP Participants

We obtained student demographic and test data from CPS. The test data were from the ISAT, the state assessments in mathematics and reading for grades three through eight and in science for grades four and seven. Table II.1 shows the average ISAT scores for math and reading by grade level, as well as the corresponding standard deviations, for the study baseline; data are from March 2007. As the table shows, the scores rise with grade level, suggesting that the tests are scored on a developmental scale. That is, the 11-point difference between the average fourth-grade math score and the average fifth-grade math score, for example, can be thought of as a single year of growth, so one point represents 1/11, or 0.09, school years of growth, equivalent to roughly 3.5 weeks of learning in a school year with 38 weeks. We did not have science data for 2007.

Table II.1. Properties of the Baseline ISAT Test Scores by Subject and Grade Level, March 2007

	Reading		Math	
Grade	Mean	Standard Deviation	Mean	Standard Deviation
All grades	222.8	27.8	236.9	30.0
Grade 4	202.5	25.9	213.6	25.9
Grade 5	212.1	25.7	224.6	26.8
Grade 6	222.5	23.5	235.9	25.8
Grade 7	229.7	24.3	244.6	25.9
Grade 8	242.0	23.0	259.7	24.4

Note: Data pertain to 93,657 students in 308 schools.

The standard deviation within a grade level can also be used to interpret the size of score differences. For example, a standard deviation of 26 points (the value for grade four math) means that a one-point difference corresponds to 1/26th, or 0.04, standard deviation units. If test scores are normally distributed, then this is the same as a difference between performance at the 50th and the 52nd percentiles.[12]

In addition, we obtained data on student background information, such as race, gender, free/reduced-price lunch eligibility, enrollment status, and disability or special education status. CPS provided these data for the 2006–2007 through 2009–2010 school years.

[11] Among teachers in noncharter elementary schools, response rates were 79 percent overall, 84 percent for cohort 3, 82 percent for cohort 4, and 75 percent for matched comparison teachers.

[12] More precisely, an effect size of 0.0384 corresponds to a change in the cumulative normal probability of 0.015 if the starting probability is 0.500.

We also obtained administrative data on teachers' credentials, years of service in the district, and teaching assignments from the CPS Talent Office. Data cover the 2005–2006 through 2010–2011 school years.

Finally, CPS provided us with the following data on Chicago TAP: teacher scores on the Skills, Knowledge, and Responsibilities classroom observation rubric performance; payouts by teacher; and scores on a program review that tells each Chicago TAP school how well they have been implementing the program over the current school year.

D. Sample Characteristics

Tables II.2 through II.4 show the characteristics of the schools in the study; they indicate characteristics of students at baseline and of teachers during the study period. The statistics are presented by school group, with students and teachers in Chicago TAP schools compared to their counterparts in control and matched comparison schools. We present tests of statistical significance of the difference between Chicago TAP and each comparison group. Readers should be aware that statistical significance is not the same as policy relevance. It may be reassuring to note a great degree of similarity in the observable student and teacher characteristics of the Chicago TAP and comparison groups, but that similarity is not necessary for unbiased estimation of the experimental impacts of Chicago TAP, because we control for observable differences through regression adjustment. What is required is that the groups be similar in terms of the *unobserved* determinants of student achievement growth and teacher retention, such as the level of motivation among students or teachers, which are not accounted for in the data. Properly implemented random assignment ensures that there should be no *systematic* differences between treatment and control groups in terms of observed or unobserved factors. Readers should exercise caution in interpreting the quasi-experimental findings because these unobserved determinants of student achievement and teacher retention may be confounded with Chicago TAP status, leading to bias of unknown direction and magnitude.

Table II.2 shows the balance across school groups in terms of student demographics and baseline test scores.[13] The only statistically significant difference between Chicago TAP schools (cohorts 1 and 3) and control schools (cohorts 2 and 4) was the difference in the percentage of students eligible for free or reduced-price lunch. We do not find any statistically significant differences between the combined Chicago TAP group (cohorts 1, 2, 3, and 4) and matched comparison schools.

We measured teacher characteristics using the teacher survey and CPS administrative records. In the survey sample, none of the differences between Chicago TAP and control group teacher characteristics (shown in Table II.3) was statistically significant. The only significant difference between Chicago TAP and the comparison group was the difference in the percentage of teachers who have a master's degree or higher.

[13] These student characteristics are presented to illustrate the types of schools included in the study, but they are not identical to the characteristics of the students whose 2009–2010 school year test scores were analyzed. We tabulated statistics for those students from the impact analysis sample and found similar results to those presented here.

Using the administrative data sample, we found a few statistically significant differences between Chicago TAP and matched comparison schools in teacher characteristics (Table II.4). The percentage of teachers holding National Board certification was higher in Chicago TAP schools than in comparison schools. Chicago TAP schools had lower percentages of teachers in academic but nontested grades or subjects and late-career teachers (having more than 24 years of service in the district) than comparison schools.

Table II.2. Baseline School Characteristics by School Group (percentage unless otherwise noted)

Characteristic of School's Students	Chicago TAP (Cohorts 1& 3) Mean	Control (Cohorts 2& 4) Mean	Difference	Chicago TAP (Cohorts 1–4) Mean	Matched Comparison (to Cohorts 1–4) Mean	Difference
Race/Ethnicity						
African American	91.7	89.4	2.4	95.1	92.3	2.9
Hispanic/Latino	7.3	9.2	-1.9	4.3	6.6	-2.3
White	0.3	0.7	-0.4	0.4	0.7	-0.2
Eligible for free/ reduced-price lunch	97.3	94.1	3.3*	95.5	94.6	0.9
Limited English proficient	2.7	3.8	-1.1	1.0	1.5	-0.5
Special education (has Individualized Education Program)	13.3	14.4	-1.1	13.4	13.4	0.0
Over age for grade	36.3	32.3	4.0	45.9	46.7	-0.7
School average achievement (scale points)						
ISAT Reading	213.3	211.0	2.2	211.0	211.6	-0.6
ISAT Math	224.6	222.8	1.8	222.3	223.1	-0.8

Note: N = 5,176 cohort 1 and 3 students in 17 schools; 5,596 cohort 2 and 4 students in 17 schools; 11,331 cohort 1, 2, 3, and 4 students in 39 schools; and 31,068 matched comparison students in 99 schools. Data for cohorts 3 and 4 pertain to the student body in the school during the 2008–2009 school year. Data for cohorts 1 and 2, for the combined Chicago TAP group (cohorts 1–4), and for matched comparison schools pertain to the student body in the schools during the 2006–2007 school year.

* Difference is statistically significant at the 10 percent level.
** Difference is statistically significant at the 5 percent level.
*** Difference is statistically significant at the 1 percent level.

Table II.3. Teacher Characteristics by School Type, 2009–2010 Survey Sample (percentages except where noted)

	Means			Difference in Means	
Characteristic	Chicago TAP (Cohort 3)	Control (Cohort 4)	Comparison to Cohort 3	Chicago TAP-Control	Chicago TAP-Comparison
Master's degree or higher	75.5	73.3	65.3	2.2	10.2**
Alternative certification	17.3	16.1	17.1	1.2	0.2
National board certification	4.3	5.9	2.9	-1.6	1.4
Taught tested subject/grade	38.3	44.5	38.4	-6.1	-0.1
Experience (years)	11.9	13.0	12.8	-1.1	-0.9
<5 years of experience	18.3	14.9	17.0	3.4	1.3
Age (years)	41.3	42.1	43.6	-0.8	-2.3
Male	9.4	12.4	15.4	-3.0	-6.0
African American	66.1	53.9	53.7	12.3	12.4
Hispanic/Latino	4.9	15.4	16.1	-10.4	-11.2
Attended CPS as a student	56.2	58.2	58.3	-1.9	-2.1

Note: N = 140 cohort 3 teachers in 8 schools, 144 cohort 4 teachers in 8 schools, and 244 matched comparison teachers in 16 schools.

* Difference is statistically significant at the 10 percent level.
** Difference is statistically significant at the 5 percent level.
*** Difference is statistically significant at the 1 percent level.

Table II.4. Teacher Characteristics by School Group, 2009–2010, Administrative Data Sample (percentage unless otherwise noted)

Characteristic of School's Students	Chicago TAP (Cohorts 1–3) Mean	Matched Comparison (to Cohorts 1–3) Mean	Difference
Master's degree or higher	66.8	62.5	4.3
National Board certification	5.7	2.9	2.8*
Taught academic subject, tested grade/subject	37.4	34.8	2.6
Taught academic subject, nontested grade/subject	46.3	50.2	-4.0*
Years of service	10.5	11.6	-1.1
Percentage <5 years of service	31.2	26.8	4.4
Percentage 5–24 years of service	62.6	64.5	-1.9
Percentage >24 years of service	6.2	8.8	-2.6*

Note: N = 612 Chicago TAP teachers in 21 schools and 2,082 matched comparison teachers in 77 schools.

* Difference is statistically significant at the 10 percent level.
** Difference is statistically significant at the 5 percent level.
*** Difference is statistically significant at the 1 percent level.

III. IMPLEMENTATION

In order to interpret the study's impact estimates, it is important to understand what Chicago TAP schools implemented and how faithful they were to the Chicago TAP model. This chapter aims to address this question. We used district administrative records to describe teacher payouts. We also used district-supplied ratings of implementation fidelity, which were produced by reviewers who visited each Chicago TAP school, and summarized internal evaluation reports produced by CPS. The other source of information came directly from the teachers. We conducted two teacher surveys. The first, in spring 2008, asked teachers in new Chicago TAP schools (cohort 1) and non-TAP schools (cohort 2/control and matched comparison) about their experiences with mentoring, professional development, and compensation. Our surveys also measured attitudes about factors that could be affected by Chicago TAP. We repeated the survey in spring 2010, focusing on teachers in the newest cohort of Chicago TAP schools (cohort 3) and their control group and a matched comparison group as well.

A. Teacher Payouts, TAP Fidelity, and Teacher Attitudes

1. Payouts

For performing their extra duties, mentor teachers receive an additional $7,000 per year and lead teachers an additional $15,000.[14] Program documents indicate that in the first year of implementing Chicago TAP, the pool for teacher performance bonuses was supposed to support an average bonus of $2,000 per teacher based on value added to student achievement and observed classroom performance (Chicago Board of Education and Chicago Teachers Union 2007; Chicago Public Education Fund n.d.). In subsequent years, the target average payout was supposed to rise to $4,000 per teacher. Principals can earn up to $5,000 each year based on the quality of program implementation and schoolwide value added. Other school staff can receive up to $500 in the first year and $1,000 in subsequent years based on schoolwide value added.

Data provided to the authors by CPS suggest that the teacher payouts averaged less than the original target amounts (Table III.1). We found that the average performance bonus payout for a school's first year of Chicago TAP implementation was approximately $1,100 per teacher in the first three years of district rollout (2007–2008, 2008–2009, and 2009–2010), rising to $1,400 for new Chicago TAP implementers in 2010–2011, with a maximum payout of less than $2,700 in any year. For a school's second and third years of implementation, the average payout was approximately $2,500, with a maximum payout of $6,400, although in the fourth year of implementation, the average payout was lower, about $1,900 with a maximum less than $4,600.

Table III.1 shows the mean and range of teacher payouts under Chicago TAP, but another aspect of the teacher payout distribution is the degree to which the individual teacher performance determined payouts relative to schoolwide performance. If most of the variation occurs *between* schools rather than within schools, it suggests that Chicago TAP effectively offered school-level performance bonuses rather than individual teacher bonuses. To measure this variation, we

[14] For the final year of program rollout, Chicago TAP added a "lead plus" position in which a lead teacher supports two other TAP schools in addition to his or her home Chicago TAP school. Lead plus teachers receive an additional $5,000 per year, for a total stipend of $20,000.

computed the intraclass correlation coefficient (ICC), which can be interpreted as the proportion of variation in teacher payouts that occurs between schools relative to total variation (between and within schools). These ICCs are presented in Table III.2 by year and by cohort. An ICC of 1.0 means that every teacher within each school had the same payout. An ICC of 0.0 means that average payouts were the same for all schools. We multiplied the ICCs by 100 so they can be read as percentages. The results suggest that more than two-thirds of the variation occurred between schools in the first year of rollout in the district (2007–2008), but the between-school percentage declined across years, with the exception of cohort 1 in the fourth year (2010–2011). In the final two years, most of the variation in payouts occurred within schools.

Table III.1. Average Performance- Based Payouts Under Chicago TAP by Cohort and Year

	School Group			
Year	Cohort 1	Cohort 2	Cohort 3	Cohort 4
2007–08	$1,078 (Range = $0 to $2,045)	n.a.	n.a.	n.a.
2008–09	$2,622 (Range = $0 to $6,320)	$1,066 (Range = $0 to $2,458)	n.a.	n.a.
2009–10	$2,415 (Range = $0 to $6,123)	$2,402 (Range = $0 to $6,400)	$1,091 (Range = $0 to $2,250)	n.a.
2010–11	$1,945 (Range = $0 to $4,527)	$2,478 (Range = $0 to $5,400)	$2,480 (Range = $0 to $5,457)	$1,409 (Range = $0 to $2,675)

Note: Data pertain to CPS elementary schools; N = 5–10 schools in each cohort/year table cell. Shading of the cells represents the number of years implementing Chicago TAP ranging from 1 year (lightest) to 4 years (darkest).

n.a. = not applicable.

Table III.2. Percentage of Variation in Teacher Payouts Occurring Between Chicago TAP Schools, by Cohort and Year

	School Group			
Year	Cohort 1	Cohort 2	Cohort 3	Cohort 4
2007–08	70	n.a.	n.a.	n.a.
2008–09	17	55	n.a.	n.a.
2009–10	15	34	38	n.a.
2010–11	31	23	28	28

Note: Data pertain to CPS elementary schools; N=5–10 schools in each cohort/year cell. Each number is an intraclass correlation statistic, representing the percentage of variation between schools, with the remaining percentage being within schools. Shading of the cells represents the number of years implementing Chicago TAP ranging from 1 year (lightest) to 4 years (darkest).

n.a. = not applicable.

The student achievement component of teacher performance-based payouts was originally supposed to be based on schoolwide value added in a school's first program implementation year and a combination of schoolwide and classroom-level value added in subsequent years (Chicago Public Education Fund n.d.). In practice, however, the classroom-level value-added component was not implemented. Only school-level value added was used in the first two years of district rollout (2007–2008 and 2008–2009). Starting with the third rollout year (2009–2010), CPS used school- and school-grade level value added in calculating performance bonuses for teachers in schools that were in their second or later year of program implementation. According to Chicago TAP staff, CPS did not implement the classroom-level value-added component because the data that were needed to reliably link students and teachers at the classroom level were not available.

2. Program Review Scores (TAP Fidelity Ratings)

Every spring, NIET conducts site visits to TAP schools to verify that they are implementing the program according to the organization's standards. Schools are rated along several dimensions and given a summary score ("cumulative program review score") that describes their implementation. The average scores by year and cohort are shown in Table III.3. In the first two years of rollout, NIET gave the Chicago TAP schools in this study an average score of approximately 3 on a five-point scale, where a 5 represents "the fullest, most complete, and high quality level of implementation," according to program review reports supplied by CPS to the authors. NIET program review scores averaged 2.7 in the third rollout year.

During the summer following the third rollout year (August 2010), NIET informed CPS that the district had not implemented the TAP system. NIET indicated that elements of TAP had been introduced, but TAP implementation had not been "rigorous," according to a letter supplied by NIET to the authors.

For the final year of rollout, CPS staff, rather than NIET, conducted the program reviews. During that year the average scores were nearly two points higher than in the previous year, as shown in Table III.3.

3. Teacher Attitudes Reported in CPS Evaluation Reports

CPS conducted an internal evaluation of Chicago TAP aimed at providing formative feedback to program staff. The year 1 evaluation report (Foster 2009) used teacher surveys and focus groups to describe perceptions of the program among staff in Chicago TAP schools during the 2007–2008 school year and to document the degree to which staff found the program helpful and implemented it faithfully. The study reported that teachers generally had positive perceptions of Chicago TAP but it raised concerns about the communication of program structure and expectations. The majority of teachers participated in the cluster activities that focused on teaching them new skills (through, for example, a demonstration by an expert teacher or feedback from a colleague or mentor). The CPS implementation study also found that teachers reported Chicago TAP coaching to be more frequent than coaching offered in the year prior to implementation, and that the professional development delivered through the program was perceived as more effective than other forms of professional development.

Table III.3. Average Program Review Score by Cohort and Year

	School Group			
Year	Cohort 1	Cohort 2	Cohort 3	Cohort 4
2007–08[a]	2.9 (Range = 2.5 to 3.4)	n.a.	n.a.	n.a.
2008–09[a]	3.1 (Range = 2.2 to 3.8)	3.2 (Range = 2.8 to 3.7)	n.a.	n.a.
2009–10[a]	2.5 (Range = 1.9 to 3.3)	2.8 (Range = 2.0 to 3.5)	2.6 (Range = 2.2 to 3.0)	n.a.
2010–11[b]	4.5 (Range = 3.5 to 5.0)	4.4 (Range = 2.7 to 5.0)	4.7 (Range = 3.7 to 5.0)	4.5 (Range = 3.2 to 5.0)

Note: Data pertain to CPS elementary schools; N=5–10 schools in each cohort/year cell. Shading of the cells represents the number of years implementing Chicago TAP ranging from 1 year (lightest) to 4 years (darkest).

n.a. = not applicable.

[a] Program reviews were conducted by NIET.
[b] Program reviews were conducted by CPS.

In a follow-up year 3 study (Crown 2010), CPS reported a generally favorable teacher reaction to Chicago TAP implementation, with more enthusiasm for professional development and less for performance-based pay. The study found that teachers in TAP schools "assign favorable ratings to the impact of TAP on their schools and their teaching." The study cited coaching, training, and support as the key to Chicago TAP's success. Teachers reported that cluster group meetings occur weekly as planned and last at least 45 minutes in "most, but not all" Chicago TAP schools. Chicago TAP teachers were found to "overwhelmingly" endorse the observation and coaching system, which they found "very useful." However, performance-based pay was seen to have "a minimal impact on how teachers view their jobs, and a moderate impact on teachers' motivation to improve their performance." According to the teacher survey, the first two cohorts of Chicago TAP teachers identified both improvements and challenges over time, and expressed the view that it takes time for the program to become established.

The current study differs from the NIET and CPS efforts to measure implementation in that it incorporates data from non-TAP schools to provide additional context. This practice allows us to describe implementation and impacts relative to the norm for the district or for district schools that might have implemented Chicago TAP but did not.

B. Implementation in Chicago TAP and Non-TAP Schools Based on Teacher Surveys

To assess the first year under Chicago TAP for the most recent cohort of schools to begin the program, we compared how teacher development and compensation practices in Chicago TAP schools differ from practices normally implemented in CPS schools. For Chicago TAP, or "treatment," schools, we used cohort 3 schools, which were randomly assigned to begin the program in fall 2009; for non-TAP, or "control," schools, we used cohort 4 schools, which were randomly assigned to delay implementation of the program until fall 2010. Using specific practices as outcomes, we calculated regression-adjusted means for the treatment and control schools. The control school means enabled us to characterize the counterfactual condition—that is, the experiences that would have occurred in the absence of Chicago TAP. We performed t-tests to

assess the extent to which practices in Chicago TAP schools differed significantly from practices in non-TAP schools. Similar findings were obtained when we compared practices in cohort 3 schools to practices in a matched comparison sample of non-TAP schools. Those results are presented in Appendix B (Tables B.1–B.7). Results for an earlier administration of the survey, conducted in 2008 with cohorts 1 (Chicago TAP), 2 (control), and a matched comparison group, were largely similar to those presented here. Those findings can be seen in an earlier report from this study (Glazerman et al. 2009).

1. Mentoring, Leadership, and Feedback

Overall, we found mentoring, leadership, and feedback in both Chicago TAP and non-TAP schools, but Chicago TAP schools tended to provide more mentoring support for teachers. Compared to control teachers, treatment teachers reported spending more time receiving guidance from an advisor. Veteran teachers in treatment schools were more likely than veteran teachers in control schools to provide mentoring support for other teachers. Observation and feedback were generally about as common in non-TAP as in Chicago TAP schools.

a. Mentoring Received

Chicago TAP incorporates mentoring into the regular school day through ongoing classroom support provided by master teachers (known as lead teachers in Chicago) and mentor teachers. Teachers meet weekly in small cluster groups led by lead or mentor teachers; the goal of these meetings is for teachers to collaborate on improving their instruction and increasing student achievement. Mentor teachers are also assigned specific traditional-classroom teachers whose professional development they are meant to foster (NIET 2008).

According to teachers, mentoring was prevalent in both Chicago TAP and non-TAP schools (see Table III.4). Treatment teachers were more likely than control teachers to report having at least one advisor from whom they received professional advice and direct assistance with their teaching duties, but the difference was not statistically significant. We did, however, find several other meaningful differences suggesting that teachers in Chicago TAP schools received significantly more mentoring support than teachers in non-TAP schools.

There were significant differences in the type of individuals from whom teachers receive advice and assistance. Compared to control teachers, treatment teachers were more likely to receive guidance from an individual they characterized as a mentor, a lead teacher, or a principal. Eighty-six percent of treatment teachers indicated that their main advisor worked only in their school, compared to 66 percent of control teachers. Having a mentor in the building, which is the aim of the Chicago TAP model, might provide more opportunities for assistance on demand than having itinerant mentors. The main advisors of treatment teachers were also significantly more likely to be full-time teachers and to receive release time from classroom teaching in order to perform their mentoring duties. These findings are consistent with the Chicago TAP model, in which mentor and lead teachers are given release time to work with traditional-classroom teachers in their schools. Other significant differences relative to main advisors in control schools were that main advisors in Chicago TAP schools were less likely to be district specialists and more likely to be from a teacher licensing, certification, or preparation program.

Table III.4. Mentoring Received

Outcome	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error
Received professional advice and assistance in teaching duties from an advisor (percentage)	96.7	85.6	11.1	7.68
Had an advisor who was a(n) . . . (percentage)				
Mentor	56.9	10.9	46.0***	4.74
Literacy coach	46.2	61.5	-15.3	13.74
Math coach	24.2	29.7	-5.5	14.01
Lead teacher	77.2	9.0	68.2***	4.53
Principal	63.9	33.9	30.0***	8.08
Assistant or vice principal	48.8	34.6	14.1	11.22
Peer	42.1	33.8	8.3	6.85
Had a main advisor who was a . . . (percentage)				
Full-time teacher	60.5	37.5	23.0***	7.00
Person who works in teacher's school only	85.6	65.9	19.6**	8.31
Person who works in more than one school	4.9	13.6	-8.8**	3.56
Teacher with release time	44.5	26.1	18.4***	6.49
Person with no classroom teaching	60.7	58.5	2.3	11.08
Principal or school administrator	18.0	22.7	-4.7	5.23
School-based specialist	43.5	43.6	-0.1	10.30
District specialist	3.3	10.3	-7.0	3.87
Person from a teacher licensing, certification, or preparation program	22.9	17.5	5.3	3.17
Time spent with main advisor				
Frequency of scheduled meetings (number per week)	1.4	0.9	0.5**	0.19
Duration of each scheduled meeting (minutes)	68.2	39.2	29.0***	6.83
Duration of informal contact (minutes per week)	79.0	51.6	27.5*	14.66
Frequency of total contact (minutes per week)	178.7	95.9	82.7***	22.72
During most recent full week, scheduled time main advisor spent . . . (minutes)				
Observing teacher's teaching	28.0	18.6	9.4**	4.37
Meeting with teacher one-on-one	27.9	17.1	10.8**	4.95
Meeting with teacher together with other teachers	43.5	28.8	14.7*	7.10
Modeling a lesson	20.6	6.4	14.2***	3.53
Coteaching a lesson	10.8	5.9	4.9	3.42
Teacher received useful feedback from main advisor (percentage)	88.1	85.9	2.2	4.87

Note: N = 247 to 283 teachers per outcome.

[a] Means are regression adjusted.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

Teachers in Chicago TAP schools reported more frequent and longer meetings and activities with their main advisor. On average, treatment teachers had 1.4 scheduled meetings per week with their main advisor compared to 0.9 scheduled meetings per week for control teachers, with the average meeting for treatment teachers lasting 29 minutes longer. Both one-on-one and small-group meetings with main advisors were longer for treatment teachers than control teachers during the most recent full week of teaching. Compared to control teachers, treatment teachers also spent more scheduled time in the most recent full week being observed teaching by their main advisor and having their main advisor model a lesson. Chicago TAP also increased the amount of informal

contact teachers had with their main advisor, with treatment teachers reporting nearly 30 more minutes of informal contact per week than control teachers. In total, Chicago TAP teachers averaged nearly three hours of scheduled and informal contact with their main advisor per week, compared to an average of 96 minutes of total contact per week for control teachers.

Both Chicago TAP and non-TAP teachers tended to regard the feedback they received from their main advisor as useful. Eighty-eight percent of treatment teachers and 86 percent of control teachers reported receiving useful feedback, a difference that was not statistically significant.

b. Leadership Roles Held

Chicago TAP offers teachers opportunities to take on leadership responsibilities and earn extra pay without having to leave the classroom entirely. Teachers can become mentor or lead teachers who serve on the Chicago TAP leadership team responsible for the overall implementation of Chicago TAP, analyze student data, and develop academic achievement plans. In addition, mentor and lead teachers support the professional development of traditional-classroom teachers, known as career teachers. Responsibilities of these teacher-leaders include leading cluster groups, observing and evaluating career teachers, team teaching with colleagues, and modeling lessons. Mentor teachers provide day-to-day mentoring and coaching to career teachers. Sharing leadership and authority with the principal, lead teachers are also responsible for overseeing the professional development of both mentor and career teachers (NIET 2008).

Under the Chicago TAP model, mentor and lead teachers receive release time from classroom teaching in order to fulfill their leadership responsibilities. Chicago TAP schools are expected to provide mentor teachers with one to five hours of student-free time per week outside of cluster meetings. The model recommends that lead teachers teach two hours per day and devote the remainder of their work day to lead teacher responsibilities.

Measuring the impact of lead and mentor teachers is complicated by the fact that these positions have no clear analogue in non-TAP schools. Therefore, our approach was to measure for each school the amount of leadership and mentoring provided by teachers who could plausibly have played similar roles as lead or mentor teachers. We focused on veteran teachers, whom we defined as having at least five years of experience as a head classroom teacher. This experience cutoff roughly approximates the minimum experience levels required to become mentor or lead teachers for the first year of Chicago TAP's implementation. Chicago TAP requirements for these teacher-leader roles include a minimum of four years of teaching experience for mentor teachers; lead teachers must have at least six years of successful teaching, with at least four years as a classroom teacher (NIET 2008). If one mentor teacher is assigned to each group of 8 career teachers, and one lead teacher is assigned to each group of 15 career teachers, then one might expect about 17 percent of all teachers to be providing leadership in a Chicago TAP school. When the sample is restricted to veteran teachers, one might expect the percentage to be higher. The goal of the analysis is to estimate that percentage for Chicago TAP and non-TAP schools.

Consistent with the career path component of the Chicago TAP model, veteran teachers in Chicago TAP schools were more likely than their control group counterparts to provide mentoring support for other teachers in a range of topics (see Table III.5). Twenty-five percent of veteran Chicago TAP teachers reported providing "formal mentoring services" to teachers in their schools, compared to 15 percent of veteran control teachers. Though a variety of topics was covered in mentoring activities in both Chicago TAP and non-TAP schools, veteran Chicago TAP teachers were significantly more likely to report covering certain topics, such as devising strategies for

teaching literacy, setting instructional goals and determining ways to achieve them, and preparing lesson plans or other instructional activities.

Table III.5. Mentoring Provided (teachers with at least five years of experience)

Outcome	Chicago TAP Mean[a]	Control Mean[a]	Difference	Standard Error
Provided formal mentoring services (percentage)	24.9	15.0	9.9**	4.38
Mentoring topics included . . . (percentage)				
Strategies for teaching literacy	23.8	8.2	15.5***	4.11
Strategies for teaching math	10.6	5.3	5.3	2.80
Strategies for teaching other subjects	11.7	8.6	3.1	4.43
Increasing content area knowledge	16.3	10.4	5.9	4.64
Selecting or adapting curriculum materials	16.8	6.7	10.1***	3.21
Teaching or aligning curriculum to meet state or district standards	20.0	10.2	9.9**	4.22
Aligning local curriculum assessment to state standards	13.8	8.2	5.6	3.60
Setting instructional goals and determining ways to achieve them	21.8	10.5	11.3***	4.33
Preparing students for standardized tests	16.0	8.1	7.9*	4.37
Using assessments to inform teaching	19.9	11.9	8.0**	4.07
Preparing lesson plans or other instructional activities	20.1	8.1	12.0***	4.41
Providing differentiated instruction to meet student needs	19.4	10.1	9.3**	4.77
Received release time for mentoring (percentage)	3.6	0.7	2.9***	0.90
Release time for mentoring (hours per week)	1.4	0.8	0.6	0.72
Mentoring outside of specified contract hours (hours per week)	1.8	0.8	1.1	0.71
Teachers mentored (number)	2.2	0.8	1.3**	0.50
Frequency of scheduled meetings (number per week per teacher)	0.3	0.3	0.0	0.12
Duration of each scheduled meeting (minutes)	13.9	7.6	6.3*	3.35
Informal contact with all teachers (minutes per week)	68.0	38.0	29.9	28.19
Total contact with all teachers (minutes per week)	182.9	43.1	139.7**	63.83
Mentoring activities included . . . (percentage)				
Observing teaching	5.6	1.4	4.2***	1.20
Meeting with teachers one-on-one	24.2	9.9	14.3***	4.52
Meeting in small groups or clusters	4.9	1.5	3.4***	0.91
Modeling a lesson	24.0	6.5	17.5***	5.06
Coteaching a lesson	3.6	0.9	2.8***	0.83
Writing evaluations	17.5	3.5	14.1***	5.20
During most recent full week, scheduled time spent . . . (minutes)				
Observing teaching	39.7	41.4	-1.7	31.53
Meeting with teachers one-on-one	37.2	19.2	18.0	11.16
Meeting in small groups or clusters	32.5	15.2	17.3*	9.17
Modeling a lesson	24.2	12.8	11.4	11.03
Coteaching a lesson	16.9	8.0	8.8	8.36
Writing evaluations	45.1	6.4	38.7**	13.74

Note: N=215 to 234 teachers per outcome

[a] Means are regression adjusted.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

We also found significant differences in the time spent, and the specific activities conducted, as a mentor. Veteran Chicago TAP teachers were significantly more likely than veteran control teachers to receive release time from their regular professional duties to perform their mentoring, although the amount of release time received did not differ significantly between the two groups of teachers. Chicago TAP also affected the number of teachers mentored, with veteran Chicago TAP teachers mentoring about two teachers on average, compared to one teacher for veteran control teachers. Compared to veteran control teachers, veteran Chicago TAP teachers had longer scheduled meetings with each teacher on average and had more total contact each week with all the teachers they mentored.

According to findings on mentoring activities, veteran Chicago TAP teachers were significantly more likely to have recently spent time meeting with teachers in small groups and writing evaluations. Compared to veteran control teachers, veteran Chicago TAP teachers spent on average 17 more minutes of scheduled time during the most recent full week meeting in small groups or clusters, and 39 more minutes writing evaluations. We also found that veteran Chicago TAP teachers were more likely to observe other teachers, meet with teachers one-on-one, model a lesson, and coteach a lesson as part of their mentoring responsibilities, but none of the differences in scheduled time spent on these activities during the most recent week of full teaching was statistically significant.

We found few significant differences between Chicago TAP and non-TAP schools in the percentage of veteran teachers having leadership roles or responsibilities other than those responsibilities like mentoring and observation that are explicitly part of the Chicago TAP model (see Table III.6).[15] Thirty-nine percent of veteran Chicago TAP teachers and 45 percent of veteran control teachers reported having such leadership tasks, a difference that was not statistically significant. Veteran Chicago TAP teachers were significantly more likely than veteran control teachers to serve as a lead teacher. We also found that veteran Chicago TAP teachers were more likely than control teachers to receive a pay increase in association with their nonmentoring leadership roles and responsibilities. Other reported differences in leadership roles and responsibilities, primarily in areas like governance and decision making, were not statistically significant.

c. Observation and Feedback

As part of establishing instructionally focused accountability, the Chicago TAP model calls for observations of teachers conducted by the Chicago TAP leadership team, which consists of the principal, lead teachers, and mentor teachers. During the first year of implementation, the program model allows for practice observations during the first semester and prescribes at least two official observations during the second semester; the official observations are used to determine performance-based compensation (Chicago Board of Education and Chicago Teachers Union 2007; NIET 2008).

[15] Readers should note that we conduct a large number of hypothesis tests, each of which has a probability of falsely rejecting the null hypothesis of no difference (denoted as the significance level, usually 5 percent). When conducting large numbers of hypothesis tests, it is likely that at least some relationships will appear "statistically significant" purely by chance. For example, at a 5 percent significance level, one in 20 independent test results will appear statistically significant even if there is no underlying relationship. Therefore, isolated significant results are suggestive but not conclusive evidence of a relationship.

We found that in Chicago TAP schools, there were more frequent observations by teachers in leadership roles than occurred in control schools (see Table III.7). During the 2009–2010 school year, a mentor, coach, or lead teacher observed teachers in treatment schools more than three times, on average; in control schools, there were about two observations by comparable teacher-leaders. However, the frequency of observations by school administrators was similar in the two groups.

We did not find evidence that Chicago TAP affected how often teachers received feedback. In particular, there were no statistically significant differences between Chicago TAP and control teachers in how often they were given feedback in the following contexts: as part of a formal evaluation, outside of an evaluation, or specifically pertaining to lesson plans.[16]

Table III.6. Other Leadership Roles and Responsibilities (teachers with at least five years of experience)

Outcome	Chicago TAP Mean[a]	Control Mean[a]	Difference	Standard Error
Had other leadership roles or responsibilities beyond mentoring (percentage)	39.3	44.7	-5.5	7.57
Other leadership roles included . . . (percentage)				
Being a lead teacher	2.2	0.8	1.4***	0.48
Being a department head or chair	5.2	7.7	-2.5	2.46
Being a grade-level lead teacher	8.6	13.0	-4.4	5.03
Being on a school improvement team	15.6	20.9	-5.4	5.07
Being on a schoolwide committee/task force	12.6	14.2	-1.6	6.32
Other leadership responsibilities included . . . (percentage)				
Setting school policies	9.7	17.8	-8.1	6.80
Developing curriculum	14.3	15.1	-0.8	4.19
Reviewing/selecting curriculum	15.2	15.6	-0.4	5.73
Providing input on improving facilities/technology	10.9	15.3	-4.4	5.91
Providing professional development activities	20.5	17.3	3.2	4.83
Developing standards	10.2	8.7	1.5	5.34
Evaluating teachers	0.7	0.0	0.7	0.44
Associated with these other leadership roles and responsibilities, received . . . (percentage)				
Credit toward certification	1.2	1.0	0.1	0.48
Pay increase	16.1	1.5	14.6***	5.06

Note: N = 228 to 232 teachers per outcome.

[a] Means are regression adjusted.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

[16] Responses were top-coded at five occurrences during the school year.

Table III.7. Observation and Feedback

Outcome	Chicago TAP Mean[a]	Control Mean[a]	Difference	Standard Error
Frequency of observation (number in 2009–2010)				
Observation by principal or assistant principal	2.9	3.1	-0.1	0.32
Observation by mentor, coach, or lead teacher	3.2	2.4	0.8*	0.39
Frequency of feedback (number in 2009–2010)				
Feedback as part of a formal evaluation	2.5	2.3	0.2	0.25
Feedback outside of a formal evaluation	2.6	2.3	0.3	0.27
Feedback on lesson plans	2.2	2.5	-0.3	0.41

Note: N = 277 to 281 teachers per outcome.

[a] Means are regression adjusted.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

2. Professional Development

Chicago TAP offers teachers "ongoing applied professional growth" through school-based professional development during the school day. Through weekly cluster meetings, as well as other interactions among lead and mentor teachers and career teachers, the program provides opportunities to collaborate on improving the quality of instruction and to learn new research-based instructional strategies for increasing academic achievement.

We found few significant differences in professional development received by treatment and control teachers (see Table III.8). In both groups, most teachers participated in professional development activities addressing a range of topics. We did not find any statistically significant differences in topic areas covered.

The majority of teachers in both groups characterized their professional development activities as useful and reported having incorporated what they learned into their teaching. A significantly higher percentage of treatment than control teachers reported being more satisfied with professional development in the 2009–2010 school year than in previous years. We did not find a pattern of significant differences between Chicago TAP and non-TAP schools in the extent to which teachers received compensation or benefits in association with professional development activities.

The lack of impacts found for professional development suggests that cluster group meetings, which were an integral part of Chicago TAP, may have been characterized by teachers as mentoring and not as professional development. Alternatively, it may be the case that similar types of support were provided in schools that were not implementing Chicago TAP.

Table III.8. Professional Development Received

Outcome	Chicago TAP Mean[a]	Control Mean[a]	Difference	Standard Error
Participated in professional development activities that addressed . . . (percentage)				
Strategies for teaching literacy	98.9	99.5	-0.6	0.41
Strategies for teaching math	83.3	76.8	6.6	5.74
Strategies for teaching other subjects	66.1	65.4	0.6	6.06
Increasing content area knowledge	87.4	84.5	2.9	4.97
Selecting or adapting curriculum materials	67.4	69.2	-1.8	7.00
Teaching or aligning curriculum to meet state or district standards	81.4	78.1	3.3	5.52
Aligning local or teacher-developed curriculum assessment to state standards	70.2	71.4	-1.2	7.39
Setting instructional goals and determining ways to achieve them	78.2	80.7	-2.5	5.13
Preparing students for standardized tests	70.2	71.0	-0.8	7.10
Using assessments to inform teaching	88.8	89.2	-0.5	3.97
Preparing lesson plans or other instructional activities	77.4	77.1	0.4	6.08
Providing differentiated instruction to meet student needs	86.7	88.9	-2.3	3.40
Responded that professional development in 2009–2010 . . . (percentage)				
Was useful to their teaching	83.8	85.0	-1.2	5.15
Was more satisfactory than in previous years	34.7	22.3	12.4**	5.35
Had been implemented in their teaching	92.0	91.3	0.7	2.73
In association with professional development, received . . . (percentage)				
Scheduled nonteaching time in contract year	83.2	72.1	11.1**	4.74
Other release time from teaching	46.1	37.2	8.8	7.19
Stipend	65.6	66.4	-0.8	7.32
Fee reimbursement	19.0	18.7	0.2	5.78
Travel or expense reimbursement	6.1	14.9	-8.8**	3.81
Course credits toward certification	51.2	48.5	2.7	4.07
Pay increase	25.5	20.5	5.1	4.87
Recognition or higher ratings on an annual teacher evaluation	24.2	19.3	4.8	6.40

Note: N = 270 to 283 teachers per outcome.

[a] Means are regression adjusted.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

3. Compensation

The Chicago TAP model can affect teacher pay through two routes: (1) multiple career paths (bonuses for serving as mentor or lead teacher) and (2) performance-based compensation (bonuses for scoring high marks on classroom observations and/or classroom- and school-level value added). Chicago TAP lead and mentor teachers receive an additional $15,000 and $7,000, respectively, as compensation for assuming more responsibility. Performance-based compensation provides bonuses to teachers who demonstrate their skills through classroom evaluations and who increase their students' academic achievement growth over the course of the year. In the first year of the program's implementation, 25 percent of the performance award was to be based on teacher performance as assessed through classroom observations and 75 percent on schoolwide student

achievement growth. Payments were expected to average $500 per teacher based on classroom observations and up to $1,500 per teacher based on value added to student achievement growth (NIET 2008).

At the time the teacher survey was administered, Chicago TAP teachers had not yet received performance bonuses from the program. We describe below teacher expectations about compensation.

We found that teacher reports were consistent with the program having been implemented as intended. Consistent with Chicago TAP's emphasis on multiple career paths, Chicago TAP teachers were significantly more likely than control teachers to expect additional compensation for leadership (see Table III.9). Nineteen percent of Chicago TAP teachers expected to receive additional pay for leadership roles and responsibilities, compared to 4 percent of control teachers. The average annual amount of leadership pay expected by Chicago TAP teachers exceeded that of non-TAP teachers by $1,412.

We also found significant differences regarding nonleadership pay, with the largest differences occurring in areas emphasized by Chicago TAP. For a teacher incentive to work effectively, teachers must be aware that they are eligible to receive pay conditional on their performance. Seventy-eight percent of Chicago TAP teachers reported being eligible for additional compensation based on instructional performance or student achievement, compared to 14 percent of control teachers. Expectations of actually receiving such compensation differed as well: 44 percent of treatment teachers expected to receive additional compensation for instructional performance or student achievement growth, compared to 3 percent of control teachers. Differences in eligibility for and expectations of compensation for other nonleadership reasons were smaller and in most cases not statistically significant.

The amount of nonleadership compensation expected differed significantly for Chicago TAP and non-TAP schools as well. On average, Chicago TAP teachers expected to receive an annual $871 as additional compensation for nonleadership reasons, compared to $180 in additional pay expected by control teachers.

4. Teacher Attitudes

As a program offering performance-based pay, Chicago TAP can affect the climate of the school in many ways. It can create competition or jealousy, and thereby potentially undermine collegiality. Alternatively, it can build collegiality by rewarding the group's performance, or improve morale by bringing colleagues together, either in shared activities or in the larger effort to implement a new program. To assess Chicago TAP's effect on school climate, we included questions on the teacher survey about teachers' satisfaction with and attitudes toward their school.

Teacher survey responses suggested that Chicago TAP did not change school climate: the majority of teachers in both Chicago TAP and non-TAP schools reported collaborative, supportive environments (see Table III.10). Seventy-nine percent of Chicago TAP teachers and 83 percent of control teachers reported being satisfied with the supportive atmosphere among faculty and collaboration with colleagues. The majority of both groups agreed with the statement that their principal worked to create a sense of community at their school. In neither case was the difference between the groups statistically significant.

Table III.9. Compensation

Outcome	Chicago TAP Mean[a]	Control Mean[a]	Difference	Standard Error
Academic-year base salary ($)	62,137	62,029	107	1,277.49
Base salary included leadership compensation (percentage)	12.4	11.0	1.4	3.91
Additional compensation was expected for leadership (percentage)	19.3	4.4	14.9***	4.22
Expected amount of additional compensation for leadership ($)	1,767	355	1,412***	383.23
Eligible for additional nonleadership compensation (percentage)	81.7	37.9	43.7***	5.90
Eligible for additional nonleadership compensation based on . . . (percentage)				
Instructional performance	71.7	11.0	60.7***	7.34
Student achievement growth	64.6	11.9	52.7***	6.40
Instructional performance or student achievement growth	78.2	14.1	64.1***	6.35
Subject matter taught	19.2	10.2	9.0*	4.98
Student population taught	13.2	10.5	2.7	3.74
Professional development	32.5	32.2	0.2	6.06
University courses taken	14.4	20.3	-5.9	5.08
Expected or had received additional nonleadership compensation (percentage)	53.9	13.6	40.3***	5.44
Expected or had received additional nonleadership compensation based on . . . (percentage)				
Instructional performance	35.8	2.4	33.5***	5.81
Student achievement growth	30.3	1.3	28.9***	4.99
Instructional performance or student achievement growth	44.7	3.1	41.6***	5.77
Subject matter taught	0.0	0.0	0.0	0.00
Student population taught	0.0	0.0	0.0	0.00
Professional development	15.7	9.0	6.8*	3.91
University courses taken	0.0	0.0	0.0	0.00
Expected amount of additional nonleadership compensation ($)	871	181	690***	159.91
Expected additional compensation from an outside job (percentage)	11.0	4.6	6.4*	2.79

Note: N = 240 to 274 teachers per outcome.

[a] Means are regression adjusted.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

When we examined other teacher attitudes, we found few significant differences between Chicago TAP and non-TAP schools (see Table III.10). Treatment teachers were significantly more likely than control teachers to report being satisfied with their salary and benefits, and fewer treatment than control teachers agreed with the statement that their principal is strongly committed to shared decision making. But other differences were not statistically significant, and among both treatment and control teachers, positive attitudes about their principals and other aspects of teaching were prevalent.

Table III.10. Teacher Attitudes

Outcome	Chicago TAP Mean[a]	Control Mean[a]	Difference	Standard Error
Satisfied with . . . (percentage)				
Supportive atmosphere/collaboration with colleagues	78.6	83.2	-4.6	4.75
Administration support	76.1	79.7	-3.6	7.21
Policies/practices input	74.4	80.1	-5.7	6.07
Classroom autonomy	88.6	86.2	2.4	4.24
Professional development opportunities	87.9	85.9	2.0	3.53
Caliber of colleagues	79.0	85.0	-6.1	5.37
Salary and benefits	90.4	84.6	5.7**	2.35
Leadership opportunities	79.0	80.6	-1.6	4.29
School policies	72.7	72.1	0.6	5.31
District policies	57.0	54.5	2.5	6.13
Agreed that the principal . . . (percentage)				
Works to create a sense of community	73.6	79.9	-6.3	7.17
Is strongly committed to shared decision making	73.9	84.0	-10.2*	5.99
Promotes parent/community involvement	85.8	86.4	-0.6	5.58
Supports and encourages risk taking	71.6	70.8	0.8	6.84
Is willing to make changes	88.7	87.3	1.4	2.99
Strongly supports most changes	81.2	85.0	-3.8	4.49
Encourages trying new instructional methods	88.7	88.8	0.0	4.48

Note: N = 270 to 278 teachers per outcome.

[a] Means are regression adjusted.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

This page has been left blank for double-sided copying.

IV. IMPACTS ON STUDENT ACHIEVEMENT

According to CPS, Chicago TAP (2009) was designed to support and develop high-quality teaching, which in turn would boost student learning. Consequently, students' ISAT scores are the main outcomes of interest for the study. We focus on tested grades and subjects: math and reading in grades four through eight as well as science in grades four and seven. We examined outcomes in each year of the program's four-year rollout in the district. This chapter explains our findings.

A. Experimental Evidence

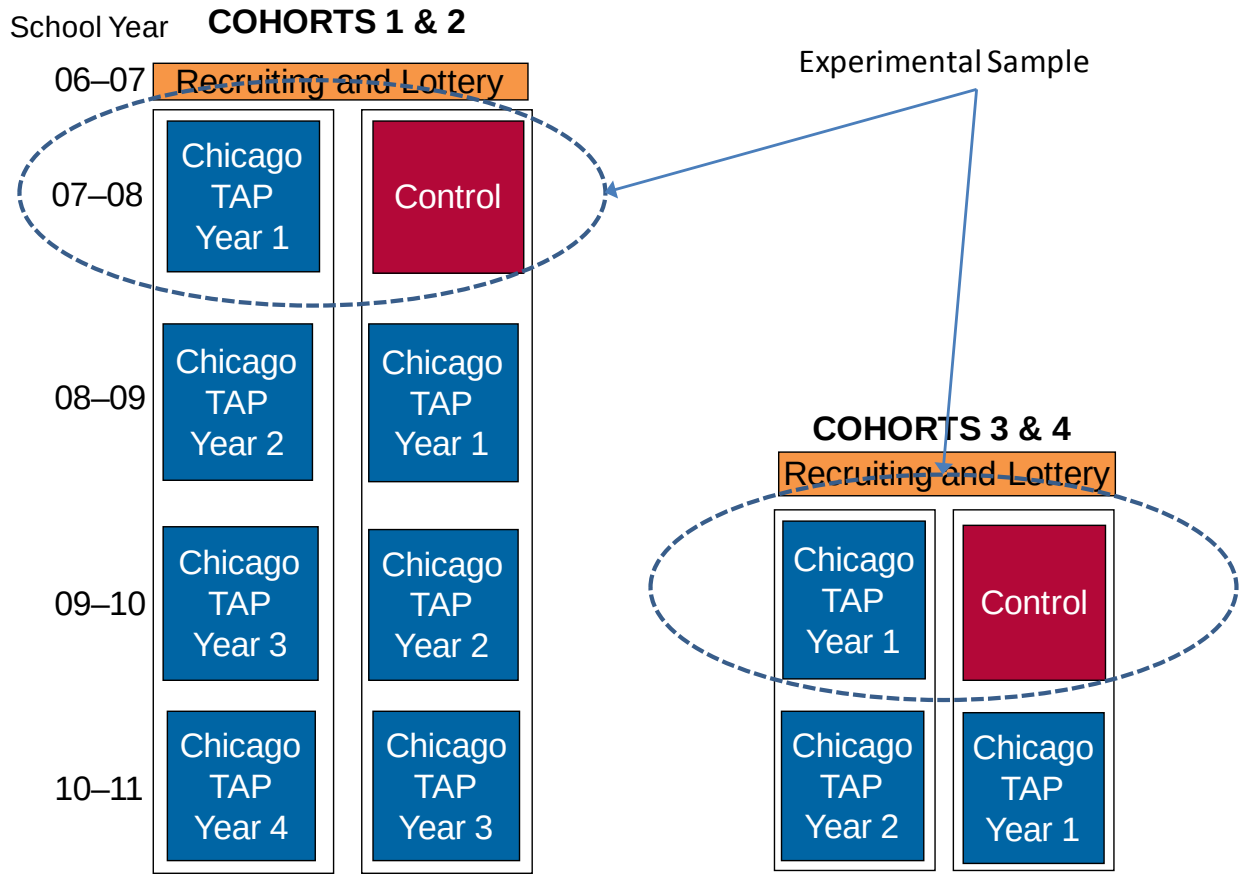
As described in Chapter II, we used a lottery to assign schools randomly to either a treatment group, which would implement Chicago TAP right away, or a control group, which would delay implementation by one year. The random assignment design allowed us to generate unbiased estimates of the impact of Chicago TAP in the first year of implementation by comparing outcomes for the two groups. Tests were administered in March of each year. We ran two lotteries, one in spring 2007 assigning cohorts 1 and 2, and one in spring 2009 assigning cohorts 3 and 4. We combined data from spring 2008 pertaining to the schools assigned in the 2007 lottery with data from spring 2010 pertaining to schools assigned in 2009 to form our main experimental analysis data set. Thus the treatment group of schools implementing Chicago TAP consists of cohort 1 in 2008 and cohort 3 in 2010, while the control group consists of schools in cohort 2 in 2008 and cohort 4 in 2010. These groups are shown in Figure IV.1. The differences in these years between Chicago TAP and control schools represent the impacts as of spring of the first year of implementing Chicago TAP.

Overall, we found that in March of the first year of implementation, test scores for students in Chicago TAP schools were statistically indistinguishable from those of students in control schools (Table IV.1). The observed differences were less than 3 percent of a standard deviation for math and reading, but for science the students in Chicago TAP schools scored nearly 4 ISAT points, or 14 percent of a standard deviation, higher than students in control schools.[17] However, the impact on science scores was estimated from a relatively small sample because science test data were not available before 2009 and the test was only administered in two grades (four and seven). Therefore, even an effect size of 0.14 is within the margin of error.[18]

[17] Standard deviations used to calculate effect sizes are derived from the distribution of the full sample for each outcome. The standard deviations for reading, math, and science were approximately 32, 35, and 31 points, respectively. Thus, a one-point ISAT difference translates into approximately 3 percent of a standard deviation, or 1.2 percentile points.

[18] For the test of whether the impact on science scores is different from zero the p-value, which is the probability of observing this result if the true impact were zero, was 0.117. Accordingly, a 90 percent confidence interval still includes zero (no impact).

Figure IV.1. Experimental Sample in First Year of Implementation



As noted in Chapter II, each of the estimates is derived from a regression model in which we control for students' prior-year achievement in both subjects (math and reading) and other factors that might be related to test scores. These include family poverty (eligibility for free/reduced-price lunch), special needs (whether an Individualized Education Program was in place), language (whether limited English proficient), race/ethnicity, grade level, and age (whether older than the normal age for a grade). We accounted for the clustering of students or teachers within schools by estimating robust standard errors. We refer to this as the benchmark regression model. Alternatives to the benchmark model are presented later in this chapter.

Table IV.1. Impacts on ISAT Scores by Subject, First Year of Implementation

Subject	Chicago TAP Mean[a]	Control Mean[a]	Impact	Standard Error	Effect Size
Reading	221.3	221.0	0.3	0.7	0.01
Math	233.4	234.3	-0.9	0.9	-0.03
Science	204.3	200.4	3.7	2.3	0.14

Notes: First-year outcomes measured in March. N = 7,661 students (reading), 7,656 students (math), 1,717 students (science). Data for math and reading are from spring 2008 (cohorts 1 and 2) and from spring 2010 (cohorts 3 and 4). Science data are from spring 2010 only (cohorts 3 and 4). Science scores were not available for 2008.

[a] Means are regression adjusted.

Differences are not statistically significant at the 10 percent level.

We estimated the results separately by grade level and did not find any statistically significant impacts. The results, shown in Table IV.2, are based on students in grades four through eight because testing begins at the end of grade three and a pretest was required to estimate regression-adjusted impacts. As with the full sample results, only the score differences on the science test were noticeable and those again were not statistically significant.

Table IV.2. Impacts on ISAT Scores by Subject and Grade Level, First Year of Implementation

Subject, Grade	Chicago TAP Mean[a]	Control Mean[a]	Impact	Standard Error	Effect Size
Reading					
Grade 4	201.4[b]	200.8[b]	0.6	2.3	0.03
Grade 5	211.1[b]	212.5[b]	-1.4	1.1	-0.06
Grade 6	223.5	221.3	2.1	1.8	0.09
Grade 7	227.9	229.1	-1.2	1.5	-0.05
Grade 8	238.4	236.9	1.4	0.9	0.08
Math					
Grade 4	209.4	211.0	-1.6	2.1	-0.07
Grade 5	219.9	222.5	-2.7	2.0	-0.11
Grade 6	233.8	233.4	0.3	2.1	0.01
Grade 7	242.9	243.9	-1.0	1.8	-0.04
Grade 8	255.8	255.2	0.6	1.8	0.03
Science					
Grade 4	188.4	184.7[b]	3.7	4.0	0.15
Grade 7	218.7	214.7	3.9	2.7	0.18

Note: First-year outcomes measured in March. N = 1,381 to 1,678 students per grade in math and reading, 840 to 876 students per grade in science.

[a] Means are regression adjusted.
[b] Average score is classified as "Below Standards." All other scores are in the category "Meets Standards."

Differences are not statistically significant at the 10 percent level.

Because the ISAT is vertically scaled, the scores should be higher for each grade level, and the scores within each grade can be compared to state-set criteria for whether the students are meeting standards for their grade. The average scores shown in Table IV.2 were in the category for "meets standards" for every grade and subject except reading in grades four and five (Chicago TAP and control schools) and science in grade four (in control schools only), which were below standards.[19]

We tested whether the experimental results were robust to different methods and assumptions. To test the sensitivity of the findings, we reestimated the test score impacts several times, with each new model making one change to the benchmark model used in Table IV.1. The sensitivity results are summarized in Tables IV.3, IV.4, and IV.5 for reading, math, and science scores, respectively. For the sensitivity tests, we estimated impacts with the following variations: separate pretest effects estimated by grade, a nonlinear relationship (quadratic and cubic) between pretest and posttest, a model controlling for pretest in the same subject (reading pretest for reading posttest and math pretest for math or science posttest) but not the "opposite" subject, use of limited or no covariates, outcomes standardized within grade by subtracting the districtwide mean and dividing by the

[19] The range of scores that defines each category can be found in the 2011 guide to the ISAT issued by the state (Illinois State Board of Education 2010).

districtwide standard deviation, and use of a specification that treated the school effect as a random variable. One further variation for estimating impacts involved use of specifications that corrected for pretest measurement error in one of three ways: by calculating gain score as the posttest minus pretest and using that as the outcome instead of posttest; by using the opposite subject pretest as an "instrument" for the same subject pretest; and by using a measurement error correction method that accounted for test score reliability being less than 1.0 (since the true reliabilities were not known, we used 0.9 and 0.8). For the science results (Table IV.5), we did not estimate a gain model because the pretest was only available in a different subject, math.

Table IV.3. Sensitivity of Impact on ISAT Reading Scores, First Year of Implementation

Model Specification	Chicago TAP-Control Difference	Standard Error	Effect Size	Sample Size (students)
Benchmark	0.3	0.7	0.01	7,661
Covariates				
Separate pretest effect by grade	0.6	0.8	0.02	7,661
Pretest squared and cubed	0.2	0.9	0.01	7,661
Omit math pretest	0.3	0.8	0.01	7,684
No pretest: grades 3–8	1.0	0.9	0.03	10,065
No pretest: grades 4–8 only	1.3	1.0	0.05	8,248
Pretest only (no other covariates)	0.1	0.8	0.00	7,661
Scores standardized within grade (z-score)	0.00	0.03	0.01	7,661
Alternative variance estimation method				
Random effects (RE)	0.3	0.7	0.01	7,661
RE with school characteristics	0.1	0.7	0.00	7,661
Measurement error correction				
Instrumental variables	0.3	0.8	0.01	7,661
Gain model	-0.3	0.8	-0.02	7,684
Errors-in-variables model, reliability = .9	0.1	0.8	0.01	7,684
Errors-in-variables model, reliability = .8	-1.0	1.0	-0.04	7,684

Note: First year outcomes measured in March.

Differences are not statistically significant at the 10 percent level.

Under these alternative models the reading and math impact estimates remain small and statistically insignificant. The science impact estimates, which were positive but insignificant in the benchmark model, become statistically significant under select alternative models (Table IV.5). Specifically, the manner in which we adjust for prior achievement has an influence on the size of the estimate. When we omitted a control for prior reading scores the impact estimate was closer to 5 ISAT points, or 17 percent of a standard deviation, which is large enough to be statistically significant (in other words, too large to have been plausibly generated by chance).

Table IV.4. Sensitivity of Impact on ISAT Math Scores, First Year of Implementation

Model Specification	Chicago TAP-Control Difference	Standard Error	Effect Size	Sample Size (students)
Benchmark	-0.9	0.9	-0.03	7,656
Covariates				
Separate pretest effect by grade	-0.6	1.0	-0.02	7,656
Pretest squared and cubed	-0.6	1.0	-0.02	7,656
Omit reading pretest	-0.8	0.9	-0.03	7,672
No pretest: grades 3–8	0.1	1.2	0.00	10,071
No pretest: grades 4–8 only	0.3	1.3	0.01	8,250
Pretest only (no other covariates)	-0.9	1.0	-0.03	7,656
Scores standardized within grade (z-score)	0.00	0.03	-0.04	7,656
Alternative variance estimation method				
Random effects (RE)	-0.9	1.2	-0.03	7,656
RE with school characteristics	-1.2	1.2	-0.04	7,656
Measurement error correction				
Instrumental variables	-0.8	0.9	-0.03	7,656
Gain model	-0.7	1.0	-0.02	7,917
Errors-in-variables model, reliability = .9	-1.0	0.9	-0.07	7,672
Errors-in-variables model, reliability = .8	-2.0	2.1	-0.04	7,685

Note: First-year outcomes measured in March.

Differences are not statistically significant at the 10 percent level.

Table IV.5. Sensitivity of Impact on ISAT Science Scores, First Year of Implementation

Model Specification	Chicago TAP-Control Difference	Standard Error	Effect Size	Sample Size (students)
Benchmark	3.7	2.3	0.14	1,717
Covariates				
Separate pretest effect by grade	4.2	2.9	0.15	1,717
Pretest squared and cubed	3.8	2.4	0.14	1,717
Omit reading pretest	4.7*	2.4	0.17	1,723
No pretest	5.4*	3.1	0.20	1,807
Pretest only (no other covariates)	3.2	2.4	0.12	1,717
Scores standardized within grade (z-score)	0.10	0.09	0.16	1,717
Alternative variance estimation method				
Random effects (RE)	2.8	2.5	0.10	1,717
RE with school characteristics	3.2	2.8	0.11	1,717
Measurement error correction				
Instrumental variables	4.7*	2.4	0.17	1,717
Errors-in-variables model, reliability = .9	4.7*	2.5	0.25	1,723
Errors-in-variables model, reliability = .8	-2.0	2.4	-0.08	1,723

Note: First-year outcomes measured in March.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

The findings above (Tables IV.1 through IV.5) focused on 2008 and 2010 only because those were the only years during which a randomized control group was not implementing Chicago TAP. Because standardized testing takes place in March of each year, seven months into the nine-month school year, the analysis necessarily focuses on just the first seven months of the first year of implementation, a period during which the schools might not have had a chance to fully implement and become accustomed to the program.

To understand the longer-term effects of Chicago TAP, it was important to examine results from schools that had been implementing Chicago TAP for a longer period. The random assignment component of the study design does not allow for comparisons of Chicago TAP to non-Chicago TAP schools for longer periods of implementation because the control groups began implementing Chicago TAP in the second school year after their initial assignment.

However, we did produce experimental comparisons of schools that had different amounts of experience implementing Chicago TAP at any point in time. We compared schools with one versus two years of implementation (cohort 1 versus cohort 2 in 2009 and cohort 3 versus cohort 4 in 2011), with two versus three years of implementation (cohort 1 versus cohort 2 in 2010), and with three versus four years (cohort 1 versus cohort 2 in 2011). The results, shown in Table IV.6, suggest that there was no advantage associated with having spent one extra year implementing Chicago TAP in any of the three subjects tested. The only statistically significant finding was that math scores were lower for Chicago TAP schools in their second year of implementation than scores of Chicago TAP schools still in their first year, a difference of 1.6 ISAT points.

Table IV.6. Effect of One Additional Year of Chicago TAP Implementation on March ISAT Scores

Implementation Years and Subject	More Chicago TAP Mean[a]	Less Chicago TAP Mean[a]	Difference	Standard Error	Effect Size	Sample Size
Second year vs. first year[b]						
Reading	220.6	221.0	-0.5	0.6	-0.02	7695
Math	234.4	236.0	-1.6*	0.9	-0.05	7680
Science	203.1	204.0	-0.9	1.6	-0.03	1723
Third year vs. second year[c]						
Reading	220.8	220.5	0.3	1.1	0.01	3,123
Math	234.7	235.5	-0.8	1.5	-0.03	3,131
Science	200.1	199.4	0.7	3.0	0.03	1,227
Fourth year vs. third year[d]						
Reading	222.4	223.5	-1.2	1.1	-0.04	2,908
Math	236.7	238.7	-2.0	1.5	-0.07	2,909
Science	203.8	205.4	-1.6	2.4	-0.06	1,104

Note: Outcomes measured in March of each year.

[a] Means are regression adjusted.
[b] "More Chicago TAP" means pertain to cohort 1 in 2009 and cohort 3 in 2011. "Less Chicago TAP" means pertain to cohort 2 in 2009 and cohort 4 in 2011.
[c] "More" and "less" Chicago TAP means refer to cohorts 1 and 2, respectively, in 2010.
[d] "More" and "less" Chicago TAP means refer to cohorts 1 and 2, respectively, in 2011.

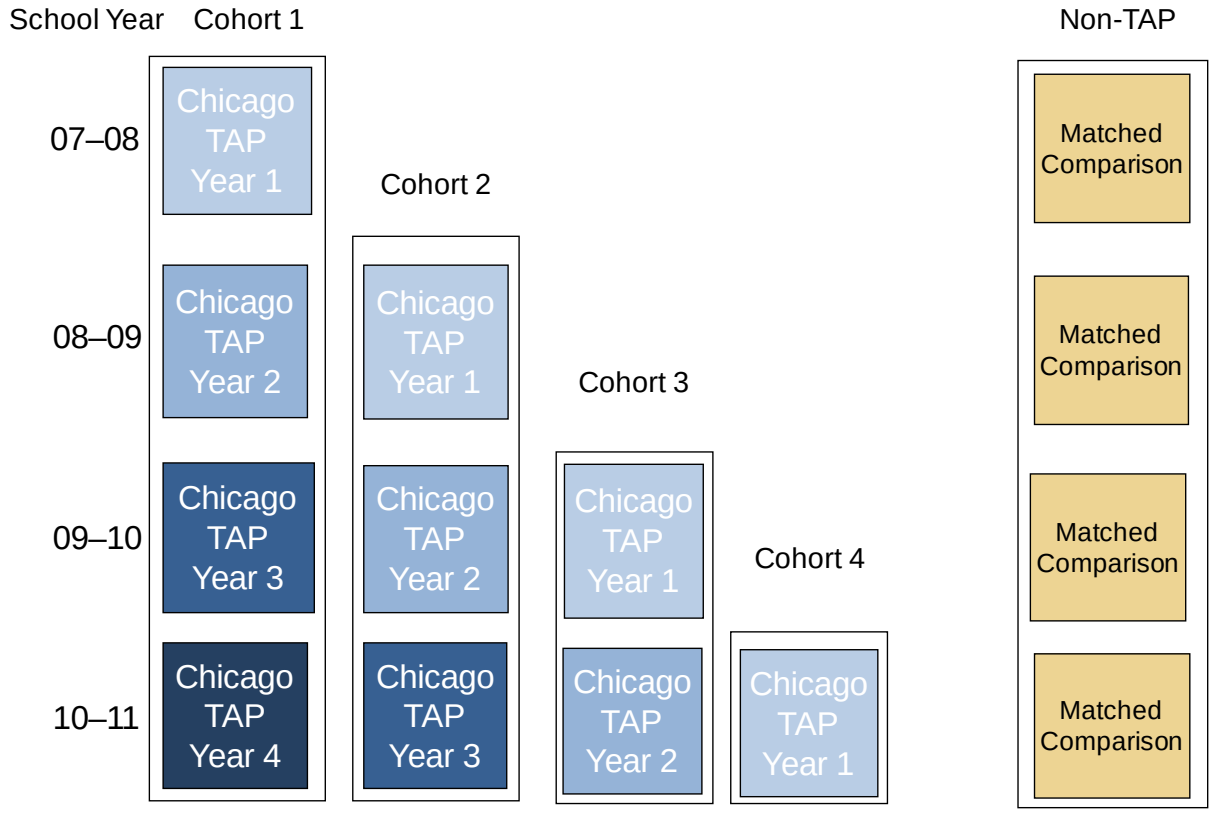
* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

B. Quasi- Experimental Evidence

All of the previous results are experimental, meaning that they focus on schools that had been randomly assigned to their Chicago TAP cohort status. Most of the results are focused on the period immediately following random assignment, even though the schools continued implementing Chicago TAP until 2010–2011. We also conducted quasi-experimental analyses that compared Chicago TAP schools to matched comparison schools that had not been assigned to implement Chicago TAP at all. Such an analytic approach allowed us both to take advantage of the more than 300 additional schools in Chicago that could serve as potential points of comparison, and to make use of several years of data, rather than being constrained to the first year of Chicago TAP implementation. In order to account for the underlying differences between schools that choose to adopt Chicago TAP and those that do not, we used propensity score matching (see Chapter II and Appendix A for more detailed information on the methodology).

Using a longitudinal data set with test scores from 2007–2008 to 2010–2011, we compared test scores of students in each Chicago TAP school in each year to test scores of students in matched comparison schools. This aspect of the study design is shown in Figure IV.2. The results are shown in Table IV.7.

Figure IV.2. Quasi- Experimental Design



Note: Shading indicates years of experience implementing Chicago TAP, ranging from one year (lightest shading) to four years (darkest shading).

Table IV.7. Summary of Quasi- Experimental Findings: Impacts on March ISAT Scores

	Reading		Math		Science	
Model	Impact	Standard Error	Impact	Standard Error	Impact	Standard Error
1. Overall Chicago TAP effect	-0.1	0.3	0.1	0.4	1.1	0.8
2. Chicago TAP effects by year of implementation (school experience)						
First year (all four cohorts)	-0.2	0.4	0.4	0.5	1.3	1.2
Second year (cohorts 1–3)	-0.1	0.5	-0.1	0.6	0.1	1.0
Third year (cohorts 1 & 2)	0.4	0.6	0.2	0.8	1.9	1.7
Fourth year (cohort 1)	-0.7	0.9	-1.3	1.4	2.1	3.6
3. Chicago TAP effects by school year (district experience)						
Year 1: 2007–2008 (cohort 1)	-0.6	0.8	0.2	1.1	NA	NA
Year 2: 2008–009 (cohorts 1 & 2)	-0.4	0.6	0.3	0.9	-0.4	1.7
Year 3: 2009–2010 (cohorts 1–3)	1.3*	0.7	1.0	0.9	1.6	1.9
Year 4: 2010–2011 (all four cohorts)	-0.7	0.5	-0.5	0.7	0.3	1.2
4. Chicago TAP effects by cohort						
Cohort 1 (all four years)	-0.4	0.4	-0.4	0.6	1.7	1.6
Cohort 2 (2009-2011)	0.5	0.5	1.2*	0.6	0.7	1.0
Cohort 3 (2010-2011)	0.3	0.6	-0.4	0.7	1.3	1.4
Cohort 4 (2011 only)	-1.6***	0.6	-0.1	1.0	0.0	1.8
Sample size (student-year combinations)	83,214		83,125		29,364	
Sample size (school-grade-year combinations)	2,092		2,092		764	

Note: Analyses use nearest-five-neighbors matching. NA = data not available.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

The top row of Table IV.7 provides the summary results and suggests that the overall differences between Chicago TAP and non-TAP schools are statistically insignificant for reading, math, and science. We also estimated the relationship between Chicago TAP implementation and test scores by year of implementation (model 2), to test the hypothesis that as schools spend more time implementing Chicago TAP they will be more successful. These results do not support such a hypothesis. The impact on reading, for example, is less than one ISAT point and statistically indistinguishable from zero whether the school had been implementing Chicago TAP for one, two, three, or four years. For math and science, the impact estimates were also statistically insignificant for any level of school experience with Chicago TAP. The point estimates are larger (positive and negative) for more years of implementation, but these effects are based on one or two cohorts of schools and are likely due to chance because they are not statistically significant.

We also tested the hypothesis that Chicago TAP becomes more effective as the district overall has more experience implementing it (Table IV.7, model 3). In 2007, the district had just hired new staff to implement Chicago TAP in the first 10 schools. By 2011, the program had been implemented in more than 40 schools in the district. If the program had become more effective over time, the impacts would become stronger in later years. The results show that impacts were largest in

the third year, 2009– 2010, with the estimate being significant for reading (1.3 ISAT points, which is equivalent to 5 percent of a standard deviation) and insignificant for math and science. The estimated science effect was 1.7 ISAT points, but because science tests were taken only by fourth- and seventh-grade students and the scores were available only in 2009 and later, the estimate was based on a smaller sample, which makes it more difficult to detect an impact unless it is very large. A standard error of 1.9 points suggests that the impact estimate would have to have been larger than three points for it to be considered statistically significant at the 10 percent level.

Before having seen the data, we predicted that test score impacts of Chicago TAP would grow over time for schools and for the district as a whole (Glazerman et al. 2007), but we showed above that schools with more years implementing Chicago TAP did not demonstrate higher test scores than those with fewer years of experience, and that impact estimates for the fourth year of rollout for the district (2010–2011) were lower than for the third year (2009–2010). A possible explanation for these results might be the strengths or weaknesses of the particular schools that happened to be part of each cohort implementing the program. To test this explanation, we estimated impacts separately by cohort, taking into account every year for which a given cohort was implementing Chicago TAP. These results are also presented in Table IV.7 (model 4), which suggests that the Chicago TAP effect was not statistically significant for 10 of the 12 hypothesis tests conducted. The two exceptions were the estimated impact on reading scores for cohort 4, which was negative (-1.6 points), and on math scores for cohort 2, which was positive (1.2 points). The lack of a clear pattern suggests that despite these two exceptions, there was no obvious cohort effect that drives the results.

Tables IV.8 through IV.10 demonstrate the degree to which the quasi-experimental findings are robust to alternative matching methods. We used several methods, but present results for two alternatives, discussed in Chapter II. One alternative, called caliper matching, selects all potential comparison schools whose predicted probability of being a Chicago TAP school (propensity score) is within a certain distance (radius) of an actual Chicago TAP school's propensity score. We report results for one particular caliper size (radius), namely 0.025. Results for other radii are available from the authors. The other method, called kernel density estimation, selects every potential matching school, but assigns weights to each school according to the distances between that school and each Chicago TAP school in terms of their propensity score.

Table IV.8. Sensitivity of Quasi- Experimental Findings to Matching Method: Impacts on ISAT Reading Scores

	NN5 (benchmark)		Caliper (r = .025)		Kernel	
Model	Impact	Standard Error	Impact	Standard Error	Impact	Standard Error
1. Overall Chicago TAP effect	-0.1	0.3	-0.1	0.3	-0.2	0.3
2. Chicago TAP effects by year of implementation (school experience)						
First year (all four cohorts)	-0.2	0.4	-0.2	0.4	-0.3	0.4
Second year (cohorts 1–3)	-0.1	0.5	-0.2	0.5	-0.3	0.5
Third year (cohorts 1& 2)	0.4	0.6	0.4	0.6	0.3	0.6
Fourth year (cohort 1)	-0.7	0.9	-0.7	0.9	-0.7	0.9
3. Chicago TAP effects by school year (district experience)						
Year 1: 2007–2008 (cohort 1)	-0.6	0.8	-0.6	0.7	-0.4	0.7
Year 2: 2008–2009 (cohorts 1 & 2)	-0.4	0.6	-0.6	0.6	-0.8	0.6
Year 3: 2009–2010 (cohorts 1–3)	1.3*	0.7	1.4**	0.6	1.4**	0.6
Year 4: 2010–2011 (all four cohorts)	-0.7	0.5	-0.7*	0.4	-0.8*	0.4
4. Chicago TAP effects by cohort						
Cohort 1 (all four years)	-0.4	0.4	-0.4	0.4	-0.5	0.4
Cohort 2 (2009–2011)	0.5	0.5	0.5	0.5	0.4	0.5
Cohort 3 (2010–2011)	0.3	0.6	0.3	0.6	0.3	0.6
Cohort 4 (2011 only)	-1.6***	0.6	-1.6***	0.6	-1.7***	0.6
Sample size (student-year combinations)	83,214		415,105		417,857	
Sample size (school-grade-year combinations)	2,092		7,834		7,893	

Note: NN5 = nearest five neighbors.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

In terms of overall effects of Chicago TAP across all years and cohorts (model 1 in each table), the alternative matching methods produced the same findings for math and reading—no significant impacts of Chicago TAP. The overall impact of Chicago TAP on science scores, however, was positive and significant for one of the matching methods; for caliper matching, the impact was 1.5 ISAT points, equal to about 5 percent of a standard deviation, or two percentile points from the median score (e.g., the difference between the 50th and 52nd percentile). For the other matching method (kernel density estimation), the impact estimate was slightly smaller and not significant.

Both the caliper and kernel density estimators appear to use more data (larger sample) than the original matching method, but the tradeoff is that they may use schools that are not as closely matched. When we used a smaller radius of caliper to match schools—in other words, restricted the analysis to better matches—the results were similar to the large-radius results, with nonsignificant math and reading impacts, but a significant science impact of 1.7 ISAT points or 1.8 points, both significant at the 10 percent level. The smaller-radius caliper matches produced closer matches, but to use these approaches we had to drop some Chicago TAP schools from the analysis altogether because no non-TAP schools matched to them. For space reasons, these additional results are not

shown in the table. We based our choice of the nearest-five-neighbors method on the quality of match observed prior to examining any outcomes. See chapter II and Appendix A for more detail.

In terms of the school experience (model 2), district experience (model 3), and cohort-specific effects (model 4), the inferences drawn in terms of the effects' sign (positive or negative) and statistical significance were all the same regardless of matching method used, with just the two exceptions mentioned above.

For reading scores (Table IV.8), the estimated impact in year 4 was slightly more precisely estimated using the caliper matching method, and the point estimate was slightly higher using the kernel density matching method; the impact therefore becomes statistically significant at the 10 percent level for both (impact = -0.7 and -0.8, respectively).

For math scores (Table IV.9), the year 3 Chicago TAP effect, which was not significant under the benchmark model with nearest-five-neighbors matching, becomes positive and statistically significant (impact = 1.8) using a caliper of 0.025. All of the other conclusions in terms of sign and statistical significance are the same.

Table IV.9. Sensitivity of Quasi- Experimental Findings to Matching Method: Impacts on ISAT Math Scores

	NN5 (benchmark)		Caliper (r = .025)		Kernel	
Model	Impact	Standard Error	Impact	Standard Error	Impact	Standard Error
1. Overall Chicago TAP effect	0.1	0.4	0.1	0.3	0.0	0.4
2. Chicago TAP effects by year of implementation (school experience)						
First year (all four cohorts)	0.4	0.5	0.3	0.5	0.2	0.5
Second year (cohorts 1–3)	-0.1	0.6	0.0	0.6	-0.1	0.6
Third year (cohorts 1 & 2)	0.2	0.8	0.4	0.8	0.3	0.8
Fourth year (cohort 1)	-1.3	1.4	-1.3	1.4	-1.3	1.4
3. Chicago TAP effects by school year (district experience)						
Year 1: 2007–2008 (cohort 1)	0.2	1.10	-0.7	1.0	-0.8	1.0
Year 2: 2008–2009 (cohorts 1 & 2)	0.3	0.85	-0.2	0.8	-0.1	0.8
Year 3: 2009–2010 (cohorts 1–3)	1.0	0.93	1.8**	0.8	1.4	0.9
Year 4: 2010–2011 (all four cohorts)	-0.5	0.66	-0.7	0.6	-0.6	0.6
4. TAP effects by cohort						
Cohort 1 (all four years)	-0.4	0.6	-0.5	0.5	-0.6	0.5
Cohort 2 (2009–2011)	1.2*	0.6	1.3**	0.6	1.1*	0.6
Cohort 3 (2010–2011)	-0.4	0.7	-0.2	0.7	-0.3	0.7
Cohort 4 (2011 only)	-0.1	1.0	-0.2	0.9	-0.2	0.9
Sample size (student-year combinations)	83,125		414,438		417,190	
Sample size (school-grade-year combinations)	2,092		7,831		7,890	

Note: NN5 = nearest five neighbors.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

For science scores, the overall Chicago TAP effect was statistically significant when we used the caliper matching methods, as discussed above, but the more detailed findings did not change. None of the school experience effects, district experience effects, or cohort effects was statistically significant using any of the matching methods. Following the same format as that of the previous two tables, Table IV.10 shows results for three of those matching methods, including the benchmark model (five nearest neighbors), caliper matching with a radius of 0.025, and kernel density matching.

Table IV.10. Sensitivity of Quasi-Experimental Findings to Matching Method: Impacts on ISAT Science Scores

	NN5 (benchmark)		Caliper (r = .025)		Kernel	
Model	Impact	Standard Error	Impact	Standard Error	Impact	Standard Error
1. Overall Chicago TAP effect	1.1	0.8	1.5**	0.7	1.1	0.7
2. Chicago TAP effects by year of implementation (school experience)						
First year (all four cohorts)	1.3	1.2	1.7	1.1	1.3	1.1
Second year (cohorts 1–3)	0.1	1.0	0.5	0.9	0.1	0.9
Third year (cohorts 1 & 2)	1.9	1.7	2.4	1.6	1.9	1.6
Fourth year (cohort 1)	2.1	3.6	2.5	3.6	2.1	3.6
3. Chicago TAP effects by school year (district experience)						
Year 1: 2007–2008 (cohort 1)	NA	NA	NA	NA	NA	NA
Year 2: 2008–2009 (cohorts 1 & 2)	-0.4	1.7	-0.3	1.5	0.0	1.5
Year 3: 2009–2010 (cohorts 1–3)	1.6	1.9	1.6	1.6	1.4	1.7
Year 4: 2010–2011 (all four cohorts)	0.3	1.2	0.7	1.1	0.4	1.1
4. Chicago TAP effects by cohort						
Cohort 1 (all four years)	1.7	1.6	2.2	1.6	1.8	1.6
Cohort 2 (2009–2011)	0.7	1.0	1.1	0.9	0.7	1.0
Cohort 3 (2010–2011)	1.3	1.4	1.6	1.3	1.2	1.3
Cohort 4 (2011 only)	0.0	1.8	0.4	1.8	0.0	1.8
Sample size (student-year combinations)	29,364		125,315		125,984	
Sample size (school-grade-year combinations)	764		2,376		2,390	

Note: NN5 = nearest five neighbors. NA = data not available.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

Finally, we tested the sensitivity of our findings to choice of a particular regression model. We reestimated the overall quasi-experimental Chicago TAP effects using a variety of different regression models, each of which makes slightly different assumptions about the control variables, or defines the study sample a bit differently. These are the same robustness tests used for the experimental findings presented in Tables IV.3–IV.5. The results, presented in Table IV.11, show that the findings reported in the summary table are robust. That is, we confirmed the lack of detectable impacts on reading, math, and science scores.

Table IV.11. Sensitivity of Quasi- Experimental Results to Model Specification: Impacts on ISAT Scores by Subject

Model	Reading			Math			Science		
	Impact	Standard Error	Sample Size	Impact	Standard Error	Sample Size	Impact	Standard Error	Sample Size
Benchmark	-0.1	0.3	83,214	0.1	0.4	83,125	1.1	0.8	29,364
Covariates									
Separate pretest effect by grade	-0.2	0.3	83,214	0.0	0.4	83,125	0.6	1.0	29,364
Separate pretest effect by year	0.0	0.3	83,214	0.3	0.4	83,125	1.0	0.8	29,364
Separate pretest effect by year and grade	-0.1	0.5	83,214	0.1	0.6	83,125	0.5	1.0	29,364
Pretest squared and cubed	-0.1	0.3	83,214	0.0	0.4	83,125	1.0	0.8	29,354
Scores standardized within grade	0.0	0.0	83,214	0.0	0.0	521,234	0.0	0.0	159,250
No opposite subject pretest	0.0	0.3	83,214	0.0	0.4	83,429	0.8	0.8	29,464
No pretest, grades 4–8	0.0	0.4	89,070	0.2	0.6	89,101	0.6	1.0	31,262
No pretest, grades 3–8	-0.4	0.4	108,703	-0.2	0.6	108,711	NA	NA	NA
No covariates	-0.1	0.3	83,214	0.0	0.4	83,125	1.0	0.8	29,364
Alternative variance estimation method									
Random effects (RE)	0.0	0.3	83,214	0.0	0.3	83,125	1.2	0.7	29,364
RE with school characteristics	-0.1	0.3	83,214	0.0	0.3	83,125	1.1	0.7	29,364
Measurement error correction									
Instrumental variables	0.0	0.3	83,214	0.0	0.4	83,125	0.8	0.8	29,364
Errors in variables model, reliability = .9	0.0	0.3	83,499	0.0	0.4	83,429		not estimated	
Errors in variables model, reliability = .8	0.2	0.7	83,614		not estimated			not estimated	

Note: Analyses use nearest-five-neighbors matching. Sample sizes are given in student-years. NA = data not available.

Differences are not statistically significant at the 10 percent level.

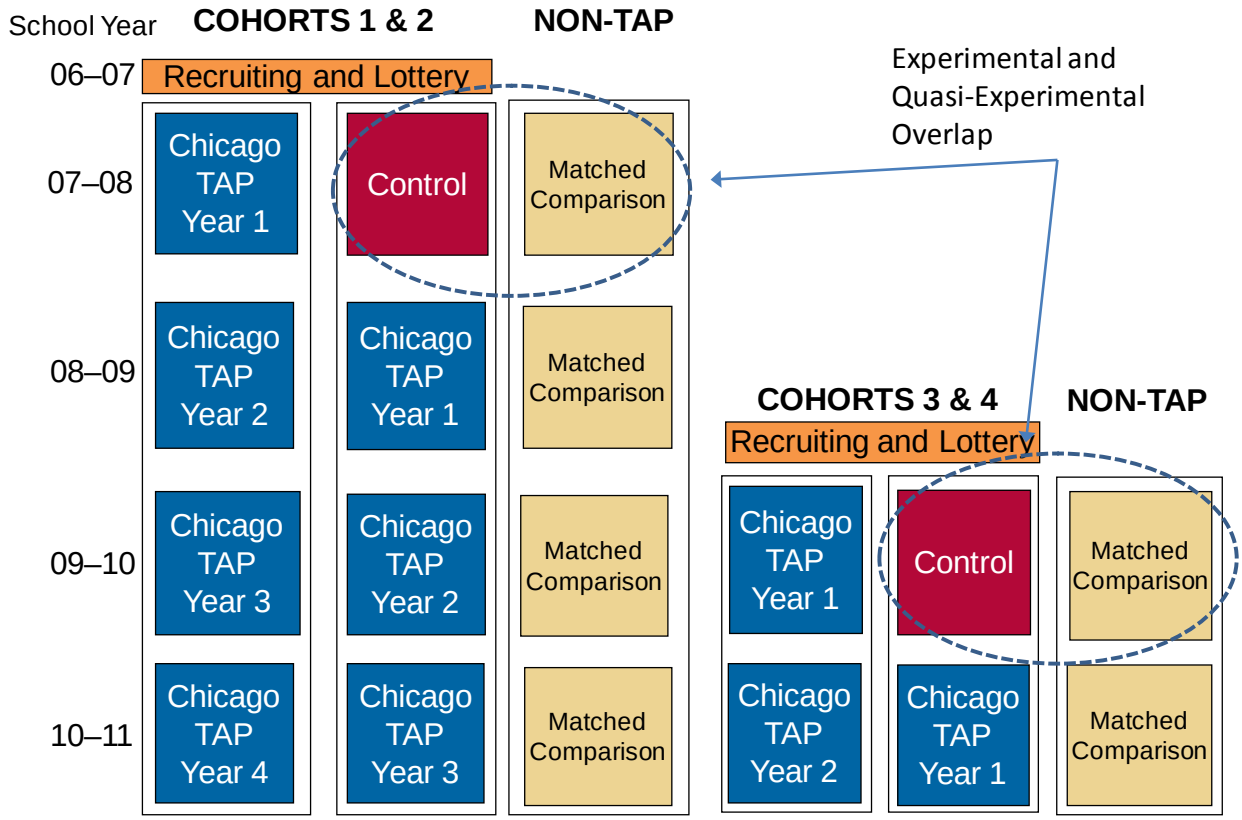
C. Reconciling the Experimental and Quasi- Experimental Evidence

As discussed in Chapter II, the experimental findings have the advantage of eliminating any systematic bias but the disadvantage of capturing only the first year of Chicago TAP implementation. Meanwhile, the quasi-experimental findings, which do capture the entire Chicago TAP experience, require us to assume that we have measured all of the important factors that are related to both Chicago TAP implementation and test score outcomes.

Recognizing these limitations, we built into the study design an overlap, where we can compare outcomes for the randomized control group directly to outcomes for the matched comparison group. If the matching produces outcomes similar to those produced by the randomization, then we can be more confident in the quasi-experimental findings, which capture the full range of Chicago TAP implementation experiences. Figure IV.3 illustrates how this comparison fits into the overall study design.

Table IV.12 shows the results of estimating the regression-adjusted difference in outcomes between the control group and the corresponding comparison group, each of which is being used to estimate the same counterfactual outcomes—those that would have been achieved had the Chicago TAP schools not implemented the program. (The treatment group in both cases is the same). The results suggest that the differences between the two counterfactual estimates were small (less than

Figure IV.3. Comparing Experimental Control Group to the Quasi- Experimental Comparison Group



one ISAT point) and not significant, except for the math scores from spring 2008.[20] In that year, students in the matched comparison schools scored two or three points lower, depending on which matching method we used, than students in the randomized control group schools. If we assume that the scores of the randomized control group are an exact representation of the counterfactual outcome (as opposed to an approximation), then it would follow that the matched comparison group outcomes are underestimated by two or three points, and thereby overstate the size of the true impacts. However, when we look at the other period of overlap, represented by the 2010 results comparing the control group (cohort 4) to the matched comparison group, there is no significant difference between the experimentally and quasi-experimentally estimated counterfactual outcomes, consistent with the quasi-experimental estimates being unbiased.

Table IV.12. Matched Comparison Group vs. Randomized Control Group: Differences in March ISAT Scores by Subject

	Reading		Math		Science	
Year and Matching Method	Comparison Minus Control	Standard Error	Comparison Minus Control	Standard Error	Comparison Minus Control	Standard Error
2007–2008 (cohorts 1 & 2)						
Nearest neighbors	-0.6	0.6	-2.8***	0.9	NA	NA
Caliper (r = 0.025)	-0.4	0.5	-2.0**	0.9	NA	NA
Kernel density	-0.4	0.5	-2.0**	0.9	NA	NA
2009–2010 (cohorts 3 & 4)						
Nearest neighbors	-0.1	0.6	-0.9	0.9	0.5	1.3
Caliper (r = 0.025)	-0.1	0.6	-0.7	0.8	-0.2	0.8
Kernel density	0.1	0.6	0.3	0.9	2.1	1.6

Note: NA = data not available.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

It is difficult to determine which of the overlapping design results, the interpretation of upward bias or the interpretation of no bias, to generalize to the rest of the quasi-experimental analysis. There are two obvious choices. One is to accept all of the quasi-experimental estimates as the best available information and assume no bias. The other is to assume that the quasi-experimental estimates of math impacts are upwardly biased by approximately two ISAT points, in which case even the few examples of positive impact are likely overstated. In either case, we would conclude that Chicago TAP was not successful in raising student test scores over the time horizon covered by this study.

[20] The difference for science scores in 2010 is also greater than one ISAT point for one of the matching estimators (difference = 2.1 ISAT points for science using kernel density matching), but that result is imprecisely estimated and not significant. In other words, the science results are based on fewer grades, thus a smaller sample; thus we cannot consider them to be a good test of the matching method against the experimental benchmark.

This page has been left blank for double-sided copying.

V. IMPACTS ON TEACHER RETENTION

Chicago TAP is hypothesized to help schools retain their best teachers by rewarding performance, providing professional development and leadership opportunities, and creating a career ladder. In this chapter we examine the impacts of Chicago TAP on the rate at which teachers were retained by their schools from year to year and over longer periods of time.

For the retention analysis, the matched comparison sample—not the randomized control group—is the most credible benchmark to use for the Chicago TAP sample. As discussed in Chapter II, we randomly assigned schools to either a treatment group that implemented Chicago TAP right away or to a control group that delayed implementation by one year. Teachers in control schools knew that their school would be adopting Chicago TAP soon, and that knowledge might have influenced the career plans of the schools' teachers. For that reason, the randomized control group is a contaminated source of information on outcomes such as retention that may be influenced by future Chicago TAP participation. Therefore, we rely only on the matched comparison sample of non-TAP schools. These comparison schools were identified through propensity score matching as being similar to Chicago TAP schools on preintervention measures of retention, student achievement, and other variables (see Chapter II and Appendix A for further details on the matching methodology).

We estimate impacts by comparing teacher retention rates observed in Chicago TAP schools to those observed in matched comparison schools that did not implement Chicago TAP. The retention rates in the matched comparison schools are used to approximate the retention rates that would have been observed in the absence of Chicago TAP; these non-TAP retention rates reflect economic conditions, layoffs, and other policies that affected schools throughout the district.

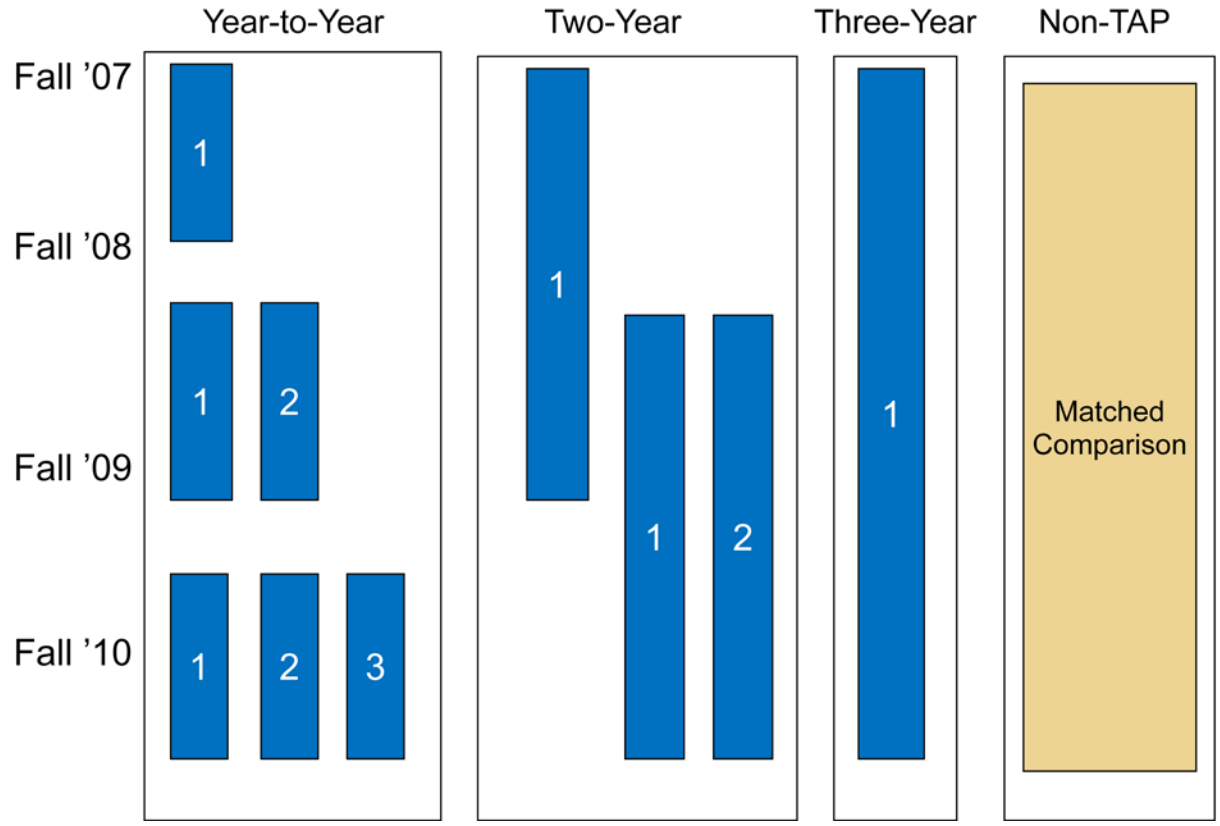
Unfortunately, we are unable to estimate the impact of Chicago TAP on the *quality* of teachers retained because we lack reliable measures of teacher quality for both Chicago TAP and non-TAP schools. Nevertheless, the teacher retention rate itself is of policy interest because of the costs associated with turnover. In addition to the financial costs associated with hiring replacement teachers, high teacher turnover may contribute to teacher shortages, lead to out-of-field teaching, and hinder the development of cohesive learning communities within schools, all factors that can adversely affect student achievement.

A. Impacts on Teacher Retention over Time

We measured retention by obtaining teaching assignments for the late fall of each school year from district administrative records. A classroom teacher was considered retained if he or she worked in the same school in fall of the base year and fall of the follow-up years. The 2007–2008 school year marked the start of a four-year period during which Chicago TAP would be in place, so we focus on teacher retention throughout that period. We did not examine retention of teachers beyond that period because between the third and fourth year of implementation, the district indicated that Chicago TAP would not be continued (Ahmed 2010). Thus, we examined the retention of teachers from fall 2007 through fall 2010. Over that period, we were able to observe teachers' year-to-year retention (fall 2007–fall 2008, fall 2008–fall 2009, or fall 2009–fall 2010), retention over two years (fall 2007–fall 2009 or fall 2008–fall 2010), or retention over the three years (fall 2007–fall 2010). We could also estimate separate impacts by Chicago TAP cohort. This was helpful in case we wanted to understand the relationship between the number of years of implementation and retention.

The possible retention outcomes by duration (one-, two-, or three-year retention), year, and Chicago TAP cohort are presented in Figure V.1. For each outcome, we estimated the impact of Chicago TAP by comparing retention in the Chicago TAP schools represented in each blue box to retention in the matched comparison schools. Matching was done using the nearest-five-neighbors method discussed in Chapter II and in Appendix A. When the retention outcome was available for more than one cohort, we combined across cohorts to increase statistical power. We also assessed the robustness of the overall impacts to the choice of matching method by presenting estimates produced using alternative matching strategies. Because Chicago TAP is a school-specific program, we focus primarily on Chicago TAP's impact on schools' retention of teachers ("school retention"); we present findings for the retention of teachers in the district ("district retention") in Appendix B (Tables B.8–B.11 and Figure B.1).

Figure V.1. Retention Outcomes by Year, Duration, and Cohort



Note: Each blue box represents a group of classroom teachers in schools of the indicated cohort (1, 2, or 3) followed from fall of the base year to fall of the follow-up year.

We found positive, statistically significant impacts of Chicago TAP on year-to-year retention rates for the first two years of program rollout (Table V.1). In particular, we estimated that 85 percent of teachers in fall 2007 would have returned to their schools in fall 2008 had they been in Chicago TAP, compared to 77 percent of teachers had they been in non-TAP schools, indicating an impact of eight percentage points. For fall 2008–fall 2009, we found a three-percentage-point impact, with an average retention rate of 83 percent for teachers in Chicago TAP schools compared to 80 percent of teachers in non-TAP schools. The estimated Chicago TAP-comparison difference for fall 2009–fall 2010 was smaller in magnitude and not statistically significant.

Table V.1. Impacts on School Retention Rates (percentage)

Retention Period	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error	Sample Size (teachers)
One- Year Rates					
Fall 2007–fall 2008 (cohort 1)	85.2	77.2	7.9***	3.01	1,102
Fall 2008–fall 2009 (cohorts 1 & 2)	82.8	79.6	3.2*	1.83	1,686
Fall 2009–fall 2010 (cohorts 1, 2, & 3)	81.2	80.9	0.3	1.81	2,694
Two- Year Rates					
Fall 2007–fall 2009 (cohort 1)	77.8	59.5	18.3***	4.87	881
Fall 2008–fall 2010 (cohorts 1 & 2)	71.2	67.9	3.3	2.76	1,879
Three- Year Rates					
Fall 2007–fall 2010 (cohort 1)	67.0	55.5	11.5***	4.21	781

Note: School retention rate is defined as the percentage of classroom teachers in fall of the base year who worked at the same school in fall of the follow-up year.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

To provide another perspective on retention, we also compared the two-year and three-year teacher retention rates of Chicago TAP and non-TAP schools (Table V.1). We found that for teachers present at the start of rollout in the district, Chicago TAP had a positive, statistically significant impact on two-year school retention, defined as the percentage of teachers returning in fall 2009 to the schools where they taught in fall 2007. The average two-year retention rate for teachers in Chicago TAP schools exceeded that of teachers in comparison schools by 18 percentage points (78 percent versus 60 percent). When measured from fall 2008, however, the estimated impact on two-year retention rates—three points—was not statistically significant. Similar to the two-year retention finding for teachers present for district rollout of the program, we found a statistically significant impact of Chicago TAP on the three-year retention rate; 67 percent of fall 2007 teachers in Chicago TAP schools taught in the same school three years later compared to 56 percent of teachers in non-TAP schools. In other words, teachers in Chicago TAP schools in fall 2007 were about 20% more likely than teachers in comparison schools to be in those same schools three years later.

We hypothesized that teachers in tested grades and subjects would behave differently from those in nontested grades and subjects because only teachers in tested grades and subjects had a direct effect on value-added measures that help determine performance-based compensation. We found evidence to support this hypothesis, but the pattern of subgroup findings was not consistent across the retention outcomes (Table V.2). Estimated impacts for teachers in tested grades and subjects and teachers in academic but nontested grades or subjects tended to be positive. Findings were more varied for teachers in "other" nonacademic subjects; Chicago TAP-comparison differences for this subgroup were positive for half of the retention rates we examined and negative for the other half. The relative magnitudes of the estimates across the three subgroups and the statistical significance of the findings varied as well.

Table V.2. Impacts on School Retention Rates, by Teaching Assignment (percentage)

Teaching Assignment	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error	Sample Size (teachers)
One- Year Rates					
Fall 2007–fall 2008 (cohort 1)					
All teachers	85.2	77.2	7.9***	3.01	1,102
Academic subjects, tested grades/subjects	89.2	84.2	5.0*	2.85	323
Academic subjects, nontested grades/subjects	90.9	79.7	11.2***	3.25	497
Other	92.8	82.6	10.2**	4.17	164
Fall 2008–fall 2009 (cohorts 1 & 2)					
All teachers	82.8	79.6	3.2*	1.83	1,686
Academic subjects, tested grades/subjects	86.5	82.3	4.2	3.14	551
Academic subjects, nontested grades/subjects	88.1	83.8	4.3*	2.24	777
Other	83.0	80.3	2.7	3.89	239
Fall 2009–fall 2010 (cohorts 1, 2, & 3)					
All teachers	81.2	80.9	0.3	1.81	2,694
Academic subjects, tested grades/subjects	80.6	83.0	-2.4	3.18	903
Academic subjects, nontested grades/subjects	86.5	83.6	3.0	2.16	1,290
Other	75.2	76.3	-1.1	4.70	385
Two- Year Rates					
Fall 2007–fall 2009 (cohort 1)					
All teachers	77.8	59.5	18.3***	4.87	881
Academic subjects, tested grades/subjects	82.0	70.3	11.6*	6.05	278
Academic subjects, nontested grades/subjects	88.8	78.0	10.8**	5.29	399
Other	78.8	52.3	26.5**	12.89	114
Fall 2008–fall 2010 (cohorts 1 & 2)					
All teachers	71.2	67.9	3.3	2.76	1,879
Academic subjects, tested grades/subjects	75.9	72.8	3.1	4.43	587
Academic subjects, nontested grades/subjects	79.7	73.8	6.0*	3.17	888
Other	57.7	59.3	-1.6	6.48	264
Three- Year Rates					
Fall 2007–fall 2010 (cohort 1)					
All teachers	67.0	55.5	11.5***	4.21	781
Academic subjects, tested grades/subjects	77.9	56.0	21.8***	5.85	221
Academic subjects, nontested grades/subjects	73.2	65.8	7.4	4.75	366
Other	29.4	55.6	-26.3**	10.80	106

Note: School retention rate is defined as the percentage of classroom teachers in fall of the base year who worked at the same school in fall of the follow-up year. Teaching assignment category is missing for 118 teachers for fall 2007–fall 2008, 119 teachers for fall 2008–fall 2009, 116 teachers for fall 2009–fall 2010, 90 teachers for fall 2007–fall 2009, 140 teachers for fall 2008–fall 2010, and 88 teachers for fall 2007–fall 2010.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

We also examined teacher experience subgroups, defined by years of service in CPS (Table V.3). Although we did find differences among experience subgroups, there was not a consistent pattern in the findings across outcomes. For the one-year rates, Chicago TAP-comparison differences tend to be positive for both early- and mid-career teachers and negative for late-career teachers. For the two- and three-year rates, estimated impacts were consistently positive for all three experience subgroups. As with the teaching assignment subgroups, the relative magnitudes of the estimates across the experience subgroups and the statistical significance of the findings varied with the particular period examined.

Table V.3. Impacts on School Retention Rates, by Years of Service (percentage)

Years of Service	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error	Sample Size (teachers)
One- Year Rates					
Fall 2007–fall 2008 (cohort 1)					
All teachers	85.2	77.2	7.9***	3.01	1,102
<5 years	76.4	67.6	8.8*	4.84	371
5–24 years	89.3	82.4	6.9***	2.48	634
>24 years	100.0	95.8	4.1	3.60	97
Fall 2008–fall 2009 (cohorts 1 & 2)					
All teachers	82.8	79.6	3.2*	1.83	1,686
<5 years	96.8	95.0	1.7	1.28	538
5–24 years	86.6	85.3	1.3	2.36	1,010
>24 years	81.0	83.6	-2.6	5.68	138
Fall 2009–fall 2010 (cohorts 1, 2, & 3)					
All teachers	81.2	80.9	0.3	1.81	2,694
<5 years	73.3	70.5	2.8	4.43	731
5–24 years	85.9	86.0	-0.1	1.78	1741
>24 years	61.6	76.8	-15.2*	9.16	222
Two- Year Rates					
Fall 2007–fall 2009 (cohort 1)					
All teachers	77.8	59.5	18.3***	4.87	881
<5 years	77.8	64.0	13.8**	6.63	301
5–24 years	86.1	70.9	15.2**	6.16	495
>24 years	90.8	64.2	26.6***	9.63	85
Fall 2008–fall 2010 (cohorts 1 & 2)					
All teachers	71.2	67.9	3.3	2.76	1,879
<5 years	68.3	59.3	9.0	6.06	540
5–24 years	76.3	74.4	1.9	3.50	1,165
>24 years	54.8	50.8	4.0	12.26	174
Three- Year Rates					
Fall 2007–fall 2010 (cohort 1)					
All teachers	67.0	55.5	11.5***	4.21	781
<5 years	65.4	48.8	16.6*	8.94	247
5–24 years	71.2	60.0	11.3**	4.55	469
>24 years	64.0	40.3	23.7	33.12	65

Note: School retention rate is defined as the percentage of classroom teachers in fall of the base year who worked at the same school in fall of the follow-up year.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

We caution that these subgroup findings were based on relatively small samples and were sensitive to estimation decisions such as whether to allow into the potential comparison sample schools known to have closed in future years. The findings may reflect chance differences that were not caused by Chicago TAP.

As a robustness check, we estimated differences in the overall retention rates between Chicago TAP and comparison schools using alternative methods to construct the non-TAP matched comparison groups. The school retention findings presented above were estimated using the nearest-five-neighbors method, which we found did the most to reduce differences between Chicago TAP and non-TAP schools on observable preintervention characteristics. The sign of the estimated Chicago TAP-comparison differences was generally consistent across alternative matching methods, but the magnitude and statistical significance of the estimates varied with the method used (Table V.4). We found that single nearest neighbor matching was very unstable.

The estimates presented above provide suggestive evidence of a positive impact of Chicago TAP on school retention. We found positive, statistically significant impacts for teacher retention rates measured from the beginning of the Chicago TAP rollout, including the one-year rate (fall 2007–fall 2008), the two-year rate (fall 2007–fall 2009), and the three-year rate (fall 2007–fall 2010). These three retention rates, however, can be measured only for cohort 1. Impacts of Chicago TAP for retention rates that incorporate other cohorts tend to be positive but smaller in magnitude and not statistically significant. Figure V.2 presents the impact estimates separately by year, duration, and cohort. The pattern of findings suggests that teacher retention rates tend to be higher in Chicago TAP schools than in non-TAP schools; however, while half of the impact estimates were at least seven percentage points in magnitude and statistically significant, the other half were five points or fewer in magnitude and were not statistically significant. It is possible that the treatment effects were stronger for some groups of schools than others—for reasons we cannot explain with our data—or that the matching method worked differently for some cohorts than others. The findings broken down by cohort suggest that impacts were strongest for cohort 1. The one opportunity we had to observe the retention behavior of teachers in cohort 3, from fall 2009 to fall 2010, indicated that Chicago TAP teachers were retained at lower rates than their comparison group counterparts, but the difference was not statistically significant.

We also examined district retention rates, defined as the rate at which teachers returned to CPS even if they changed schools. Increased district retention rates were not an explicitly hypothesized impact of Chicago TAP, but it could be true that the program strengthens teachers' bonds with the district overall. However, we did not find consistent evidence of an impact of Chicago TAP on district retention. Among the main district retention outcomes examined using the nearest-five-neighbors matching method, none of the Chicago TAP-comparison differences for teachers overall was statistically significant except for the fall 2007–fall 2009 retention rate. In the few cases in which statistically significant impacts emerged for particular subgroups, or for a cohort-specific impact in a particular year, there was not a clear pattern of findings. Overall district retention impacts were generally similar when estimated using alternative matching algorithms. The results are shown in Appendix B (Tables B.8–B.11 and Figure B.1).

Table V.4. Impacts on School Retention Rates, Sensitivity Analysis by Matching Method (percentage)

Matching Method	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error	Sample Size (teachers)
One- Year Rates					
Fall 2007–fall 2008 (cohort 1)					
Nearest neighbor	84.0	85.5	-1.5	2.83	499
Nearest 5 neighbors	85.2	77.2	7.9***	3.01	1,102
Caliper match (r = 0.005)	86.9	80.7	6.3***	2.43	2,944
Caliper match (radius = 0.025)	85.7	80.9	4.8**	1.90	11,739
Kernel density	86.6	82.4	4.2***	1.62	12,040
Fall 2008–fall 2009 (cohorts 1 & 2)					
Nearest neighbor	82.3	79.4	2.9	2.84	661
Nearest 5 neighbors	82.8	79.6	3.2*	1.83	1,686
Caliper match (r = 0.005)	83.5	80.3	3.3**	1.59	2,678
Caliper match (r = 0.025)	83.4	81.5	1.9	1.46	5,823
Kernel density	84.0	81.8	2.2	1.37	11,422
Fall 2009–fall 2010 (cohorts 1, 2, & 3)					
Nearest neighbor	81.9	82.7	-0.8	3.56	974
Nearest 5 neighbors	81.2	80.9	0.3	1.81	2,694
Caliper match (r = 0.005)	81.1	81.2	-0.1	2.01	3,290
Caliper match (r = 0.025)	81.4	80.7	0.7	1.62	11,320
Kernel density	81.4	81.1	0.3	1.69	11,340
Two- Year Rates					
Fall 2007–fall 2009 (cohort 1)					
Nearest neighbor	89.0	30.4	58.6***	6.29	322
Nearest 5 neighbors	78.4	60.6	17.8***	5.16	871
Caliper match (radius = 0.005)	78.5	65.5	12.9***	4.40	1,697
Caliper match (radius = 0.025)	78.1	66.7	11.4***	3.37	5,775
Kernel density	78.8	67.3	11.5***	3.07	11,413
Fall 2008–fall 2010 (cohorts 1 & 2)					
Nearest neighbor	71.4	70.4	1.0	2.61	631
Nearest 5 neighbors	71.2	67.9	3.3	2.76	1,879
Caliper match (r = 0.005)	71.6	66.5	5.2*	2.73	2,352
Caliper match (r = 0.025)	71.2	68.6	2.6	1.84	10,819
Kernel density	71.2	68.8	2.4	1.94	11,373
Three- Year Rates					
Fall 2007–fall 2010 (cohort 1)					
Nearest neighbor	60.0	64.0	-3.9	3.50	265
Nearest 5 neighbors	67.0	55.5	11.5***	4.21	781
Caliper match (r = 0.005)	68.8	53.2	15.6***	3.87	1,190
Caliper match (r = 0.025)	68.4	55.8	12.7***	2.43	10,008
Kernel density	66.6	58.3	8.3**	3.31	11,279

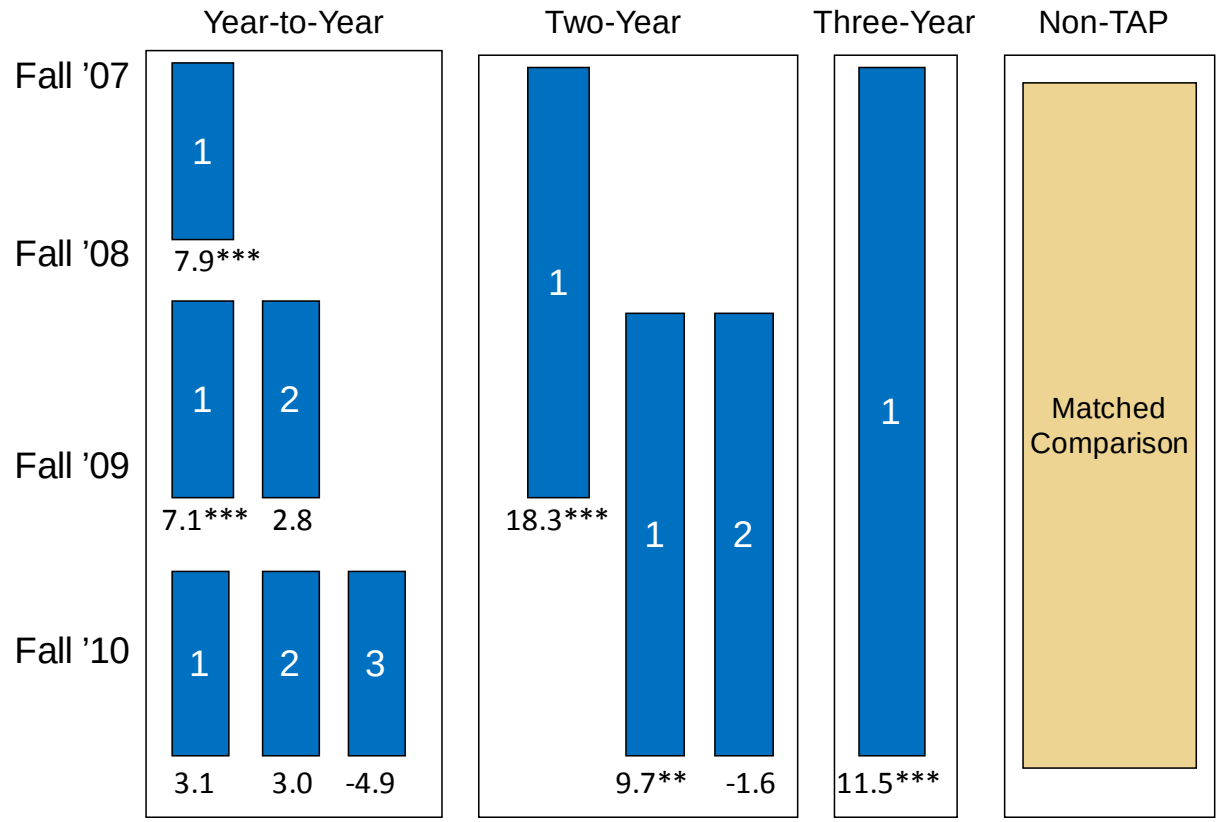
Note: School retention rate is defined as the percentage of classroom teachers in fall of the base year who worked at the same school in fall of the follow-up year. Shaded cells are based on the nearest-five-neighbors method, which produced the closest match between Chicago TAP and non-TAP schools in diagnostic tests performed before the outcomes were examined.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

We further explored retention by examining the impacts of Chicago TAP on the pattern of teacher mobility, focusing on where the movers went. Table V.5 reports percentages of Chicago TAP and nearest-five-neighbors comparison teachers by their follow-up year destinations. We generally found significant differences between the mobility patterns of Chicago TAP and comparison schools, driven largely by differences in the percentages of teachers staying at the same school or moving to other CPS elementary schools. These results are not regression adjusted and may reflect differences in background characteristics that are unrelated to Chicago TAP.

Figure V.2. Impacts on School Retention Rates by Year, Duration, and Cohort



Note: Each blue box represents a group of teachers in schools of the indicated cohort (1, 2, or 3) followed from fall of the base year to fall of the follow-up year. Numbers below the blue boxes are impacts of Chicago TAP on school retention rates estimated using the nearest-five-neighbors matching method to form comparison groups.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

Table V.5. Impacts on Teacher Mobility by Destination (percentage)

Destination	Chicago TAP Mean	Comparison Mean	Difference
One- Year Mobility			
Fall 2007–fall 2008 (cohort 1)***			
Stayed at same school	81.7	75.0	6.7
Moved to a Chicago TAP school	3.4	0.6	2.8
Moved to a comparison school	0.0	1.0	-1.0
Moved to other CPS elementary school	3.4	7.3	-3.9
Moved to a CPS high school	1.3	1.3	0.0
Moved to a citywide or central office position	0.0	1.7	-1.7
Left CPS	10.2	13.1	-2.9
Fall 2008–fall 2009 (cohorts 1 & 2)			
Stayed at same school	81.0	77.1	3.9
Moved to a Chicago TAP school	1.3	0.8	0.5
Moved to a comparison school	1.3	1.5	-0.2
Moved to other CPS elementary school	3.6	4.7	-1.1
Moved to a CPS high school	0.5	0.4	0.2
Moved to a citywide or central office position	1.8	3.7	-1.9
Left CPS	10.5	11.8	-1.3
Fall 2009–fall 2010 (cohorts 1, 2, & 3)*			
Stayed at same school	79.6	79.3	0.3
Moved to a Chicago TAP school	1.0	0.6	0.4
Moved to a comparison school	0.2	1.3	-1.2
Moved to other CPS elementary school	2.1	3.1	-1.0
Moved to a CPS high school	0.2	0.4	-0.2
Moved to a citywide or central office position	1.5	1.1	0.4
Left CPS	15.5	14.2	1.4
Two- Year Mobility			
Fall 2007–fall 2009 (cohort 1)***			
Stayed at same school	73.3	61.7	11.5
Moved to a Chicago TAP school	2.7	2.2	0.5
Moved to a comparison school	0.0	2.6	-2.6
Moved to other CPS elementary school	4.3	9.2	-4.9
Moved to a CPS high school	1.1	1.4	-0.3
Moved to a citywide or central office position	2.1	3.7	-1.5
Left CPS	16.6	19.3	-2.7
Fall 2008–fall 2010 (cohorts 1 & 2)*			
Stayed at same school	68.9	67.5	1.5
Moved to a Chicago TAP school	2.2	1.0	1.2
Moved to a comparison school	0.8	2.1	-1.3
Moved to other CPS elementary school	3.5	6.4	-2.9
Moved to a CPS high school	0.8	0.7	0.1
Moved to a citywide or central office position	1.9	1.0	0.9
Left CPS	21.9	21.3	0.6
Three- Year Mobility			
Fall 2007–fall 2010 (cohort 1)*			
Stayed at same school	64.5	56.4	8.1
Moved to a Chicago TAP school	1.8	1.8	0.0
Moved to a comparison school	0.0	2.2	-2.2
Moved to other CPS elementary school	4.2	8.2	-3.9
Moved to a CPS high school	1.2	2.1	-0.9
Moved to a citywide or central office position	1.2	1.3	-0.1
Left CPS	27.1	28.1	-1.0

Note: N = 235 Chicago TAP and 867 comparison teachers for fall 2007–fall 2008; 390 Chicago TAP and 1,296 comparison teachers for fall 2008–fall 2009; 612 Chicago TAP and 2,082 comparison teachers for fall 2009–fall 2010; 187 Chicago TAP and 694 comparison teachers for fall 2007–fall 2009; 370 Chicago TAP and 1,509 comparison teachers for fall 2008–fall 2010; and 166 Chicago TAP and 615 comparison teachers for fall 2007–fall 2010. We conducted a chi-square test to determine the statistical significance of the difference in the distributions for each retention outcome.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

B. Descriptive Analysis of Skills, Knowledge, and Responsibilities (SKR) Scores by Mobility Status

Policymakers care not only about the teacher retention rate, but also about the quality of teachers retained. If Chicago TAP is successful in rewarding effective teachers, then ineffective teachers might prefer to leave the school and more effective teachers might prefer to stay on longer than they would otherwise. All else equal, one would expect such an improvement in the mix of teachers to result in more-effective teachers accounting for a higher proportion of the teacher workforce remaining in Chicago TAP schools than would have remained in the absence of Chicago TAP. Unfortunately, we could not compare the effectiveness of teachers in Chicago TAP schools to that of teachers in non-TAP schools because we lack reliable measures of teacher effectiveness that can be used in both Chicago TAP and non-TAP schools.

However, we did examine scores obtained by teachers in Chicago TAP schools on a classroom observation rubric known as Skills, Knowledge, and Responsibilities (SKR). As part of establishing instructionally focused accountability, the Chicago TAP model calls for observations of teachers by the principal, lead teachers, and mentor teachers, all of whom undergo training and certification in using the SKR rubric. SKR scores are assigned based on observed classroom performance in four domains: designing and planning instruction, learning environment, instruction, and responsibilities. Each domain is rated on a five-point scale, with 1 indicating "needs improvement," 3 indicating "proficient," and 5 indicating "exemplary." The composite SKR scores are used in determining performance-based compensation.

Table V.6 reports average SKR scores among three groups of teachers based on their movement between a baseline and follow-up period: stayers, who remained in the same school; movers, who moved to another CPS school or to a citywide or central office position within CPS; and leavers, who left CPS. The average SKR score across all teachers in Chicago TAP schools on a five-point scale ranged from 3.0 to 3.4 depending on the period examined. For all periods, the standard deviation of the average SKR score was typically about 0.6 points, implying that about half the population scored within 0.4 points of the average.

We did not find evidence that mobility affected the composition of teachers who remain in Chicago TAP schools. For three of the six periods examined, the differences in average SKR scores across groups were not statistically significant.[21] For the three periods for we which we detected statistically significant differences, there was not a consistent pattern of average SKR scores across groups. During the one-year period from fall 2009 to fall 2010 and the two-year period from fall 2007 to fall 2009, teachers who left their school but remained in CPS (movers) had higher average scores than stayers or leavers. From fall 2008 to fall 2010, however, stayers and movers had the same average score; their average score exceeded that for leavers. Each group had an average score within 0.4 points of the other two groups in all six periods.

We emphasize that this descriptive analysis does not provide causal evidence of Chicago TAP's impact on teacher quality. Because SKR scores were not available for non-TAP schools, we cannot

[21] We conducted an F-test using analysis of variance and failed to reject the hypothesis that the three average SKR scores were the same.

estimate what the SKR scores of stayers, movers, and leavers would have been in the absence of Chicago TAP.

Table V.6. SKR Scores by Mobility Status (points)

Mobility Period	Stayers	Movers	Leavers	Sample Size (teachers)
One- Year Mobility				
Fall 2007–fall 2008 (cohort 1)	3.4	3.2	3.2	221
Fall 2008–fall 2009 (cohorts 1 & 2)	3.1	2.9	3.0	357
Fall 2009–fall 2010 (cohorts 1, 2, & 3)***	3.0	3.1	2.8	560
Two- Year Mobility				
Fall 2007–fall 2009 (cohort 1)	3.3	3.5	3.1	176
Fall 2008–fall 2010 (cohorts 1 & 2)**	3.1	3.1	2.9	338
Three- Year Mobility				
Fall 2007–fall 2010 (cohort 1)	3.2	3.4	3.1	158

Note: Mobility status is defined by the movement of classroom teachers from the base period to the follow-up period: stayers remained in the same school; movers moved to another CPS school or to a citywide or central office position within CPS; leavers left CPS. SKR scores are on a five-point scale, with 5 being the highest score. For each mobility period, we conducted an analysis of variance F-test of the hypothesis that the three average SKR scores across mobility groups were the same.

* We reject the equality of mean SKR scores across mobility groups at the 10 percent level.
** We reject the equality of mean SKR scores across mobility groups at the 5 percent level.
*** We reject the equality of mean SKR scores across mobility groups at the 1 percent level.

This page has been left blank for double-sided copying.

VI. SUMMARY AND DISCUSSION OF FINDINGS

The findings of the study can be summed up in terms of its three main questions: (1) How was Chicago TAP implemented? (2) What was its impact on student achievement? (3) What was its impact on teacher retention? We summarize the findings for each of these in turn and provide some concluding thoughts.

A. Implementation

The current study was aimed at estimating the impacts of Chicago TAP on student and teacher outcomes, but the impact estimates must be interpreted in light of what was implemented. Based on evidence from several data sources, we conclude that Chicago TAP changed the way teachers were evaluated and compensated, their mentoring, and the career advancement opportunities in the schools that adopted the program. We conducted two surveys of teachers participating in Chicago TAP as well as teachers who had not yet begun participating in the program and teachers in schools that were not in the program at all. We found that in terms of mentoring, weekly meetings, career opportunities, and compensation, teacher reports were consistent with the goals of the program, and that Chicago TAP teachers' experiences were sufficiently different from those of teachers in non-TAP schools to conclude that real change occurred.

Program review scores provide evidence on the extent to which Chicago TAP schools implemented the program with fidelity to the national TAP model. The average program review score was around 3 out of 5 for each cohort of schools in each of the first three years of rollout. NIET concluded that elements of TAP had been introduced, but TAP implementation had not been "rigorous." After the third year of rollout, NIET informed CPS that the district had not implemented the TAP system. In the final year of rollout, when CPS conducted the program reviews instead of NIET, the average scores were higher than 4 out of 5.

Two areas where program implementation did not occur as initially planned were the small size of the average performance-based payouts and the absence of a teacher-level value-added component to the teacher evaluation formula. The average teacher payouts in our data were never higher than $1,500 in the first year of implementation and never higher than $2,700 in subsequent years of implementation. The highest performance-based payout anyone ever received was $6,400. For a teacher earning $50,000 annually, these average payouts would represent 3 and 5 percent of salary, respectively, with the maximum just under 13 percent. Given that a CPS teacher with a bachelor's degree and no experience might earn $50,000 to start, the Chicago TAP performance incentives in percentage terms would be much smaller for teachers with more education and experience, and consequently higher salary.[22] In each year of the program, individual teacher performance was measured using the SKR rubric. Value-added measures—those that use test score growth controlling for differences in student background—were calculated at the school level only until the final two years of Chicago TAP, when school-grade level estimates were also used. In other words, the program did not use teacher-level value-added measures in any of its payout formulas.

[22] According to the CPS website (www.cps-humanresources.org/careers/salary.htm, accessed January 5, 2012), the starting 38.6 week salary for a teacher with a bachelor's degree was $50,577 in 2010–2011.

B. Impacts on Test Scores

The changes that occurred in Chicago TAP schools did not translate into positive impacts on test scores in any consistent way. We found evidence of both positive and negative test score impacts in selected subjects, years, and cohorts of schools, but overall there was no detectable impact on math, reading, or science achievement that was robust to different methods of estimation. For example, impacts on science scores overall (across years and cohorts) were positive, but not statistically significant unless we used one particular matching method that excluded some Chicago TAP schools from the analysis.

1. District Learning

We tested the hypothesis that impacts on test scores might be small initially, but would grow as the school district became more familiar and experienced with Chicago TAP. According to this hypothesis, the staff operating the program would grow and refine their ability to support school principals and teachers over time. However, the evidence did not suggest that impacts grew over the four years in which the program was rolled out in Chicago. The impacts on scores in all three subjects (reading, math, and science) were largest in the third year (and statistically significant in reading), but fell in the fourth year to zero or negative and not statistically significant.

2. School Learning

We tested another hypothesis about start-up effects, which is that impacts on test scores would grow as individual *schools* became more accustomed to the program. Again, we found no consistent evidence of a learning effect from either the experimental or quasi-experimental analyses. We observed schools with one, two, three, and four years of experience implementing Chicago TAP but the years of experience had no relationship to size of the impact estimate.

3. Other Explanations

We considered other explanations and found nothing to suggest that there was a hidden positive impact of Chicago TAP on student achievement. For example, we considered the possibility that certain groups of schools, regardless of when they started or how much experience they had, were simply more successful with Chicago TAP than others, but we found no impact for any of the four cohorts examined on their own. There was evidence of a significant negative impact on reading scores for cohort 4 and a significant positive impact on math scores for cohort 2, but we should exercise caution in interpreting such results. The probability of erroneously labeling an estimate as significant rises as the number of hypothesis tests increases. In this case, we conducted 12 hypothesis tests (three subjects for four cohorts).

C. Impacts on Retention

The other major outcome besides student achievement was teacher retention. Preventing large losses of teachers is important because it helps a school maintain its culture and avoid disruption. We hypothesized that teachers would find aspects of Chicago TAP appealing and would want to continue teaching in their schools at a higher rate than they would have if their school had not participated in Chicago TAP.

We found that teachers who were working in Chicago TAP schools in 2007 did return in each of the following three years at higher rates than teachers in comparable non-TAP schools. When we looked at teachers who were working in schools that started in later years, some of the impact estimates were not statistically significant. We also found some evidence that impacts on retention were stronger for subgroups of teachers, such as those with less experience, but the pattern was not consistent. The overall conclusion was that Chicago TAP most likely had an impact, inducing teachers to stay longer in their schools, but the impacts were not uniform or universal across years, cohorts, and subgroups of teachers.

To understand the retention findings in more depth we examined the destination of teachers who did not return to their schools, and we tabulated for teachers in Chicago TAP schools the observed performance of stayers, movers (those who left their school but stayed within CPS), and leavers. We did not find much evidence that Chicago TAP helped prevent teachers from leaving the district. Much of the attrition from study schools was due to movement between schools. When we examined the SKR scores of stayers, movers, and leavers, we failed to find evidence that attrition came from the top or bottom of the teacher performance distribution in Chicago TAP schools. Comparable data were not available for non-TAP schools, but if a similar pattern held in all schools, we could conclude that teacher attrition was not changing the composition of the schools very dramatically. Thus the chief benefit of a positive impact on retention appears to be through reduction in the disruptive effect of turnover and of having to find replacement teachers.

D. Conclusions

Implementation of Chicago TAP increased the amount of mentoring, promotion opportunity, and compensation relative to non-TAP schools, and these increases alone may have translated into making Chicago TAP schools a more desirable place to continue working, as evidenced by the positive impacts on retention. However, these changes did not in turn result in higher student achievement, often considered the "bottom line" for education interventions, within the specified four-year time frame.

Despite the real changes in Chicago TAP schools noted in the implementation section, some key aspects of performance-based pay were not really tested: (1) there was no individual teacher-level component of test-based accountability; and (2) the rewards may not have been sufficiently meaningful and differentiated—they fall just below the thresholds suggested by the U.S. Department of Education, for example, in its guidance to applicants to the most recent round of the Teacher Incentive Fund competition.[23]

Chicago TAP was only partially successful in achieving its goals. The program can be credited with improved retention outcomes for some of its schools, but it did not have a noticeable positive impact on student achievement over the four-year rollout in Chicago. This result provides a caution to funders investing in future programs in terms of what to expect over a four-year period. However, designers of new policies might consider how to change selected program elements to produce more favorable outcomes in the future.

[23] The *Federal Register* notice (http://edocket.access.gpo.gov/2010/pdf/2010-12218.pdf, accessed January 5, 2012) provides an example of a meaningful and differentiated compensation system as one with an average payout of 5 percent of average teacher salary with a possible payout of 15 percent for top-performing teachers and principals, or at least levels that are "high enough to create change in the behavior of current and prospective teachers and principals."

This page has been left blank for double-sided copying.

REFERENCES

Ahmed, Azam. "CPS Teacher Compensation Fails to Pay Off." June 1, 2010. Available at [http://archive.chicagobreakingnews.com/2010/06/cps-teacher-compensation-fails-to-pay-off.html]. Accessed December 28, 2011.

Chicago Board of Education and Chicago Teachers Union. "Memorandum of Understanding Between the Chicago Board of Education and Chicago Teachers Union, Local No. 1, AFT, AFL-CIO." July 2007. Available at [http://www.cpef.org/ctap/files/CPS_CTU_MOU.pdf]. Accessed January 19, 2009.

Chicago Public Education Fund. "FAQs." Available at [http://www.cpef.org/ctap/faqs.php]. Accessed December 19, 2008.

Chicago TAP. "Chicago TAP: Recognizing Excellence in Academic Leadership." 2009. Available at [http://www.chicagotapschools.org/]. Accessed March 3, 2010.

Crown, Laurel. "Interim Report for the Year 3 Internal Evaluation of Chicago TAP: Findings from the Fall, 2009 Chicago TAP Teacher Survey." Chicago: Chicago Public Schools, March 2010. Available at [https://research.cps.k12.il.us/resweb/DownLoaderAdv?dir=program_evaluation&file=Fall09TAPSurvey_report_FINAL.pdf]. Accessed February 29, 2012.

Daley, Glenn, and Lydia Kim. "A Teacher Evaluation System That Works." Working paper. Santa Monica, CA: National Institute for Excellence in Teaching, August 2010.

Foster, Jessica. "Implementation Findings for the Chicago TAP Program: 2007–08." Chicago: Chicago Public Schools, February 2009. Available at [https://research.cps.k12.il.us/resweb/DownLoaderAdv?dir=program_evaluation&file=Chicago_TAP_Year_1_Report_07-08.pdf]. Accessed February 29, 2012.

Glazerman, Steven, Dan Levy, and David Myers. "Nonexperimental Versus Experimental Estimates of Earnings Impacts." *The Annals of the American Academy of Political and Social Science,* April 2003, pp. 3-42.

Glazerman, Steven, Allison McKie, Nancy Carey, and Dominic Harris. "An Evaluation of the Teacher Advancement Program (TAP) in the Chicago Public Schools: Study Design Report." Washington, DC: Mathematica Policy Research, November 2007.

Glazerman, Steven, Allison McKie, and Nancy Carey. "An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year One Impact Report." Washington, DC: Mathematica Policy Research, April 2009.

Glazerman, Steven, and Allison Seifullah. "An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year Two Impact Report." Washington, DC: Mathematica Policy Research, June 2010.

Glazerman, Steven, Sarah Senesky, Neil Seftor, and Amy Johnson. "Design of an Impact Evaluation of Teacher Induction Programs." Washington, DC: Mathematica Policy Research, January 2006.

Hudson, Sally. "The Effects of Performance-Based Teacher Pay on Student Achievement." Undergraduate thesis. Stanford, CA: Stanford University, July 30, 2010.

Illinois State Board of Education. "Interpretative Guide 2010 Illinois Standards Achievement Test: Reading Mathematics Science Writing." 2010. Available at [http://www.isbe.state.il.us/assessment/pdfs/ISAT_Interpr_Guide_2010.pdf.] Accessed January 6, 2011.

McEntegart, Damian J. "The Pursuit of Balance Using Stratified and Dynamic Randomization Techniques: An Overview." *Drug Information Journal*, Vol. 37, 2003, pp. 293–308.

NIET (National Institute for Excellence in Teaching). "Teacher Advancement Program Implementation Manual: Customized for Chicago Public Schools." March 2008. Available at [http://www.cpef.org/ctap/files/Implementation_Manual.pdf]. Accessed November 16, 2010.

Schacter, John, Tamara Schiff, Yeow Meng Thum, Cheryl Fagnano, Micheline Bendotti, Lew Solmon, Kimberly Firetag, and Lowell Milken. "The Impact of the Teacher Advancement Program on Student Achievement, Teacher Attitudes, and Job Satisfaction." Santa Monica, CA: Milken Family Foundation, November 2002.

Schacter, John, Yeow Meng Thum, Daren Reifsneider, and Tamara Schiff. "The Teacher Advancement Program Report Two: Year Three Results from Arizona and Year One Results from South Carolina TAP Schools." Santa Monica, CA: Milken Family Foundation, March 2004.

Solmon, Lewis, J. Todd White, Donna Cohen, and Deborah Woo. "The Effectiveness of the Teacher Advancement Program." Santa Monica, CA: National Institute on Excellence in Teaching, April 2007.

Springer, Matthew, Dale Ballou, and Art (Xiao) Peng. "Impact of the Teacher Advancement Program on Student Test Score Gains: Findings from an Independent Appraisal." National Center on Performance Incentives Working Paper 2008-19. 2008. Available at http://www.performanceincentives.org/data/files/news/PapersNews/Springer_et_al_2008.pdf.] Accessed March 3, 2010.

APPENDIX A

PROPENSITY SCORE MATCHING

This page has been left blank for double-sided copying.

We identified nearly 400 CPS K–8 elementary schools to serve as potential matched comparison schools for the Chicago TAP schools in the study (cohorts 1, 2, 3, and 4). To form the best possible comparison group from among these schools, we employed several propensity score matching methods (algorithms). Using the same set of propensity scores, each algorithm selects a different set of comparison schools and generates a corresponding set of weights. The goal was to find a comparison group with a close fit to the Chicago TAP schools under study, as judged by preintervention measures of the school characteristics that are related to study outcomes (baseline covariates). Because we conducted analyses over several years, we rematched in each year using the same baseline covariates but a different set of Chicago TAP schools, depending on which ones were implementing the program in that year. We did this separately for the test score analysis and for the retention analysis because the set of Chicago TAP schools in each analysis differed slightly in any given year. The procedures are described in Chapter II. This appendix provides additional detail on the matching algorithms and how we selected one to use as a default for presentation purposes.

The process of matching schools and selecting a comparison group consisted of several steps. First, we first applied filters to the universe of CPS K–8 elementary schools in each year to create a set of potential comparison schools that matched the Chicago TAP schools on basic characteristics. We then included the Chicago TAP schools and the potential comparison schools in a logistic regression model that used a number of matching variables to predict the probability of being selected into the Chicago TAP finalist pool. Using the predicted probabilities ("propensity scores") from this regression, different matching algorithms were then applied to create alternative comparison groups. Finally, we conducted diagnostics to select one matching algorithm to highlight in the presentation of results. Below we provide further details about this process.

1. Exact Match Criteria/Filters

Before estimating propensity scores, we first selected as potential comparison schools those that met the following basic criteria:

- The school was open during the 2006–2007 school year and had data from that year.
- The school had at least five of the six tested grades, which were grades three through eight.
- At least 50 percent of the school's students were low income (defined as being eligible for free/reduced-price lunch). To be eligible for Chicago TAP, the school had to serve at least 75 percent low-income students. We used percentage of low-income students as a matching variable, but allowed the possibility of schools below the eligibility threshold matching with schools above it. This restriction prevents matching to schools with especially low percentages of low-income students.
- For teacher retention analysis only: The school was not a charter school. Teaching assignments for charter schools were not available in the administrative data provided by CPS.
- For teacher retention analysis only: The school was not selected for Chicago TAP. We did not want the possibility of a comparison school's staff knowing it would be

implementing the program in the future, as such knowledge could affect teachers' decisions to return to the school.[24] (Note that for the test score analysis, we allowed a future Chicago TAP school to serve as a comparison school during a year before they implemented Chicago TAP.)

2. Matching Variables

We matched schools on variables that were measured before the rollout of Chicago TAP, including preintervention measures of the outcomes of interest: student test scores and teacher retention.[25] We standardized spring 2007 math and reading ISAT scores within grade to have a common mean and standard deviation by grade (zero and one, respectively) and then averaged across grades for each school. Standardizing the test scores reduces the influence of having different proportions of students in different grade levels. Retention rates were based on CPS human resources data and expressed as the percentage of classroom teachers in fall 2005 returning as classroom teachers in fall 2006 to the same school. Separately, we measured retention for teachers who were in their first four years of service in the district and those who had 5 to 24 years of service. We did not take into account retention rates for teachers close to retirement age. This group could not be stably estimated, nor does retention at this stage have the same interpretation as novice and midcareer retention.

We also used 2006–2007 student demographic information, including total school enrollment, enrollment squared (to improve matches for very small or large schools), and race/ethnicity. Because we observed that most Chicago TAP schools could be categorized as nearly all African American, with a few nearly all Hispanic or mixed, we collapsed school race/ethnicity into a small number of categories to emphasize substantive, rather than minute, qualitative differences. If less than one-third of a school's students were African American, it was given a value of 1; if one-third to two-thirds were African American, it had a value of 2; and more than two-thirds had a value of 3. (This is equivalent to rounding the fraction of African American students to the nearest third). We coded the percentage of Hispanic students in the same way. We also used the percentages of students who were low income (eligible for free/reduced-price lunch), in special education (had an Individualized Education Program), and limited English proficient. Finally, we used indicators of whether the school had made adequate yearly progress toward goals under No Child Left Behind and, for the student achievement analysis, whether the school was a charter school.

All of the matching variables were used in a logistic regression to estimate the theoretical probability for selecting a pool of Chicago TAP schools to enter the program. The predicted probability is the propensity score. We examined the score distributions and selected or reweighted potential comparison schools to form the best possible comparison group.

---

[24] Two exceptions occurred to this retention analysis filter. We allowed cohort 3 and 4 schools to be comparison schools for cohort 1 schools in 2007–2008. Because teachers in cohort 3 and 4 schools did not learn of their future Chicago TAP status until spring 2009, fall 2007 to fall 2008 retention could not have been affected by knowledge of future Chicago TAP status. We also allowed the noncharter cohort 4 replacement school to be a comparison school for cohort 1 and 2 schools in 2008–2009 because the school was not announced as a replacement school until spring 2010; the school's future Chicago TAP status could not have affected fall 2008 to fall 2009 retention decisions.

[25] For a small number of schools that expanded by more than two grades between 2006–2007 and 2008–2009, we used a later baseline. All variables except teacher retention for these schools were measured in 2008–2009; teacher retention was measured as the percentage of teachers in fall 2007 who returned to the school in fall 2008.

3. Matching Algorithms

There are several alternative algorithms for selecting a comparison group, each of which has advantages and disadvantages. The nearest-neighbor method is probably the most intuitive because it is analogous to a balanced random assignment experiment and gives each Chicago TAP school a fixed number of comparison schools (albeit with some counting more than once because of replacement). The propensity score was used to rank all the schools sequentially along a number line. We formed two nearest-neighbor comparison groups: one using the single nearest neighbor to each Chicago TAP school and the other using the nearest five neighbors. We selected the nearest neighbors with replacement so that each Chicago TAP school in each year was matched to the non-TAP schools whose propensity scores were closest to the Chicago TAP school, regardless of whether the non-TAP schools were also among the nearest neighbors of any other Chicago TAP school. Non-TAP schools matched with more than one Chicago TAP school received proportionally more weight in the analysis.

Another algorithm is called the caliper method because we define a fixed distance (the radius of a caliper) in terms of propensity score from each Chicago TAP school and select all comparison schools that fall within that distance. The radius size for the caliper is arbitrary and involves a trade-off between the quality and quantity of matches. A larger radius captures more comparison schools, but a smaller one captures schools that are more closely matched. We examined radii of different lengths and used the ones that rendered superior matches in terms of the matching variables described above. With very small radii we found that some Chicago TAP schools had no matches; rather than discard these schools, we used larger radii.

Finally, we used kernel density matching, which uses the full set of comparison schools but allows the weights to vary with distance from Chicago TAP schools. For each Chicago TAP school, the weight corresponding to each comparison school is smaller as the distance from the Chicago TAP school is greater.[26]

4. Diagnostics

Most of the matching algorithms produced similar results, but we had to select one to simplify the presentation. We sought to make a selection based on the quality of the matches, which could be measured prior to seeing any outcome data. This task was complicated by the fact that we used propensity score matching to form comparison groups for each cohort and year and did so separately for the test score and retention analyses. Matching this many times meant that an algorithm that produced the closest balance in baseline covariates for one match might not do so for another. We decided to select one algorithm for all analyses, provided that it generated the greatest balance for key matches and acceptable balance for all other matches.

We chose the nearest-five-neighbors approach because among the algorithms for which all Chicago TAP schools had matches, this algorithm reduced the initial Chicago TAP-comparison differences more than the others. In particular, the nearest-five-neighbors approach most often had the smallest mean standardized bias across all the matching covariates across the student

[26] The magnitude of the weight is based on the probability density function (PDF) for the normal distribution, which looks like a bell-shaped curve sitting on a number line, centered on the propensity score for each Chicago TAP school. The weight is proportional to the height of the curve (the kernel) of the normal PDF.

achievement analyses, where the standardized bias for each covariate is the difference between the Chicago TAP and comparison group sample means as a percentage of the square root of the average of the sample variances in both groups.

Table A.1 reports the sample sizes and mean standardized biases considered in selecting our preferred matching method. We first determined the number of Chicago TAP and comparison schools that could be matched for the student achievement analysis each year using each method. In presenting our findings, we preferred to highlight a matching method that always included all Chicago TAP elementary schools implementing the program; we therefore eliminated from consideration methods that failed to produce matches for all Chicago TAP schools in any year.[27] Of the remaining methods, we chose the one that most often had the smallest standardized mean bias, the nearest-five-neighbors method.

[27] As shown in the main report, we assess the robustness of our findings to choice of matching method by also reporting results estimated using alternative matching methods that sometimes discard Chicago TAP schools.

Table A.1. Mean Standardized Bias and Sample Sizes for Student Achievement Analyses, by Matching Method

Matching Method	Mean Standardized Bias (percentage)	Sample Sizes (schools)	
		Chicago TAP	Comparison
2007–2008 (cohort 1)			
Nearest neighbor	16.8	9	9
Nearest 5 neighbors	13.6	9	34
Caliper match (radius = 0.005)	12.3	9	105
Caliper match (radius = 0.025)	8.4	9	385
Kernel density	13.4	9	391
2008–2009 (cohorts 1 & 2)			
Nearest neighbor	15.2	18	18
Nearest 5 neighbors	8.9	18	77
Caliper match (radius = 0.005)	3.3	17[a]	207
Caliper match (radius = 0.025)	3.9	17[a]	375
Kernel density	7.1	18	379
2009–2010 (cohorts 1, 2, & 3)			
Nearest neighbor	12.7	25	25
Nearest 5 neighbors	4.0	25	97
Caliper match (radius = 0.005)	2.6	22[a]	180
Caliper match (radius = 0.025)	2.9	24[a]	368
Kernel density	3.7	25	369
2010–2011 (cohorts 1, 2, 3, & 4)			
Nearest neighbor	8.5	33	28
Nearest 5 neighbors	2.9	33	100
Caliper match (radius = 0.005)	4.6	29[a]	148
Caliper match (radius = 0.025)	4.7	31[a]	357
Kernel density	2.3	32[a]	359

Note: Matching methods shown in red were not chosen as the preferred method because they did not produce a match for at least one Chicago TAP school in at least one year. The nearest-five-neighbors method, shown shaded, had the smallest mean standardized bias among the methods that produced matches for all Chicago TAP schools in every year. Williams Elementary (serving grades pre-K–five) and Williams Middle (serving grades six–eight) were selected for Chicago TAP cohort 3; these two schools were combined as a single "virtual" school in the analyses.

[a] At least one Chicago TAP school did not have a match and was discarded from the sample.

This page has been left blank for double-sided copying.

APPENDIX B

SUPPLEMENTAL TABLES

This page has been left blank for double-sided copying.

Table B.1. Mentoring Received (matched comparison)

Outcome	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error
Received professional advice and assistance in teaching duties from an advisor (percentage)	96.7	75.2	21.6***	3.21
Had an advisor who was a(n) . . . (percentage)				
Mentor	57.4	8.7	48.7***	3.32
Literacy coach	43.8	42.9	0.9	11.03
Math coach	23.4	22.2	1.2	10.04
Lead teacher	75.0	11.9	63.1***	5.06
Principal	64.7	35.5	29.2***	6.78
Assistant or vice principal	47.5	31.6	15.8	8.34
Peer	40.9	32.2	8.7	5.76
Had a main advisor who was a . . . (percentage)				
Full-time teacher	61.0	36.2	24.8***	4.08
Person who works in teacher's school only	83.7	55.3	28.4***	6.47
Person who works in more than one school	5.5	14.2	-8.7***	3.29
Teacher with release time	45.9	24.0	21.9***	7.02
Person with no classroom teaching	60.9	39.6	21.3***	6.93
Principal or school administrator	20.6	19.9	0.8	5.85
School-based specialist	43.2	34.9	8.4	7.00
District specialist	3.0	5.3	-2.3	1.76
Person from a teacher licensing, certification, or preparation program	21.3	9.0	12.3***	3.77
Time spent with main advisor				
Frequency of scheduled meetings (number per week)	1.4	0.8	0.6***	0.13
Duration of each scheduled meeting (minutes)	66.7	39.9	26.8***	6.26
Duration of informal contact (minutes per week)	76.3	48.4	27.9*	14.05
Frequency of total contact (minutes per week)	173.1	86.6	86.5***	19.41
During most recent full week, scheduled time main advisor spent . . . (minutes)				
Observing teacher teaching	27.7	15.6	12.0***	4.20
Meeting with teacher one-on-one	27.3	23.5	3.8	6.34
Meeting with teacher together with other teachers	40.8	26.8	14.0*	6.86
Modeling a lesson	20.6	6.2	14.4***	3.25
Coteaching a lesson	10.2	5.7	4.4	3.83
Received useful feedback from main advisor (percentage)	88.7	91.3	-2.6	2.85

Note: N = 304 to 382 teachers per outcome.

[a] Means are regression adjusted.

* Chicago TAP-control difference is statistically significant at the 10 percent level.
** Chicago TAP-control difference is statistically significant at the 5 percent level.
*** Chicago TAP-control difference is statistically significant at the 1 percent level.

Table B.2. Mentoring Provided (teachers with at least five years of experience, matched comparison)

Outcome	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error
Provided formal mentoring services (percentage)	27.5	22.6	4.9	5.04
Mentoring topics included . . . (percentage)				
Strategies for teaching literacy	25.9	18.0	7.9*	4.68
Strategies for teaching math	16.8	17.0	-0.1	4.47
Strategies for teaching other subjects	16.3	17.0	-0.7	4.83
Increasing content area knowledge	17.3	18.6	-1.3	5.57
Selecting or adapting curriculum materials	19.0	20.0	-1.0	3.94
Teaching or aligning curriculum to meet state or district standards	21.0	17.6	3.4	4.49
Aligning local curriculum assessment to state standards	17.6	16.4	1.2	3.70
Setting instructional goals and determining ways to achieve them	26.0	20.1	5.8	4.92
Preparing students for standardized tests	18.9	14.4	4.5	4.32
Using assessments to inform teaching	23.7	20.8	2.9	4.36
Preparing lesson plans or other instructional activities	22.4	19.5	2.9	4.52
Providing differentiated instruction to meet student needs	23.6	21.2	2.4	4.84
Received release time for mentoring (percentage)	17.3	7.3	10.0***	2.97
Release time for mentoring (hours per week)	1.4	0.2	1.2***	0.41
Mentoring outside of specified contract hours (hours per week)	1.9	0.9	1.0	0.67
Teachers mentored (number)	2.1	0.8	1.3***	0.44
Frequency of scheduled meetings (number per week per teacher)	0.4	0.6	-0.2	0.12
Duration of each scheduled meeting (minutes)	13.8	10.0	3.8	3.15
Informal contact with all teachers (minutes per week)	69.0	26.6	42.4*	20.90
Total contact with all teachers (minutes per week)	196.8	74.5	122.3**	55.89
Mentoring activities included . . . (percentage)				
Observing teaching	26.3	18.6	7.6*	4.49
Meeting with teachers one-on-one	27.6	19.8	7.8	5.13
Meeting in small groups or clusters	23.9	11.2	12.7***	3.57
Modeling a lesson	25.2	18.9	6.4	4.73
Coteaching a lesson	18.2	13.3	4.9	3.73
Writing evaluations	19.6	10.9	8.7*	4.57
During most recent full week, scheduled time spent . . . (minutes)				
Observing teaching	51.0	29.8	21.2**	9.80
Meeting with teachers one-on-one	38.9	19.8	19.2**	9.21
Meeting in small groups or clusters	31.1	11.3	19.8**	7.98
Modeling a lesson	23.8	16.0	7.9	10.12
Coteaching a lesson	16.5	8.6	7.9	7.73
Writing evaluations	41.2	5.7	35.4***	11.99

Note: N = 291 to 315 teachers per outcome.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

Table B.3. Other Leadership Roles and Responsibilities (teachers with at least five years of experience, matched comparison)

Outcome	Chicago TAP Mean[a]	Control Mean[a]	Difference	Standard Error
Had other leadership roles or responsibilities beyond mentoring (percentage)	37.1	48.3	-11.1*	5.95
Other leadership roles included . . . (percentage)				
Being a lead teacher	14.2	12.8	1.4	2.55
Being a department head or chair	0.8	1.6	-0.8*	0.46
Being a grade-level lead teacher	9.2	20.9	-11.7*	6.13
Being on a school improvement team	15.6	22.4	-6.8	5.15
Being on a schoolwide committee/task force	11.1	16.2	-5.2	5.14
Other leadership responsibilities included . . . (percentage)				
Setting school policies	8.8	10.1	-1.3	3.69
Developing curriculum	13.4	20.0	-6.7*	3.96
Reviewing/selecting curriculum	16.3	22.8	-6.6**	3.27
Providing input on improving facilities/technology	10.5	16.0	-5.5	4.67
Providing professional development activities	22.0	15.3	6.7**	3.32
Developing standards	8.3	9.9	-1.6	4.18
Evaluating teachers	1.1	0.1	1.0***	0.38
Associated with these other leadership roles and responsibilities, received . . . (percentage)				
Credit toward certification	6.3	7.0	-0.7	2.44
Pay increase	16.3	1.6	14.7***	4.09

Note: N = 303 to 311 teachers per outcome.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

Table B.4. Observation and Feedback (matched comparison)

Outcome	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error
Frequency of observation (number in 2009–2010)				
Observation by principal or assistant principal	3.0	2.5	0.5	0.28
Observation by mentor, coach, or lead teacher	3.3	1.9	1.4***	0.29
Frequency of feedback (number in 2009–2010)				
Feedback as part of a formal evaluation	2.5	2.1	0.4	0.23
Feedback outside of a formal evaluation	2.7	2.1	0.6**	0.25
Feedback on lesson plans	2.2	2.1	0.1	0.39

Note: N = 374 to 378 teachers per outcome.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

Table B.5. Professional Development Received (matched comparison)

Outcome	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error
Participated in professional development activities that addressed . . . (percentage)				
Strategies for teaching literacy	94.0	94.1	-0.1	2.62
Strategies for teaching math	82.7	68.6	14.2***	5.40
Strategies for teaching other subjects	65.8	63.3	2.5	4.08
Increasing content area knowledge	100.0	100.0	0.0	0.00
Selecting or adapting curriculum materials	68.7	68.6	0.0	5.52
Teaching or aligning curriculum to meet state or district standards	80.6	78.3	2.3	5.80
Aligning local or teacher-developed curriculum assessment to state standards	69.1	70.2	-1.1	6.72
Setting instructional goals and determining ways to achieve them	77.6	81.0	-3.4	3.51
Preparing students for standardized tests	72.6	72.7	0.0	6.15
Using assessments to inform teaching	87.9	88.0	0.0	3.83
Preparing lesson plans or other instructional activities	76.6	75.1	1.5	3.39
Providing differentiated instruction to meet student needs	86.0	88.4	-2.4	3.58
Responded that professional development in 2009–2010 . . . (percentage)				
Was useful to their teaching	84.0	85.1	-1.1	4.66
Was more satisfactory than in previous years	34.8	37.6	-2.8	3.60
Had been implemented in their teaching	100.0	100.0	0.0	0.00
In association with professional development, received . . . (percentage)				
Scheduled nonteaching time in contract year	83.0	68.6	14.4***	4.20
Other release time from teaching	45.1	48.5	-3.4	5.13
Stipend	67.5	64.2	3.3	6.52
Fee reimbursement	19.1	21.3	-2.2	5.46
Travel or expense reimbursement	6.1	14.1	-8.1**	4.09
Course credits toward certification	50.3	49.6	0.7	5.64
Pay increase	26.5	22.1	4.4	4.58
Recognition or higher ratings on an annual teacher evaluation	25.0	27.8	-2.9	4.44

Note: N = 367 to 383 teachers per outcome.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

Table B.6. Compensation (matched comparison)

Outcome	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error
Academic-year base salary ($)	61,958	61,792	166	1,398.65
Base salary included leadership compensation (percentage)	12.2	13.3	-1.1	2.89
Additional compensation for leadership was expected (percentage)	22.3	7.0	15.3***	3.37
Expected amount of additional compensation for leadership ($)	1,797	188	1,609***	406.57
Eligible for additional nonleadership compensation (percentage)	81.5	39.7	41.8***	5.56
Eligible for additional nonleadership compensation based on . . . (percentage)				
Instructional performance	71.0	6.8	64.2***	6.74
Student achievement growth	64.0	8.0	56.0***	5.64
Instructional performance or student achievement growth	78.1	9.2	68.9***	6.02
Subject matter taught	19.3	6.5	12.7***	3.87
Student population taught	13.4	7.6	5.8	3.59
Professional development	32.8	31.5	1.2	6.64
University courses taken	13.0	11.7	1.3	3.90
Expected or had received additional nonleadership compensation (percentage)	54.3	20.4	33.9***	5.41
Expected or had received additional nonleadership compensation based on... (percentage)				
Instructional performance	35.3	0.8	34.5***	6.08
Student achievement growth	28.1	1.0	27.1***	5.19
Instructional performance or student achievement growth	43.0	1.1	41.9***	5.86
Subject matter taught	0.4	0.0	0.3*	0.18
Student population taught	0.0	0.0	0.0	0.00
Professional development	16.3	15.2	1.1	4.50
University courses	0.0	0.0	0.0	0.00
Expected amount of additional nonleadership compensation ($)	880	239	641***	176.09
Expected additional compensation from an outside job (percentage)	12.1	14.7	-2.7	4.40

Note: N = 334 to 370 teachers per outcome.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

Table B.7. Teacher Attitudes (matched comparison)

Outcome	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error
Satisfied with . . . (percentage)				
Supportive atmosphere/collaboration with colleagues	79.7	81.9	-2.2	4.78
Administration support	76.3	79.3	-3.0	7.55
Policies/practices input	73.9	75.7	-1.8	6.60
Classroom autonomy	88.5	90.9	-2.5	3.71
Professional development opportunities	88.4	86.7	1.7	3.02
Caliber of colleagues	77.7	84.6	-6.9	6.19
Salary and benefits	88.2	80.0	8.2***	3.15
Leadership opportunities	79.2	81.9	-2.7	4.50
School policies	72.7	74.2	-1.5	6.11
District policies	55.4	57.4	-2.0	7.89
Agreed that the principal . . . (percentage)				
Works to create a sense of community	74.7	82.2	-7.6	7.58
Is strongly committed to shared decision making	74.3	79.9	-5.6	6.99
Promotes parent/community involvement	86.0	88.0	-2.1	4.72
Supports and encourages risk taking	71.9	79.9	-7.9	6.81
Is willing to make changes	87.6	84.7	2.9	5.36
Strongly supports most changes	80.8	80.8	0.1	5.83
Encourages trying new instructional methods	89.6	90.1	-0.5	4.40

Note: N = 370 to 379 teachers per outcome.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

Table B.8. Impacts on District Retention Rates (percentage)

Retention Period	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error	Sample Size (teachers)
One- Year Rates					
Fall 2007–fall 2008 (cohort 1)	92.3	89.6	2.7	2.26	1,102
Fall 2008–fall 2009 (cohorts 1 & 2)	91.2	89.7	1.5	1.40	1,686
Fall 2009–fall 2010 (cohorts 1, 2, & 3)	85.8	87.3	-1.5	1.60	2,694
Two- Year Rates					
Fall 2007–fall 2009 (cohort 1)	86.6	81.9	4.7**	2.08	881
Fall 2008–fall 2010 (cohorts 1 & 2)	81.1	79.7	1.4	2.47	1,879
Three- Year Rates					
Fall 2007–fall 2010 (cohort 1)	75.9	72.1	3.8	2.62	781

Note: District retention rate is defined as the percentage of classroom teachers in fall of the base year who worked in the district in fall of the follow-up year.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

Table B.9. Impacts on District Retention Rates, by Teaching Assignment (percentage)

Teaching Assignment	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error	Sample Size (teachers)
One- Year Rates					
Fall 2007–fall 2008 (cohort 1)					
All teachers	92.3	89.6	2.7	2.26	1,102
Academic subjects, tested grades/subjects	93.4	92.5	0.9	2.29	323
Academic subjects, nontested grades/subjects	96.2	91.5	4.8***	1.81	497
Other	98.8	97.2	1.5	1.81	164
Fall 2008–fall 2009 (cohorts 1 & 2)					
All teachers	91.2	89.7	1.5	1.40	1,686
Academic subjects, tested grades/subjects	93.4	90.6	2.7	1.95	551
Academic subjects, nontested grades/subjects	92.9	92.3	0.6	1.72	777
Other	94.9	94.3	0.6	2.96	239
Fall 2009–fall 2010 (cohorts 1, 2, & 3)					
All teachers	85.8	87.3	-1.5	1.60	2,694
Academic subjects, tested grades/subjects	83.4	88.8	-5.4*	3.10	903
Academic subjects, nontested grades/subjects	90.5	88.5	2.0	1.86	1,290
Other	84.1	90.9	-6.8*	3.55	385
Two- Year Rates					
Fall 2007–fall 2009 (cohort 1)					
All teachers	86.6	81.9	4.7**	2.08	881
Academic subjects, tested grades/subjects	85.8	85.5	0.3	3.77	278
Academic subjects, nontested grades/subjects	92.7	90.5	2.1	1.88	399
Other	100.0	99.9	0.1	0.31	114
Fall 2008–fall 2010 (cohorts 1 & 2)					
All teachers	81.1	79.7	1.4	2.47	1,879
Academic subjects, tested grades/subjects	83.0	82.9	0.1	3.88	587
Academic subjects, nontested grades/subjects	85.4	84.5	0.9	2.46	888
Other	78.1	75.5	2.6	6.71	264
Three- Year Rates					
Fall 2007–fall 2010 (cohort 1)					
All teachers	75.9	72.1	3.8	2.62	781
Academic subjects, tested grades/subjects	81.0	68.4	12.6**	5.10	221
Academic subjects, nontested grades/subjects	78.8	80.2	-1.5	3.16	366
Other	78.7	79.0	-0.3	6.94	106

Note: District retention rate is defined as the percentage of classroom teachers in fall of the base year who worked in the district in fall of the follow-up year. Teaching assignment category is missing for 118 teachers for fall 2007–fall 2008, 119 teachers for fall 2008–fall 2009, 116 teachers for fall 2009–fall 2010, 90 teachers for fall 2007–fall 2009, 140 teachers for fall 2008–fall 2010, and 88 teachers for fall 2007–fall 2010.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

Table B.10. Impacts on District Retention Rates, by Years of Service (percentage)

Years of Service	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error	Sample Size (teachers)
One- Year Rates					
Fall 2007–fall 2008 (cohort 1)					
All teachers	92.3	89.6	2.7	2.26	1,102
<5 years	89.2	87.1	2.1	3.28	371
5–24 years	94.2	92.8	1.4	1.89	634
>24 years	100.0	99.7	0.3	0.63	97
Fall 2008–fall 2009 (cohorts 1 & 2)					
All teachers	91.2	89.7	1.5	1.40	1,686
<5 years	98.9	98.8	0.1	0.40	538
5–24 years	94.9	93.8	1.1	1.21	1,010
>24 years	81.2	86.0	-4.8	5.93	138
Fall 2009–fall 2010 (cohorts 1, 2, & 3)					
All teachers	85.8	87.3	-1.5	1.60	2,694
<5 years	81.5	82.6	-1.0	4.20	731
5–24 years	89.8	91.4	-1.6	1.56	1,741
>24 years	67.2	77.4	-10.2	8.64	222
Two- Year Rates					
Fall 2007–fall 2009 (cohort 1)					
All teachers	86.6	81.9	4.7**	2.08	881
<5 years	90.5	93.5	-3.0**	1.20	301
5–24 years	97.8	93.8	4.0*	2.21	495
>24 years	88.3	77.2	11.1	9.81	85
Fall 2008–fall 2010 (cohorts 1 & 2)					
All teachers	81.1	79.7	1.4	2.47	1,879
<5 years	83.7	76.1	7.7**	3.54	540
5–24 years	84.3	85.5	-1.3	2.91	1,165
>24 years	55.2	51.5	3.8	12.14	174
Three- Year Rates					
Fall 2007–fall 2010 (cohort 1)					
All teachers	75.9	72.1	3.8	2.62	781
<5 years	75.9	71.1	4.8	5.33	247
5–24 years	83.1	76.6	6.5	4.64	469
>24 years	58.7	50.7	8.0	32.14	65

Note: District retention rate is defined as the percentage of classroom teachers in fall of the base year who worked in the district in fall of the follow-up year.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

Table B.11. Impacts on District Retention Rates, Sensitivity Analysis by Matching Method (percentage)

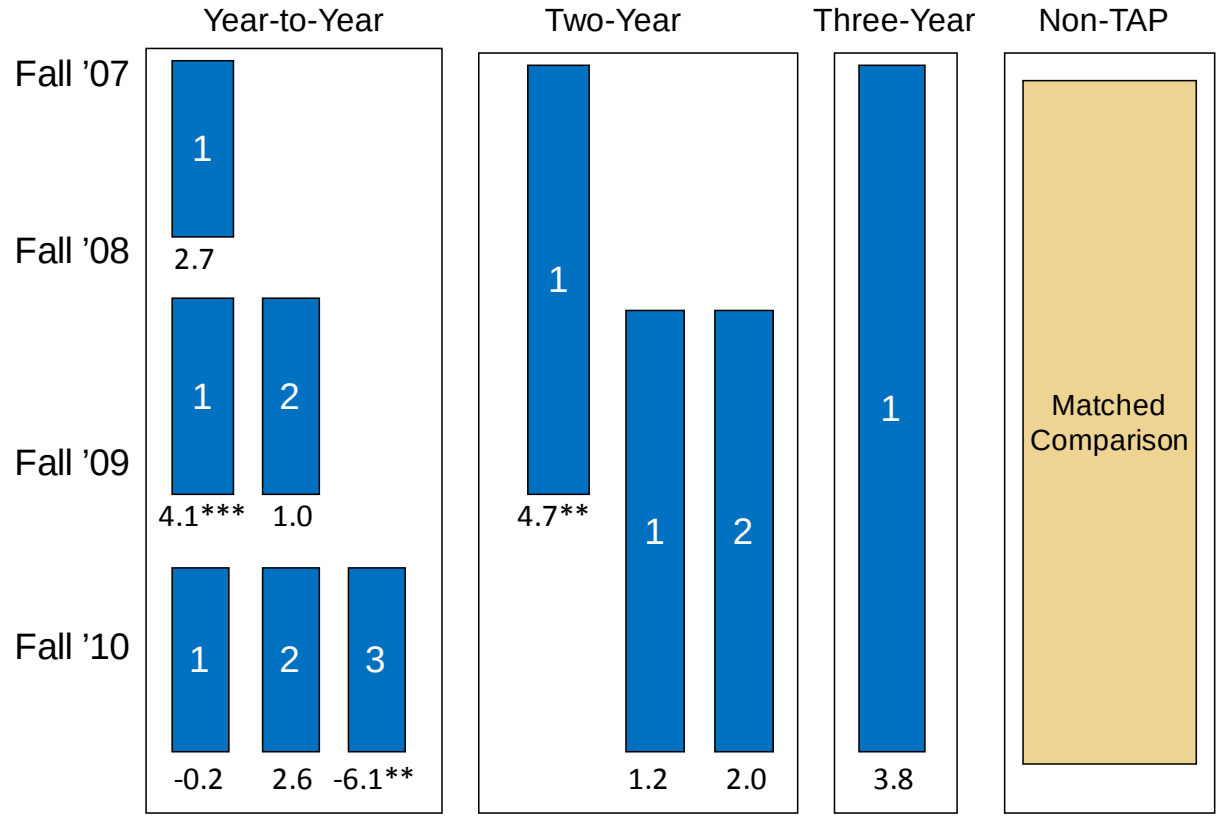
Matching Method	Chicago TAP Mean[a]	Comparison Mean[a]	Difference	Standard Error	Sample Size (teachers)
One- Year Rates					
Fall 2007–fall 2008 (cohort 1)					
Nearest neighbor	93.6	94.4	-0.9	2.47	499
Nearest 5 neighbors	92.3	89.6	2.7	2.26	1,102
Caliper match (r = 0.005)	93.7	91.4	2.3	1.90	2,944
Caliper match (radius = 0.025)	92.8	91.8	1.1	1.85	11,739
Kernel density	93.0	91.9	1.1	1.69	12,040
Fall 2008–fall 2009 (cohorts 1 & 2)					
Nearest neighbor	92.4	90.5	1.9	2.46	661
Nearest 5 neighbors	91.2	89.7	1.5	1.40	1,686
Caliper match (r = 0.005)	91.4	89.9	1.5	1.31	2,678
Caliper match (r = 0.025)	91.4	90.6	0.8	1.17	5,823
Kernel density	91.5	90.6	0.9	1.11	11,422
Fall 2009–fall 2010 (cohorts 1, 2, & 3)					
Nearest neighbor	87.3	89.2	-1.9	2.00	974
Nearest 5 neighbors	85.8	87.3	-1.5	1.60	2,694
Caliper match (r = 0.005)	85.7	88.4	-2.6	1.69	3,290
Caliper match (r = 0.025)	85.9	87.8	-1.9	1.45	11,320
Kernel density	85.9	87.7	-1.8	1.49	11,340
Two- Year Rates					
Fall 2007–fall 2009 (cohort 1)					
Nearest neighbor	88.1	78.2	9.9***	3.32	332
Nearest 5 neighbors	86.6	81.9	4.7**	2.08	881
Caliper match (radius = 0.005)	87.0	83.2	3.8**	1.94	1,707
Caliper match (radius = 0.025)	86.8	82.8	3.9***	1.30	5,785
Kernel density	87.0	83.4	3.7***	1.17	11,423
Fall 2008–fall 2010 (cohorts 1 & 2)					
Nearest neighbor	82.0	78.1	3.9	3.51	631
Nearest 5 neighbors	81.1	79.7	1.4	2.47	1,879
Caliper match (r = 0.005)	81.0	79.4	1.7	2.40	2,352
Caliper match (r = 0.025)	81.0	80.2	0.8	2.17	10,819
Kernel density	81.0	80.0	1.0	2.19	11,373
Three- Year Rates					
Fall 2007–fall 2010 (cohort 1)					
Nearest neighbor	74.2	71.3	2.9	3.75	265
Nearest 5 neighbors	75.9	72.1	3.8	2.62	781
Caliper match (r = 0.005)	76.9	71.2	5.7*	3.01	1,190
Caliper match (r = 0.025)	75.9	73.6	2.3	1.76	10,008
Kernel density	75.1	74.5	0.6	1.87	11,279

Note: District retention rate is defined as the percentage of classroom teachers in fall of the base year who worked in the district in fall of the follow-up year. Shaded cells are based on the nearest-five-neighbors method, which produced the closest match between Chicago TAP and non-TAP schools in diagnostic tests performed before the outcomes were examined.

[a] Means are regression adjusted.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

Figure B.1. Impacts on District Retention Rates by Year, Duration, and Cohort



Note: Each blue box represents a group of teachers in schools of the indicated cohort (1, 2, or 3) followed from fall of the base year to fall of the follow-up year. Numbers below the blue boxes are impacts of Chicago TAP on school retention rates estimated using the nearest-five-neighbors matching method to form comparison groups.

* Chicago TAP-comparison difference is statistically significant at the 10 percent level.
** Chicago TAP-comparison difference is statistically significant at the 5 percent level.
*** Chicago TAP-comparison difference is statistically significant at the 1 percent level.

MATHEMATICA
Policy Research

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC

Mathematica® is a registered trademark of Mathematica Policy Research