Abstract Title Page

Title:

The Impact of Teacher Evaluation Reform on Student Learning:
Success and Challenges in Replicating Experimental Findings With Non-Experimental Data

Authors and Affiliations:

Jennie Y. Jiang, University of Chicago Consortium on Chicago School Research
Lauren Sartain, University of Chicago Consortium on Chicago School Research
Susan E. Sporte, University of Chicago Consortium on Chicago School Research
Matthew P. Steinberg, University of Pennsylvania

Abstract Body

Background / Context:

One of the most persistent and urgent problems facing education policymakers is the provision of highly effective teachers in all of our nation's classrooms. Of all school-level factors related to student learning and achievement, the student's teacher is consistently the most important (Goldhaber 2002; Rockoff 2004; Rivkin, Hanushek, and Kain 2005). Even with substantial within-school variation in teacher effectiveness (Rivkin, Hanushek, and Kain 2005; Aaronson, Barrow, and Sander 2007), historically teacher evaluation systems have inadequately differentiated teachers who effectively improve student learning from lower-performing teachers. In Chicago from 2003 to 2006, for example, nearly all teachers (93 percent) received performance evaluation ratings of "Superior" or "Excellent" (based on a four-tiered rating system) while at the same time 66 percent of CPS schools failed to meet state proficiency standards under Illinois' accountability system (The New Teacher Project 2007).

Only recently have policy efforts begun to address alternative methods for evaluating teacher performance. Increasingly, state and local education agencies are replacing traditional teacher evaluation approaches in order to incorporate multiple methods of assessing and evaluating teachers. This move has been largely influenced by the federal Race to the Top (RTTT) competition. The 2010 RTTT competition emphasized more rigorous performance evaluation through the use of multiple measures of teacher performance as well as the incorporation of multiple teacher ratings categories to differentiate teacher effectiveness. These evaluations use student test score data to estimate a teacher's idiosyncratic contribution to student learning (so called "value-added" metrics). However, a majority of teachers – upwards of 70 percent nationally – teach in grades or subjects in which standardized achievement exams are not administered, and therefore will not have a value-added score (Watson, Kraemer, and Thorn 2009). As a result, qualitative classroom observation-driven measures of teacher performance remain critically important components of teacher evaluation.

When considering evidence of the success of school reform initiatives, we often look for randomized control trials to provide the highest standard in estimating effects. However, in most cases, researchers must rely on non-experimental data to evaluate programs. In this proposed presentation, we are able to look at the impact of teacher evaluation systems on student learning using both experimental data and non-experimental data. We hope to acknowledge the challenges that researchers face, while providing both types of evidence from CPS.

Purpose / Objective / Research Question / Focus of Study:

We seek to answer the following research questions about two waves of teacher evaluation reform in Chicago, a pilot focused on rigorous classroom observations (2008-10) and a fully implemented evaluation system that incorporates information from classroom observations and student assessment (2012-13 to present):

1. What does experimental evidence tell us about the effect teacher evaluation can have on school-level performance in mathematics and reading in elementary schools?
2. What does experimental evidence tell us about how teacher evaluation can differentially impact schools with different characteristics (for example, are there greater impacts in lower- or higher-achieving schools)?

3. Can estimates of these impacts be replicated across different teacher evaluation programs and using experimental and non-experimental designs?

Setting:
The research in this proposal is conducted in Chicago Public Schools.

Population / Participants / Subjects:
In 2012-13, CPS employed approximately 23,000 teachers. We focus in particular on the approximately 470 elementary schools, which have an enrollment of about 250,000 students each year. Chicago students are likely to be from low-income families (87 percent), and 42 percent are African American and 44 percent are Latino.

Intervention / Program / Practice:
Teacher evaluation reform in Chicago occurred in two waves. First, the Excellence in Teaching Pilot (EITP) was implemented in about 100 elementary schools in 2008-10. The emphasis of EITP was on improving the classroom observation process, and it did not formally include the use of student assessment data or value-added measures. Teachers were observed two times per year using the Charlotte Danielson Framework for Teaching, and administrators and teachers met before and after these observations to discuss instructional practices in detail. Another significant component of EITP was principal training, as this approach to teacher evaluation was new for many administrators. The training for principals emphasized the collection of evidence, rather than opinions, about what was happening in the classroom. After the observation, the principal was expected to match his or her classroom observation notes to the Danielson Framework rubric in order to rate teacher performance in ten areas of instructional practice. Despite early signs of success of the pilot, the district ended the mandatory participation of schools in EITP at the end of the 2009-10 school year due to a change in district leadership.

In 2010, the Illinois state legislature passed the Performance Evaluation Reform Act (PERA), which required all districts overhaul their teacher evaluation systems. In response, the second wave of teacher evaluation reform in Chicago was wider in reach both in the scope and requirements of the evaluation system, as well as the number of principals and teachers who were affected by the rollout. In 2012-13, all of Chicago's nearly 600 non-charters schools implemented REACH. The REACH evaluation system included a classroom observation process that is very similar in design to the EITP classroom observations, allowing for continuity for schools that were already using the Danielson Framework. As with many new teacher evaluation systems, REACH also includes indicators based on student achievement data. (See appendix figure 1 for more details about how teachers' evaluation ratings are calculated.) In order to build off previous work on teacher evaluation, the district also structured principal training to be similar to that principals received under EITP. As with EITP, principals and teachers were generally positive about the new REACH evaluation system, though teachers were more cautious about the use of student assessments in evaluation ratings (see appendix table 5 for survey evidence of perceptions of REACH).

Research Design:
In the spirit of the conference's theme of replication, we present two contrasts in understanding the effect of teacher evaluation systems on student learning. We can think about Chicago Public

Schools as implementing two waves of teacher evaluation. The first wave implementation from 2008-10 was staggered, and it allows for us to provide estimates of the impact of teacher evaluation on student learning through an experimental design. However, the second wave of teacher evaluation reform in Chicago occurred in all elementary and high schools simultaneously beginning in the 2012-13 school year. While experimental design is often touted as the gold standard in social science research, it is often the case – and especially in the world of teacher evaluation – that districts want all of their schools to implement reforms at the same time. In these cases, researchers must do the best they can with non-experimental data to identify the effects on student learning. In our presentation, we hope to highlight the successes and challenges in replicating experimental research in a non-experimental data context.

First wave of implementation. In 2008-10, CPS unveiled the Excellence in Teaching Pilot (EITP) – this effort took place prior to the national emphasis on improving teacher evaluation. Forty-four elementary schools in four of CPS's subdistricts (called instructional areas) launched the EITP initially in 2008-09. These schools were selected randomly, allowing us to estimate the impact of this improved classroom observation system by comparing the treatment schools to an extremely similar group of control schools. It should be noted that the control schools were late adopters; they initiated the pilot a year behind the treatment schools in 2009-10. (This staggered rollout means that after the first year of EITP, we lose the experimental design, which also poses specific challenges that we will discuss in our presentation.)

To estimate the impact of the teacher evaluation pilot on a school's math and reading achievement among our sample of CPS elementary schools, we estimate variants of the following basic model:

$$(1) \qquad Y_i = \beta_0 + \beta_1(\text{Pilot}_i) + X_i`\Gamma + \theta_g + \varepsilon_i \,,$$

where Y_i is a school achievement outcome for school *i*; $Pilot_i$ is an indicator variable that equals one if school *i* was randomly assigned to participate in the pilot in the 2008-09 school year, and zero otherwise; **X** is a vector of school covariates; and ε_i is a random error term. Because the randomization was done at the instructional area level, we also include area fixed effects (θ_g) to account for the block design of the experimental study.

Second wave of implementation. In response to state legislation in Illinois, CPS unveiled REACH Students in the 2012-13 school year. Given the legislative timeline, REACH involved immediate take-up across all schools in the district, which means that we can no longer rely on the experimental design as with EITP. This challenge is by no means unique to this specific reform in CPS – it is something researchers often face: how to estimate the true impact of a program on an outcome of interest. Complicating matters further, CPS implemented a number of reforms in the 2012-13 school year, in particular a longer school day and Common Core standards. If we see improvements in student learning, it is difficult to disentangle the effects of REACH from the effects of, say, the longer school day or Common Core.

In the case of REACH, the district did not roll out implementation, so all schools began to use the new evaluation system at the same time. Implementation of this nature poses challenges to identifying the impact of REACH on student learning, which we will discuss in this presentation.

We propose using econometric analysis techniques to overcome these shortcomings. In particular, we will a differences-in-differences approach combined with propensity score matching. We will use the EITP elementary schools as the counterfactual group and identify similar elementary schools that did not implement EITP. Then, we will identify the difference in performance between the two groups before and after the implementation of REACH.

Data Collection and Analysis:

Data for this presentation consist of CPS administrative, personnel, and test score data from the 2005-06 school year to the 2012-13 school year. Administrative data collected on students include basic demographic information, such as gender and race/ethnicity, as well as information on poverty level and students with special education needs. These data allow us to check for covariate balance among the student characteristics across schools in the EITP groups (see appendix for evidence of covariate balance in tables 1 and 2), as well as to obtain more precise estimates of treatment effects. Further, the inclusion of student demographic characteristics enables us to identify heterogeneous treatment effects by school composition. The outcome of interest is student achievement. Students in Illinois take the Illinois Standards Achievement Test (ISAT) in reading and mathematics in grades 3-8. The administration of the ISAT typically occurs in March of each school year. This presentation uses ISAT data from 2005-06 to 2012-13.

Teacher personnel data include teacher-level data about tenure status, years of experience in the district, demographic information, as well as level of education attained and certification status. These data are particularly important for the REACH analysis, as we may be able to exploit variation in the proportion of non-tenured teachers in a school to construct a dosage measure for each school. In the 2012-13 school year, REACH consisted of formal evaluation for non-tenured teachers and only informal evaluation for tenured teachers.

Findings / Results:

With EITP, we find that at the end of the first year of implementing the teacher evaluation pilot, Cohort 1 schools improved student achievement in reading by 0.10 standard deviations compared to the Cohort 2 schools. The gap in math is about half the size (0.05 standard deviations), though the effect is not statistically significant. (See appendix tables 3 and 4.) We also find that the pilot did not impact all schools in the same way. More advantaged schools (i.e., schools that were high achieving prior to implementation, schools with lower rates of student poverty) tended to benefit the most from EITP. This finding suggests that an intervention such as teacher evaluation requires high levels of capacity in the school building in order to affect student learning. Analysis of the REACH data is still ongoing.

Conclusions:

The findings about the effects of teacher evaluation based on the EITP experiment show both promise and caution. In addition, they indicate areas of study that will be further investigated through quasi-experimental methods in the subsequent district-wide implementation of REACH. Taken together, they can point to a replicable structure of a smaller scale pilot project, studied through experiment, followed by a larger implementation, studied through quasi-experimental models.

Appendices

Appendix A. References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*,25(1), 95-135.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Goe, L, Bell, C, & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D. (2002). The mystery of good teaching. *Education Next*, 2(1), 50-55.

Kane, T. J., Taylor, E.S., Tyler, J.H., & Wooten, A.L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, 46(3), 587-613.

Rivkin, S.G., Hanushek, E.A., &. Kain, J.F. (2005). Teachers, schools and academic achievement. *Econometrica*, 73(2), 417-458.

Rockoff, J.E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2), 247–252.

Taylor, E,S., & Tyler, J.H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 102(7), 3628-3651.

The New Teacher Project. (2007). *Teacher hiring, assignment, and transfer in Chicago Public Schools.* Brooklyn, NY: The New Teacher Project.

Watson, J.G., Kraemer, S.B. & Thorn, C.A. (2009). *The Other 69 Percent*. Washington, DC: Center for Educator Compensation Reform, U.S. Department of Education, Office of Elementary and Secondary Education.

Appendix B. Tables and Figures

Table 1. Baseline Characteristics for EITP Elementary Schools

School Characteristic	Cohort 1 Mean (SD)	Cohort 2 Mean (SD)	P-Value of Difference
Enrollment	448.2 (209.8)	497.8 (296.0)	0.268
% Female	.49 (.027)	.50 (.031)	0.729
% African-American	.60 (.401)	.67 (.374)	0.255
% Hispanic	.24 (.287)	.20 (.254)	0.511
% White	.11 (.175)	.08 (.131)	0.343
% Asian	.05 (.091)	.05 (.089)	0.916
% IEP	.13 (.043)	.14 (.059)	0.455
% FRPL	.82 (.232)	.83 (.207)	0.978
Math Achievement	.036 (1.083)	-.032 (.930)	0.826
Reading Achievement	-.008 (1.106)	.007 (.906)	0.829

Notes. Mean (standard deviation) of school characteristics for the 2008-09 school year. There are 44 Cohort 1 schools and 49 Cohort 2 schools in the sample. Math and Reading achievement are for the 2008 ISAT (given in the spring of the 2007-08 school year) and are standardized within sample. *%IEP* is the proportion of students in a school in receipt of an individualized education plan; *%FRPL* is the proportion of students receiving free or reduced-price lunch. Since elementary schools were randomly assigned to treatment within one of four local instructional areas, the p-value of difference of means is adjusted for area fixed effects. Coefficients statistically significant at the *10 percent, **5 percent and ***1 percent levels.

Table 2. Teacher Characteristics for EITP Elementary Schools

Teacher Characteristic	Cohort 1 Mean (SD)	Cohort 2 Mean (SD)	P-Value of Difference
Number of Teachers	28.4 (11.3)	29.9 (15.7)	0.451
% Female	.85 (.077)	.85 (.082)	0.988
% African-American	.38 (.300)	.42 (.284)	0.385
% Hispanic	.11 (.155)	.11 (.125)	0.932
% White	.46 (.199)	.42 (.207)	0.301
% Asian	.04 (.057)	.04 (.051)	0.674
Age (years)	44.4 (4.18)	44.9 (4.09)	0.593
Experience (years)	11.8 (2.63)	12.1 (2.74)	0.602
Master's Degree (%)	.59 (.124)	.60 (.122)	0.606
National Board Certification (%)	.03 (.042)	.02 (.035)	0.283
Tenured (%)	.76 (.119)	.76 (.151)	0.743

Notes. Mean (standard deviation) of teacher characteristics for the 2008-09 school year. There are 44 Cohort 1 schools and 49 Cohort 2 schools in the sample. Teacher tenure information is unavailable for two Cohort 1 and four Cohort 2 schools. *Experience* is the number of years a teacher has taught in CPS (not including any experience outside of the district). Since elementary schools were randomly assigned to treatment within one of four local instructional areas, the p-value of difference of means is adjusted for area fixed effects. Coefficients statistically significant at the *10 percent, **5 percent and ***1 percent levels.

Table 3. Impact of EITP on Math Achievement

	Year 1				Year 2				Year 3			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Pilot	.136 (.2109)	.107 (.1826)	.127 (.1310)	.054 (.0542)	.128 (.2110)	.114 (.1832)	.146 (.1380)	.080 (.0816)	.110 (.2111)	.100 (.1836)	.120 (.1368)	.066 (.0894)
School Characteristics			X	X			X	X			X	X
Baseline Math Achievement				X				X				X
Area FE		X	X	X		X	X	X		X	X	X
# of Schools	93	93	93	93	92	92	92	92	92	92	92	92
R^2	.0046	.2669	.6942	.9463	.0041	.2689	.7030	.9100	.0030	.2652	.6550	.8450

Notes. Coefficients (with robust standard errors) reported are in standard deviation units and represent the intent-to-treat effect of the teacher evaluation Pilot on math achievement. Year 1 effects for the 2008-09 school year; Year 2 effects for the 2009-10 school year; and Year 3 effects for the 2010-11 school year. School Characteristics include enrollment, gender, race/ethnicity, the proportion of special education students and the proportion of students receiving free or reduced-price lunch. *Baseline Math Achievement* is a school's pre-treatment math achievement (standardized) for the 2007-08 school year. Coefficients statistically significant at the *10 percent, **5 percent and ***1 percent levels.

Table 4. Impact of EITP on Reading Achievement

	Year 1				**Year 2**				**Year 3**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Pilot	.087 (.2116)	.060 (.1871)	.107 (.1246)	.099 ** (.0463)	.089 (.2119)	.078 (.1853)	.104 (.1257)	.115 * (.0685)	.094 (.2116)	.087 (.1813)	.091 (.1262)	.1200 (.0803)
School Characteristics			X	X			X	X			X	X
Baseline Reading Achievement				X				X				X
Area FE		X	X	X		X	X	X		X	X	X
# of Schools	93	93	93	93	92	92	92	92	92	92	92	92
R^2	.0019	.2335	.7226	.9620	.0020	.2581	.7466	.9292	.0022	.2894	.6971	.8803

Notes. Coefficients (with robust standard errors) reported are in standard deviation units and represent the intent-to-treat effect of the teacher evaluation Pilot on reading achievement. Year 1 effects for the 2008-09 school year; Year 2 effects for the 2009-10 school year; and Year 3 effects for the 2010-11 school year. School Characteristics include enrollment, gender, race/ethnicity, the proportion of special education students and the proportion of students receiving free or reduced-price lunch. *Baseline Reading Achievement* is a school's pre-treatment reading achievement (standardized) for the 2007-08 school year. Coefficients statistically significant at the *10 percent, **5 percent and ***1 percent levels.

Table 5. Teacher Perceptions of the REACH Evaluation

My Voice My School Teacher Survey
May 2013

Overall Evaluation Process

	Agree or Strongly Agree
Teacher evaluation at this school is fair	76.62%
The criteria on which I am evaluated is fair	73.89%
The teacher evaluation process at this school encourages my professional growth	75.50%
I have professional conversations with my principal that are focused on instruction	80.88%
Overall, I am satisfied with the teacher evaluation process at this school	71.55%

Evaluator

	To Some Extent or To a Great
My evaluator is able to accurately assess my instruction	87.86%
My evaluator knows my strengths and weaknesses as a teacher	84.36%
My evaluator is fair and unbiased	86.77%
My evaluator supports my professional growth	88.65%
My evaluator knows what is going on in the classroom	80.68%

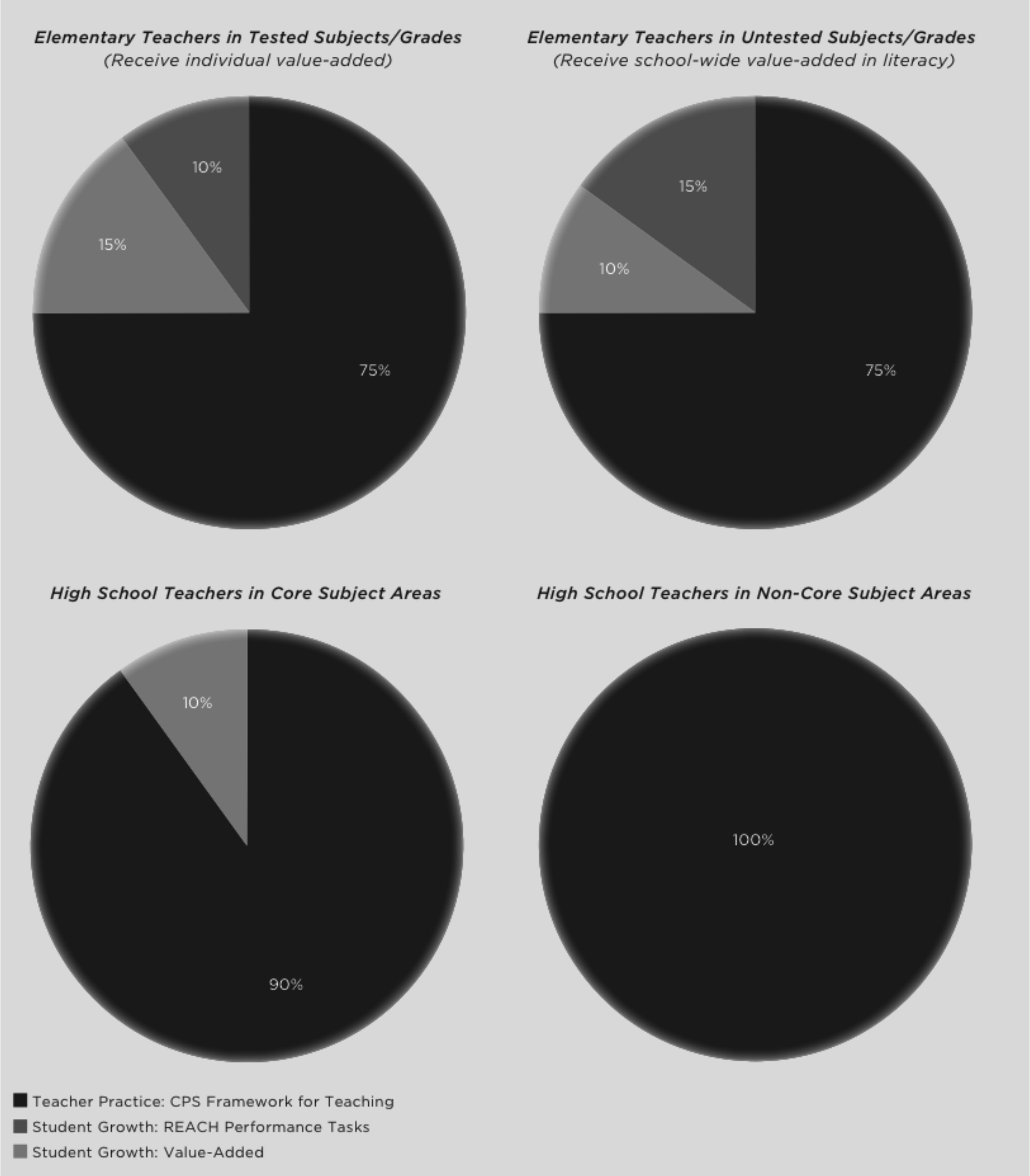
Feedback, Standardized Tests, and Assessments

	Somewhat Useful or Very Useful
How useful is your evaluator’s feedback for your instruction?	67.03%
Performance Tasks	67.20%
ISAT	45.58%
NWEA-MAP	69.18%
EPAS	41.92%
Assessments developed by teachers in your school or department	83.42%

(n = 19,417)
Note: These figures do not include missing values.

Figure 1. Components of REACH Summative Score

In 2012-2013 a teacher's REACH summative evaluation score was comprised of a teacher practice score and up to two measures of student growth. The teacher practice component includes classroom observations completed by a certified administrator utilizing the CPS Framework for Teaching, an observation rubric based on the Danielson Framework for Teaching. The student growth component may include two measures– a value-added measure and growth on district-developed performance tasks that are customized by grade level and subject area.



Source: Chicago Public Schools