PERFORMANCE

ACHIEVEMENT

QUALITY OF INSTRUCTION

PROFESSIONAL DEVELOPMENT

ACCOUNTABILITY

Rethinking Teacher Evaluation
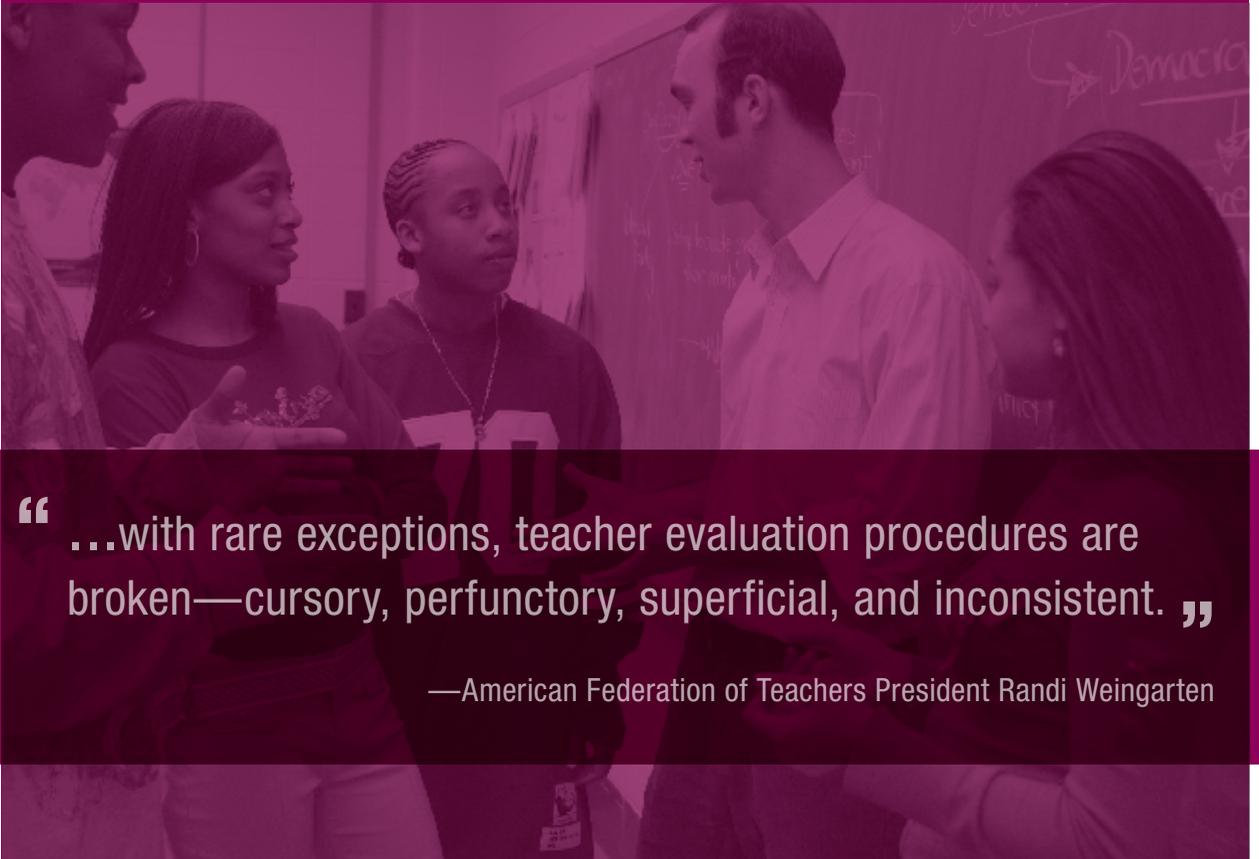
FINDINGS FROM THE FIRST YEAR OF THE EXCELLENCE IN TEACHING PROJECT IN CHICAGO PUBLIC SCHOOLS

RESEARCHERS: Lauren Sartain, Sara Ray Stoelinga, Eric Brown, Stuart Luppescu, and Kavita Kapadia Matsko
AUTHORS: Lauren Sartain, Sara Ray Stoelinga, and Emily Krone

Teacher evaluation has become a subject of increased emphasis and contentious debate nationwide. The issue has gained urgency as schools and districts have come under increased pressure to raise achievement and the public demands more information about the effect individual teachers have on student learning. Teacher evaluation policies raise fundamental questions about what constitutes effective instruction and whether those practices can be fairly measured. They also tend to be highly politicized because they involve issues central to the collective bargaining agreements between teachers' unions and school districts: compensation, hiring and firing, and career advancement.

There is a growing consensus that the way most states and districts across the country evaluate teachers fails to improve student learning or teacher practice. In a recent opinion article, American Federation of Teachers President Randi Weingarten acknowledged that "with rare exceptions, teacher evaluation procedures are broken—cursory, perfunctory, superficial, and inconsistent."[1]

Research confirms that most evaluation systems are ineffective. They typically fail to provide teachers with the information they need to make timely and effective improvements in their instructional practice.[2] Often, they rely upon a single observation by a principal, who is minimally trained as an evaluator.[3] At the same time, many evaluation tools are seen as subjective, and most tools do not differentiate between strong instruction and weak, rendering evaluation meaningless.[4]

Of particular concern, most evaluation systems fail to identify or facilitate the removal of low-performing teachers. A 2005 report by the Illinois Small Newspaper Group found that 83 percent of the state's school districts had never rated a tenured teacher as "unsatisfactory."[5] School systems as diverse as Denver, Chicago, Atlanta, and San Francisco rarely dismiss low-performing teachers—often less than 1 percent of teachers in any given year.[6]

Policymakers and others have responded to flaws in the current system by demanding that districts start using data on student academic growth to evaluate teachers. The U.S. government advanced this agenda by

> "There is a growing consensus that the way most states and districts across the country evaluate teachers fails to improve student learning or teacher practice."

1 Weingarten (2010).

2 McLaughlin (1990); Searfross & Enz (1996).

3 Haefele (1993).

4 Haefele (1993); McLaughlin (1990).

5 Small Newspaper Group (2005).

6 Darling-Hammond (1996); Eisner (1992); Van Sciver (1990); Wise et al. (1984); The New Teacher Project (2007).

requiring states competing for $4.35 billion in federal Race to the Top Funds to remove any existing legal barriers to linking student achievement data to teacher evaluations.[7] States and districts have responded. New legislation in Illinois, for example, requires all districts to implement standards-based teacher evaluation systems with a student achievement indicator.[8]

Yet, researchers have raised a number of questions about whether student achievement data can be used fairly or accurately for purposes of teacher evaluation.[9] Others have noted that achievement data alone cannot provide teachers with the information they need to improve their practice. Recognizing these limitations, the federal government and many states have specified that student test score data should be just one of a variety of measures used to evaluate teachers.[10] Other measures would likely include some form of classroom observation, which in turn has generated new demand for tools that principals and others can use to judge whether effective teaching is taking place.

The Charlotte Danielson Framework for Teaching, which attempts to delineate the observable components of effective teaching, is perhaps the most well-known example of such a tool. Districts including Chicago, Cincinnati, and Las Vegas have adopted the Framework to structure teacher evaluation. The Teacher Advancement Program (TAP), a widely implemented performance-pay and leadership development system for teachers, also uses an evaluation rubric based on the Framework.

A team of researchers from the Consortium on Chicago School Research (CCSR) at the University of Chicago is studying the implementation of the Danielson Framework in Chicago Public Schools (CPS) and providing real-time, objective feedback to the district on its new pilot teacher evaluation program, the Excellence in Teaching Project. This policy brief describes the first year of implementation in CPS and highlights key early findings and policy implications from the study. The findings presented are relevant for policymakers contemplating how best to support the design and development of effective teacher evaluation systems. They are particularly important for districts seeking valid, reliable ways to measure and evaluate the complex activity of teaching.

7   U.S. Department of Education (2009).

8   Performance Evaluation Reform Act, Illinois Public Act 096-0861.

9   Education Week (2009).

10 Duncan (2009).
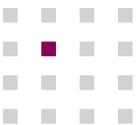
Teacher Evaluation in Chicago Public Schools

The evaluation system in the Excellence in Teaching Project is the proposed replacement for a checklist that has been used in CPS for 30 years. On the checklist, principals label various components of teaching as a "strength" or a "weakness". After filling out the checklist, principals assign an overall rating to teachers: Unsatisfactory, Satisfactory, Excellent, or Superior. The form does not include criteria to define a strength or weakness, and some of the components are ambiguous. Further, there is no guidance on how the checklist relates to a teacher's final evaluation rating.

A 2007 report from The New Teacher Project on CPS teacher hiring, assignment, and transfer policies revealed that neither principals nor teachers perceived the checklist system to be meaningful or fair. The report also demonstrated that the checklist system does not lead to the identification or removal of low performing teachers. In fact, very few teachers were identified as Unsatisfactory (0.3 percent) or even just Satisfactory (7 percent).[11]

A joint committee with representatives from CPS and the Chicago Teachers' Union (CTU) worked for three years to develop the Excellence in Teaching Project. The committee members chose the Charlotte Danielson Framework for Teaching to guide classroom observations and conversations around instruction. As the initiation of the pilot neared, the district-union joint committee broke down due to a disagreement about a separate issue related to teachers' contracts. CPS leadership proceeded with implementation, and in 2008-09, schools in the evaluation pilot were required to use both the Danielson Framework and the checklist in their schools simultaneously.

The first year of the evaluation pilot, 2008-09, included 44 elementary schools. Principals received extensive professional development, including three days of training in the summer and four half-day professional development sessions throughout the year. Principals also met monthly to discuss the evaluation process. Support for teachers was less extensive, consisting of two school-based sessions that provided an overview of the Charlotte Danielson Framework.

> ### ❭ BY THE NUMBERS

91%

of CPS teachers received a "superior" or "excellent" evaluation rating in 2007–08

66%

of CPS schools failed to meet state standards that same year

Source: The New Teacher Project (2009).

---

11 The New Teacher Project (2007).

> FACT:

States and districts using the Danielson Framework include: Prince George's County (MD), Hillsborough County Public Schools (Tampa, FL), Cincinnati Public Schools (OH), Clark County School District (Las Vegas, NV), Idaho public schools

The Charlotte Danielson Framework

Charlotte Danielson's Framework for Teaching divides teaching into four domains: Planning and Preparation, Classroom Environment, Instruction, and Professional Responsibilities. The focus of this study is on the two observable domains, Classroom Environment and Instruction. Principals must provide one rating for each of the following components:

Domain 2: The Classroom Environment	Domain 3: Instruction
Creating an Environment of Respect and Rapport	Communicating with Students
Establishing a Culture for Learning	Using Questioning and Discussion Techniques
Managing Classroom Procedures	Engaging Students in Learning
Managing Student Behavior	Using Assessment in Instruction
Organizing Physical Space	Demonstrating Flexibility and Responsiveness

Principals choose one of four levels of performance for each of the components:

- Unsatisfactory: Teaching is below the standard of "do no harm" and requires immediate intervention.

- Basic: Teacher understands the components of teaching, but implementation is sporadic.

- Proficient: Teacher has mastered the work of teaching.

- Distinguished: Teacher has established a community of learners with students assuming responsibility for their own learning.

The Study Design

The Consortium on Chicago School Research is conducting a multi-year study of the district's Excellence in Teaching Project. Our year-one work, which is the subject of this brief, explores the reliability of the Framework, principal and teacher perceptions of the Framework, and how the Framework is being implemented at the school level. Our second-year work will explore the validity of the framework, that is, whether the Framework actually measures what it claims to measure.

The study design in year one (2008-09) used "matched" observations to test the Framework's reliability—whether two people watching the same teacher will rate that teaching the same way. External observers and school administrators conducted classroom observations at the same time; however, they assigned Framework ratings independently. Quantitative data for year one of the study included joint observation data available for 277 matched observations. Qualitative data consisted of 39 principal interviews and 25 teacher interviews.

In the second year of the pilot (2009-10), the number of participating schools expanded to 100. However, principals in the second cohort received significantly less training than the first cohort. At the same time, principals became responsible for evaluating all teachers in their buildings. In 2008-09, the sample of observed teachers contained mostly new teachers, whereas in 2009-10, the sample includes new and veteran teachers. This has implications for our study, as well as for implementation. It may be the case that the second year of implementation will have different results due to these factors. The story of the Chicago evaluation pilot is still a work in progress, as is our study.

The complete report of the year one study of the Excellence in Teaching Project may be found at **http://ccsr.uchicago.edu/ publications/Joyce_TE_yr1_finaldoc.pdf.**

■ ■ ■

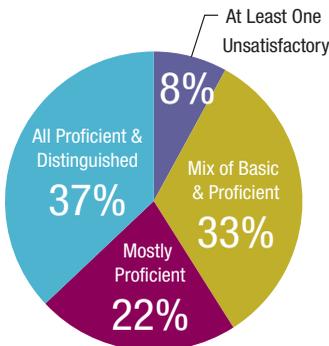*re·li·abil·i·ty*
Function: noun

The extent to which an experiment, test, or measuring procedure yields the same results on repeated trials.

Source: Merriam-Webster.com

Danielson Framework Ratings

Pre-tenured teachers received a much wider range of ratings under the new Framework than under the old CPS checklist system.

N=95 pre-tenured teachers in 44 pilot schools; 2008-09 Framework ratings from principals

Key Findings

1. **Overall, principals and trained experts use the rating scale consistently.** To understand the reliability of the Framework, principals and highly trained external observers conducted simultaneous classroom observations but assigned Framework ratings independently. Considered in aggregate, there is no significant difference between the ratings given by principals and those given by external observers. However, there are some individual differences in rater severity—both among principals and observers. That is, across the board, principals and external observers generally agree; however, some individual principals are more severe (30 percent) or more lenient (16 percent) than the external observers. In addition, principals and observers use the rating scale the same way from one observation to the next. That is, severe principals generally gave low ratings to all of their teachers, and lenient principals generally gave high ratings to all of their teachers.

2. **More teachers were identified as low-performing under the new evaluation system.** In previous years, only 0.3 percent of teachers in CPS had been rated as unsatisfactory. However, 8 percent of teachers in this sample received at least one unsatisfactory rating on the Framework. Unsatisfactory practice is characterized as doing harm to students.

3. **Principals found four areas of instruction to be particularly challenging to evaluate.** On three aspects of instruction, principals consistently gave ratings that were lower than those of the external observers; on one component, they consistently assigned higher ratings. Principals rated the following areas of teaching lower: communicating with students, using assessment in instruction, and organizing physical space. Principals were more likely to rate teachers higher on student engagement in learning than observers. The inconsistency in ratings for this component is particularly notable since engaging students in learning is the most important component, or what Danielson refers to as "the heart of the Framework."[12]

4. **Principals had no trouble identifying unsatisfactory teaching practices.** However, when using the high end of the scale, principals inflated their ratings across all ten observable components. That is, principals and

---

12 Danielson (2007).

external observers agreed about unsatisfactory practice, but principals were much more likely than external observers to identify instruction as distinguished. Principals acknowledged this tendency, pointing to the need to preserve relationships with teachers who had previously received the highest possible evaluation rating.

5. **Just over half of the principals were highly enthusiastic about the evaluation process.** Fifty-seven percent of principals had positive attitudes about the Framework and their conferences with teachers, perceived teacher buy-in as high, and said they saw changes in instructional practice stem from the evaluation system. A little less than half (43 percent) of the principals were characterized by mixed to mostly negative attitudes about both the Framework and the conferences. These principals generally said that they were "already doing" evaluation in the "right way" and were more likely to suggest that they "just knew" if teachers were good or bad. They also were less likely to believe that changes in instructional practice had happened as a result of participation in the evaluation process and placed teacher evaluation at the low end of priorities compared to their other responsibilities.

Implications

In the first year of the Excellence in Teaching Project, CPS leaders took significant steps toward revitalizing teacher evaluation in Chicago. The district chose a tool that defined instructional practice, striving to establish a common definition of good teaching along a developmental continuum. They hoped to promote, structure, and improve conversations between principals and teachers and focused squarely on instructional improvement. The pilot program reveals some areas of promise and some areas of concern for policymakers to consider.

• **The Danielson Framework has potential for improving teacher evaluation systems.** Our study of the early implementation of the Excellence in Teaching Project indicates that the Charlotte Danielson Framework is a reliable tool for identifying low-quality teaching. This suggests that it is an appropriate tool for fairly identifying teachers in need of supports or sanctions. In addition, principals were generally positive about using the Danielson Framework. Principal and teacher buy-in

"The thing I like about the Framework is it actually makes you cognizant of what behaviors constitute excellence in teaching, and then holds you accountable for actually doing those behaviors."
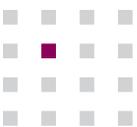
—CPS Principal

> ## BY THE NUMBERS
> # 50
> The number of
> hours of professional
> development
> principals received
> on using the new
> evaluation system.

is critical for the success of any initiative. This is especially true for efforts aiming to identify low-quality instruction, remove ineffective teachers, make more informed decisions about staffing schools, and, ultimately, improve teaching and thereby student learning.

- **To realize the Danielson Framework's potential as an evaluation tool, ongoing training and support for principals is necessary.** CPS provided high-quality, ongoing professional development and support for principals in the first year of the pilot; yet, principals still struggled to rate some areas of instruction consistently. Even with high levels of training and support, there still will be challenges when using a tool like the Danielson Framework for teacher evaluation. Because evaluating instruction is complex, continued training and meaningful supports are vital to ensure that evaluation tools are fair and useful. If scale-up to a larger number of schools does not include training and support that is intensive and ongoing, there are likely to be problematic inconsistencies in the use of the Framework by principals. At the same time, principal turnover and the difficulty of providing extensive training when an initiative expands to all schools in a large district pose legitimate challenges.

- **There may be challenges in using observational tools for high-stakes decisions.** The consequences of inconsistent application of the Danielson Framework become clear when we discuss using ratings for evaluation purposes. Inconsistencies in the way that principals rate some components of the Framework and differences in severity pose significant challenges for evaluation. For instance, a principal who is a severe rater may have detrimental effects on the careers of borderline teachers in that school. On the other hand, lenient principals may keep teachers who should otherwise be removed due to low performance.

- **Successful implementation of a rigorous evaluation system requires changing the way practitioners and district leaders think about teacher evaluation.** While introducing a high-quality teacher evaluation tool is an important step in revamping evaluation practices, changing the evaluation process also requires a long-term shift in the way people think about teacher evaluation. While the majority of principals in the first year were highly engaged and enthusiastic, a little less than half of the principals had more mixed or negative perceptions. Many of the more negative

principals revealed attitudes and assumptions about evaluation (for instance, "just knowing" if a teacher is good) that need to be addressed if teacher evaluation practices are to improve. Truly transforming teacher evaluation relies upon finding ways to shift perceptions among principals who do not see the value in deeper evaluation practices.

Conclusion

It is important to note that our analysis and findings come very early in the implementation of the pilot project, which continues to grow. Nevertheless, our preliminary analyses reveal areas of particular promise for states and districts contemplating a redesign of their evaluation systems. In order to improve evaluations based on classroom observations schools and districts need tools that are both reliable and valid. In the Chicago pilot, the overall consistency of ratings from principals and trained observers suggest the Danielson Framework does provide reliable information about the type of instruction taking place in classrooms.

In spring 2011, we will release another policy brief focused on the validity of the Framework. A valid Framework accurately measures the teaching practices that lead to student learning. Thus, our year-two report will investigate the relationship between Framework ratings and student outcomes. These findings should advance our understanding of the link between academic achievement (student outputs) and instructional practice (teacher inputs).

"In the Chicago pilot, the overall consistency of ratings from principals and trained observers suggest the Danielson Framework provides reliable information about the type of instruction taking place in classrooms."

References

Danielson, C. (2007)
Enhancing professional practice: A framework for teaching (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Duncan, A. (2009)
Partners in reform: Remarks of Arne Duncan to the National Education Association. U.S. Department of Education. http://www2.ed.gov/news/speeches/2009/07/07022009.html.

Darling-Hammond, L. (1996)
What matters most: A competent teacher for every child. *Phi Delta Kappan* vol. 78, no. 3: 193-200.

Eisner, E.W. (1992)
Education reform and the ecology of schooling. *Teachers College Record* 93, 4: 610-627.

Fredman, T. (2003)
Development of the Oklahoma teacher enhancement program (OTEP) P-16 teacher evaluation. Oklahoma City: Oklahoma State Regents for Higher Education.

Haefele, D.L. (1993)
Evaluating teachers: A call for change. *Journal of Personnel Evaluation in Education* vol. 7, no. 1: 21-31.

Learning Point Associates. (2010)
Measurement of student growth: Emerging trends reflected in the state phase 1 Race to the Top applications. http://www.learningpt.org/pdfs/RttT_Measurement.pdf.

McLaughlin, M.W. (1990)
Embracing contraries: Implementing and sustaining teacher evaluation. In *The new handbook of teacher evaluation*, eds. J. Millman and L. Darling-Hammond. Newbury Park, CA: Sage.

Milanowski, A., & Kimball, S.M. (2003)
The framework-based teacher performance assessment systems in Cincinnati and Washoe. Madison, WI: CPRE-UW Working Paper Series.

Performance Evaluation Reform Act. (2010)
Laws of Illinois, 2010. Illinois Public Act 096-0861.

Searfross, L., & Enz, B.J. (1996)
Can teacher evaluation reflect holistic instruction? *Educational Leadership* vol. 53, no. 6: 38-41.

Small Newspaper Group. (2005)
The hidden costs of tenure series. http://thehiddencostsoftenure.com/

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

The New Teacher Project. (2009)
The widget effect. http://widget effect.org/district-reports/chicago/.

United States Department of Education. (2009)
Resources for Race to the Top. http://www2.ed.gov/programs/racetothetop/index.html.

Van Sciver, J. (1990)
Teacher dismissals. *Phi Delta Kappan* vol. 72, no. 4: 318-319.

Viadero, D. (2009)
'Value added' gauge of teaching probed. *Education Week*, July 15, Print edition.

Weingarten, Randi. (2010)
A new path forward for public education. *The Huffington Post*, January 15, Posted on-line.

Wise, A., Darling-Hammond, L., McLaughlin, M., & Bernstein, H. (1984)
Teacher evaluation: A study of effective practices. Santa Monica, CA: Rand.

■ ■ ■

Acknowledgements

| 1313 East 60th Street | Chicago, Illinois 60637 | T 773-702-3364 | F 773-702-2010 |

Our Mission

The Consortium on Chicago School Research (CCSR) at the University of Chicago conducts research of high technical quality that can inform and assess policy and practice in the Chicago Public Schools. We seek to expand communication among researchers, policy makers, and practitioners as we support the search for solutions to the problems of school reform. CCSR encourages the use of research in policy action and improvement of practice, but does not argue for particular policies or programs. Rather, we help to build capacity for school reform by identifying what matters for student success and school improvement, creating critical indicators to chart progress, and conducting theory-driven evaluation to identify how programs and policies are working.