Evaluation of the Excellence in Teaching Pilot

Year 1 Report to the Joyce Foundation

Lauren Sartain
Sara Ray Stoelinga
Eric Brown

The Consortium on Chicago School Research
at the University of Chicago

8/14/09

We gratefully acknowledge the Joyce Foundation for their support of this critical and timely work, and especially John Luczak for his assistance and thoughtful advice in this project. We thank Chicago Public Schools, especially the Excellence in Teaching Project staff: Sheri Frost Leo, Sheila Cashman, Amy Silverman, Cindy Moyer, and Nicole Cox-Lofton. Without their assistance, this work would not be possible. Also critical to this study is the participation of pilot principals and teachers—we thank them for generously sharing their time, schools, and perspectives with us. Thanks to John Easton for providing the study design and getting this work off the ground. We continue to benefit immensely from John's vision in this project. We also acknowledge the advice and on-going support of Audrey Soglin from the Consortium for Educational Change. Our evaluation committee also contributed immensely to the direction of this study, and we thank them and the committee chair, Larry Stanton. Larry also played a big part in the initial planning stages of the study. Numerous CCSR staff members reviewed this report and provided thorough comments, including Elaine Allensworth, Penny Bender Sebring, and David Stevens. We express deep gratitude to Kavita Kapadia and Frances Miller who participated in the data collection and analysis for this project. We also appreciate the technical assistance of Stuart Luppescu who conducted the multi-facet Rasch analysis for this report, and George Karabatsos from the University of Illinois-Chicago. For information on this study, contact Lauren Sartain lsartain@ccsr.uchicago.edu.

Executive Summary

Teacher evaluation is a timely topic. School districts across the country are shifting from evaluations that lack a focus on instruction to systems that hold teachers to high standards and provide structured guidance for teachers to improve their practice. National attention has also been focused on teacher evaluation, and, in Chicago, the district has taken steps to move towards a more rigorous approach in this area. The Chicago Public Schools' (CPS) Excellence in Teaching initiative is a pilot program in select elementary schools. The pilot initiative provides a continuum of teacher practice in the Charlotte Danielson Framework forTeaching, as well as structured conferences between principals and teachers. In implementing the pilot evaluation system, CPS is working towards three goals:

1. Provide a common definition of effective teaching for all schools
2. Guide meaningful discussion and collaboration around teaching practice
3. Direct continuous advancement of teachers along a continuum, helping teachers to have greater impact on student learning

Throughout the 2008-09 school year, CPS and the Consortium on Chicago School Research (CCSR) have worked together closely to understand the implementation of the initiative. With funding from the Joyce Foundation, CCSR is conducting a formative and summative evaluation of the district's Excellence in Teaching initiative. In our evaluation work, we are investigating the following research questions:

1. What are the technical properties—including reliability and validity—of the evaluation tool itself?
2. How do principals perceive the utility of the new teacher evaluation system? Does it help achieve the stated goals?
3. How do teachers perceive the utility, fairness, and helpfulness of the new evaluation system? What components of the system are the most or least helpful? Are the pre- and post-conferences useful?
4. What supports are in place in the pilot year, are they effective, and how well are the goals and procedures communicated across the evaluation system?
5. Does the new system have the desired effect at the school level, including shaping the professional development, professional culture, teacher hiring, the quality of teaching, and student learning?

To learn about these aspects of the pilot, we are using a variety of statistical approaches to examine Framework reliability using classroom observation ratings given by administrators and external observers. In addition, we have conducted interviews with teachers and principals to learn more about implementation. In Year 2 of the study, we will be looking at validity of the Framework (i.e., how Framework ratings relate to value-added measures) and focusing on how the evaluation system relates to school change, as well as continuing to explore reliability and implementation issues.

Technical Properties of the Framework

Reliability. Using two statistical techniques (multi-facet Rasch analysis and multilevel logistical modeling), the major findings are that the Framework has the potential to identify strong and weak teachers reliably. However, there are some components where Framework raters (e.g., administrators, outside observers) need additional support, particularly around the middle of the scale: the basic and proficient levels of performance. We suggest that principals receive focused training on the following Domain 3 (Instruction) components:

- 3a Communicating With Students
- 3c Engaging Students in Learning
- 3d Using Assessment in Instruction

Framework Component Difficulty. We were also able to use the ratings data to determine which Framework components are hardest for teachers to demonstrate proficiency, as well as those where teachers are more likely to excel. The most difficult Framework components were 3b (Using Questioning and Discussion Techniques) and 3c (Engaging Students in Learning). The areas where teachers received the highest ratings were 2e (Organizing Physical Space) and 2a (Creating an Environment of Respect and Rapport). In general, teachers struggled more with Instruction components than with Classroom Environment components.

Potential for Summative Ratings. In Year 1 of the pilot, the Framework was used for formative evaluation, though principals could also use the data they collected to inform their efficiency ratings of teachers. However, there was no formal recommendation from the district on how to use Framework ratings to provide a summative evaluation of teachers. One possible criterion that the district and the union discussed was identifying low-performing teachers as those who received unsatisfactory Framework ratings. According to this benchmark, 8% of teachers in our sample would be identified as low-performing—a considerable increase from the 0.3% of teachers who received unsatisfactory efficiency ratings in 2007-08. We explore the use of other benchmarks in the report. In short, where the benchmark is set has a huge impact on the fraction of teachers identified as low-performing and is a human capital concern.

Principal and Teacher Perceptions

Framework Perceptions. Principals and teachers positively commented on the quality of the Framework and its ability to accurately measure teacher performance. Principals identified the components they found most difficult to rate, and teachers discussed if and how participation in the evaluation process had influenced their instructional practice. In general, both teachers and principals were overwhelmingly positive about the Framework, although there were a few areas of concern. Overall, 84% of principals and 100% of teachers expressed mostly positive or mixed perceptions about the Framework.

Conference Perceptions. Principals and teachers were asked a series of questions about the pre- and post-observation conferences. Each principal or teacher was asked if they participated in pre- and post-conferences, the format and length of these conferences, and their perceptions of the process. While both principals and teachers were very positive about the Framework, teachers were generally

less positive about the conferences. However, the concerns were mostly about the time commitment required and implementation difficulties. Overall, 85% of principals and 88% of teachers expressed mostly positive or mixed perceptions about the conferences.

Training Perceptions. Principals and teachers were asked about their perceptions of the training they received for the evaluation pilot. As discussed in the overview section of the report above, principals received three days of summer training, four half-day professional development sessions, and participated in Area-based professional learning communities where they discussed the evaluation process with other principals. Teachers received two workshops for a total of 3.5 hours of training. Teacher workshops occurred at the beginning of the year and again in mid-to-late fall. The majority of principals' and teachers' opinions concerning the training they were received mostly positive. However, there were a few teachers who could not remember or did not attend the training.

Principal Typology: Categorizing Principal Buy-In

Using the principal interviews, we identified five concepts that could inform us about principal buy-in: 1) Framework attitudes, 2) conference attitudes, 3) evaluation attitudes, 4) description of teacher buy-in, and 5) description of changes in instructional practices. From these concepts, we sorted principals into 4 types:

1. *Paradigm Shift* (15%): The small group of PS principals reported a "paradigm shift" in their perception of evaluation that was a direct result of participating in the pilot study. Paradigm shift principals talked about how they realized their evaluation had been subjective in the past or how they perceived their teachers were of higher quality in a certain area.
2. *High Enthusiasm* (40%): Both the PS and HE principals tended to describe their teacher buy-in as high, and agreed that they saw changes in instructional practice between the two observations they attributed across observations. The PS and HE principals were most likely to report substantive changes in instructional practice.
3. *Mixed Emotions* (28%): Their negativity tended to focus on the perception that this was an additional initiative, layered on top of countless existing programs and initiatives that were already in their schools, and that they did not have time for the labor intensive evaluation approach.
4. *Low Enthusiasm* (15%): LE principals were mostly negative about the Framework and conferences, stated that they were already doing the type of evaluation in the new system or that they "just knew" teachers' abilities. This group was characterized by a perception of a lack of influence of the evaluation system on instructional practice and described their teachers' buy-in as low to medium.

Overall, the district should be encouraged that over half of the pilot principals were extremely positive about the initiative. Remember that this is Year 1 of a rigorous evaluation that requires much more from principals in terms of time and evaluation competence.

Introduction

This report summarizes the results of the first year of the evaluation of the Excellence in Teaching pilot being undertaken in the Chicago Public Schools (CPS). The evaluation is being conducted by researchers from the Consortium on Chicago School Research (CCSR) at the University of Chicago, using funding from the Joyce Foundation. The evaluation aims to answer the following research questions:

1. What are the technical properties—including reliability and validity—of the evaluation tool itself?
2. How do principals perceive the utility of the new teacher evaluation system? Does it help achieve the stated goals?
3. How do teachers perceive the utility, fairness, and helpfulness of the new evaluation system? What components of the system are the most or least helpful? Are the pre- and post-conferences useful?
4. What supports are in place in the pilot year, are they effective, and how well are the goals and procedures communicated across the evaluation system?
5. Does the new system have the desired effect at the school level, including shaping the professional development, professional culture, teacher hiring, the quality of teaching, and student learning?

We begin with an overview of the Excellence in Teaching pilot, the Charlotte Danielson Framework for Teaching, and the CCSR evaluation design. We then describe the results of the first year of our study. We first explore the technical properties of the evaluation tool (research question 1), analyzing ratings data using descriptive and statistical techniques. We then turn to research questions 2-4, drawing upon analysis of principal and teacher interviews. Question 5 is not included in this report but rather will be part of the focus of the Year 2 evaluation work.

Teacher Evaluation in Chicago Public Schools: The Checklist System

In order to understand the pilot evaluation system, as well as the comparisons we make throughout the report of the current and pilot evaluation systems, we must first describe the context of teacher evaluation in the district. Known informally, and referred to in this report, as the checklist system, the current teacher evaluation scheme in CPS has been in place since for the last 30 years. The checklist (formally the Classroom Teacher Visitation Form) covers three aspects of teaching: instruction, school environment, and professional and personal standards. For each of the components that comprise those three aspects of teaching, principals choose one category from strength, weakness, and does not apply that best represents teacher performance. (See Appendix E for the entire checklist.) However, there are no criteria to define a strength or weakness, and some of the components themselves are outdated and/or ambiguous. Further, there is no guidance on how the checklist relates to a teacher's final summative evaluation rating (called the summative efficiency rating).

Per its contract with the Chicago Teachers Union (CTU), CPS requires that teachers be observed by an administrator two times per year. One of the observations must be conducted by the principal. Formal observations are to be followed up by a conference within 10 days of the observation. There is a distinction to be made between classroom observations and formal evaluation. While classroom observations must occur every year, formal evaluation occurs only every other year for most tenured

teachers (i.e., those with efficiency ratings of Excellent or Superior) and every year for non-tenured teachers and tenured teachers with a Satisfactory efficiency rating. At the time of formal evaluation, principals assign a summative efficiency rating to teachers. Efficiency ratings can be one of four categories:

- Unsatisfactory: Observations indicate an overall level of performance which is unacceptable. Identified major weakness or weaknesses have not been corrected by the teacher.
- Satisfactory: Observations indicate a generally acceptable, average level of performance.
- Excellent: Observations indicate an overall level of performance which is of higher than average quality. Exhibits potential and desire to strengthen level of performance.
- Superior: Observations indicate an overall level of performance which is outstanding. Teacher has very positive effect upon students and upon the school environment.

As with the checklist, the district does not provide a rubric for each of these efficiency ratings, so there is no clear definition of what it means to be an excellent teacher versus a superior teacher. The only guidance principals receive is the language provided above.

The New Teacher Project (TNTP) came out with a report on CPS's teacher hiring, assignment, and transfer policies—part of this report was an analysis of teacher evaluation practices (TNTP, 2007). The TNTP work concludes that neither principals nor teachers see the checklist system as meaningful or fair. Additionally, the checklist system does not lead to the identification and/or removal of low-performing teachers. In fact, teachers were extremely rarely identified as Unsatisfactory (0.3%) or even Satisfactory (7%), which means that 93% of the district's teachers were Excellent or Superior according to the checklist evaluation system.[1]

Around the district, practitioners held the view that a Satisfactory efficiency rating really meant Unsatisfactory. As an indicator of the negative light in which Satisfactory ratings were viewed, many teachers who received Satisfactory efficiency ratings grieved them, and some principals did not have the evidence to justify giving teachers this low rating. Some teachers also claimed that their principals never observed their teaching or never held post-observation conferences. All of these factors contributed to a negative culture surrounding teacher evaluation in the district.

Over three years, CPS and CTU worked together in a joint committee to solve the problems surrounding teacher evaluation and to develop a better system.[2] After looking at three classroom observation rubrics, the joint committee decided to move forward with the Charlotte Danielson Framework for Teaching. As the start of the 2008-09 school year neared, the joint committee disbanded due to a disagreement regarding the principals' ability to non-renew non-tenured teachers. The CTU's position was that non-tenured teachers who went through a rigorous evaluation like the Danielson Framework and who proved that their teaching was successful according to that Framework should not

[1] There are no Unsatisfactory teachers in our sample. In this report, we refer to teachers with Satisfactory efficiency ratings as having low ratings. Teachers with Excellent or Superior efficiency ratings are said to have high ratings. This language choice will help to differentiate between the summative efficiency ratings and the Danielson ratings, or levels of performance: unsatisfactory, basic, proficient, and distinguished.

[2] The CPS-CTU joint committee included five members from CPS and five members from CTU. Audrey Soglin, the Executive Director of the Consortium for Educational Change, facilitated the committee work.

be eligible for non-renewal. CPS felt that principals' non-renewal privilege should be protected. Therefore, the checklist remained intact, and the Framework was inserted into a section of the checklist designated Local School Unit Criteria. As such, both the checklist and the Framework were being implemented simultaneously in pilot schools in 2008-09.

Moving Towards Rigorous Evaluation: The Excellence in Teaching Pilot

Before the joint committee disbanded, the Excellence in Teaching pilot was developed over three years by CPS and CTU. The committee that designed the pilot had several goals it hoped the new evaluation system could achieve: 1) to provide a common definition of good teaching, 2) to promote, structure, and improve conversations between principals and teachers about instruction, and 3) to promote stronger teacher practice along a defined continuum.

The work of CPS and CTU to revitalize teacher evaluation mirrors efforts that are underway in other districts and states. As the efforts of school reform have turned toward instruction, teacher evaluation has become a topic of increased focus and contentious debate (Brandt, 1996). Teacher evaluation practices have been criticized as being outdated and rewarding teacher-centered forms of instruction (Sclan, 1994). Evaluation processes are lacking in accountability for low-performing teachers who are unwilling to change and at the same time criticized for not providing useful and actionable feedback to teachers trying to improve their practice (McLaughlin, 1990; Searfross & Enz, 1996).

Changes around school organization have provided another impetus for efforts to revise the process of teacher evaluation. The push for professional learning communities that enable schools to be more collaborative environments with shared decision making suggests the need for formative evaluation that is integrated into teachers' everyday practice (Sclan, 1994). There is a recognition among practitioners and researchers alike that a single observation by the principal with limited follow up does not provide teachers with the information and insight they need to assess their practice or to make timely and effective improvements (Peterson, 1990). Further, principals are generally untrained as evaluators and teacher evaluation tools lack validity (Haefele, 1993).

The challenges of effectively evaluating teachers also exist at the district level across the country. Many school districts do not have clear goals of teacher evaluation (Gitlin & Smyth, 1990). In particular, stated and written goals for teacher evaluation at the district level do not always distinguish between the formative purposes (e.g., providing feedback to teachers, mentoring, and training) and summative purposes (e.g., making decision about promotion, removal, or contract renewal) for teacher evaluation (Gitlin & Smyth, 1990; Latham & Wexley, 1982). An important consideration in teacher evaluation is the ineffectiveness of district evaluation systems in identifying and removing low performing teachers (Haefele, 1993; McLaughlin, 1990). Studies of school systems as diverse as Delaware, Chicago, Atlanta, and San Francisco revealed shockingly low teacher removal rates—often less than 1% of teachers in any given year (Darling-Hammond, 1996; Eisner, 1992; Van Sciver, 1990; Wise et al, 1984; TNTP, 2007).

The impetus for improved teacher evaluation, in part, stems from national organizations and national policy. The National Council for Accreditation of Teacher Education (NCATE), in cooperation with the Interstate New Teacher Assessment and Support Consortium (INTASC), requires that teacher education programs be able to provide performance data on current and graduated teacher candidates. INTASC has created a set of standards to define the knowledge, dispositions, and performances critical

for beginning teachers (INTASC, 1992). At the same time, state laws have begun to stipulate the requirement for standards-based teacher evaluation approaches in states such as California and Ohio (Henemen & Milanowski, 2003). Further, in order to be eligible to receive federal Race to the Top funds, states must show that they can report the number and percentage of teachers rated at each level of performance in each district's teacher evaluation system, which indicates that the Department of Education is focused on teacher evaluation (U.S. Department of Education, 2009).

These pressures for changes in teacher evaluation have resulted in the spread of improved systems throughout the United States. Examples come from diverse settings throughout the country. The STARSS system (Standards, Teaching, Accountability, Reflection, & Support System) was piloted in San Francisco beginning in 2000. By 2003-04, 30 schools were participating in the pilot (San Francisco Unified School District, 2000). Similarly, the state of Oklahoma launched the Oklahoma Teacher Enhancement Program (OTEP) in 2003 (Fredman, 2003). Las Vegas unveiled its Performance Evaluation Reports approach in 1999 and Tulsa introduced the Multi-Track Teacher Assessment system in 2000 (Smith, 2003). Cincinnati Public Schools has implemented its Teacher Evaluation System, which is based on the Danielson Framework and includes a peer evaluation component (Milanowski & Kimball, 2003). All of these rigorous systems set the stage for the Chicago reform.

As previously mentioned, the CPS-CTU joint committee agreed on a consistent definition of excellent teaching and selected Charlotte Danielson's Framework for Teaching to provide a common language for professional conversations between principals and teachers about improving teaching. It includes four major domains, each of which has multiple components. Based on rubrics that describe each level of performance, teachers receive one of four ratings for each component: unsatisfactory, basic, proficient, or distinguished.

Although important to the new teacher evaluation process, the rating tool is only one part of a larger and more extensive system. Additional features include pre- and post-observation conferences between principals and teachers; teacher documentation of their practices and evidence of student progress; assistance to low rated teachers to help improve practices; alignment of professional development resources around the evaluation; the availability of expert evaluators to assure consistency across schools; and support—plus accountability—to help principals implement the system fairly and thoroughly.[3]

As the evaluation system is broader than the tool itself, so are the district's goals for the process. These aspirations include improving teaching and learning in the school, developing a stronger professional learning climate among teachers and the principal, and developing a constructive, rather than punitive, climate around teacher evaluation.

The first year of the evaluation pilot, 2008-09, included 44 elementary schools in four CPS Areas.[4] Half of the schools within each of the four Areas were randomly assigned to participate in the pilot while the other half serve as the control group. Principals and teachers were provided with a variety of professional development and supports. Principals received one 3-day summer training, four half-day professional development sessions throughout the year, and Area meetings with fellow principals to discuss the evaluation process that took the form of professional learning communities

[3] Though part of the initial CPS-CTU plans, the targeted support for low-performing teachers was not implemented as part of the pilot.

[4] One of the pilot elementary schools grieved the pilot, so it is not included in our sample. As such, the pilot was implemented in 43 elementary schools.

(PLCs). Teachers received two school-based professional development sessions that provided an overview of the Charlotte Danielson Framework, totaling 3.5 hours of support.

The ideal observation process using the Charlotte Danielson Framework consisted of administrators conducting two classroom observations across the year. Each observation was to include: 1) a pre-observation conference (15-25 minutes), 2) the observation (a lesson, 30-60 minutes), 3) administrators matching their classroom observation notes to the Framework rubric in order to choose a level of performance for each of 10 components (45 minutes), and 4) a post-observation conference (20-30 minutes). The specified lengths of time of each aspect were provided by CPS.

In the observation, the administrators and external raters focused on Domains 2 and 3 (those which Charlotte Danielson notes are seen in a classroom observation). Domain 2 focuses on the classroom environment and components such as interactions, culture for learning, classroom management, student behavior, and physical space. Domain 3 focuses on instruction and components such as communication, questions/discussion, engagement, assessment, and flexibility.

Framework Reliability and Validity: Descriptive and Statistical Analysis

The 2008-09 school year was an off year for evaluation of tenured teachers who received high efficiency ratings in 2007-08. The teachers in our sample were drawn from those eligible for formal evaluation in Year 1 of the pilot—non-tenured teachers and tenured teachers with low prior efficiency ratings. A small number of tenured teachers with high prior efficiency ratings were also included in the sample.[5]

The evaluation design relies upon the use of "matched" observations. Three external observers were hired to observe classrooms alongside administrators in pilot schools. In this study, external observers and school administrators conducted classroom observations at the same time; however, they assigned Framework ratings independently without discussing the observed lesson. The purpose of this was to explore the technical properties of the Danielson Framework, checking for inter-rater reliability and gathering information about the patterns of ratings by component. This matched observation approach also serves to identify components that appear to be problematic for raters. Throughout the

[5] The sample of teachers was chosen based on tenure status as of the beginning of August 2008. Several factors complicate the use of the tenure status variable. Tenure status changes on a teacher's anniversary date, so if the teacher was a PAT1 (the designation for a first year teacher) at the beginning of August and had his/her anniversary date later in August, he/she was a PAT2 (the designation for a second year teacher) at the beginning of the school year. Further, a teacher's tenure status can change during the school year if that teacher was hired during the school year. As a result, many of the teachers in our sample reached their anniversary dates during that period between the beginning of August (when the sample was chosen) and the beginning of the school year. This means that we have some teachers in our sample who are tenured teachers with high efficiency ratings in 2008—teachers who do not undergo formal evaluation in 2008-09. These teachers are generally those who we had identified as PAT3 at the beginning of August, reached their anniversary date before the first observation, and are therefore in their fourth year. Another implication regarding the fluidity of tenure status is that a teacher can be observed as a PAT1, for example, in Round 1 and as a PAT2 in Round 2. Due to the complexity of this issue, we used teacher tenure in the middle of the year—December 2008—to denote tenure status in this report. Most of the classrooms observed (37%) were those of PAT2 teachers.

report, the external observers are referred to simply as "observers." In the section that follows, we summarize the results of analyses focused on this rating data.

Sample Description

The data used in this Year 1 report come from administrator and external observer ratings provided to CCSR by CPS on July 6, 2009. At that time, there were joint observation data available for 277 matched observations with principal ratings missing for 10% of the observations.[6] Of the matched observations, 147 come from Round 1 (September-December 2008) and 130 come from Round 2 (November 2008-April 2009). The joint observation period began September 18, 2008 and ended March 6, 2009. As of March 6, principals should have given all probationary appointed teachers (PATs) their efficiency ratings. PATs are teachers who have not yet received tenure status.

Table 1 provides a descriptive overview of the observation data used in this report. Most of the observations come from Area 2, which is the largest of the four pilot CPS Areas. Most of the lessons observed were in mathematics or English language arts (language arts, reading, and writing). Nearly one-third of the classrooms were grades 3-5, and the majority of classrooms are from the primary grades (preK-5). About one-third of the classrooms are from the middle grades (6-8). Multiple grade classrooms were generally special education or resource classrooms.

Few tenured teachers are part of the study—only about one-quarter of the sample (27%). The reason for this is because tenured teachers are formally evaluated in CPS every other year. This school year, 2008-09, was an off year for tenured teacher evaluation, except for teachers who received a low efficiency rating in 2007-08. PATs, however, undergo formal evaluation in the district until they have received tenure status.

[6] There is no statistical difference in observer ratings for the observations where the principal submitted data and where the principal ratings are missing.

Table 1. Description of the Lessons Observed (N=277 observations)

CPS Area					
2		*8*		*13*	*16*
125 (45%)		52 (19%)		29 (10%)	71 (26%)
Subject Area					
Math	*ELA*	*Science*	*Social Studies*		*Other*
62 (23%)	127 (46%)	34 (12%)	8 (3%)		44 (16%)
Grade Level					
PK	*K-2*	*3-5*	*6-8*		*Multiple*
14 (5%)	72 (26%)	84 (30%)	56 (20%)		51 (18%)
Tenure Status					
PAT1		*PAT2*	*PAT3*		*Tenured*
34 (12%)		104 (38%)	64 (23%)		75 (26%)
Efficiency Rating (2008-09)[7]					
None		*Satisfactory*	*Excellent*		*Superior*
36 (13%)		63 (23%)	117 (42%)		61 (22%)
Month of Observation					
September	*October*	*November*	*December*	*January*	*February*
24 (9%)	73 (26%)	59 (21%)	30 (11%)	51 (18%)	40 (14%)

Note. This table includes data from only the administrator-observer matched observations. Percentages may not add to 100 due to rounding. Also the number of observations for subject area does not add to 277 because we do not know the subject area for two of the observations, so the total sample size in the table is 275.

Ratings Data: An Overview

The Danielson Framework has four levels of performance: unsatisfactory, basic, proficient, and distinguished. Unsatisfactory is generally used to describe harmful teaching practice, while distinguished depicts teaching that is student-driven and classrooms that are truly learning communities. According to the Danielson rubric, both ratings, then, should be relatively rare. In fact, the observers only gave unsatisfactory or distinguished ratings to 4% of the observations.

Table 2 shows the overall distribution of all Framework ratings awarded to teachers in the sample. As expected, overall there are few unsatisfactory or distinguished ratings—however there is a noticeable difference between the proportion of distinguished ratings given by principals and those

[7] As mentioned, an efficiency rating is a teacher's formal evaluation rating under the checklist system. Teachers could be rated Unsatisfactory, Satisfactory, Excellent, or Superior. Unsatisfactory and Satisfactory (low) efficiency ratings were rare with 93% of ratings in 2007-08 of Excellent or Superior (high).

given by observers. Generally when a principal gave a distinguished rating, the observer saw the lesson as proficient. This difference is explored further in the statistical analysis section of the report.

Table 2. Overall Distribution of Framework Ratings (N=5,659 ratings)

Level of Performance	***Overall***	***Principal***	***Observer***
Unsatisfactory	2%	3%	2%
Basic	31%	32%	29%
Proficient	60%	53%	67%
Distinguished	7%	12%	2%

Domain 2: The Classroom Environment

Figure 1 shows the distribution of ratings for 277 joint observations for Domain 2, The Classroom Environment, which includes the following components of teaching:

- 2a. Creating an Environment for Respect and Rapport
- 2b. Establishing a Culture for Learning
- 2c. Managing Classroom Procedures
- 2d. Managing Student Behavior
- 2e. Organizing Physical Space

Figure 1 contains five sets of paired columns. The left column in each pair (labeled PR/AP) represents the administrator ratings, and the right (labeled CO for classroom observer) summarizes the observer ratings. The numbers in the columns are percentages. To read the graph, the leftmost column shows the distribution of ratings—the percent of unsatisfactory, basic, proficient, and distinguished ratings—that administrators gave in component 2a (Creating an Environment of Respect and Rapport). In this component, principals gave 1% of observations unsatisfactory, 21% basic, 60% proficient, and 17% distinguished. The column to the right of it shows the distribution of observer ratings for the same lessons—4% unsatisfactory, 18% basic, 76% proficient, and 3% distinguished. For this component, principals and observers gave a similar proportion of combined unsatisfactory and basic ratings, while principals used the distinguished level of performance much more frequently than observers did.

Administrators and observers generally assigned the same proportion of unsatisfactory/basic ratings across the Domain 2 components. The exception is with component 2e (Organizing Physical Space). Observers were likely to give a proficient rating (84%) in 2e, while administrators were considerably more critical. Principals raised concerns about this component in interviews, suggesting that component 2e was difficult to rate. For example, one principal stated:

> I have struggled with physical space [2e]—exactly what is good, what is bad? I have looked at the description. I did note for use of technology, the overhead. But is that really what they are looking for when they say technology use?

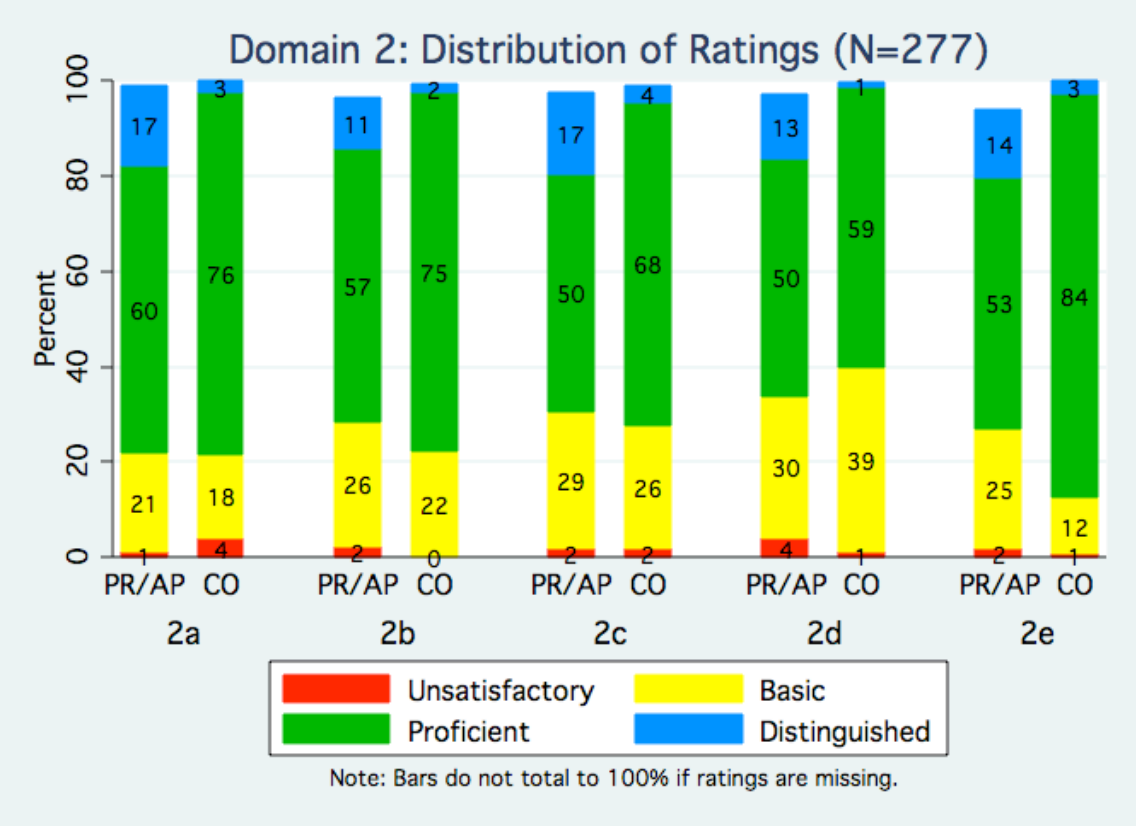


Figure 1. Distribution of Administrator and Observer Ratings for Domain 2 (N=277 joint observations)

Another possible hypothesis is that some principals may be penalizing teachers for less than ideal classroom spaces, which could explain the basic ratings. In addition, there are a number of observations where the principal chose not to assign a rating (6%, or 17 observations).

One area of concern for all Domain 2 components is the distinction between proficient and distinguished ratings. Just like with unsatisfactory ratings, external observers rarely reported distinguished teaching. Principals were much more likely to report distinguished practice, indicating they had seen it in 11% (2b. Creating an Environment of Respect and Rapport) to 17% (2c. Managing Classroom Procedures) of the lessons observed. Analysis of principal interviews provides some clues to the reasons for this. In the first place, principals expressed concern that the "distinguished" rating felt unattainable for teachers and thus might be discouraging. These principals expressed that they used high ratings to motivate teachers, rather than objectively describing instructional practice. This suggests that principals might assign distinguished ratings to proficient teaching in order to motivate or reward teachers. Principals also revealed that they, at times, assigned higher ratings than were warranted in order to preserve relationships:

> I am not going to get into a fight between these two things [proficient and distinguished] because what good does it do? You just ruin your relationship with the teacher. It would be much better for me to coach them than explain the differences between proficient and distinguished.

Further investigation is needed to deepen our understanding of the reasons that principals are more likely to assign distinguished ratings than observers. This will be an area of attention in the Year 2 evaluation work.

Domain 3: Instruction

Figure 2 shows the distribution of administrator and observer ratings in Domain 3: Instruction, which includes the following components of teaching:

- 3a. Communicating With Students
- 3b. Using Questioning and Discussion Techniques
- 3c. Engaging Students in Learning
- 3d. Using Assessment in Instruction
- 3e. Demonstrating Flexibility and Responsiveness

For both administrators and observers, there are more unsatisfactory and basic ratings in Domain 3 (Instruction) than Domain 2 (Classroom Environment). In fact, over one-third of the classrooms were rated as less than proficient on all components of instruction except component 3a (Communicating With Students). We might expect such a pattern since half of the teachers in our sample are in their first two years of teaching—instruction seems to be more difficult than establishing the classroom environment.

In Domain 3 (Instruction), there are notable differences in the proportion of unsatisfactory/basic ratings given by administrators versus observers. The component with the largest difference in ratings is 3c (Engaging Students in Learning). In 3c, 38% of administrator ratings fall below the proficient level compared to 49% of observer ratings. Unlike some of the other components in this domain, administrators were more generous in ratings than were observers. With component 3a (Communicating With Students), administrators rated lessons lower than proficient more often than observers did (32% compared to 22%). Component 3b (Using Questioning and Discussion Techniques) also has systematic rating differences. With 3b, administrators (47%) gave out more unsatisfactory/basic ratings than did observers (38%).

As in Domain 2, administrators were more likely than observers to assign distinguished ratings, though generally the difference is not as large as in Domain 2. Administrators' assignment of distinguished ratings ranged from 4% (3d. Using Assessment in Instruction) to 16% (3a. Communicating With Students). The percent of lessons that observers rated as distinguished ranged from 0% to 4%, varying by component.

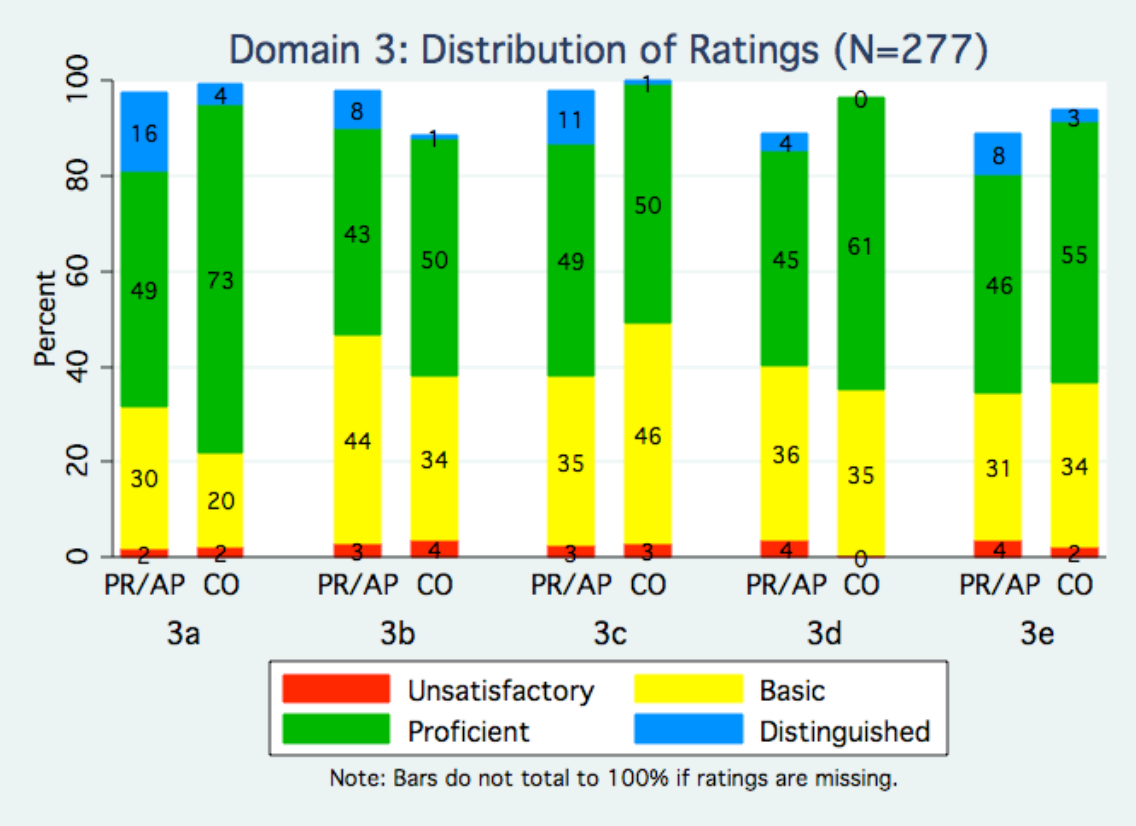


Figure 2. Distribution of Administrator and Observer Ratings for Domain 3 (N=277 joint observations)

Comparison to Prior Efficiency Ratings From the Checklist System

Figure 3 shows the distribution of all Framework ratings sorted by the 2007-08 efficiency rating of the teacher given under the checklist system. It is important to note that there are two types of teachers that fall into the category labeled "No Rating." These are teachers who are either new to the district, or teachers who successfully grieved their previous efficiency ratings. Remember that most of the teachers who successfully grieved their efficiency ratings had been given a low efficiency rating, while new teachers simply did not have a prior efficiency rating because they were not teaching in CPS in 2007-08.

Principals gave more unsatisfactory and basic Framework ratings to teachers with low efficiency ratings (i.e., Satisfactory efficiency ratings) than observers did (55% compared to 42%). Regardless of the rater, teachers with low efficiency ratings received the lowest Framework ratings of the four groups of teachers. At the other end of the spectrum, principals awarded more distinguished Framework ratings to teachers with high efficiency ratings than did observers.[8] It is important to note that the external observers did not know the teachers' efficiency ratings.

[8] Note that the 2007-08 efficiency rating is an important characteristic when looking at statistical inconsistencies between principal and observer ratings that we discuss later in this report.

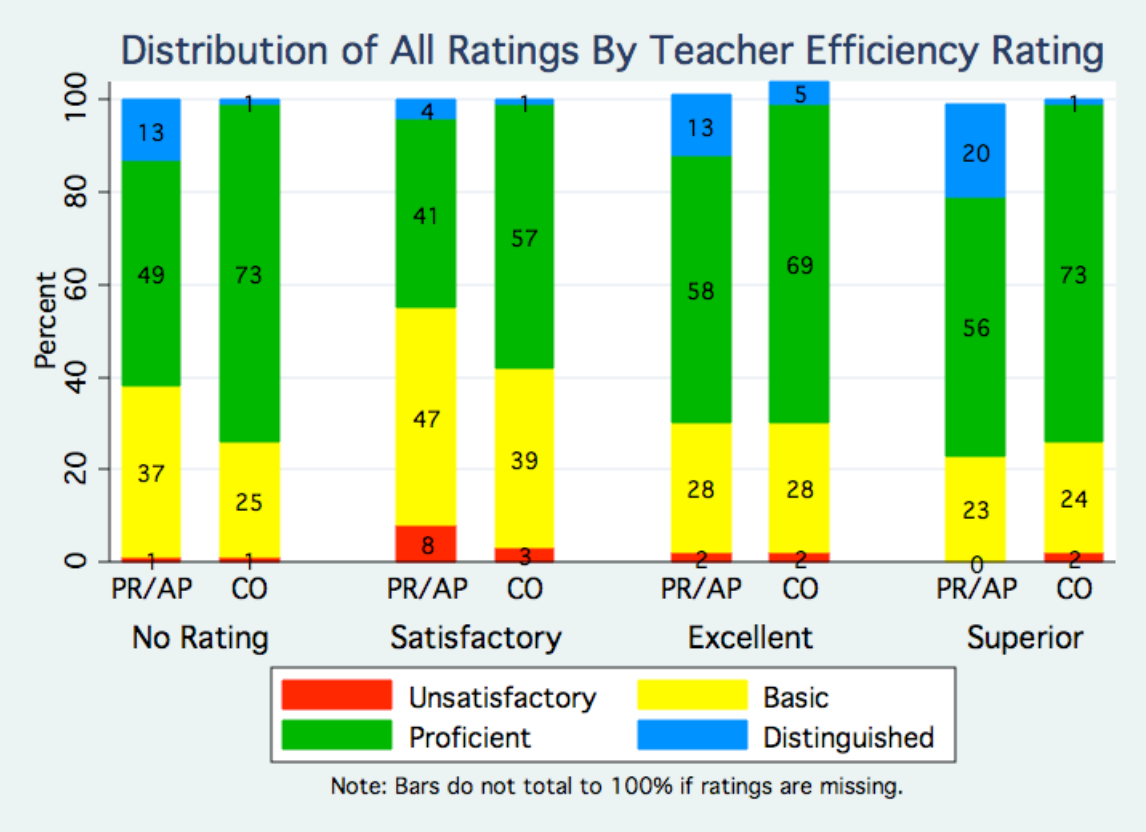


Figure 3. The Overall Distribution of Framework Ratings by Teachers' 2007-08 Efficiency Ratings (N=5,659 ratings from 277 joint observations)

Did Administrators and Observers Award the Same Ratings to the Same Lessons?

While the Framework ratings distribution graphs from the previous sections provide a snapshot of the ratings awarded, they do not indicate whether the administrator and the observer gave the same rating on a single observation. Table 3 shows administrator-observer agreement on individual components, indicating the percentage of observations in which the administrator gave the same rating as the observer (labeled as an exact match).

The second column of Table 3 contains the percent of observations where the administrator and the observer gave the same lesson an identical rating while ignoring missing data. For this column, if either the administrator or the observer did not give the lesson a rating, this observation is not included in the calculation. If, however, both the administrator and the observer did not rate the lesson, this counts as an exact match. The third column is the percentage of exact matches, including missing ratings. If an administrator rated a lesson but the observer did not (and vice-versa), this is tabulated as a mismatch and lowers the percentage of exact matches. The result is that the percentages in column 3 are lower than those in column 2. (For more detail about missing ratings, see Appendix A.)

Components with a large difference between the percentages in columns 2 and 3 are components where there are a lot of observations with missing ratings. For example, components 3d and 3e have the most missing data. In interviews, principals often cited these components as things they often did not see during a classroom observation. For instance one principal stated: "The hardest part for me was I went back to what I wrote in the classroom and I just didn't have evidence for some

components. Assessment, I couldn't rate that. Or questioning, it didn't happen in that lesson." Principals wanted clarification about what to do if they did not observe a component in a lesson.

Table 3. Exact Match of Administrator and Observer Ratings for a Lesson (N=277 observations)

Component	***% Exact Match (no missing)***	***% Exact Match (missing)***	***Rating >1 level off (N)***
2a) Environment of respect and rapport	56%	55%	5%
2b) Culture for learning	60%	57%	3%
2c) Classroom procedures	49%	47%	4%
2d) Student behavior	52%	51%	4%
2e) Physical space	54%	51%	3%
3a) Communication	53%	52%	3%
3b) Questioning and discussion techniques	51%	44%	5%
3c) Student engagement	49%	48%	6%
3d) Assessment in instruction	55%	48%	3%
3e) Flexibility and responsiveness	52%	43%	4%

From column 2, the percentages range from 49-60%. The component with the highest percentage of exact matches (60%) is 2b (Establishing a Culture for Learning), while the lowest (49%) are components 2c (Managing Classroom Procedures) and 3c (Engaging Students in Learning). On average, there is an exact match for 54% of the lessons on Domain 2 components and for 52% of the lessons on Domain 3 components.

Column 4 shows the percentage of observations (total N=277) where the administrator's rating was more than one level off from the observer's rating, denoted as an extreme mismatch. This can happen in two cases: 1) one rating is unsatisfactory and the other is proficient/distinguished, or 2) one rating is distinguished and the other is unsatisfactory/basic. Of note here is that there are very few observations with an extreme mismatch between administrator and observer ratings. If the cell is shaded blue, that means the administrators rate higher than the observers in most of the cases; and yellow indicates that there is not a clear pattern. For none of the components, do observers generally give higher ratings than administrators in these cases with extreme mismatch of ratings.

The table shows mismatches per component. If we look at mismatches per observation (not shown in the table), 21% of observations had at least one component where the administrator and observer were off by more than one level of performance. Of these observations, 73% were cases where

the administrator awarded a higher rating than the observer did. The largest number of extremely mismatched ratings occurs in 3c (Engaging Students in Learning) (N=15 or 6%). The fewest extremely mismatched ratings occurs in 2e (Organizing Physical Space) and 3d (Using Assessment in Instruction) (N=7 or 3%).

The following is one example of the evidence for 3c (Engaging Students in Learning) when an observer rated the lesson as basic and the administrator rated the lesson as distinguished. The phonics lesson is in reading for the upper elementary grades and took place in the late fall.

> Observer evidence for a basic rating. Whole group instruction—the students correct a sentence together on white board. Students copy sentence with its corrections from the board. A lot of time (approx. 15 min.) was spent going over the sentence and having the students copy from the board. Students go to desks and have a discussion about word families. Students volunteer to give letters to complete the words written on the board. Lesson lacks rigor for older higher-functioning students (they should already know 3-letter words by 4th grade).
>
> Principal evidence for a distinguished rating. The teacher is constantly providing the students with compliments and redirecting their negative behavior with positive compliments ("Everyone eyes up here" or "What do we do when…?"). The students know where their supplies are on the table. The students follow the cues as to when to make transition from one subject to another. The teacher works with the students in groups and one-on-one.

As expected, the observer's evidence is focused on the content of 3c; she describes the activities and assignments, the intellectual engagement, and the structure and pacing of the lesson. The principal, however, provides evidence that relates to other components. For example, "the teacher is constantly providing the students with compliments" applies to 2a (Creating an Environment of Respect and Rapport), and "the students know where their supplies are on the table" is 2c (Managing Classroom Procedures). The matching of evidence to Framework components is a critical step in assigning a Framework level of performance and an area in which principals may need more support. A preliminary look at the classroom observation evidence indicates that this pattern generally holds true across the extreme mismatches; however, part of our Year 2 evaluation work will include a systematic textual analysis of principal and observer evidence.

Tables 4 and 5 contain the cross tabulations of administrator ratings with observer ratings at the component level for 3c (Engaging Students in Learning) and 3d (Using Assessment in Instruction). We provide these tables to show around which levels of performance the most disagreement occurs. We choose to highlight two components where our statistical analysis (discussed later in this report) indicates that there are inconsistencies in the ways that principals and observers are giving ratings, especially around the basic and proficient levels.

For component 3c, just under half of the observations were exact matches (49%), and there were very few missing ratings (only 6 observations). In Table 4, the exact matches are denoted in bold italics. Of the 137 observations that were not exact matches, most often (in 71% of the mismatches) administrators and observers disagreed over whether an observation should be rated basic or proficient. For example, in 22% of the observations administrators gave proficient ratings when the observer gave

basic ratings (the blue cell in Table 4). In another 13% of the cases, the observer gave proficient ratings when the administrator gave basic ratings (the pink cell in Table 4). In general, the administrator rated higher than the observer in mismatched observations, though not always. In interviews, principals were concerned about the difference between the basic and proficient levels of performance; it was an area where they wanted more guidance.

Table 4. Component 3c (Engaging Students in Learning): Administrator-Observer Agreement

Observer Ratings	***Administrator Ratings***				
	Unsatisfactory	Basic	Proficient	Distinguished	Total
Unsatisfactory	***2*** ***(.7%)***	5 (2%)	1 (.4%)	0 (0%)	8 (3%)
Basic	1 (.4%)	***56*** ***(20%)***	60 (22%)	10 (4%)	128 (46%)
Proficient	4 (1%)	37 (13%)	***74*** ***(27%)***	5 (2%)	139 (50%)
Distinguished	0 (0%)	0 (0%)	0 (0%)	***2*** ***(.7%)***	2 (.7%)
Total	7 (3%)	98 (35)%	135 (49%)	31 (11%)	271[9] (100%)

Table 5 provides a detailed look at administrator-observer agreement for another component, 3d (Using Assessment in Instruction). Unlike with 3c, there are a lot of missing ratings for 3d. Either the administrator or observer (or both) did not enter a rating for 3d for 14% of the observations (N=39). In most of the observations for which the 3d rating was missing (74%), the administrator did not enter a rating even though the observer did.

Table 5. Component 3d (Using Assessment in Instruction): Administrator-Observer Agreement

Observer Ratings	***Administrator Ratings***				
	Unsatisfactory	Basic	Proficient	Distinguished	Total
Unsatisfactory	***0*** ***(0%)***	1 (.4%)	3 (0%)	0 (0%)	1 (.4%)
Basic	5 (2%)	***45*** ***(19%)***	35 (15%)	3 (1%)	88 (37%)
Proficient	4 (2%)	52 (22%)	***87*** ***(37%)***	6 (3%)	149 (63%)
Distinguished	0 (0%)	0 (0%)	0 (0%)	***0*** ***(0%)***	0 (0%)
Total	9 (4%)	98 (41%)	122 (51%)	9 (4%)	238[10] (100%)

[9] Note that for six observations either the principal or the observer did not submit a rating, so the total matched observations for 3c is 271 instead of 277.

[10] Note that for 39 observations either the principal or the observer did not submit a rating, so the total matched observations for 3d is 238 instead of 277.

With component 3d, just over half of the observations were exact matches (56%). In Table 5, the exact matches are denoted in bold italics. Of the 106 observations that were not exact matches, most often (in 82% of the mismatches) administrators and observers disagreed over whether an observation should be rated basic or proficient. For example, in 15% of the observations administrators gave proficient ratings when the observer gave basic ratings (the blue cell in Table 5). In another 22% of the cases, the observer gave proficient ratings when the administrator gave basic ratings (the pink cell in Table 5). In general, the administrator rated lower than the observer in mismatched observations, though not always.

In sum, most disagreement occurred around the basic and proficient levels of performance. (This disagreement is also highlighted in the statistical analysis in the report.) In addition, when it comes to the basic/proficient disagreement, sometimes principals rate higher than observers, while other times observers rate higher.

How the Ratings Change Over Time

Table 6 shows the percent of proficient/distinguished ratings from both administrator AND observer ratings for Round 1 AND Round 2 observations. The data include only teachers for whom the administrator has submitted both Round 1 and Round 2 ratings, which accounts for 254 observations of 127 teachers.

Observer ratings increased significantly between rounds for three components:[11]

- 2d. Managing Student Behavior
- 3d. Using Assessment in Instruction
- 3e. Demonstrating Flexibility and Responsiveness

Administrator ratings, however, showed a significant increase for only one component: 3b (Using Questioning and Discussion Techniques). One possible explanation for the limited change in ratings is that there was a relatively short amount of time between Round 1 and Round 2 observations. On average, 12 weeks of school passed between observations. It is also worth noting that principals had on-going professional development to help them use the Framework and learn more about the components and the levels of performance. Therefore, it is reasonable to say that principals were refining their assignment of Framework ratings throughout the year. We would expect such adjustments to occur in the first year of implementation.

Despite the short timeframe between observations, the reason for improved ratings in questioning may be due to a focused effort on the part of principals. A large proportion of principals reported increased focus on questioning because of their use of the Framework. More than half of interviewed principals, when asked to talk about the Framework or specific components, talked about

[11] While there is a significant drop in observer ratings for Component 2b (Establishing a Culture for Learning), this is an area where the observers received additional training in the middle of the observation schedule. The drop in ratings suggests that the Round 1 observer ratings for 2b may have been inflated. It does not necessarily mean that teacher performance in this area got worse over time.

3b (Using Questioning and Discussion Techniques). Principals expressed astonishment at how low a level teachers' questions were. For instance, once principal stated:

> Using the Framework made me realize, we really have to focus on higher order thinking questions because the questions are just way too basic, they are low level. Because sometimes I think as teachers we just want the kids to give us the right answer and that makes us feel like we have taught them something. I started saying, oh, I am hearing basic here, you know? And this could be someone in the past that I had been rating Superior!

The majority of these principals stated that they explicitly made improving questioning a goal for the second observation, which might explain the improved ratings.

Table 6. Trends in Proficient/Distinguished Ratings Over Time (N=127 teachers/256 observations)

Component	***Observer Ratings***		***Administrator Ratings***	
	Round 1	***Round 2***	***Round 1***	***Round 2***
2a) Environment of respect and rapport	76%	84%	78%	78%
2b) Culture for learning	83%	73%**	74%	69%
2c) Classroom procedures	69%	75%	69%	66%
2d) Student behavior	53%	68%*	66%	64%
2e) Physical space	88%	86%	74%	71%
3a) Communication	76%	80%	68%	71%
3b) Questioning and discussion techniques	55%	61%	49%	57%**
3c) Student engagement	48%	52%	59%	67%
3d) Assessment in instruction	56%	74%**	51%	60%
3e) Flexibility and responsiveness	53%	71%**	63%	62%

Note. Asterisks indicate a significant difference between Round 1 and Round 2 ratings. *** p<.01, ** p<.05, *p<.10.

Statistical Analysis of Ratings Data

To understand the Framework tool and how raters assigned Framework ratings, we used two methods: multi-facet Rasch measurement (MFRM) and hierarchical linear modeling (HLM). The MFRM analysis is a technique similar to Rasch analysis that allows us to investigate topics such as component

difficulty and rater severity (Linacre, 1994). We also looked at how reliably the Framework ratings in aggregate measure overall teaching ability, controlling for rater severity. HLM is a hierarchical regression analysis technique that allows us to look at how various principal and teacher characteristics may affect the way a principal rates a lesson (Bryk & Raudenbush, 1992). This technique allows us to group Framework ratings within individual teachers. A more detailed explanation of these methods and how they were applied to our data is included in Appendix A.

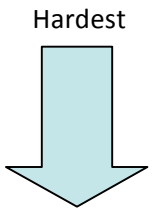
Reliability of the Framework: Multi-Facet Rasch Measurement

The MFRM analysis allowed us to investigate influences on teacher ratings in and among six categories: teacher, component, rater (includes the three external observers and each individual principal), observation round, subject area, and CPS Area. The model calculates the probability that a teacher will get a particular rating taking into consideration these categories, or facets, including rater severity. The model also provides us with a measure of rater severity for each of the observers and principals.

The individual teacher measures of teaching ability generated in this analysis combine all of the Framework component teacher ratings for an individual teacher. We find that the individual teacher measures are highly reliable (reliability=.94, separation=3.92). When we use reliability in this context, we mean that teachers with estimated measures of high teaching ability actually were more successful in the classroom during the observed lesson than teachers with estimated measures of low teaching ability. Reliability also means that we have faith that the teacher measures can be reproduced. Again, remember that teacher measures were created by aggregating the principal and observer Framework ratings for each teacher. The separation indicates the ratio of signal to noise, which means that the instrument has about 4 times as much signal as noise. In other words, the Framework ratings combine to produce reliable measures of overall teacher ability.

Another important finding is that it makes sense to treat each of the components as distinct aspects of teaching (reliability=.97, separation=5.59). In this case, reliability is a measure of how well we can separate each of the components from each other, and the reliability of the Framework components is high (i.e., close to 1). We are also able to determine a hierarchy of components. We can order the components in terms of difficulty (see Figure 4). In other words, the components at the top of the figure were those on which it was most difficult to receive a higher rating. From the order of the component difficulties, we can conclude that the component hierarchy is a reasonable one. For example, it is logical that a skill such as 3b (Using Questioning and Discussion Techniques) is more difficult for a teacher to master than 3a (Communicating With Students), or that components in the Instruction domain are more difficult to master than those in the Classroom Environment domain. Because the components arrange into an order that is in accordance with our conceptual expectations and the reliability measure is high, there is considerable indication that the Framework components are valid.

Figure 4 lists the components in order of difficulty. The Domain 2 (Classroom Environment) components are in green font, while the Domain 3 (Instruction) components are purple. In general, the Domain 3 components are more difficult for teachers than are Domain 2 components. The hardest component is 3b (Using Questioning and Discussion Techniques) and the easiest component is 2e (Organizing Physical Space). The MFRM analysis shows that each of these components is measuring a unique aspect of teaching. In other words, the Framework components are discrete.



Framework Component
Using questioning and discussion techniques
Engaging students in learning
Using assessment in instruction
Demonstrating flexibility and responsiveness
Managing student behavior
Managing classroom procedures
Establishing a culture for learning
Communicating with students
Creating an environment of respect and rapport
Organizing physical space

Figure 4. Ranking of Components According to Difficulty

Another finding from the MFRM analysis pertains to how severe or lenient the raters are. We found that there are differences in rater severity—both among principals and among observers. However, on the whole, there are no significant differences between rater types (i.e., external observer versus administrator). Moreover, each principal is assigned a measure of severity, and we categorized the principals in terms of their severity. High severity principals were two standard errors more severe than the most severe external observer, while low severity principals were two standard errors more lenient than the least severe external observer. All the remaining principals were average in terms of severity. Table 7 shows the distribution of principal severity. Again, keep in mind that these differences in severity did not interfere with overall ratings—there is no significant difference between principals and observers (t=.615).

Table 7. Categories of Principal Severity

Level of Severity	***Percent of Principals (N=42)***
Severe	30%
Average	53%
Lenient	16%

Despite differences in severity among raters, the hierarchy of the components is consistent. No matter how severe a rater is, for example, 3b always receives the lowest ratings and 2e receives the highest ratings. But for a more severe rater, the ratings are lower than for a more lax rater (and similarly for other components).[12]

[12] In addition, there are indications that some raters are treating some of the components as more difficult or less difficult than expected. However, the number of such significantly large unexpected interactions is very small, amounting to only about 10% of the total number of rater-by-component interactions.

Finally, we can say that both principals and observers are using the rating scale (the levels of performance) consistently with one another and within themselves. For each of the levels of performance, the fit statistics are very close to 1, which indicates that the rating scale is consistent. Given this finding, another way to think about rater severity is that principals who consistently give lower ratings than observers (i.e., severe principals) maintain the component difficulty shown in Figure 4, but their rating scale shifts up so that it is harder to get higher ratings from that principal. Because principals are using the rating scale consistently—despite severity—for severe principals, the rating scale is shifted up for all of their teachers. It is generally not the case that an individual principals shift the scale up for some teachers and down for other teachers.

In sum, the major takeaways from the MFRM analysis are the following:

- Overall individual measures for teachers created from all of the principal and observer Framework ratings are reliable—after controlling for rater severity.
- There are differences among individual raters in terms of severity.
- Despite these differences, there is a hierarchy in terms of component difficulty—one that, on face value, makes sense conceptually.
- Individual principals and observers are using the rating scale, or levels of performance, consistently.

Inconsistencies in Framework Ratings: Hierarchical Modeling

The MFRM analysis allowed us to explore the potential reliability of the Framework tool itself after controlling for rater severity. The model generated estimated teaching ability measures for the individual teachers in the sample by aggregating all Framework ratings across components and raters for these teachers. The following hierarchical modeling analysis allowed us to look at the ratings data in a different way. To complement the MFRM analysis, we used hierarchical modeling to understand systematic differences between principal ratings and observer ratings—both overall and component-by-component. While the MFRM analysis tells us something about the reliability of the Framework tool, the hierarchical modeling analysis will help the district understand where it needs to target future professional development for principals and for Framework users in general.

The analysis of Framework rating data included a three-model approach. Each hierarchical logit model looks at a different step in the rating scale:[13]

- Model 1 compares the likelihood of getting an unsatisfactory rating to getting a basic/proficient/distinguished rating. This model focuses on rater effects at the low end of the ratings scale.
- Model 2 compares the likelihood of getting a proficient/distinguished rating to getting an unsatisfactory/basic rating. This model focuses on rater effects in the middle of the ratings scale, which is where most of the ratings are.
- Model 3 compares the likelihood of getting an unsatisfactory/basic/proficient rating to getting a distinguished rating. This model focuses on rater effects at the high end of the ratings scale.

[13] Another approach we could have taken is to use an ordered logit model. However, the ordered model assumes that the distance between ratings on the scale is equal. We believe that the distance from unsatisfactory to basic, and also from proficient to distinguished, is much larger than that from basic to proficient, so we used the three model approach.

For each of these models, the overall structure of the HLM is ratings within teachers. This means that we are taking all of the Framework ratings that the principals and observers assigned and clustering them together for each of the teachers in the sample. At Level 1 of the models, the outcome is the Framework rating, and the explanatory variables all relate to the observation—the Framework component, the round of the observation, and an overall main rater indicator variable or interactions between the component and the rater. The interactions are what allow us to identify any consistent differences between principals and observers. Level 2 of the model includes variables related to principal and teacher characteristics. If we find rater effects for any of the components in Level 1 of the model, we can try and explain those differences with the variables in Level 2. Possible Level 2 covariates include 2007-08 teacher efficiency rating, tenure status, CPS Area (2, 8, 13, and 16), whether or not the teacher is designated as a special education teacher, subject area, grade level, proxy for student achievement at a school, and whether or not a principal is on first contract. We start by looking at the models that focus on the low and high ends of the rating scale (models 1 and 3), and then we explore the middle two categories in more detail (model 2).

In Model 1—the model looking at the unsatisfactory level of performance—only 2% of all Framework ratings (N=126) were unsatisfactory. The results from this model indicate that principals and observers are consistently identifying low-level instruction in general. While there is no significant overall rater effect, there are two components where principals and observers are rating inconsistently: 2d (Managing Student Behavior) and 3d (Using Assessment in Instruction). With both of these components, principals are more likely to use the unsatisfactory rating than are observers. Otherwise, principals and observers are generally agreeing when they see unsatisfactory teaching.

In Model 3—the model looking at the distinguished level of performance—only 7% of all Framework ratings (N=387) were distinguished. While larger than the number of unsatisfactory ratings, there were still very few distinguished ratings awarded. The results from this model indicate that there are large overall rater effects—when it comes to giving the distinguished rating, principals are much more likely than observers to rate teaching as distinguished. This trend holds for each of the 10 Framework components with the largest disparities in ratings in components 3c (Engaging Students in Learning) and 2d (Managing Student Behavior).

Model 2 is where we will focus most of our attention. Most of the ratings (91%) were basic or proficient, so we want to look closely at the distinction between those levels of performance.

Table 8 shows for each Framework component the odds that a principal gives a proficient/distinguished rating compared to the odds that an observer gives a proficient/distinguished rating (i.e., the odds ratio). The odds ratio can be interpreted in the following way:

- For an odds ratio that equals 1, the likelihood that a principal gives a proficient/distinguished rating is the same as the likelihood that an observer gives a proficient/distinguished rating. In other words, there is no consistent difference in the way principals and observers give ratings.
- For an odds ratio that is greater than 1, the likelihood that a principal gives a proficient/distinguished rating is higher than the likelihood that an observer gives a proficient/distinguished rating. In other words, the principal is likely to give consistently higher ratings than the observer. An odds ratio of 2 means that the principal is twice as likely to give higher ratings.

- For an odds ratio that is less than 1, the likelihood that a principal gives a proficient/distinguished rating is lower than the likelihood that an observer gives a proficient/distinguished rating. In other words, the principal is likely to give consistently lower ratings that the observer. An odds ratio of .5 means that the principal is half as likely to give higher ratings.

In the table, the asterisks indicate components where there is a significant difference in the way principals and observers assign ratings. For components 2e (Organizing Physical Space), 3a (Communicating With Students), and 3d (Using Assessment in Instruction), the principal consistently assigns lower ratings than the observer. In other words, the principal will often assign a basic rating when the observer assigns a proficient rating. With component 3c (Engaging Students in Learning), however, the principal consistently rates higher than the observer—the principal will often assign a proficient rating when the observer assigns a basic rating.

Table 8. How Principals Rate Components Compared to Observers

Component	***Odds Ratio***	***Explanatory Variables***
2a) Environment of respect and rapport	0.96	--
2b) Culture for learning	0.71	--
2c) Classroom procedures	0.85	--
2d) Student behavior	1.21	--
2e) Physical space	0.39***+	1. 2008-09 efficiency ratings 2. Tenure status 3. Student achievement++
3a) Communication	0.62***	None
3b) Questioning and discussion techniques	0.86	--
3c) Student engagement	1.46**	1. 2008-09 efficiency ratings
3d) Assessment in instruction	0.67**	None
3e) Flexibility and responsiveness	0.96	--

Note. Asterisks indicate for which components there are systematic differences between principal and observer. *** p<.01, ** p<.05. + Indicates that we eliminate the difference in rating with the explanatory variables. ++ We use a proxy of student achievement in our analysis. We took the average 2008 3rd grade ISAT reading score for all of the pilot schools and then divided them into quartiles. The model included indicator variables for the quartiles.

For these four components, we try to provide some explanation for the inconsistency in ratings.

Component 2e. For component 2e (Organizing Physical Space), we are able to explain all differences in principal and observer ratings with the use of three explanatory variables: teachers' prior efficiency ratings, teacher tenure status, and a proxy for student achievement. The effect of these three variables on component 2e is described below.

- Efficiency ratings: Compared to observers, principals were more likely to give teachers with low 2007-08 efficiency ratings lower Framework ratings. Principals were also likely to give teachers with high 2007-08 efficiency ratings higher Framework ratings when compared to observers.[14]
- Tenure status: Compared to observers, principals gave higher ratings to PAT2 teachers than to tenured teachers.[15]
- Student achievement: Compared to observers, principals in the top two quartile schools rated lower than principals in the bottom two quartiles of schools.

Component 3c. For component 3c (Engaging Students in Learning), we are not able to eliminate ratings differences, but we are able to explain the difference partially with 2007-08 efficiency ratings. As with 2e, compared to observers, principals were more likely to give teachers with low 2008-08 efficiency ratings lower Framework ratings. Principals were also likely to give teachers with high 2007-08 efficiency ratings higher Framework ratings when compared to observers.

Components 3a and 3d. For components 3a (Communicating With Students) and 3d (Using Assessment in Instruction), we are not able to explain the differences in principal and observer ratings. These two components (and especially 3d, which caused problems at the low end of the rating scale as well) require further analysis. Two approaches that we plan to take in the next year are 1) to ask targeted questions about problematic components to learn more about how principals and teachers define these aspects of teaching, and 2) to conduct a systematic textual analysis of the evidence provided by principals and observers to see if the two parties are capturing similar information.

To summarize the findings from the HLM models (see Table 9), principals and observers are pretty consistently identifying low-level teaching. However, at the high end of the ratings scale, principals are more likely to call teaching distinguished than observers are. The overwhelming majority of the ratings are basic or proficient, and there are a few components that are problematic for raters. While we cannot explain all of the ratings differences, a consistent contributing factor is the teacher's 2007-08 efficiency rating. Compared to observers, principals give teachers with low prior efficiency ratings significantly lower Framework ratings than teachers with high prior ratings.

In sum, the major takeaways from the HLM analysis are the following:

- Principals and observers generally agree on low-level teaching—the exceptions being with 2d (Managing Student Behavior) and 3d (Using Assessment in Instruction) where principals are consistently more likely to give unsatisfactory ratings.
- Most of the rating inconsistencies occur around the basic/proficient levels of performance. Principals need more support in establishing norms around using these levels of performance and also in forming a shared definition of these components of teaching:
    - 2e (Organizing Physical Space)
    - 3a (Communicating With Students)

[14] With efficiency ratings, it is important to note that the principal who gave the Framework ratings was the same principal who gave the efficiency rating in the prior school year. The only exception is if the teacher switched schools or if the teacher was new to the district (in which case he/she would not have a prior efficiency rating). Further, the observer did not have knowledge of a teacher's prior efficiency rating.

[15] Half of the tenured teachers in this sample had low prior efficiency ratings of Satisfactory compared to 7% of all teachers in CPS.

- 3c (Engaging Students in Learning)
- 3d (Using Assessment in Instruction)

- At the distinguished level, principals consistently award these high ratings more often than observers. This holds across all components. Part of this may be due to the simultaneous implementation of the Framework and the checklist and, therefore, a principal's desire to maintain a teacher's prior high efficiency rating.

Table 9. Summary of Rating Inconsistencies and Possible Contributing Factors

Overall Rater Effect	***Component-Level Rater Effects***	***Principal: High or Low***	***Explanation***
Model 1: The Low End of the Rating Scale			
no	2d. Managing Student Behavior	low	***2008-09 efficiency ratings*** Principal first contract Subject area
	3d. Using Assessment in Instruction	low	***2008-09 efficiency ratings*** Special education teacher Student achievement
Model 2: The Middle of the Rating Scale			
no	2e. Organizing Physical Space	low	***2008-09 efficiency ratings*** Teacher tenure status Student achievement
	3a. Communicating With Students	low	None of the covariates
	3c. Engaging Students in Learning	high	***2008-09 efficiency ratings***
	3d. Using Assessment in Instruction	low	None of the covariates
Model 3: The High End of the Rating Scale			
yes	All 10 Components	high	***2008-09 efficiency ratings*** CPS area Subject area Teacher tenure status Grade level Student achievement

Putting Together the MFRM and HLM Analyses

The MFRM analysis demonstrates that the Framework tool itself can be used reliably to create aggregate measures of teacher performance. However, the HLM analysis reveals that there are some areas where principals need more support in applying Framework ratings to teaching—particularly with the basic and proficient levels of performance. One conjecture is that those principals who have been

identified through the MFRM analysis as being significantly severe or those who are significantly lenient are driving the differences that we found in the HLM analysis.

Using the Framework for Summative Evaluation

Charlotte Danielson intended for the Framework to be used for formative purposes, contending that the most important aspect of this observation work is the conversation about instructional improvement that occurs between principals and teachers. Understandably, the CPS-CTU joint evaluation committee sought to find a way to use the Framework for both summative and formative evaluation with the agreement that the checklist system was broken in that it was not adequately identifying low-performing instruction nor was it a factor in facilitating professional, instruction-focused conversations between principals and teachers.

During the design and implementation of the pilot evaluation system, CPS and CTU discussed the possibility of using the Framework ratings for summative evaluation purposes. For instance, criteria could be established using ratings on the Framework for the non-renewal, removal, or mandatory professional development for teachers. Part of this discussion included differentiated criteria based on tenure status—the logic being that more seasoned teachers should be more highly skilled teachers than newer teachers. For PATs, the possible standard for being identified as low performing was receiving any unsatisfactory Framework rating. Tenured teachers would be identified as low performing if they received any unsatisfactory Framework rating or more than one basic Framework rating per domain.

This section explores the application of different criteria for identifying low-performing teachers using the Framework. The analysis provides answers to questions like: If different criteria were applied, how many teachers would be rated as low performing? What percentage of teachers in the sample would be identified as eligible for supports or sanctions?

Figure 5 shows how many teachers in the sample meet four different ratings benchmarks. The teachers are divided into three groups:

- PATs—these are teachers in the first three years in a CPS classroom.
- Tenured Satisfactory—these are tenured teachers who received a low (Satisfactory) efficiency rating in the 2008-09 school year.[16] Though labeled Satisfactory, it is commonly accepted that these teachers are not performing at a high level.
- Tenured Plus—these are teachers who receive a high (Excellent or Superior) efficiency rating in the 2008-09 school year. Generally, these are teachers in their fourth year in the classroom.[17]

In Figure 5, the first circle shows the number of teachers with principal Framework ratings from Round 2 observations. Keep in mind that Round 2 observations took place between November 2008 and February 2009, so it is possible that these teachers improved by the end of the school year. The leftmost

[16] It should be noted that this group was oversampled, as they are the only tenured teachers who underwent formal evaluation in the 2008-09 school year.

[17] The tenured teacher sample is not representative of the greater population of tenured teachers in CPS. Further the number of tenured teachers in our sample is small.

circle includes the total number of teachers: 130. The second set of circles shows how these teachers split among PAT, tenured Satisfactory (tenured teachers with low 2007-08 efficiency ratings), and tenured plus (teachers with high 2007-08 efficiency ratings)—73% of these teachers are PATs. The remaining set of circles shows the number of teachers who meet the four ratings benchmarks, based only on administrator ratings, which are described below.

1) No unsatisfactory Framework ratings—these are the teachers who did not receive any unsatisfactory ratings. 92% of the teachers in the sample meet this benchmark, identifying 8% of teachers as low-performing.

2) Fewer than 5 basic Framework ratings—these are the teachers who met benchmark 1 and who, at a maximum, had 4 basic ratings. 64% of the teachers in the sample meet this benchmark, identifying 36% of teachers as low-performing.

3) Fewer than 3 basic Framework ratings—these are the teachers who met benchmarks 1 and 2 and who, at a maximum, had 2 basic ratings. 52% of the teachers in the sample meet this benchmark, identifying 48% of teachers as low-performing.

4) Only proficient and/or distinguished Framework ratings—these are the teachers without any unsatisfactory or basic ratings. 34% of the teachers in the sample meet this benchmark, identifying 66% of teachers as low-performing.

In short, where the benchmark is set has a huge impact on the fraction of teachers identified as low-performing and is a human capital concern. This is exacerbated by the inconsistencies in the way that principals use the Framework for some of the components. However, an important point is that any set of the discussed criteria for using the Framework for summative evaluation does a better job of identifying struggling teachers than the checklist system. This is important in terms of being able to offer struggling teachers the support they need in order to improve. Further, even some of the tenured teachers identified as high performing using the checklist receive multiple basic Framework ratings, suggesting that the district's Excellent and Superior teachers have room to improve their teaching.

Additional considerations on this topic are included in the section on implications.

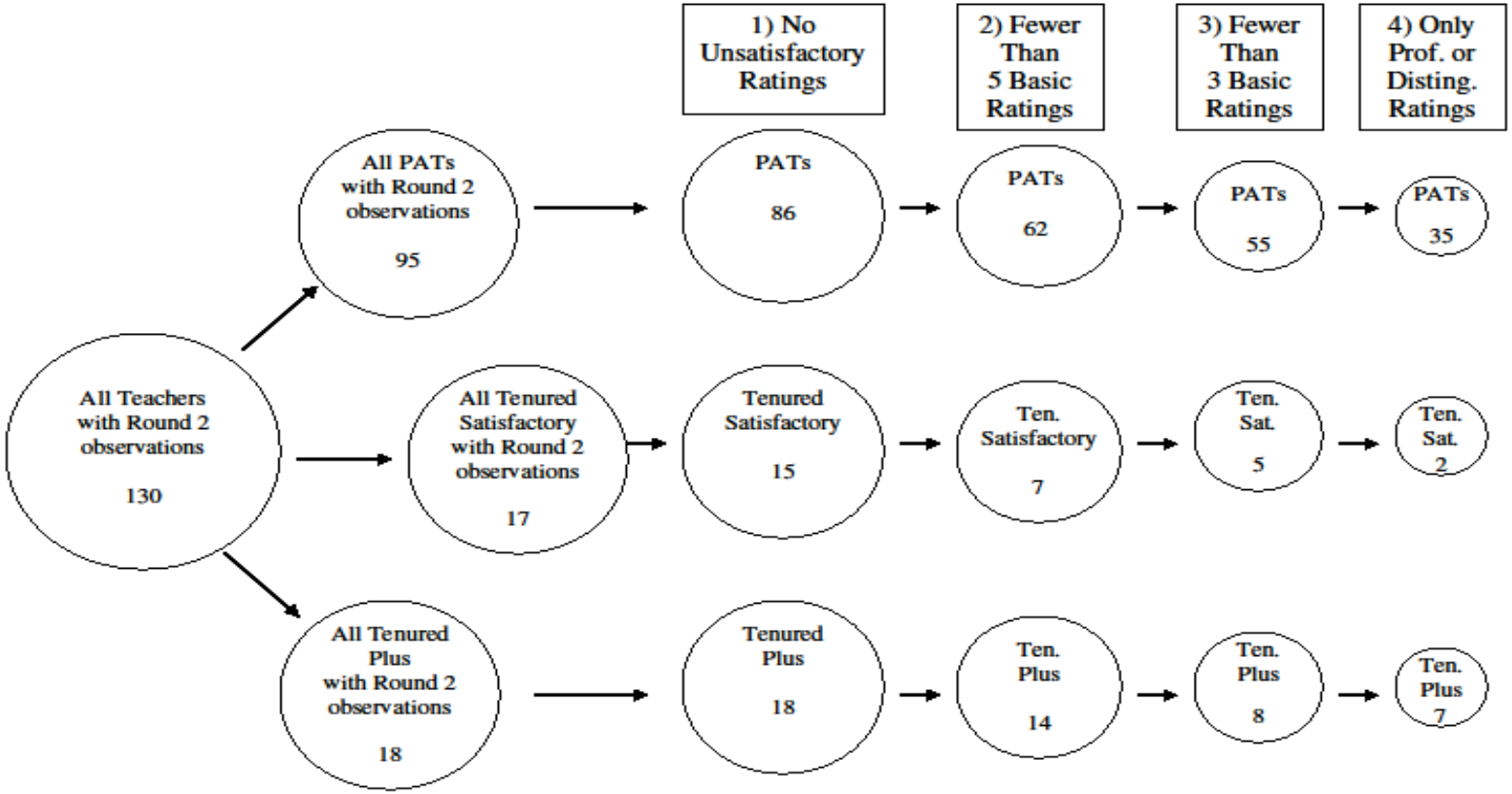


Figure 5. Applying Various Ratings Benchmarks to the Teachers (Principal Round 2 Observation Ratings Only)

Exploring Principal and Teacher Perceptions: Interview Analysis

In this section, we explore research questions 2 through 4, which focus on the perceptions of principals and teachers regarding their participation in the Excellence in Teaching pilot study. Principals and teachers were asked questions about the following: a) the training and professional development they received on the Charlotte Danielson Framework for Teaching, b) their perceptions of the Danielson Framework, c) their roles in implementing the evaluation system, d) the pre- and post-conferencing process, and e) their perception of school change that resulted from implementing the Danielson Framework. For a detailed explanation of the qualitative methodology, see Appendix A.

Perceptions of the Charlotte Danielson Framework

Principals and teachers were asked a series of questions about their perceptions of the Charlotte Danielson Framework. Both groups commented on the quality of the Framework and its ability to accurately measure teacher performance. Principals identified the components they found most difficult to rate, and teachers discussed if and how participation in the evaluation process had influenced their instructional practice. In general, both teachers and principals were overwhelmingly positive about the Framework, although there were a few areas of concern. (See Table 10 for an overview of principal and teacher perceptions about the Framework.)

Table 10. Principal and Teacher General Attitudes about the Framework[18]

	Mostly Positive	Mixed	Mostly Negative
Principal Attitudes (N=39)	69%	16%	16%
Teacher Attitudes (N=26)	65%	35%	0%

Principal Attitudes

When prompted to discuss their general impressions, 69% of principals from our sample expressed mostly positive statements about the Danielson Framework, 16% of principals offered mixed statements, and 16% of principals were mostly negative. According to principals, the Framework focused their observations and allowed them to be more concrete in coaching teachers on instruction. Principals also perceived that using the Framework encouraged teachers to be more reflective about their instructional practices and more receptive to constructive criticism. For instance, one principal opined:

> When you're a good teacher or a great teacher or a distinguished teacher, a lot of the Framework you do, but you do not recognize that you are actually doing it. The thing I like about the Framework is it actually makes you cognizant of what behaviors constitute excellence in teaching, and then it holds you accountable for actually doing those behaviors.

[18] Note that percentages for principal attitudes do not sum to 100% due to rounding.

Overall, principals believed the Danielson Framework is intuitive as an evaluation tool because it identifies and delineates several aspects of exemplary teaching and instruction. Additionally, the fact that principals noted substantial changes across observations indicated that teachers addressed areas that needed improvement.

However, some principals had criticisms of the Danielson Framework as an evaluation tool. Some principals believed the components were not mutually exclusive and complicated the process of appropriately documenting evidence. A few principals also struggled with which level of performance best described teaching practice, specifically for those instances where a teacher fell somewhere in between "basic" and "proficient." For example, one principal stated:

> I would go in and try to write everything down just so I could remember what evidence I did see that could take a teacher from unsatisfactory to basic to proficient to distinguished. Then it kind of got muddied because I really could not put my finger on the evidence for teachers who were borderline. They are basic, but maybe there are things in the lesson that are proficient.

Virtually all interviewed principals also acknowledged that the implementation of the new evaluation system came with an increased time commitment. Most principals (N=36 or 92%) stated that using the Charlotte Danielson Framework and having meaningful, reflective conversations with teachers in the pre- and post-conferences took significantly more time than their previous evaluations using the CPS checklist system. "All of the pieces of the process take longer; the conferences, the observation, the write-up, DS2…it is much more labor-intensive," one principal stated, echoing the sentiments of her colleagues. The interpretation of this time commitment also varied across principals. The majority of principals "thought the increased time was worth it, because improving instruction is our most important role" (N=26 or 67%). A smaller number of principals were less positive, stating that "time constraints will make this system impossible with a whole school staff," and "big schools will never be able to make this work." Additional analyses are needed to determine the influence of school size on implementation and school staff perceptions of the evaluation system.

Teacher Attitudes

Teacher feedback on the Danielson Framework was overwhelmingly positive. Of the 26 teachers interviewed, none of them expressed mostly negative views of the Framework. Sixty-five percent or 17 of 26 teachers held mostly positive attitudes toward the Framework, while the remaining 35% (7 teachers) expressed mixed views. Teachers generally agreed that the Framework and its accompanying rubric provided them with concrete and specific standards on which to base their instructional practice. Furthermore, they felt that the Framework challenged teachers to reflect, and to self-evaluate. Negative teacher views regarding the use of the Framework were focused on challenges in implementation rather than on the tool itself. Pervasive attitudes can be summarized by one teacher's characterization of her general impressions of the Framework:

> I appreciate it. A lot. To be honest, my first year teaching, I received a pretty high [efficiency] rating even though I felt like my principal didn't really know what was going on in my classroom. If you want to raise teacher quality, I think you need to be assessed appropriately, so I think that this is a step in the right direction.

Furthermore, teachers were positive about the fact that levels of performance delineated in the Framework outlined attainable standards for improving instructional practice. Teachers also appreciated the personalized feedback received from administrators during observations. For example, one teacher stated:

> Definitely reflection. It's also structured. It has a form. You have to show what you're going to do. It makes you think about planning, as well as reflection. We all write up lesson plans, but God knows if anybody follows them. And it allows time to actually sit down with the principal and assistant principal related to your practice. You get to close the door, turn off the noise, and actually sit and talk, which is really, really nice.

In terms of perceived weaknesses of the Danielson Framework, some teachers thought the rubric was too subjective, the paperwork was redundant or confusing, and a typical lesson often does not encompass every component delineated in the Framework. Finally, like principals, teachers also pointed to time as a significant constraint. One teacher stated:

> Just because of demands of the system and working in a school, I feel like it's something that if there wasn't time to really follow-up, and kinda do self-study on and keep up with, that it would be difficult to have a large scale change. I feel like teachers could be evaluated pretty rigorously by it, but not really have the time to deal with the areas if they need improvement.

Perceptions of Conferences

Principals and teachers were asked a series of questions about the pre- and post-observation conferences. Each principal or teacher was asked if they participated in pre- and post-conferences, the format and length of these conferences, and their perceptions of the process. While both principals and teachers were very positive about the Framework, teachers were generally less positive about the conferences. However, the concerns were mostly about the time commitment required and implementation difficulties. (See Table 11 for an overview of principal and teacher perceptions about the conferences.)

Table 11. Principal and Teacher General Attitudes About the Conferences

	Mostly Positive	***Mixed***	***Mostly Negative***
Principal Attitudes (N=39)	67%	18%	15%
Teacher Attitudes (N=26)	46%	42%	12%

Principal Attitudes

Principals were sorted into three categories based on their conference attitudes: mostly positive, mixed, and mostly negative. Twenty-six principals were categorized as mostly positive (66%); 7 were categorized as mixed (19%); and 6 exhibited mostly negative attitudes about conferences (15%).

Principals' positive comments about their participation in conferences clustered around 5 main topics: 1) teachers were better prepared for conferences and lessons, 2) clearer expectations, 3) improved relationships, 4) more reflective discussion, and 5) improved focus on instruction.

The first two topics, better preparation and clearer expectations, focused primarily on pre-conferences, while the remaining three were discussed in relationship to both pre- and post-conferences. Roughly half of the principals (N=19 or 49%) suggested that the use of the pre-conference led to better preparation on the part of the teachers. "It made them plan. It made them think," one principal stated. "We talked together about the lesson, and she revised on the spot, making the planning process deeper and more reflective," another principal stated.

A little less than half of the principals (N=17 or 44%) stated that use of the pre-conference led to clearer expectations in the observations. This was perceived by principals to be a two-way street, in which teachers could communicate their intentions and clarify their approach, and principals could frame the purposes and foci of their observations. "In the pre-conference, the teacher and I dialogue about specific feedback, on what he or she is looking for from me and on what I am looking for from her," one principal explained. Another principal described how a pre-kindergarten teacher had explained her method of informal assessment of her students, the questions she would be asking them to assess their understanding. This information was critical to the principal's understanding of the lesson. "If I hadn't had the pre-conference with her and she had not told me in so much detail exactly what she was doing, I would not have known that," the principal explained. Pre-conferences may also help alleviate the concern that a single lesson may not touch on all aspects of the Framework.

When asked about the reasons for their positive perceptions of the Framework, about a third of principals expressed that participating in pre- and post-conferences using the Charlotte Danielson Framework led to improved relationships with their teachers (N=14 or 36%). "Using the Danielson [framework] to frame the conversation removed the subjectivity a bit," one principal stated. "It made our conversation more professional and I think it improved my relationships with my teachers," another principal contributed. One principal even discussed how use of the Framework in the conferences helped to improve a particularly contentious relationship she had with a teacher. The teacher was non-renewed as a PAT. "I think the post-conference improved the relationship between myself and this particular teacher," she stated. "We sat down and looked at my documentation and we went through it, and it was just clear, I am not going to be able to renew you. It wasn't a problem. He was even still personable!"

More than half of principals stated that the pre- and post-conferences using the Charlotte Danielson Framework led to more reflective discussion (N=23 or 59%). "Conversations were deepened because the Framework has explicit goals for improving instruction," one principal stated. About half of these principals made direct comparisons to the conversations with teachers in previous evaluations using the CPS checklist. One principal said about one of his teachers: "She didn't see the value of it last year, but this year…I don't know if we ever would have had that conversation before."

When asked about their positive perceptions of the Framework, about a quarter of principals thought that use of the Danielson Framework led to an increased focus on instruction in pre- and post-conferences (N=10 or 26%).[19] In particular, principals reflected on conferences they had conducted in the past, suggesting that the Danielson Framework changed the "content and tone" of the discussion. "The conversation is entirely different," one principal explained. "My conversation before was 'you were

[19] Note that we only asked this question explicitly in the second round of interviews.

tardy,' 'you didn't turn in your lesson plans,' all those kinds of things. Now I think this conversation is about good instruction."

Principals' negative comments about their use of pre- and post-conferences focused on three main topics: 1) time constraints, 2) need for training around conducting the conferences, and 3) negotiating the coach versus evaluator role.

Time constraints were a significant challenge in the use of pre- and post-conferences, reported by more than 85% of principals (N=34 or 87%). Even principals who were mostly positive about the conferencing component acknowledged it was a time-consuming process. "You start to add up the time for a staff, and this can be several hours of conferencing per person," one principal, who was mostly positive about conferences, stated. When asked if conferences, particularly the addition of the pre-conference was worth the additional time allotted, that principal responded: "It is absolutely worth it. The most important thing principals do is support instructional improvement." This viewpoint was echoed throughout by the principals who were mostly positive about conferences, despite their opinions concerning the length of the time commitment.

Principals in the mixed or mostly negative groups were less accepting of the increased time commitment that came from the use of the Danielson Framework and the addition of the pre-conference requirement. "Teachers are overwhelmed by the pre and post using this Framework," one principal explained. "It is too much time, too intense, and unnecessary." Another principal similarly stated, "Does the Board think this is all we have to do?" This negative view of the time commitment associated with the conferences represented 23% of interviewed principals (N=9).

An additional challenge with conferences identified by interviewed principals was their lack of certainty about how to frame the conversations and on a related note, the lack of training available to them about how to conduct the conferences. About one third of principals spoke about this challenge (N=14 or 36%). "I'm not sure if I'm asking the right questions to bring the teacher to that reflective state that we want them to be in," one principal stated. Another principal suggested that teachers did not necessarily know how to have the reflective conversation either: "Since I have a lot of new teachers, they're not sure how to do it. I'm not having that reflective conversation—I'm more leading, teaching, and directing." About half of the principals who talked about not knowing how to frame reflective conversations explicitly wished for training in this area. "The training…I don't think there's a lot of emphasis on the pre-observation piece of it or the post-observation piece," one principal stated.

A final challenge identified by principals was negotiating the roles of instructional coach versus evaluator. The use of the Danielson Framework was perceived by principals to change the nature of the conversation to focus more explicitly on instruction. While this change was positive, principals perceived that it put them in a "coaching" role, they perceived to be in conflict with their role as an evaluator. About one quarter of interviewed principals reflected on this challenge (N=9 or 23%). One principal explained: "I think in a situation like that, well, I am a coach and I am an evaluator. I think if I was not evaluating her she would have been more honest about the changes she would have made to make her lesson better rather than try to convince me." This was typical of remarks by principals who perceived this tension.

Teacher Attitudes

About 88% of teachers interviewed held mixed (N=11) or mostly positive (N=12) attitudes toward the conferences. The remaining 12% (N=3) expressed mostly negative views of the conferences. Negative opinions were focused on difficulties in implementation rather than the conferences per se.

Teachers were also asked to provide specific examples of the strengths and weaknesses of the conferencing process. Fifty-four percent of teachers stated that conferences provided a basis for the improvement of instruction and student learning. Many teachers also noted the conferences encouraged teachers to reflect on the quality and style of their instruction. The pre-conferences also helped to clarify what would be observed and helped quell anxiety about the evaluation process. While some teachers suggested that completing paperwork for the conferences was cumbersome, many teachers also found this aspect of the evaluation useful for lesson planning.

The negative sentiments expressed by teachers in regards to the Danielson Framework generally stemmed from logistical or implementation issues. Some teachers also felt conferences were not long enough to discuss all the pertinent details highlighted by the Framework during a classroom observation. One factor underlying teacher perceptions is illustrated by examining the frequency of pre- and post- observation conferences. As indicated in Table 12, only 69% of teachers (N=18) reported participating in pre-observation conferences before *each* of their formal observations, 19% of teachers reported having a pre-observation conference before *only one* of their observations, 12% report *never experiencing* a pre-observation conference. The same inconsistency is evident in the occurrence of post-observation conferences. In 73% of cases, teachers reported they participated in *two* post-observation conferences; the remaining 27% participated in the post-observation conference following some of their formal observations.

Table 12. Teachers' Reports of Conference Participation (N=26)

Conferences Occurred	***Always***	***Sometimes***	***Never***
Pre-conference	69%	19%	12%
Post-conference	73%	27%	0%

From our current data, it is difficult to assess whether the inconsistency lies with principal or teacher accounts on the frequency of the pre- and post-conferences. Additional data collection and analyses are needed to better understand the relationship between the implementation and format of pre- and post-conferences and teacher perceptions of their value.

Perceptions of Training

Principals and teachers were asked about their perceptions of the training they received for the evaluation pilot. As discussed in the overview section of the report above, principals received three days of summer training, four half-day professional development sessions, and participated in Area-based professional learning communities where they discussed the evaluation process with other principals. Teachers received two workshops for a total of 3.5 hours of training. Teacher workshops occurred at the beginning of the year and again in mid-to-late fall. The majority of principals' and teachers' opinions concerning the training they were received mostly positive. (See Table 13 for an overview of principal and teacher attitudes towards the training.)

Table 13. Principal and Teacher Attitudes Towards Training[20]

Training Type	*Mostly Positive*	*Mixed*	*Mostly Negative*
Principal Attitudes			
Summer	56%	36%	8%
Half Day	76%	21%	3%
PLC	65%	19%	16%
Teacher Attitudes			
Teacher Training	65%	25%	10%

Principal Attitudes

When prompted to discuss their level of preparedness for implementing the Danielson Framework at the start of the school year, most principals felt prepared after the initial training. Most principals remarked that the three-day summer training was helpful and commented favorably on the opportunities to observe video examples, discuss their ratings with colleagues, and practice identifying where evidence fits in to the different domains of the Danielson Framework. However, principals suggested the initial training placed more emphasis on gaining a conceptual understanding of the Danielson Framework while lacking a focus on the logistics of implementation such as filling out forms or documenting evidence. One principal stated:

> I think I understood the concept. I understood very well the purpose, the big picture. That was not a tough sell. But the logistics of what this question is trying to get at, what do you put in these little blanks…I do not think there was a strong understanding of those logistics coming out of the three-day training.

To some degree, principals' concerns around implementation were addressed during subsequent half-day trainings and meetings within their professional learning communities. These mechanisms for ongoing support provided a forum for principals to discuss the evaluation process with their colleagues and the methods they used to mitigate challenges encountered during the implementation phase. The half-day trainings on the Framework were beneficial in assisting principals

[20] We did not specifically ask about the summer training in the second round of principal interviews that took place in the spring, though some principals offered comments on the topic. Further, for some of the interviews, it was difficult to determine if the principals were discussing the half-day trainings or the PLCs. For these reasons, we were not able to categorize attitudes regarding training for 14 of the principals with summer training, 5 with the half-day training, and 7 with the PLCs. These responses are not included in the table. For the teachers, 23% did not attend or did not remember the Framework training.

with logistical matters such as how to assign the evidence to the appropriate component and how to enter information in to the DS2 interface[21]. One principal stated:

> I gained a stronger understanding of the logistics and what each question is trying to get at by attending subsequent trainings that clarified how you gather evidence or what you're supposed to put in a certain box. The second training with Charlotte Danielson that focused on process was also helpful.

Teacher Attitudes

Of the twenty-six teachers interviewed, 23% (N=6) either did not attend the training or did not remember enough to comment. Of those 20 teachers who could recall the training, 65% held mostly positive attitudes with respect to the training they received, 25% held mixed views, 10% held mostly negative views. Teachers felt the training helped to clarify what principals focused on during classroom observations. One teacher stated:

> We got to view one teacher being observed [the video example], and that was good, so that gave you a bigger, better perspective on what's going to actually happen. And then, they critiqued it afterward, according to the evaluation process, so we had an opportunity to critique it too.

In particular, teachers felt the training activities involving case studies and discussion of the literature on the Danielson Framework broadened their understanding of the new evaluation system. Teachers also reported that the process of looking at the Framework and the domains was also helpful to their understanding.

However, teachers also reported shortcomings of the training, that some of the activities concerning the domains were confusing, and that there were mixed messages about what was going to transpire in terms of evaluation. For instance, it was unclear to teachers why they were chosen to be observed using the Framework, what the processes were for pre- and post-conferences and why the external observer was in their room. "I thought I was chosen because I was a bad teacher," one PAT confessed. "I thought the external observer was from the Board, evaluating whether I should get non-renewed," stated another teacher.

Principal and Teacher Recommendations for Improvement

Principals and teachers were asked specifically about how to improve the evaluation system, and many respondents offered suggestions throughout the course of the interview.

Conferences. Most of the principal suggestions around conferences focused on teachers. Some principals suggested having teachers complete the Framework as a self-analysis prior to the post-observation conference. Principals also felt that teachers need training on completing the conference forms. Teachers also noted some redundancy in the conference forms and expressed that the forms should be clarified or eliminated from the evaluation process.

[21] DS2 is the online technology into which the principals entered Framework ratings for the teachers. Principals in the district use DS2 for a variety of purposes related to staff.

Training. Principal suggestions regarding training varied. Some principals wanted the summer training condensed, while others wanted it moved closer to the start of school. Principals asked for training to be differentiated based on principal experience with scripting and clinical observation. Principals also wanted more training for teachers to expand their knowledge of the Danielson Framework. Teachers concurred more training was necessary in order to understand the Framework and the entire evaluation process. Specifically, teachers requested that the training take place throughout the year as opposed to just in the beginning and that more of the training is centered on hands-on evaluation in which they can observe other teachers using the Framework.

Support in using the Framework. Overwhelmingly, principals asked for support in the form of a co-observer. Principals want to use the Framework correctly, and they think that having somebody go through an entire observation cycle (including the conferences) with them would be helpful. They want this person to be an "expert" in the Framework. One principal suggested that a point person who can provide support with the Framework process would have been helpful. Another principal had the idea of pairing Cohort 1 principals with Cohort 2 principals so advice on logistics can be shared with colleagues just beginning to implement the Danielson Framework in their schools.

Streamlining the process. Principals also want the data entry process to be streamlined although there was no consensus on how to accomplish this objective . Some principals wanted to eliminate components, while others said that every component is important. One popular suggestion was to enter evidence in a limited way—perhaps enter evidence only for unsatisfactory/basic ratings, or enter ratings but keep a paper copy of evidence. A few principals asked for the Framework in checklist form. Other principals suggested focusing one observation on Domain 2 components and the second observation on Domain 3 components. Another suggestion was to use the full Framework for the first observation, but then to target the second observation around those components where the teacher struggled most. Some teachers also made this suggestion so that the whole process could be more manageable. Since principals were not collecting evidence for Domains 1 and 4 in a systematic way—if at all—some wanted to eliminate those domains. Perhaps these domains could be evaluated every other year.

Communication from the district. Most of the suggestions related to district communication comprised general statements about clarifying various aspects of the evaluation process utilizing the Danielson Framework. However, principals specifically sought further explanation on the role of teachers who attended the summer training. While principals believed staff participation in the summer training was extremely beneficial, most principals expressed not knowing how to best integrate these teachers into the implementation process at the school level.

Evaluation in general. The principals also made comments about the overall structure of evaluation in the district. These comments are relevant regardless of the evaluation system used. First many principals felt the timeline for observations was restrictive, and some even thought it was unfair to teachers. Principals suggested moving the timeline for starting to begin later in the year so that teachers had an opportunity to successfully establish routines in their classrooms. Some principals did not think it made sense to start the second round of observations before the winter break. Principals also thought that replacing the four-level efficiency rating with a simple “meets” or “does not meet” standard would help alleviate some of the tension between being a coach and instructional leader and being an evaluator. A few principals argued for a staggered evaluation cycle for teachers in order to make rigorous evaluation more manageable to complete with their entire staff.

Implementation of the Framework process. Throughout their interviews, teachers offered general suggestions as to how the district can better implement this new evaluation process in the

future. Teachers suggested school-wide implementation of the evaluation process would have a greater impact as opposed to conducting the evaluation with only a select group of teachers. Teachers wanted principals to observe them more frequently and while principals echoed this sentiment, they felt overwhelmed managing multiple demands on their time. One principal suggested that CPS implement a 3-minute walkthrough system to accompany the Framework observations. Teachers thought the system should be an ongoing process or an ongoing conversation and that the evaluation results (particularly the common weaknesses) be used to structure the content of professional development. Perhaps implementing a short walkthrough system in conjunction with the formal Framework observations would help to make the process feel more continual and seamless. And, as long as the checklist system remains in place, principals thought teachers needed more support in understanding the differences between the Framework and the checklist and Framework ratings and efficiency ratings.

Principal Attitudes About Evaluation

The next two descriptive sections focus exclusively on principals, exploring principal attitudes toward evaluation and their perceptions of changes in instructional practice that resulted from use of the Danielson Framework. The use of the Framework accompanied by the pre- and post-conference format provided an evaluation process framed around 10 distinct components focused on the classroom environment and instructional practice, a continuum of descriptions of these components, as well as conference structures and forms to promote dialogue about them. What were principals' perceptions of this new evaluation system and what influenced their opinions?

A little more than half of principals were categorized as having mostly positive attitudes about the evaluation (N=22 or 56%). Mostly positive principals discussed three reasons for their enthusiasm for the evaluation approach: 1) paradigm shift, 2) already doing it, and 3) new approach. These were not mutually exclusive subgroups, but rather, principals who were categorized as mostly positive generally expressed one or more of these sentiments.

A small number of principals (N=6 or 28% of mostly positive principals) noticed a significant change in their evaluation paradigm through participation in the Excellence in Teaching Pilot. Conversations with these principals revealed an "aha" moment in which the principal explained that using the Charlotte Danielson Framework impacted perceptions about their teachers' performance. In some cases, these shifts in understanding focused on specific components within the Danielson Framework. For instance, one principal stated: "The quality of questions component, when I applied it, I was shocked. Here I thought this was something our teachers did really well and using the Framework was eye-opening to the fact that my teachers were quite basic and had a long way to go." Each of these principals, however, expressed a profound change in thinking about evaluation as a result of engaging in this process:

> One of the things I learned from this was how subjective I was in evaluation in the past…using the Danielson made me realized that my evaluation was colored by my opinion in the past…whether it was being turned off by the teacher's personality or their inappropriate dress or my perception of their teaching last year, I was influenced. I realized that in this process. That's a very honest statement.

These principals who experienced such a paradigm shift expressed a high level of enthusiasm for the new evaluation process, despite increased time commitments and other challenges such as DS2.

About half of the mostly positive principals expressed that their enthusiasm for the evaluation approach was a result of the fact that they were already doing a number of the types of things included in the pilot (N=11 or 48% of mostly positive principals). Several of these principals were already doing pre-conferences before the Excellence in Teaching pilot began, and so introduction of the new system did not significantly change the time commitment. In addition, the inclusion of the pre-conference in the pilot represented school system support for something the principals already had been doing. "It was nice to see that what I had already been doing was valued by CPS," one principal explained. Principals in this group were also scripting their observations, to a certain extent. "This use of evidence, I had already been doing that and so Danielson just gave a nicer frame in which to put what I was already doing with evidence," one principal stated.

About half of mostly positive principals were enthusiastic because this was "a new, fresh approach that provided a system that worked" (N=13 or 57%). Many principals who fit into the mostly positive category "did not realize how inadequate the old system was" until "something better was provided." Principals were enthusiastic about being provided with a more effective way to evaluate teachers.

Interestingly, principals who were of mixed or mostly negative shared this reason of "already doing it" with their mostly positive counterparts. Roughly a third of principals (N=11 or 27%) fit into the mixed category. A smaller portion of principals fit into the mostly negative category (N=7 or 16%). A group of these principals across the two categories shared the perception that they were "already doing" portions of the new evaluation system before the pilot was introduced. However, this was perceived negatively. "Basically whether its Charlotte Danielson or Robin hood, we've been doing it, so okay?" stated one principal. These principals did not think that the use of the Charlotte Danielson Framework had significantly changed the process of evaluation in their school. "Other than being a lot more time, I don't see any difference," one principal expressed, echoing the attitude of many of these principals. Ten of the sixteen principals in the mixed and mostly negative categories expressed this opinion.

Another emerging theme that bounded mixed and mostly negative principals together was a general belief about evaluation. Principals that fell into the mixed and mostly negative categories often expressed that they "just know" what good teaching is, without a tool or formal process for evaluation. For example, one principal stated: "I already know what my evaluations are going to be for my teachers. I knew that in September with exception of a couple of people -they're going to be the same as last year." Similarly, another principal stated: "It really doesn't take too long to figure out what is going on in a classroom, you know? I mean I can tell within 5 minutes who is a good teacher and who is a bad one. Then just filling in the Framework with that in mind, I guess. Principals just know." Finally, principals had a "just know" response when asked about rating specific components: "I need very little evidence to know if something is good or not." Nine of the sixteen principals in the mixed and mostly negative categories expressed this opinion.

Changes in Instructional Practice

Conversations with principals allowed for the exploration of the extent to which the Excellence in Teaching Pilot had resulted in changes in instructional practice. Principals reported six themes: 1) instructional grouping, 2) assessment, 3) planning, 4) general instructional practice, 5) compliance, and 6) no influence.

A little less than half of principals reported that engaging in the new evaluation approach led teachers to make changes in instructional grouping (N=17 or 43%). These principals thought that the use of the Danielson Framework and the conferences allowed them to push for differentiated instruction. For example, one principal explained:

> We did some discussions in the pre conference about how to group students, which she followed through with during the observation...I think I was really able to push her thinking about what do you do when you separate groups out and what different types of work those kids should be doing.

A similar proportion of principals reported that their conversations and observations resulted in improvements around assessment (N=15 or 38%). Some principals suggested that the use of the Framework pushed assessment because teachers were asked to bring artifacts to the conferences, while other principals thought the Framework itself helped them to push teachers' thinking about assessment. For example, one principal talked about asking the teachers to bring samples of student work to the post-conference:

> In the post observation we were getting a lot more samples of student work…I told them I want to see sample of a high achiever, middle achiever, low achiever and let's discuss what's going on there. I think the teachers many of them felt more comfortable with definitive idea of what those artifacts, what we're looking for. Okay, that's probably what the Danielson is probably going to do for us better, and that is bring us to the idea of evidence.

The Framework included an assessment component and principals talked candidly about the manner in which this tool drove conversations about assessment. For example, one principal stated:

> She had not quite flushed out how she was going to assess the lesson that she was doing. And through our conversation then she did, she developed a little a rating scale and then in the post conference brought that back to us.

Most principals agreed the new evaluation system improved instruction because it enhanced teacher planning (N=30 or 77%). "Teachers just came in more prepared, because of the pre-conference and because the Framework outlines so clearly what we are looking for," one principal explained. Another principal expressed this with a hint of sarcasm, "He was just so proud of himself and I wanted to say, just think how good you could be if you planned all your lessons this way rather than just when I am coming!"

A little more than half of principals saw general instructional improvement from their first to their second observation that they attributed to the use of the Danielson Framework in combination with the pre- and post-conference process (N=20 or 51%). Statements that fell into this category were filled with a sense of improvement that was less specified than in the categories above. For example, one principal stated: "Her rating were higher the second time, she was improved, just better organized and on-target with her approach."

A little less than one-fifth of principals focused on compliance. These principals valued the way in which the use of the pre-conference and the Framework encouraged teachers to make a plan and follow through. "She said she was doing x and she did y, and I had her pre-conference form to point it

out. It was there in writing, so she couldn't argue," one principal stated. "I loved the fact that they had to fill out the pre- and post-conference forms for me, it forced them to at least do that to prepare for our conversation," expressed another principal. Principals whose comments fit into this category saw benefits in teacher follow-through that came from participation in the new evaluation system.

Finally, a handful of principals did not perceive there were any influences of the new evaluation system on instructional practice (N=6 or 15%). This was related to principals' perception that they were "already doing" this type of evaluation, as described in the section on general evaluation attitudes. "Good teachers are good, bad are bad," one principal explained. "The evaluation forms or approach don't change that," she continued.

Assessing Principal Engagement: A Typology of Principals

Thus far, our analysis of principal interviews has provided descriptions of principal perceptions on a number of separate topics, such as the Framework, conferences or the training they received. While these descriptions are useful to understanding individual concepts, they do not provide an indication of the overall level of principal engagement in the Excellence in Teaching pilot. In a sense, we have summarized principal perceptions of each topic in isolation from the others, without consideration of how these opinions on separate topics relate to one another.

In this section, we search for patterns in the attitudes of principals across different themes. This approach allows us to provide an overall estimate of principal engagement across the pilot schools. About how many of the principals were highly engaged in the evaluation process and how many were not, and what are the perceptions of principals at different levels of engagement? The level of overall engagement provides an important indicator of the number of pilot principals highly committed to the evaluation approach and the perceptions and attitudes that are typical of principals engaged at a high, mixed or low level.

A Typology of Principals

Principals clustered into four "types": 1) Paradigm shift (PS), 2) High Enthusiasm (HE), 3) Mixed Emotions (ME), and 4) Low Enthusiasm (LE). Table 14 summarizes the typology components and the percentage of principals that fall into each type.

The themes that clustered together to formulate the typology included: 1) Framework attitudes, 2) conference attitudes, 3) evaluation attitudes, 4) description of teacher buy-in, and 5) description of changes in instructional practices.

Paradigm shift (PS) principals made up about 15% of the sample (N=6). These principals shared many of the same attitudes as High Enthusiasm principals, who were a little more than 40% of the interviewed sample (N=16 or 42%). PS and HE principals had in common their mostly positive attitudes about both the Framework and the conferences they undertook with teachers. Both the PS and HE principals tended to describe their teacher buy-in as high, and agreed that they saw changes in instructional practice between the two observations they attributed across observations. The PS and HE principals were most likely to report changes in instructional practice that included: instructional grouping, assessment, planning and, general instructional improvement.

Table 14. Typology of Principals Participating in the Excellence in Teaching Pilot

Type	Framework	Conferences	Evaluation Attitudes	Description of teacher buy-in	Changes in Practice	Severity	Percentage of 39 Interviewed Principals
Paradigm shift Principal	Mostly positive	Mostly positive	High buy in, paradigm shift	High	Influence on instructional grouping, assessment, planning, general instructional practice	High: 3 Average: 3 Low: 0	15%
High Enthusiasm Principal	Mostly positive	Mostly positive	High buy in, already doing High buy in, new approach	High	Influence on instructional grouping, assessment, planning, general instructional practice	High: 5 Average: 9 Low: 2	42%
Mixed Emotions Principal	Mixed to mostly negative	Mixed to mostly negative	Already doing Just know	Medium to High	Influence on compliance and planning	High: 1 Average: 6 Low: 4	28%
Low Enthusiasm Principal	Mostly negative	Mostly negative	Already doing Just know	Low to Medium	No influence	High: 3 Average: 3 Low: 0	15%

The contrast that sets the PS and HE principals apart is in their evaluation attitudes. The small group of PS principals reported a "paradigm shift" in their perception of evaluation that was a direct result of participating in the pilot study. This is described in more detail in the descriptive section on evaluation attitudes above. Paradigm shift principals talked about how they realized their evaluation had been subjective in the past or how they perceived their teachers were of higher quality in a certain area. The Danielson Framework with pre- and post-conference conversations led them to see evaluation, and their teachers' strengths and weaknesses, differently. PS principals also reported a significant "aha" moment as a result of their participation in the evaluation pilot.

High Enthusiasm (HE) principals also had generally positive attitudes about using the new evaluation system. Their reasons for their buy-in, however, differed. These principals fell into the categories outlined in the descriptive section above as "already doing it," or embracing the new approach as an improvement to the old system.

Mixed Emotions (ME) principals included a little less than a third of the sample (N=11 or 28%). The ME principals are characterized by mixed to mostly negative attitudes about both the Framework and the conferences. These principals, like their Low Enthusiasm principal counterparts, generally espouse that they were "already doing" the pieces of the new evaluation system and that the use of the Framework and the addition of the pre-conference did not significantly change what they were doing in terms of evaluation. As indicated in the descriptive section on evaluation attitudes above, ME and LE principals were more likely to suggest that they "just knew" if teachers were good or bad, whether with a checklist or the Charlotte Danielson Framework.

ME principals were interesting in the patterns of their positive and negative attitudes toward the new evaluation system. Their negativity tended to focus on the perception that this was an additional initiative, layered on top of countless existing programs and initiatives that were already in their schools, and that they did not have time for the labor intensive evaluation approach. Eight of the ten ME principals talked about the challenges of managing multiple initiatives and their resulting lack of time as the primary reason for their negativity.

Interestingly, while they were negative about the number of initiatives they were managing and the time requirements of the evaluation pilot, some ME principals did perceive that changes in instructional practice had taken place as a result of use of the evaluation system. However, while PS and HE principals pointed to changes in instructional grouping, assessment, general instructional practice and planning, ME principals were more likely to emphasize benefits in compliance and planning. ME principals were enthusiastic that the new system brought teachers better prepared to conferences and that teachers followed the plan they submitted but were less likely to focus on changes in instructional practice itself. Again, this dynamic marks an interesting contrast between ME principals and the higher buy-in categories (PS and HE) on the one hand and the LE principals on the other. ME principal comments in this regard imply a logic that improved instructional practice comes from better planning, accountability and sanctions. PS and HE principal descriptions implied that improved instructional practice came from deepening teacher understandings of the process of improvement, fostered by the continuum of skills described for each component within the Framework. As such, ME principals are a group who perceive the process of instructional leadership quite differently from the PS and HE principals. Additional inquiry is needed to better understand these contrasts.

Low Enthusiasm (LE) principals were mostly negative about the Framework and conferences, stated that they were already doing the type of evaluation in the new system or that they "just knew" teachers' abilities. This group was characterized by a perception of a lack of influence of the evaluation

system on instructional practice and described their teachers' buy-in as low to medium. LE principals were about 15% of principals in the sample (6 or 15%). LE principals placed teacher evaluation at the low end of priorities compared to their other responsibilities. "I would say my first job is manager of the building, evaluating staff comes further down the list," one LE principal explained. "Hiring the right people makes evaluation a moot point," another LE principal stated. The LE principals generally did not see potential connections between the new evaluation system and the school improvement process. "How does this system impact school improvement? I guess I don't see that connection. It is just a requirement from the district and the union," one principal explained.

Several things are notable about the patterns that are evident in the typology. First, it is interesting that description of teacher buy-in and descriptions of changes in instructional practice resulting from participation in the new evaluation process seem to be related to attitudes about the Framework, conferences, and the evaluation. It is not possible to completely understand this pattern. On the one hand, it could be that positive attitudes about the Framework, conferences and the evaluation leads to improved implementation of the evaluation process, which increases teacher buy-in and positively influences deeper changes in instructional practice. On the other hand, principals who are positive about the Framework, conferences and the evaluation could perceive higher levels of teacher buy-in and changes in instructional practice.

Secondly, it is hopeful to note that the majority of principals fall in the PS and HE categories. This suggests that overall, participating principals are enthusiastic about the Framework, the conferences and the evaluation and perceive that it can lead to positive changes in teacher practice.

A final pattern emerges when we consider how these principals cluster by severity. The multi-facet Rasch analysis undertaken in the preceding section allowed us to classify principals as high, average or low severity. High severity principals were two standard errors in severity above the most severe external observer, while low severity principals were two standard errors in severity below the least severe external observer. All the remaining principals were average in terms of severity. In order to explore the relationship between the typology and the severity rating, we analyzed the number of high, average and low severity principals in each type. The results of this analysis appear in column seven of the typology table (Table 14, below).

About half of the principals in each type are average in terms of severity, meaning that the average raters are distributed somewhat evenly across the types. The high severity principals also exist in each of the four types. However, a greater proportion of principals in the PS and LE categories are high severity, where half of the principals in each type are high. Of the HE principals, a slightly smaller proportion is high severity (5 or 31%). The ME principals have the lowest proportion of high severity principals, with only 1. At the same time, the ME category has the majority of low severity principals with 4.

Several interesting patterns emerge from this analysis. First, it is notable that the high severity raters are distributed somewhat evenly across three of the four categories of principals (PS, HE, LE). It would be possible to argue that high severity principals would be more likely to be either PS or LE principals. As it turns out, the high severity principals appear in both types and the LE group has the same proportion of high, average and low severity participants as the PS type. Second, no low severity principals appear in either the PS or LE types. These two points suggest that while the principals in the PS and LE groups are divergent in important ways, the severity of ratings is not a differentiating factor. Third, the low severity principals are clustered in only two types: HE and ME. In particular, the ME type has only 1 high severity principal and 4 of the low severity principals. One possible explanation is that

ME principal uncertainty about the evaluation pilot led some of them to rate very easily. We can only speculate about these patterns at this time. Additional analyses are needed to better understand these patterns.

Implications

In this section, we explore the implications of the evidence and analyses in this report. In particular, we address areas of focus for CPS as the district moves into the second year of the Excellence in Teaching pilot. The section includes implications for 1) training, 2) implementation, 3) establishing a criteria for promotion and renewal, and 4) the second year of evaluation.

Training

Principal and teachers were very positive about the benefits of the training they received in 2008-09, and were hopeful that they would receive additional training in the coming year. Analysis of ratings data and descriptive and analytical treatment of the principal and teacher interview data identify several clear pathways for sculpting that continuing professional development to improve the implementation of the evaluation system.

Observation notes and Framework evidence. Principal interview data reveal two dimensions of preparedness that are necessary for successful use of the Framework. The first focuses on successful note taking during a classroom observation and translating these notes into evidence for ratings on components. Interviews revealed that some principals struggled with the evidence collection and interpretation process while others (who were more experienced with such analysis from previous professional or educational experience) did not. Further, the analysis of ratings data suggests that principals could benefit from training around the content of certain components. For example, principals may not only need training around placing evidence for component 3d, but they could also use training on the topic of using assessment in instruction. Targeted training in this area would improve use of the Danielson Framework.

Promoting reflective conversations. Principals expressed the need for training in how to lead the pre- and post-conference process. Principals were concerned that they were being too directive in the process and did not always know how to deepen reflections. Teachers similarly expressed uncertainty about the process and purpose of conferences. Targeted professional development on conferencing for principals and teachers would promote more reflective conversations by clarifying the process and expectations for both principals and teachers. Principals could also use support in relationship building, especially when it comes to balancing the demands of coach and evaluator.

Promoting understanding of Domains 1 and 4. Principals and teachers expressed the need for training to deepen understandings of Domains 1 and 4. Our analysis revealed that 33% of principals did not have a systematic process in place to collect evidence for these domains. Some principals ignored the two domains in the first year while those who collected evidence expressed confusion and uncertainty about whether their approach was appropriate. Further, clearer expectations around the use of Domains 1 and 4 would help principals (e.g., how often are teachers supposed to be rated using these domains, how and when are principals supposed to provide feedback to teachers regarding these domains).

Deepening teacher understandings of the Framework and the pilot. Principals and teachers both expressed the need for more training for teachers on the Framework and on the purpose, goals, and details of the Excellence in Teaching pilot. While principals and teachers were generally positive about the teacher training that was provided in the first year, they were critical of the fact that it focused almost exclusively on the Framework. Expanded teacher training that is ongoing and focused on the entire evaluation process (including pre-conference, post-conference, conference forms, etc.) is needed. Teachers should also be better informed about how the Framework and checklist systems relate to each other. They need help processing the Framework ratings in light of prior and current checklist efficiency ratings.

Digging deeper into problematic components. Our analyses have identified several components that had systematic rating differences between principals and external observers. Targeted training on each of these components is needed for principals and external observers to sharpen understandings of these components to all who are observing classrooms. In particular, Framework users need help discerning between the basic and proficient levels of performance, especially for components 2e, 3a, 3c, and 3d.

Training that provides examples of levels of performance. Principals and teachers both indicated that the video observations were extremely useful in training. The participants in the pilot would benefit from a video library that contains lessons or excerpts of instructional practice that illustrate "distinguished" or "proficient" practice on a given component. Principals and teachers expressed a desire to visualize levels of performance and thought that a more extensive video library would be very helpful for future training. In addition, having examples of distinguished practice would help principals negotiate conversations with teachers who had received high efficiency ratings in the past.

Managing the evaluation process. Principals were virtually unanimous in their statement that the new evaluation system took more time than the CPS checklist approach. However, some principals found ways to manage the time requirements. Finding ways to share successful management strategies between principals would improve implementation. This might be accomplished by having principals with successful management strategies lead professional development for other pilot principals, or through targeted discussion at PLC or half-day professional development sessions. Another possibility is to pair Cohort 1 and Cohort 2 schools such that principals who are moving into their second year using the Danielson Framework could mentor and support principals in their first year.

Implementation

This report identifies a number of positive results from the first year of implementation of the evaluation system. The majority of principals reported sticking to CPS timelines for completing observations and entered their ratings into DS2. Overall indications of engagement represented in the typology of principals reveal that the majority of principals are highly engaged in the process. The findings from this report also provide possible avenues to improve the implementation of the evaluation system. The following themes are salient.

Challenges of a significant shift in evaluation processes. The Excellence in Teaching pilot marks a significant shift in the approach, assumptions, and goals of teacher evaluation in CPS. The analysis of ratings identifies some consistent differences in the ratings of principals and external observers. Some of this variation is explained by the teachers' previous evaluation rating. This is confirmed by principals, who commented that they sometimes had to rate teachers higher than was warranted because the

teacher had been rated Excellent or Superior on the previous evaluation system. This shift to the new evaluation system will be difficult because of the higher expectations of teaching in the new system. The fact that the checklist system and the Framework system were used simultaneously in 2008-09 made this disparity even more clear and more difficult for principals to navigate. The successful implementation of the new evaluation system is dependent upon a massive paradigm shift in CPS.

Reconsidering the PAT evaluation timeline. Principals were asked if they were able to comply with the required timeline to evaluate their PATs. While many principals did comply, they thought the timeline required observations to occur too early in the school year. To have second round observations starting in November, principals were conducting first round observations very early in the school year. Principals felt that this condensed observation schedule did not allow teachers adequate time to reflect and improve. Revisiting this timeline in light of the goals of the evaluation and a focus on instructional improvement is needed.

Moving to a staggered schedule. Principals expressed reservations about the possibility of being able to evaluate all of their teachers in a school in a single year using the rigorous, and therefore time-consuming, approach represented in the Danielson Framework combined with the pre- and post-observation conference. Principals suggested moving to a staggered, rotating schedule where tenured teachers would be observed every third year, with one third of the staff being formally evaluated with the Danielson approach each year. Considering this kind of revised schedule may be a necessity to ensure that the evaluation approach is conducted thoroughly, appropriately, and productively.

Establishing an Appropriate Criteria for Promotion and Renewal

This report identifies a number of positive indications about use of the Framework for evaluation purposes. Principals and teachers are generally positive about the design of the evaluation and of the use of the Charlotte Danielson Framework. A large proportion of principals and teachers confidently expressed that the Framework has deepened their professional conversations about instruction. The majority of principals participating in the pilot are characterized as either "paradigm shift" or "high enthusiasm principals" on the typology. The findings also indicate areas of concern, particularly in using the Framework for summative purposes given difficulties in inter-rater reliability and problematic components. Several recommendations follow from those findings.

Using a Meets Standards/Does Not Meet Standards scale. It is important to have a fair system that principals and teachers can trust. Ratings and principal and teacher interview data support the use of a meets/does not meet scale rather than the efficiency rating. Our analyses have provided some indication of the numbers of teachers in the pilot sample that would be identified as low-performing using various criteria. Establishing realistic criteria for teacher non-renewal is an essential component of the success of the evaluation system. Some possible approaches might include:

Weight some of the components. Components that are of particular importance to the district might have more value in a summative evaluation system.

Temporarily down-weight problematic components. Problematic components would receive less weight until principals can receive targeted training around those components.

Use the criteria to identify teachers for professional development. Identifying teachers for supports and professional development based on their performance could represent an important step

in creating a successful evaluation system. This concept is similar to the Tenured Teacher Voluntary Assistance Plan that CPS has considered using as part of the Excellence in Teaching pilot. (See Figure 5).

The Second Year of Evaluation

The first year of evaluation was successful in several respects. About 90% of principal ratings were entered into DS2. The external observers were able to complete their observations and submit ratings and evidence in a timely and thorough manner. CCSR researchers met and communicated regularly with CPS staff members about ongoing findings and emergent issues. This was done formally, through written updates and meetings, and informally through ongoing communication. CCSR staff members were also present at several professional development trainings for principals and presented findings. For instance, principals were given reports that compared their ratings with those provided by the external observer in their classrooms. This was positively perceived by principals. The report findings also lead to some implications for the second year of evaluation and research. The following summarizes those implications.

Digging deeper into problematic components. While principal and teacher training on clarifying problematic components is essential, additional research is needed to better understand the reasons certain components are challenging. CCSR researchers will systematically analyze evidence data from the first year of observations and will integrate questions about the challenging components into interview protocols for 2009-2010. Completing a document analysis of evidence submitted by principals and evidence submitted by observers will help us to understand if principals and observers are noting the same evidence for each component. We will also be able to understand how much of the evidence is fact-based rather than opinion.

Exploring leadership distribution in evaluation. The varied extent to which principals utilized their assistant principals to evaluate teachers warrants further exploration. Out of the pilot sample, the principal conducted all evaluation observations in 14 schools. In 11 schools, the assistant principal and principal conducted some or all observations simultaneously, and in 12 schools they split the responsibilities between a principal and AP.[22] We do not yet have an understanding of how this distribution of leadership affects the implementation of the evaluation system. Additional analysis of year one data will be conducted to better understand this, and new analyses will be undertaken, both quantitative and qualitative, to broaden our understanding of these dynamics.

Investigating further the level of Framework implementation. Our typology of principals suggests that the level of implementation of the Excellence in Teaching pilot varies from school to school. To understand implementation issues, we will be using a case study approach to guide our qualitative field work in Year 2. Case studies will allow us to take a deeper look at schools that are implementing the process at different levels. We will be able to learn more about how well a school implements the Framework with other issues at the school like school climate, professional culture, and relationships in the building.

Continuing investigation of Framework reliability. Though the focus of the Year 2 quantitative work will be looking to see if Framework ratings relate to student outcomes (i.e., that the component ratings measure valid concepts) and, we will continue the reliability study on a limited scale.

[22] In two schools, the manner in which observations were distributed was unclear or missing from data.

References

Brandt, R. (1996). On a new direction for teacher evaluation. *Educational Leadership, 53,* 6, 30-33.

Bryk, A. S., and Raudenbush, S.W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage Publications, Inc.

Chicago Public Schools (2007). The new teacher project: Hiring, assignment and transfer in Chicago Public Schools. Chicago: Author.

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

Darling-Hammond, L. (1996). What matters most: A competent teacher for every child. *Phi Delta Kappan, 78*, 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.

Gitlin, A., & Smyth, J. (1990). Toward educative forms of teacher evaluation. *Educational Theory, 40,* 1, 83-94.

Haefele, D.L. (1993). Evaluating teachers: A call for change. *Journal of Personnel Evaluation in Education, 7,* 21-31.

Heneman, H.G. III, and Milanowski, A.T. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17,* 3, 171-195.

Hightower, A.M. (Ed.). (2002). *School districts and instructional renewal.* New York: Teachers College Press.

Interstate New Teacher Assessment and Support Consortium (1992). *Model standards for beginning teacher licensing and development.* Washington, D.C.: Council of Chief State School Officers.

Latham G., & Wexley, K. (1982). Increasing productivity through performance appraisal. Monterey, CA: Brooks Cole.

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago: MESA Press.

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

Peterson, K. (1990). Assistance and assessment for beginning teachers. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation.* Newbury Park, CA: Sage.

Sclan, E.M. (1994). *Performance evaluation for experienced teachers: An overview of state policies.* Washington, D.C.: ERIC Clearinghouse on Teaching and Teacher Education.

Searfross, L., & Enz, B.J. (1996). Can teacher evaluation reflect holistic instruction? *Educational Leadership, 53,* 6, 38-41.

Smith, G. (2003). Reflections on teacher evaluation: A cause for concern. *Phi Delta Kappan, 103,* 21-30.

Van Sciver, J. (1990). Teacher dismissals. *Phi Delta Kappan, 72,* 318-319.

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

U.S. Department of Education (2009). Education Department: American Recovery and Reinvestment Act of 2009. (http://www.ed.gov/policy/gen/leg/recovery/index.html).

Appendix A: Data and Methods

Quantitative Data and Methods

Sample Selection

This study benefits from a two-level stratified selection plan. At the first level, schools were randomly selected for participation in the Excellence in Teaching pilot in the 2008-09 school year. The pilot was implemented in four elementary CPS Areas (2, 8, 13, and 16), and school selection took place at the Area level. Prior to randomization into the pilot, schools with first-year principals and Fresh Start schools were removed from the sample. Then, half of the remaining schools in each of the Areas were randomly selected to implement the Danielson Framework as a teacher evaluation tool. Table A1 shows the number of schools in the treatment and control group by CPS Area.

Table A1. Number of Schools in the Treatment and Control Group by CPS Area

CPS Area	***Treatment Schools (N=43)***	***Control Schools (N=50)***
2	16	18
8	9	11
13	8	7
16	10	14

Note. One of the treatment schools refused to participate in the pilot and is not included in the treatment or control group. Also two of the treatment schools are out-of-Area AMPS, but for the purposes of our analysis these two schools are grouped with the Area where they are physically located.

At the school level, teachers were randomly selected from teachers in the pilot school who were eligible for formal evaluation in the 2008-09 school year. All non-tenured teachers (PATs who are in their first three years in the classroom) and tenured teachers with a low 2007-08 efficiency rating. This random selection took place in mid-August 2009. Because some teachers reached their anniversary date between the time of random selection and the beginning of the school year, we have a small number of tenured teachers in our sample who received high efficiency ratings in 2007-08. These are generally fourth year teachers. Table A2 shows the breakdown of the teacher sample by CPS Area and tenure status.

Table A2. Number of Teachers in the Treatment Group by CPS Area and Tenure Status

CPS Area	***All Teachers (N=155)***	***PATs (N=113)***	***Tenured Satisfactory (N=27)***	***Tenured Excellent/Superior (N=15)***
2	68	57	9	2
8	28	18	6	4
13	20	13	5	2
16	39	25	7	7

Note. Random selection of PATs and low-rated tenured teachers took place separately.

Data Collection

Principals and external observers collected classroom observation data using the Danielson Framework for Teaching (as adapted by CPS). The Framework reliability study hinges on collecting two sets of Framework ratings from two independent observers—the principal and the external observer. Both parties go into the classroom simultaneously, observe a lesson (usually 30-45 minutes), and align their evidence from the observation with the Framework to assign a level of performance for 10 components. Principals and external observers do not discuss the lesson and assign rating independently.

The external observers are three practitioners who are on loan to central office from their schools for the purposes of this study. The three external observers have specialized knowledge: one is a special education teacher with National Board certification, another is bilingual, and the third has both teaching and administrative experience. The external observers received intensive training around using the Danielson Framework, including an initial three-day training, follow-up support focused on specific components, and practice observations in actual classrooms. All of the training for external observers was in a small group setting.

As discussed in the report, the principals also received in-depth training around using the Framework, starting with a three-day summer training around the Framework content and using the Framework in a classroom observation as well as the logistics of implementing the new system. Follow-up support was provided to principals in four half-day large group trainings and regular Area-based professional learning communities. The content of the PLCs was driven by artifacts from their use of the Framework that principals brought to the sessions.

By the CPS-CTU contract, teachers must be formally observed two times per year. Formal classroom observations can begin during the third week of student attendance. For PATs, the two observations must be conducted before the evaluation conference in which the teachers are given their summative efficiency rating, which occurs before the end of the first full week of instruction in March. The deadline for tenured teachers is closer to the end of the school year (and varies depending on the school's calendar). Another contract requirement is that the observing administrator hold a conference with the teacher within 10 days of the observation. The pre-observation conference that is part of the Excellence in Teaching pilot is not a contract requirement. Table A3 shows when the classroom observations started and ended.

Table A3. Timeline for Observations

Observation Type	***Start Date***	***End Date***
Contract Requirement (PATs)	September 15, 2008	March 6, 2009
Contract Requirement (Tenured)	September 15, 2008	June 5, 2009
Study Round 1	September 18, 2008	December 15, 2008
Study Round 2	November 17, 2008	February 26, 2009

Note. One Round 2 observation occurred in April 2009.

Framework ratings for each of the 310 observations (155 teachers observed twice) were submitted by the external observers. However, we only received Framework ratings for 277 of the observations from administrators. Therefore, we are missing 10.6% of the observation data. Another complicating factor is that in some cases principals and/or external observers reported not having enough evidence for a component to award a rating for that observation. Table A4 shows how many

missing ratings we had by rater type and by Framework component for the 277 observations with data from administrators and observers.

A4. Missing Framework Ratings (N=277 observations)

Component	*Missing Overall*	*Missing Administrator*	*Missing Observer*
2a	3	3	0
2b	12	10	2
2c	10	7	3
2d	9	8	1
2e	17	17	0
3a	9	7	2
3b	38	6	32
3c	6	6	0
3d	41	31	10
3e	48	31	17

Two things are important to note about the missing ratings. First, the principal is usually more likely that the external observer to leave a component blank. Second, there is a lot of variation in the amount of missing ratings across components. Components 3d (Using Assessment in Instruction) and 3e (Demonstrating Flexibility and Responsiveness) had a high rate of missing data, especially among administrators. Component 3b (Using Questioning and Discussion Techniques) was missing more often (and at a high rate) among external observers.

The Tool: Charlotte Danielson's Framework for Teaching

The rubric that principals and observers used to gather classroom observation data is the Danielson Framework for Teaching. The Framework consists of four domains of teaching:

1. Planning and Preparation
2. The Classroom Environment
3. Instruction
4. Professional Responsibilities

While principals collected evidence and gave ratings for domains 1 and 4 of the Framework, only Domains 2 and 3 are observable in a classroom observation. Therefore, our reliability study has only focused on Domains 2 and 3. These domains each consist of five components, which we have referred to throughout this report.

For each component, the district provided a four-level rubric for principals and observers. These four levels of performance are unsatisfactory, basic, proficient, and distinguished. Using the evidence collected during the observation and the component descriptors, raters were expected to choose a level of performance for each component. Appendix D includes the complete Danielson Framework as adapted by CPS.

Many-Facet Rasch Measurement (MFRM) Analysis

A major component of this study is to determine if the Danielson Framework can be used reliably. Rather than using a simple Rasch model, which would not take into account the fact that there are many different raters or judges, we applied the MFRM method. MRFM extends the Rasch model to include additional facets. The facets included in our analysis are teacher, Framework component, rater (includes the three external observers and each individual principal), observation round, subject area, and CPS area. The MFRM model shows the probability that a teacher will get a particular rating (unsatisfactory, basic, proficient, or distinguished) taking into consideration these categories, or facets, including rater severity. The model also provides us with a measure of rater severity for each of the observers and principals. What results, then, is a measure of teacher ability controlling for rater severity.

Figure A1 shows the results of the MFRM analysis. The first column shows the linear measure in logit scale. It provides a common comparison for the other columns. The column entitled Teacher Ability shows the underlying latent teaching ability as measured by the principal and observer Framework ratings. Each dot represents one teacher, and the asterisks represent two teachers. The teachers at the top of the logit scale are those with the highest latent ability (the "best" teacher is at the top of the scale at 6), and those at the bottom of the logit scale have the lowest latent teaching ability (near -2). Using the Framework ratings, then, teaching ability in our sample has an 8-logit spread.

The Rater Severity column includes both observers and principals. Each asterisk represents two raters. As with Teacher Ability, the raters are spread across the logit scale in terms of their severity—those at the top are most lenient, and those at the bottom are the most severe. They are spread on the logit scale from 3 to -4.

The next column, Component Difficulty, shows the order of the ten Framework components—those at the upper end of the scale are most difficult for teachers (i.e., components where teachers receive the lowest ratings), and those at the lower end of the scale are the components that are easiest for teachers. Unlike underlying teaching ability, the components do not spread as widely. Though, as discussed in the report, the MFRM analysis shows that each component measures a unique teaching concept. Similarly, the Round column indicates that Round 1 ratings are "harder" than Round 2 ratings, which indicates that ratings were higher in Round 2 than in Round 1. However, there is even less spread on the logit scale with the Round variable than with the Component variable. The Subject list the subject areas with the highest ratings on the top and the lowest ratings on the bottom. Again, the spread is not large, indicating that there is not much difference in ratings by subject area.

Finally, the Framework Ratings column shows where each of the levels of performance falls on the logit scale. There is considerable spread between each of the Framework levels.

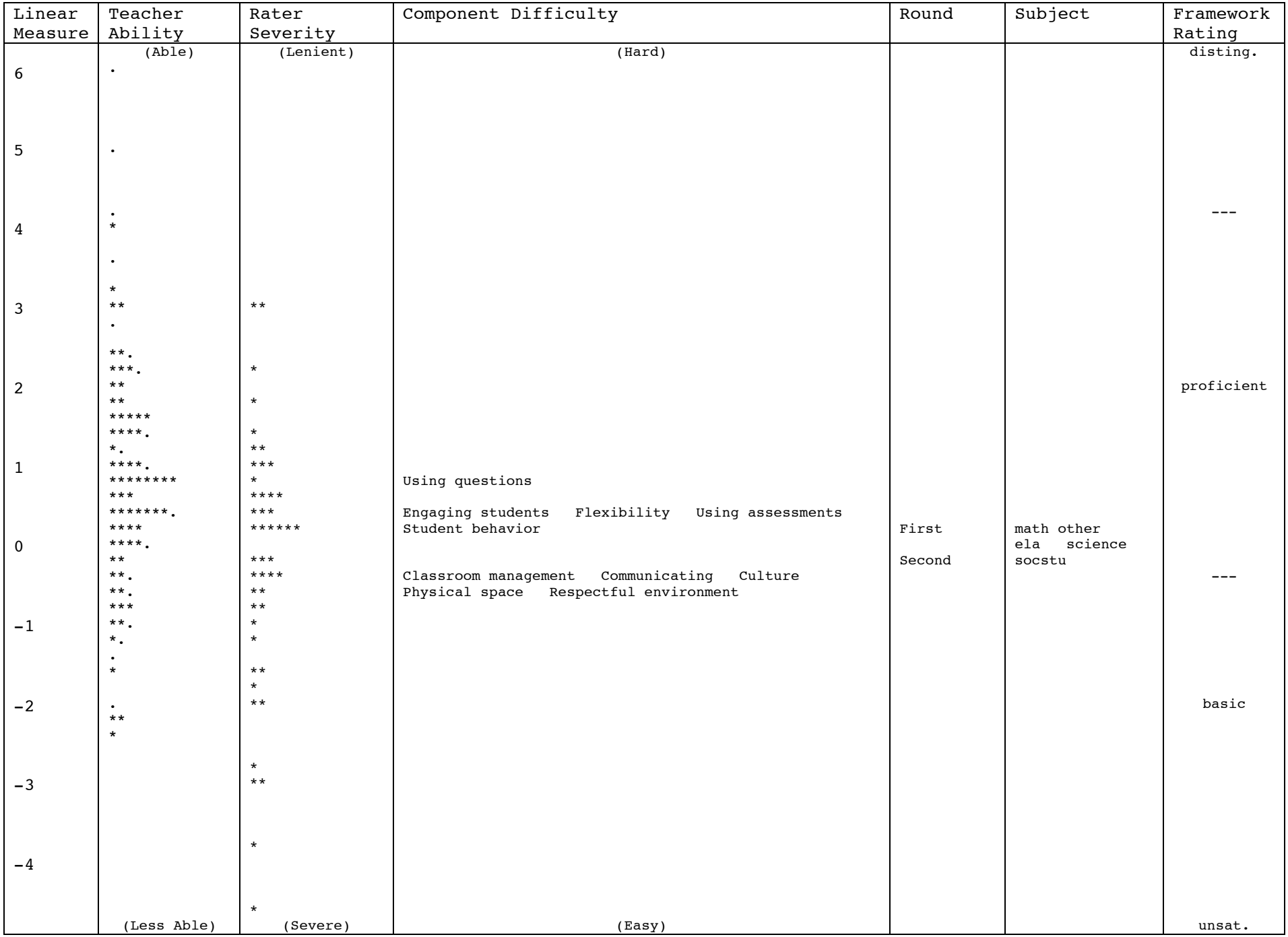


Figure A1. Multi-facet Rasch measurement results

Hierarchical Modeling

For our analyses, we used two-level hierarchical logit models, with information about the rating at Level 1 and information about the teacher, principal, and lesson at Level 2. All three of our models used binary outcomes with a binomial sampling model with a logit link function. The binary outcome variable was based on the rating.

- Model 1 compares the likelihood of getting an unsatisfactory rating to getting a basic/proficient/distinguished rating. This model focuses on rater effects at the low end of the ratings scale. The binary outcome variable equaled 1 if the rating was basic/proficient/distinguished.
- Model 2 compares the likelihood of getting a proficient/distinguished rating to getting an unsatisfactory/basic rating. This model focuses on rater effects in the middle of the ratings scale, which is where most of the ratings are. The binary outcome variable equaled 1 if the rating was proficient or distinguished.
- Model 3 compares the likelihood of getting an unsatisfactory/basic/proficient rating to getting a distinguished rating. This model focuses on rater effects at the high end of the ratings scale. The binary outcome variable equaled 1 if the rating was distinguished.

The purpose of using these three models was to investigate the possibility that principals and external observers were using the Framework ratings inconsistently. We do this in two ways: a) to check for main rater effects and b) to check for component-level rater effects. Models 1-3, then, were modified to capture both of these effects.

Main rater effects (Models 1a, 2a, and 3a). The equation used to determine if there was an overall, or main, rater effect was the following:

Level 1

Prob (Outcome = 1) = φ_{ij}

Log [φ_{ij} / (1 - φ_{ij})] = η_{ij}

η_{ij} = β_{1j} (Component 2a) + β_{2j} (Component 2b) + β_{3j} (Component 2c) + β_{4j} (Component 2d) + β_{5j} (Component 2e) + β_{6j} (Component 3a) + β_{7j} (Component 3b) + β_{8j} (Component 3c) + β_{9j} (Component 3d) + β_{10j} (Component 3e) + β_{11j} (Observation Round 1) + β_{12j} (Administrator) + ε_{ij}

Level 2

β_{pk} = γ_{p0}, for p=1 to 11

β_{12} = γ_{120} + γ_{12k} (vector of teacher characteristics)

The vector of teacher characteristics at Level 2—for both the main effects and the component-level effects—could include 2008-09 teacher efficiency rating, tenure status, CPS Area (2, 8, 13, and 16), special education, subject area, grade level, proxy for student achievement at a school, and whether or not a principal is on first contract.

Component-level rater effects (Models 1b, 2b, and 3b). The equation used to determine if there were component-level rater effects was the following:

Level 1

Prob (Outcome = 1) = φ_{ij}

Log [φ_{ij} / (1 - φ_{ij})] = η_{ij}

η_{ij} = β_{1j} (Component 2a) + β_{2j} (Component 2b) + β_{3j} (Component 2c) + β_{4j} (Component 2d) + β_{5j} (Component 2e) + β_{6j} (Component 3a) + β_{7j} (Component 3b) + β_{8j} (Component 3c) + β_{9j} (Component 3d) + β_{10j} (Component 3e) + β_{11j} (Observation Round 1) + β_{12j} (Component 2a * administrator) + β_{13j} (Component 2b * administrator) + β_{14j} (Component 2c * administrator) + β_{15j} (Component 2d * administrator) + β_{16j} (Component 2e * administrator) + β_{17j} (Component 3a * administrator) + β_{18j} (Component 3b * administrator) + β_{19j} (Component 3c * administrator) + β_{20j} (Component 3d * administrator) + β_{21j} (Component 3e * administrator) + ε_{ij}

Level 2

β_{pk} = γ_{p0}, for p=1 to 11

β_{qk} = γ_{q0} + γ_{qk} (vector of teacher characteristics), for q=12 to 20

The difference between these models and the models used to identify main rater effects is that there is an interaction between component and administrator for each of the ten components. These dummy variables equal 1 when the rating (outcome) is given for a specific component by an administrator (as opposed to an external observer). At Level 2, we are able to try to explain any significant Level 1 interactions with the same teacher characteristics as in the previous models.

The Level 1 intercept is suppressed in all of the models. Doing this allows us to compare the component coefficients and component-rater interaction coefficients absolutely, rather than to an arbitrary excluded component. In all models, the component variables are uncentered, while the other variables are grand mean centered.

The output from all of the models used in this report are in Appendix B.

Qualitative Data and Methods

Data consist of 39 principal interviews and 25 teacher interviews. Principals and teachers were interviewed using a semi-structured interview protocol and were asked questions about: (a) the professional development they and their teachers received; (b) their perceptions of the Framework; (c) their implementation of the evaluation system; (d) the pre- and post-conferences they had with their teachers and (e) their perception of school change that had resulted (or could result) from implementing the Framework.

Interviews were transcribed verbatim and codes were generated using a combination inductive and deductive approach. Deductively, a set of initial codes were created to mirror the semi-structured interview protocol. Multiple researchers used these draft themes to code the same interviews. This was undertaken both to test inter-coder reliability and to inductively generate additional codes for themes that emerged in the data that were not captured by the draft codes. The team of researchers compared the coded text and identified and clarified areas of disagreement in the coding of the shared interviews. The inductively generated codes were integrated into the final coding scheme. Coding was undertaken using Atlas ti, a qualitative analysis software.

Each of the transcribed principal and teacher interviews were coded using the tested coding scheme. Summary reports were run on each code such that all quotations assigned to each code were put together in a report. From these reports, descriptive summaries were created for each code. These descriptive summaries were combined and integrated where cross-code themes emerged. The purpose of these analytical summaries was to provide rich descriptions of principal and teacher insights, quantifying and grouping respondents wherever possible.

A second step was taken in order to better understand the clustering of attitudes evident in the coded text. For each code, principals and teachers were grouped into subsets. For instance, a grouping of “mostly positive”, “mixed”, “mostly negative” was used to summarize principals’ attitudes toward the Framework. These subgroups were determined using the data, identifying subgroups that adequately represented the natural conceptual clustering within the data. These subgroup codes were then entered into a summary matrix for each principal and teacher to look for themes and patterns across codes.

An additional set of analyses were performed to “typologize” principals who participated in the pilot evaluation. Using the summary matrix for each principal, principals were grouped based on their responses on several questions that asked about the extent to which they implemented the evaluation system and their perceptions of and attitudes toward evaluation of teachers in general. In particular, the typology process aimed to provide leaders in the CPS district about the extent to which principals in the evaluation pilot were engaging in the process, and the sophistication of these leaders to conduct teacher evaluation at a deep level.

A second round of coding was undertaken to identify additional themes that emerged within broader coding areas. This subset coding was applied to portions of the transcription that focused on the Framework, conferences, and implementation. The primary purpose of this subset coding was to expand our knowledge of important themes for the year 2 data collection and analysis plan for 2009-2010.

A final set of textual analyses were undertaken to explore the evidence data provided by principals and external observers. The primary purpose of this analysis was comparative. Evidence from the observation of the same teacher was analyzed for the principal and the external observer. This

analysis focused on components that were identified as problematic, such as 3d (Assessment) and 2e (Physical Space). By comparing two observers' description of the same instructional practice, we hope to be able to better understand the reasons for the systematic differences in rating. At the time this report was written, these textual analyses were in the preliminary stages. This will be an important focus for the ongoing and year 2 analysis plan.

Appendix B: Hierarchical Modeling Output

Models 1a-3a: The Main Rater Effect

Table B1 shows the outcome for the models that estimate the main rater effect. The outcome variable for the three models is the following:

- Model 1a: outcome equals 1 if the rating is basic or above; equals 0 if the rating is unsatisfactory
- Model 2a: outcome equals 1 if the rating is proficient or distinguished; equals 0 if the rating is basic or lower
- Model 3a: outcome equals 1 if the rating is distinguished; equals 0 if the rating is proficient or lower

All of the tables show the logit coefficient and the standard error for variables used in the six HLM models in this paper. To interpret the logit coefficient, a positive value indicates that the variable increases the likelihood of achieving the outcome, while a negative value indicates a decreased chance of attaining the outcome.

In Table B1, for example, the focus of the model is on the indicator variable Administrator. When this variable equals 1, that means the rating was given by a school administrator; when it equals 0, an external observer gave the rating. For Models 1 and 2—those looking at the likelihood of being above unsatisfactory and being above basic—the main rater effect is not significantly different from 0, which means that overall raters are using the Framework consistently. However, for Model 3—that looking at the likelihood of being distinguished—there is a highly significant positive rater effect. This means that at the high end of the Framework scale, principals are more likely to award distinguished ratings than observers. We can partially explain this variation with a number of the Level 2 covariates—though the difference is not eliminated entirely.

As has emerged in this report, a teacher's prior efficiency rating contributes to the overall rater effect, and we believe based on our principal interviews that some of this can be explained by using the Framework simultaneously with the checklist system. Principals reported often giving a teacher (who under the checklist system had a high efficiency rating) a higher Framework rating in order to maintain a good relationship with that teacher. What happens then, and what is supported by Model 3a, is that principals will give distinguished ratings when observers give proficient ratings in order to maintain the status quo.

Table B1. Hierarchical logit model output for Models 1a-3a, the main rater effect (N=5,659 ratings)

Variable	*Model 1a logit coefficient (s.e.)*	*Model 2a logit coefficient (s.e.)*	*Model 3a logit coefficient (s.e.)*
	Level 1		
Component			
Component 2a	3.77***	1.28***	-3.04***
Creating an Environment of Respect and Rapport	(.322)	(.124)	(.255)
Component 2b	4.61***	1.05***	-.353***
Establishing a Culture for Learning	(.454)	(.124)	(.273)
Component 2c	4.10***	.874***	-2.93***
Managing Classroom Procedures	(.342)	(.119)	(.251)
Component 2d	3.75***	.530***	-3.36***
Managing Student Behavior	(.347)	(.115)	(.241)
Component 2e	4.45***	1.36***	-3.15***
Organizing Physical Space	(.464)	(.131)	(.265)
Component 3a	3.92***	.980***	-2.92***
Communicating With Students	(.384)	(.118)	(.286)
Component 3b	3.40***	.186***	-3.96***
Using Questioning and Discussion Techniques	(.287)	(.112)	(.327)
Component 3c	3.69***	.251*	-3.60***
Engaging Students in Learning	(.331)	(.106)	(.292)
Component 3d	3.94***	.418**	-4.83***
Using Assessment in Instruction	(.254)	(.110)	(.348)
Component 3e	3.42***	.470***	-3.57***
Demonstrating Flexibility and Responsiveness	(.254)	(.111)	(.287)
Observation Round			
Observation Round 1	-.645***	-.224**	-.031
	(.169)	(.082)	(.862)
School Administrator Indicator			
Administrator	-.340	-.165	2.05***
	(.304)	(.118)	(.261)
	Level 2		
Covariates used to predict the main rater effect (denoted by the Level 1 Administrator variable)			
2007-08 Efficiency Rating			
2007-08 efficiency rating: Excellent	--	--	1.64***
			(.510)
2007-08 efficiency rating: Superior	--	--	1.75***
			(.581)
2007-08 efficiency rating: Missing	--	--	1.84**
			(.737)
CPS Area			
CPS Area: 16	--	--	-2.45***
			(.603)
Subject Area			
Subject area: Math	--	--	-1.01**
			(.512)
Tenure Status			
Tenure status: PAT1	--	--	-1.16*
			(.636)
Grade Level			

Variable	*Model 1a logit coefficient (s.e.)*	*Model 2a logit coefficient (s.e.)*	*Model 3a logit coefficient (s.e.)*
Grade level: 6-8	--	--	-1.88*** (.586)
Principal Contract			
Principal on first contract	--	--	-.731* (.426)
Student Achievement			
Student achievement: next to bottom quartile	--	--	-2.31*** (.660)
Student achievement: bottom quartile	--	--	-1.32* (.747)

Note. Asterisks indicate a significant effect, *** p<.01, ** p<.05, *p<.10. For brevity's sake, Level 2 covariates are only shown if the effect was significant. The excluded group for each group of variables is 2007-08 efficiency rating: Satisfactory, CPS Area: 2, subject area: ELA, tenure status: tenured, grade level: 3-5, and student achievement: top quartile.

Models 1b-3b: Component-Level Rater Effects

Table B2 shows the outcome for the models that estimate component-level rater effects. The outcome variable for the three models is the following:

- Model 1b: outcome equals 1 if the rating is basic or above; equals 0 if the rating is unsatisfactory
- Model 2b: outcome equals 1 if the rating is proficient or distinguished; equals 0 if the rating is basic or lower
- Model 3b: outcome equals 1 if the rating is distinguished; equals 0 if the rating is proficient or lower

Table B2 can be interpreted in the same way as Table B1. However, it shows the rater effects at the component level in each model instead of overall. The component 2a-rater interaction, for example, equals 1 if the outcome rating was for 2a and given by a school administrator. If the logit coefficient is positive and significant, it means that the administrator was significantly more likely to give a higher rating than the observer. If the coefficient is negative and significant, the administrator was significantly more likely to give a lower rating than the observer. The significant component-rater effects were discussed thoroughly in the HLM section of the report.

Table B2. Level 1 hierarchical logit model output for Models 1b-3b, component-level rater effects (N=5,659 ratings)

Variable	*Model 1b logit coefficient (s.e.)*	*Model 2b logit coefficient (s.e.)*	*Model 3b logit coefficient (s.e.)*
Component-Rater Interactions			
Component 2a * Principal	1.23*	-.040	2.16***
	(.701)	(.211)	(.423)
Component 2b * Principal	--	-.339*	2.03***
		(.184)	(.450)
Component 2c * Principal	-.118	-.165	1.82***
	(.703)	(.171)	(.388)
Component 2d * Principal	-1.45**	.188	2.78***
	(.606)	(.160)	(.598)
Component 2e * Principal	-1.08	-.935***	1.89***
	(1.144)	(.214)	(.503)
Component 3a * Principal	.235	-.479***	1.59***
	(.449)	(.174)	(.330)
Component 3b * Principal	.332	-.149	2.49***
	(.506)	(.171)	(.690)
Component 3c * Principal	.024	.377**	2.98***
	(.498)	(.169)	(.648)
Component 3d * Principal	-2.52**	-.398**	--
	(1.068)	(.177)	
Component 3e * Principal	-.389	-.039	1.28***
	(.485)	(.189)	(.407)

Note. Asterisks indicate a significant effect, *** p<.01, ** p<.05, *p<.10. Indicator variables for the 10 components were included in the models but are not included in this table. Component 2b was omitted from Model 1b because the observers did not give any unsatisfactory ratings on this component. Component 3d was omitted from Model 3b because the observers did not give any distinguished ratings on this component.

Tables B3 and B4 show the Level 2 covariates for the significant component-level rater effects from Models 1b and 2b. The interpretation of the coefficients is the same as for the previous tables. The findings are discussed thoroughly in the HLM section of this report.

Table B3. Level 2 for Model 1b—Explaining significant component-level rater effects (N=5,659 ratings)

Variable	***Component 2d * Principal logit coefficient (s.e.)***	***Component 3d * Principal logit coefficient (s.e.)***
Intercept	-.394	-1.70
	(.672)	(1.203)
2007-08 Efficiency Rating		
2007-08 efficiency rating: None	2.68	4.46
	(4.973)	(4.97)
2007-08 efficiency rating: Excellent	1.18	1.42**
	(.731)	(.602)
2007-08 efficiency rating: Superior	2.48***	2.98***
	(.933)	(.616)
Subject Area		
Subject area: Math	.731	--
	(1.635)	
Subject area: Social studies	2.19*	--
	(1.210)	
Subject area: Science	-1.14	--
	(1.014)	
Subject area: Other	-2.28**	--
	(.900)	
Subject area: Different subjects observed in Round 1 and Round 2	-1.090	--
	(.736)	
Principal Contract		
Principal on first contract	-1.89**	--
	(.883)	
Teacher Designated as Special Education		
Special education teacher	--	2.47**
		(.989)
Student Achievement		
Student achievement: next to top quartile	--	-2.58***
		(.739)
Student achievement: next to bottom quartile	--	-.538
		(.631)
Student achievement: bottom quartile	--	-1.60*
		(.907)

Note. Asterisks indicate a significant effect, *** p<.01, ** p<.05, *p<.10. The component-rater interaction was significant for 2a, but there were so few instances where a teacher received an unsatisfactory that we could not model the rater effect at Level 2 due to lack of variation, so it is not included in this table. The excluded group for each group of variables is 2007-08 efficiency rating: Satisfactory, subject area: ELA, and student achievement: top quartile.

Table B4. Level 2 for Model 2b—Explaining significant component-level rater effects (N=5,659 ratings)

Variable	*Component 2b * Principal logit coefficient (s.e.)*	*Component 2e * Principal logit coefficient (s.e.)*	*Component 3c * Principal logit coefficient (s.e.)*
Intercept	-.290	-.659	.378**
	(.193)	(.478)	(.166)
2007-08 Efficiency Rating			
2007-08 efficiency rating: None	--	7.77	.777
		(14.168)	(.749)
2007-08 efficiency rating: Excellent	--	.700*	.614**
		(.387)	(.261)
2007-08 efficiency rating: Superior	--	.924**	.678**
		(.450)	(.323)
CPS Area			
CPS Area: 8	.243	--	--
	(.319)		
CPS Area: 13	1.17**	--	--
	(.577)		
CPS Area: 16	-.109	--	--
	(.318)		
Tenure Status			
Tenure status: PAT1	--	.592	--
		(.605)	
Tenure status: PAT2	--	.814**	--
		(.391)	
Tenure status: PAT3	--	.679	--
		(.500)	
Principal Contract			
Principal on first contract	--	.397	--
		(.322)	
Student Achievement			
Student achievement: next to top quartile	--	-.05	--
		(.365)	
Student achievement: next to bottom quartile	--	1.26**	--
		(.529)	
Student achievement: bottom quartile	--	.661	--
		(.426)	

Note. Asterisks indicate a significant effect, *** p<.01, ** p<.05, *p<.10. While the component-rater interaction for 2b is significant, it is important to remember that this is an area where the external observers were retrained during the data collection period. Components 3a and 3d also had significant rater effects, but none of the covariates at Level 2 explained any of the variation, so they are not included in this table. The excluded group for each group of variables is 2007-08 efficiency rating: Satisfactory, CPS Area: 2, tenure status: tenured, and student achievement: top quartile.

Appendix C: Interview Protocols

Protocol
Administrator Interview Protocol
Excellence in Teaching Pilot Study

Introduction Script
Hi, my name is _____, and I'm from the Consortium on Chicago School Research at the University of Chicago. We are interested in understanding more about how the teacher evaluation system based on Charlotte Danielson's Framework for Teaching is being implemented in your school, as well as your perceptions of the framework and the conferences with your teachers. Your outlook as administrator at a pilot school is an important piece of this work, so thanks for taking the time to talk with me.
Some of the things we will discuss will ask you to reflect on your role as an evaluator of teachers and as an instructional leader, as well as on your teachers and the instruction and learning that goes on in classrooms at this school. I want to emphasize that this is a confidential interview. Your name and the name of your school will not be revealed to anyone outside of this research project. I also will not share anything that you said with anybody besides other researchers on this project. So, if an Area Instructional Officer or other district personnel asks what you said about something, I will not tell him/her. Likewise, I cannot share information that other principals or teachers provided to me. At the end of this project, we'll give a summary report of what we heard from principals and teachers without naming anyone. The only exception is if I have reason to believe that a student is being harmed or will be harmed, in which case I am obligated to take action. Your participation in this study is important, but you may request to stop at any time and without any negative repercussions. Please also feel free to ask me questions concerning this study at the time.

I would like to tape record this interview to keep track of information accurately. Is that okay with you? Also know that you can ask to turn off the recorder at any time.

Do you have any questions for me before we begin?

[TURN ON TAPE RECORDER NOW.]

This is Principal Interview Project ID # _____. You've indicated that it's okay for me to tape this interview. May we now begin?

Principal Background/School Info

DO NOT PROBE IN THIS SECTION
To start off, I'd like to learn more about your background and the school.

- How long have you been a principal in CPS? At this school?
- Have you been a principal anywhere else?
- What did you do prior to being a school administrator?
- Tell me a little bit about your student body.
- Describe your teaching staff. [E.g., experience, turnover.]

Pieces of the Pilot Evaluation System

Now I will ask you some questions pertaining to the different pieces of the pilot evaluation system, including questions about training, pre- and post-observation conferences, and the framework itself.

Training

T1. Have you been able to attend the ½-day trainings for principals? These are the trainings for all pilot principals that took place in October, November, and February, including the Charlotte Danielson presentation. *[Note: APs may not have attended any of these trainings.]*

- If so → How has this additional training been useful to you?
- If so → What would you change about these trainings?

T2. Other support that the district has in place includes monthly area-based professional learning communities. *[Note: APs do not attend the PLCs.]*

- Have you been able to attend these trainings?
  - If so → How has this additional training been useful to you?
  - If so → What would you change about these trainings?

T3. Looking back over the year, what suggestions do you have about the overall structure of the support provided to principals who are implementing the framework in their schools? This includes the initial summer training as well as the follow-up support.

Framework

[Provide administrator with a copy of the pre-observation conference form, the framework, and the post-observation form that he/she can reference during this portion of the interview.] I know you are familiar with these documents, but here are copies of the pre- and post-observation forms as well as the framework that you may want to refer to during the following questions.

F0. With about how many teachers have you gone through the framework observation process?

***F1. What are your impressions of the framework?

F2. Describe your experience using the framework in conjunction with the classroom observation.

- Did the observation provide enough information to rate the teacher in Domains 2 and 3?
- ***What components were the most difficult for you to rate? Explain why.
- Do you feel comfortable using the framework in an observation?
  - If so: About how many observations did it take before you felt comfortable?
  - If not: Describe what makes using the framework difficult for you.

F3. How have you gathered evidence for Domains 1 and 4 (i.e., the domains not focused on classroom observation)?

F4. Do the component ratings accurately reflect the teacher's level of performance? Explain.

***F5. Are there any specific components that you felt were especially important measures of teacher performance when evaluating your teachers? Why? *[Note: If principal mentions an entire Domain, prompt for specific components within that Domain.]*

F6. Is it more difficult to use the Framework with certain grade levels? Subject areas? What are the specific challenges?

Conferences

***C1. One of the goals of the new system is to facilitate professional conversations between principals and teachers. Did the pre- or post-observation conferences help to achieve this goal?

- If so → can you provide a specific example of a conversation that you had with a teacher that you feel helped to further that teacher's professional growth?
- If not → describe any barriers to professional conversations between you and your teachers.

C2. Prior to the pilot, principals were also required to conduct post-observation conferences with teachers. Can you discuss any differences in the quality of the conferences using the new framework compared to using the checklist?

C3. What did you like about the district-provided conference forms? What changes, if any, would you make?

C4. Did the first observation and conferences with your teachers lead to changes in their practice? Can you give examples of changed instructional practice from your second observations?

Implementation Fidelity

Now I will ask you some questions about the implementation of the pilot teacher evaluation system.

Implementation of the system

I1. I would like to learn about the different roles that you and your staff played in implementing the pilot.

- What role do you have in implementation of the teacher evaluation system?
- What role does your AP have in implementation of the teacher evaluation system? *[Prompt if necessary: Did your AP attend training with you?]*
- How did you determine the observation schedule this year? *[Prompt if necessary: How did you determine who to observe and when to observe them?]*
- One of your teachers attended the three-day framework training with you, correct?
    - If yes → can you describe the role of that teacher has had in the implementation of the teacher evaluation system?

I2. Have you followed the suggested CPS timeline for when to observe teachers?

[PATs: observation 1 between September 15 and November 7; observation 2 between November 13 and February 20. Tenured teachers: observation 1 between September 15 and January 23; observation 2 between January 26 and May 15.]

- If so → did this timeframe make sense for you and your teachers?
- If not → why not? What changes did you make?

***I3. Is time management an issue for you in implementing the system?

- If so → describe some of the time barriers that you face.
- If not → describe some of the strategies you use in managing your time.
- Because of this evaluation system, are you not doing things that you were doing before? *[Prompt: Are you giving up something in order to complete the observations?]*

I4. I want to ask you about some of the specific support provided to you this year.

- Describe the support provided to you by:
  - Central office staff.
  - Your AIO.
- As you look ahead to next year and using the framework process with your entire staff, what additional support will you need?

***I5. How would you describe teacher buy-in of the framework?

- What evidence do you have to support this? Can you think of specific comments teachers have made?
- Do you notice any differences in the way tenured teachers think about the framework as compared to probationary teachers?
- Did you attend the framework training provided to your teachers? If so, what are your impressions of that training?

I6. March 6th was the PAT nonrenewal deadline. Describe how you decided what efficiency ratings to give your PATs.

School Change

Now I will ask you some questions related to the district's broader goals of the evaluation system.

***SC1. The pilot teacher evaluation system has the potential to influence broad change in the district. Can you discuss how you think the new teacher evaluation system has influenced, or will influence, the following things at your school? *[Interviewer should go through each of the following points one at a time. Ask principals to provide specific examples where applicable.]*

- Professional development for teachers
- Professional culture
- Teacher hiring
- The quality of teaching
- Student learning

Closing

We've covered all of my questions. Is there anything that else you'd like to tell me about the teacher evaluation pilot before we end?

Teacher Interview Protocol
Excellence in Teaching Pilot Study
5.1.09

Introduction Script
Hi, my name is _____, and I'm from the Consortium on Chicago School Research at the University of Chicago. We are interested in understanding more about how the teacher evaluation system based on Charlotte Danielson's Framework for Teaching is being implemented in your school, as well as your perceptions of the framework and the conferences with your teachers. Your outlook as teacher at a pilot school is an important piece of this work, so thanks for taking the time to talk with me.

Some of the things we will discuss will ask you to reflect on teacher evaluation in your school, as well as professional conversations with administrators and other teachers and the teaching and learning that occurs in your classroom. I want to emphasize that this is a confidential discussion. Your name and the name of your school will not be revealed to anyone outside of this research project. I also will not share anything that you said with anybody besides other researchers on this project. So, if your principal or other district personnel asks what you've said about something, I will not tell him/her. Likewise, I cannot share information that your principal or other teachers have already provided to me. At the end of this project, we'll give a summary report of what we heard from principals and teachers without naming anyone. The only exception is if I have reason to believe that a student is being harmed or will be harmed, in which case I am obligated to take action.

I would like to tape record this interview to keep track of information accurately. Is that okay with you? Also know that you can ask to turn off the recorder at any time.

Do you have any questions for me before we begin?

[TURN ON TAPE RECORDER NOW.]

This is Teacher Interview Project ID # _____. You've indicated that it's okay for me to tape this interview. May we now begin?

Teacher background
DO NOT PROBE IN THIS SECTION

To start off, I'd like to learn more about your teaching background and your responsibilities this year.

- How many years have you been teaching in CPS? At this school?
  (If they say they've been at the same school, ask: With the same principal?)
- Have you taught anywhere else?
- What classes/grade level do you teach this year?

Features of the observation cycle
I have some questions for you about the process of being observed by your school administrator. Let's start with a broad question:

O1. About how many times have you been formally and informally observed this year by your principal or AP?

O2. How many of those observations used the Charlotte Danielson's Framework for Teaching? Were they considered formal observations?

O3. Before the observation, it was suggested that teachers participate in pre-observation conferences with administrators?

- How many pre-observation conferences did you participate in?
- Who conducted them?
- Describe that experience, how was the conference structured?
- Were you asked to bring any materials to the pre-conference?
- About how long did the pre-conference last?
- Was it helpful to you? Why or why not?

O4. After the observation, it was suggested that principals hold post-observation conferences.

- How many post-observation conferences did you participate in?
- Who conducted them?
- How soon after the observation did the post-observation conference occur?
- About how long did the post-conference last?
- Were you asked to bring any materials to the post-conference?
- Describe that experience, was the conference structured?
- What did you take away from that specific conference? *[Prompt: End results could include setting of professional goals, alignment of professional development resources around the evaluation, etc.]*

General Impressions

F1. What are your general impressions of the Charlotte Danielson Framework?
What do you like most about the framework? What do you like least?
What do you consider the protocol's strengths and weaknesses?

F2. Do you feel like the levels of performance described in the framework accurately capture your performance as a teacher? Why or why not?

F3. Has being engaged in this kind of observation cycle process with the Danielson Framework resulted in any changes in your practice?

- If so, provide specific examples of the changes.
- If not, describe any barriers to change resulting from the evaluation process in your school.

Training activities
I'd like to ask you a few questions about the training you received *at your school* for the pilot teacher evaluation system. These initial session(s) took place in the fall.

T1. Were you present at that training?
What are some of the key points that you took away from the training?

Would it be useful to have this kind of training again next fall?
If so, do you recommend any modifications?

(If it does not come up, ask: What additional district based training on the pilot evaluations system do you feel would benefit you?)

T2. Have you received additional training or support from your principal/school on this process? If yes, what kind of supports were received?

Pilot Program Goals: Impact on Overall School Climate
We are interested in learning more about how the new teacher evaluation pilot aligns with some of CPS's larger goals.

S1. DO you feel there have been any school-wide changes in instruction due to the new evaluation system? Provide specific examples.

S2. Has the new teacher evaluation system had an effect on the professional environment at your school?
Have you had any discussions with fellow teachers about the framework? Please describe these discussions.
Has the Framework influenced the nature or quality of your conversations about instruction with the principal?
If so, describe how conversations have changed.
Has the framework affected the content of professional development in your school?

S3. Describe the climate surrounding teacher evaluation at this school.
How would you describe teacher buy-in in general in your school for using the framework?
What evidence do you have to support this?
Can you think of specific comments teachers have made?

S4. Do you notice any differences in the way teachers at your school think about the framework? Clarify if necessary: Do you notice any differences in the way tenured teachers think about the framework as compared to probationary teachers?

Closing
Thanks for your comments and your time. I've gone through all of my questions, but before we end I would like to ask if there is anything else you think I should know in order to understand your perceptions of the pilot program?

[TURN OFF TAPE RECORDER NOW]

Appendix D: CPS FRAMEWORK FOR TEACHING

Domain 1: Planning and Preparation

Component	Unsatisfactory	Basic	Proficient	Distinguished
1a: Demonstrating knowledge of content and pedagogy	Teacher's plans and practice display little knowledge of the content, prerequisite relationships between different aspects of the content, or of the instructional practices specific to that discipline.	Teacher's plans and practice reflect some awareness of the important concepts in the discipline, prerequisite relations between them and of the instructional practices specific to that discipline.	Teacher's plans and practice reflect solid knowledge of the content, prerequisite relations between important concepts and of the instructional practices specific to that discipline.	Teacher's plans and practice reflect extensive knowledge of the content and of the structure of the discipline. Teacher actively builds on knowledge of prerequisites and misconceptions when describing instruction or seeking causes for student misunderstanding.
1b: Demonstrating knowledge of students	Teacher demonstrates little or no knowledge of or respect for students' backgrounds, cultures, skills, language proficiency, interests, and special needs, and does not seek such understanding.	Teacher indicates the importance of understanding and respecting students' backgrounds, cultures, skills, language proficiency, interests, and special needs, and attains this knowledge for the class as a whole.	Teacher actively shows respect for and seeks knowledge of students' backgrounds, cultures, skills, language proficiency, interests, and special needs, and attains this knowledge for groups of students.	Teacher actively shows respect for and seeks knowledge of students' backgrounds, cultures, skills, language proficiency, interests, and special needs from a variety of sources, and attains this knowledge for individual students.
1c: Setting instructional outcomes	Instructional outcomes are unsuitable for students, represent trivial or low-level learning, or are stated only as activities. They do not permit viable methods of assessment.	Instructional outcomes are of moderate rigor and are suitable for some students, but consist of a combination of activities and goals, some of which permit viable methods of assessment. They reflect more than one type of learning, but teacher makes no attempt at coordination or integration.	Instructional outcomes are stated as goals reflecting high-level learning and curriculum standards. They are suitable for most students in the class, represent different types of learning, and are capable of assessment. The outcomes reflect opportunities for coordination.	Instructional outcomes are stated as goals that can be assessed, reflecting rigorous learning and curriculum standards. They represent different types of content, offer opportunities for both coordination and integration, and take account of the needs of individual students.
1d: Demonstrating knowledge of resources	Teacher demonstrates little or no familiarity with resources, including appropriate technology to enhance own knowledge, to use in teaching, or for students who need them. Teacher does not seek such knowledge.	Teacher demonstrates some familiarity with resources, including appropriate technology available through the school or district to enhance own knowledge, to use in teaching, or for students who need them. Teacher does not seek to extend such knowledge.	Teacher is aware of the resources, including appropriate technology available through the school or district to enhance own knowledge, to use in teaching, or for students who need them.	Teacher seeks out resources, including appropriate technology in and beyond the school or district in professional organizations, on the Internet, and in the community to enhance own knowledge, to use in teaching, and for students who need them.
1e: Designing coherent instruction *Use of appropriate of data*	The series of learning experiences are poorly aligned with the instructional outcomes and do not represent a coherent structure. They are suitable for only some students.	The series of learning experiences demonstrates partial alignment with instructional outcomes, some of which are likely to engage students in significant learning. The lesson or unit has a recognizable structure and reflects partial knowledge of students and resources.	Teacher coordinates knowledge of content, of students, and of resources, to design a series of learning experiences aligned to instructional outcomes and suitable to groups of students. The lesson or unit has a clear structure and is likely to engage students in significant learning.	Teacher coordinates knowledge of content, of students, and of resources, to design a series of learning experiences aligned to instructional outcomes, differentiated where appropriate to make them suitable to all students and likely to engage them in significant learning. The lesson or unit's structure is clear and allows for different pathways according to student needs.
1f: Designing student assessment	Teacher's approach to assessing student learning contains no clear criteria or standards, lacks congruence with the instructional goals, or is inappropriate to many students. Teacher has no plans to use assessment results in designing future instruction	Teacher's plan for student assessment is partially aligned with the instructional goals, without clear criteria, and inappropriate for at least some students. Teacher plans to use assessment results to plan for future instruction for the class as a whole.	Teacher's plan for student assessment is aligned with the instructional goals, using clear criteria, is appropriate to the needs of students. Teacher uses assessment results to plan for future instruction for groups of students	Teacher's plan for student assessment is fully aligned with the instructional goals, with clear criteria and standards that show evidence of student participation in their development. Assessment methodologies may have been adapted for individuals, and the teacher uses assessment results to plan future instruction for individual students.

NOTE: The CPS Framework for Teaching has been adapted, with permission, from Charlotte Danielson's *Framework for Teaching*.

CPS FRAMEWORK FOR TEACHING

Domain 2: The Classroom Environment

Component	Unsatisfactory	Basic	Proficient	Distinguished
2a: Creating an environment of respect and rapport	Classroom interactions, both between the teacher and students and among students, are negative, inappropriate, or insensitive to students' cultural backgrounds, and characterized by sarcasm, put-downs, or conflict.	Classroom interactions, both between the teacher and students and among students, are generally appropriate and free from conflict but may be characterized by occasional displays of insensitivity or lack of responsiveness to cultural or developmental differences among students.	Classroom interactions, between teacher and students and among students are polite and respectful, reflecting general warmth and caring, and are appropriate to the cultural and developmental differences among groups of students.	Classroom interactions among the teacher and individual students are highly respectful, reflecting genuine warmth and caring and sensitivity to students' cultures and levels of development. Students themselves ensure high levels of civility among members of the class.
2b: Establishing a culture for learning	The classroom environment conveys a negative culture for learning, characterized by low teacher commitment to the subject, low expectations for student achievement, little respect for or knowledge of student's diverse cultures and little or no student pride in work.	Teacher's attempt to create a culture for learning are partially successful, with little teacher commitment to the subject, modest expectations for student achievement, some respect for or knowledge of student's diverse cultures and little student pride in work.	The classroom culture is characterized by high expectations for most students, genuine commitment to the subject by both teacher and students, respect for and knowledge of student's diverse cultures, with students demonstrating pride in their work.	High levels of student engagement and teacher passion for the subject create a culture for learning in which everyone shares a belief in the importance of the subject, and all students hold themselves to high standards of performance, for example by initiating improvements to their work. Teacher and students demonstrate high levels of respect for and knowledge of diverse student cultures.
2c: Managing classroom procedures	Much instructional time is lost due to inefficient classroom routines and procedures, for transitions, handling of supplies, and performance of non-instructional duties.	Some instructional time is lost due to only partially effective classroom routines and procedures, for transitions, handling of supplies, and performance of non-instructional duties.	Little instructional time is lost due to classroom routines and procedures, for transitions, handling of supplies, and performance of non-instructional duties, which occur smoothly.	Students contribute to the seamless operation of classroom routines and procedures, for transitions, handling of supplies, and performance of non-instructional duties.
2d: Managing student behavior	There is no evidence that standards of conduct have been established, and little or no teacher monitoring of student behavior. Response to student misbehavior is repressive, or disrespectful of student dignity.	The teacher has made an effort to establish standards of conduct for students. Teacher tries, with uneven results, to monitor student behavior and respond to student misbehavior.	Standards of conduct are clear to students, and the teacher monitors student behavior against those standards. Teacher response to student misbehavior is appropriate and respects the students' dignity.	Standards of conduct are clear, with evidence of student participation in setting them. Teacher's monitoring of student behavior is subtle and preventive, and teacher's response to student misbehavior is sensitive to individual student needs. Students take an active role in monitoring the standards of behavior.
2e: Organizing physical space	Teacher makes poor use of the physical environment, resulting in unsafe or inaccessible conditions for some students or a significant mismatch between the physical arrangement and the lesson activities.	Teacher's classroom is safe, and essential learning is accessible to most students, but the physical arrangement only partially supports the learning activities. Teacher's use of physical resources, including computer technology, is moderately effective.	Teacher's classroom is safe, and learning is accessible to all students; teacher ensures that the physical arrangement supports the learning activities, (when applicable) Teacher makes effective use of physical resources, including computer technology.	The classroom is safe, and the physical environment ensures the learning of all students, including those with special needs. Students contribute to the use or adaptation of the physical environment to advance learning. Technology is used skillfully, as appropriate to the lesson.

The building should collaboratively develop a school-wide plan to ensure positive student behavior.

NOTE: The CPS Framework for Teaching has been adapted, with permission, from Charlotte Danielson's *Framework for Teaching*.

CPS FRAMEWORK FOR TEACHING

Domain 3: Instruction

Component	Unsatisfactory	Basic	Proficient	Distinguished
3a: Communicating with students	Expectations for learning, directions and procedures, and explanations of content are unclear or confusing to students. Teacher's use of language contains errors or is inappropriate to students' diverse cultures or levels of development.	Expectations for learning, directions and procedures, and explanations of content are clarified after initial confusion; teacher's use of language is correct but may not be completely appropriate to students' diverse cultures or levels of development.	Expectations for learning, directions and procedures, and explanations of content are clear to students. Communications are appropriate to students' diverse cultures and levels of development	Expectations for learning, directions and procedures, and explanations of content are clear to students. Teacher's oral and written communication is clear and expressive, appropriate to students' diverse cultures and levels of development, and anticipates possible student misconceptions.
3b: Using questioning and discussion techniques	Teacher's questions are low-level or inappropriate, eliciting limited student participation, and recitation rather than discussion.	Some of the teacher's questions elicit a thoughtful response, but most are low-level, posed in rapid succession. Teacher' attempts to engage all students in the discussion are only partially successful.	Most of the teacher's questions elicit a thoughtful response, and the teacher allows sufficient time for students to answer. All students participate in the discussion, with the teacher stepping aside when appropriate.	Questions reflect high expectations and are culturally and developmentally appropriate. Students formulate many of the high-level questions and ensure that all voices are heard.
3c: Engaging students in learning	Activities and assignments, materials, and groupings of students are inappropriate to the instructional outcomes, or levels of understanding, resulting in little intellectual engagement. The lesson has no structure or is poorly paced. Activities, assignments, and materials are not appropriate for diverse cultures.	Activities and assignments, materials, and groupings of students are partially appropriate to the instructional outcomes, or levels of understanding, resulting in moderate intellectual engagement. The lesson has a recognizable structure but is not fully maintained. Activities, assignments, and materials are partially appropriate for diverse cultures.	Activities and assignments, materials, and groupings of students are fully appropriate to the instructional outcomes, and students' cultures and levels of understanding. All students are engaged in work of a high level of rigor. The lesson's structure is coherent, with appropriate pace. Activities, assignments, and materials are fully appropriate for diverse cultures.	Students are highly intellectually engaged throughout the lesson in higher order learning, and make material contributions to the activities, student groupings, and materials. The lesson is adapted as needed to the needs of individuals, and the structure and pacing allow for student reflection and closure. Students assist in ensuring that activities, assignments and materials are fully appropriate for diverse cultures.
*3d: Using Assessment in Instruction**	Assessment is not used in instruction, either through students' awareness of the assessment criteria, monitoring of progress by teacher or students, or through feedback to students.	Assessment is occasionally used in instruction, through some monitoring of progress of learning by teacher and/or students. Feedback to students is uneven, and students are aware of only some of the assessment criteria used to evaluate their work.	Assessment is regularly used in instruction, through self-assessment by **students,*** monitoring of progress of learning by teacher and/or students, and through high quality feedback to students. Students are fully aware of the assessment criteria used to evaluate their work.	Multiple assessments are used in instruction, through student involvement in establishing the assessment criteria, self-assessment by students and monitoring of progress by both students and teachers, and high quality feedback to students from a variety of sources.
3e: Demonstrating flexibility and responsiveness	Teacher adheres to the instruction plan in spite of evidence of poor student understanding or of students' lack of interest, and fails to respond to student questions; teacher assumes no responsibility for students' failure to understand.	Teacher demonstrates moderate flexibility and responsiveness to student questions, needs and interests during a lesson, and seeks to ensure the success of all students.	Teacher ensures the successful learning of all students, making adjustments as needed to instruction plans and responding to student questions, needs and interests.	Teacher is highly responsive to individual students' needs, interests and questions, making even major lesson adjustments as necessary to meet instructional goals, and persists in ensuring the success of all students.

***It is acknowledged that when student assessment data that accurately measure student growth are available, student learning outcomes will be addressed and incorporated into the system.**

NOTE: The CPS Framework for Teaching has been adapted, with permission, from Charlotte Danielson's *Framework for Teaching*.

CPS FRAMEWORK FOR TEACHING

Domain 4: Professional Responsibilities

Component	Unsatisfactory	Basic	Proficient	Distinguished
4a: Reflecting on Teaching	Teacher's reflection on the lesson does not provide an accurate or objective description of the event of the lesson.	Teacher's reflection provides a partially accurate and objective description of the lesson, but does not cite specific positive and negative characteristics. Teacher makes global suggestions as to how the lesson might be improved.	Teacher's reflection provides an accurate and objective description of the lesson, and cites specific positive and negative characteristics. Teacher makes some specific suggestions as to how the lesson might be improved.	Teacher's reflection on the lesson is highly accurate and perceptive, and cites specific examples that were not fully successful, for at least some students. Teacher draws on an extensive repertoire to suggest alternative strategies.
4b: Maintaining Accurate Records	Teacher's system for maintaining both instructional and non-instructional records are either non-existent or in disarray, resulting in errors and confusion.	Teacher's system for maintaining both instructional and non-instructional records is rudimentary and only partially effective.	Teacher's system for maintaining both instructional and non-instructional records is accurate, efficient and effective.	Teacher's system for maintaining both instructional and non-instructional records is accurate, efficient and effective, and students contribute to its maintenance.
*4c: Communicating with Families**	Teacher provides little or no information to families, or such communication is culturally inappropriate. Teacher makes no attempt to engage families in the instructional program.	Teacher complies with school procedures for communicating with families and makes an effort to engage families in the instructional program. But communications are not always appropriate to the cultures of those families.	Teacher communicates frequently and successfully engages most families in the instructional program. Information to families about individual students is conveyed in a culturally appropriate manner.	Teacher communicates frequently and sensitively with individual families in a culturally sensitive manner, with students participating in the communication. Teacher successfully engages families in the instructional program; as appropriate.
4d: Participating in a Professional Community	Teacher avoids participating in the job-embedded professional community or in school and district events and projects, relationships with colleagues are negative or self-serving and teacher is resistant to feedback from colleagues.	Teacher becomes involved in the job-embedded professional community and in school and district events and projects when specifically asked; relationships with colleagues are cordial. Teacher accepts, with some reluctance, feedback from colleagues.	Teacher participates actively during the in the job-embedded** professional community and maintains positive and productive relationships with colleagues. In addition, teacher welcomes feedback from colleagues.	Teacher makes a substantial contribution to the job-embedded** professional community, and assumes a leadership role with colleagues. In addition, teacher seeks out feedback from colleagues.
4e: Growing and Developing Professionally	Teacher does not participate in professional development activities, even when such activities are clearly needed for the development of teaching skills.	Teacher's participation in job-embedded professional development activities is limited to those that are convenient or are required.	Teacher engages in opportunities for job-embedded professional development that is based on a self- assessment of need.	Teacher actively pursues professional development opportunities, and makes a substantial contribution to the profession through such activities as action research and mentoring new teachers.
4f: Demonstrating Professionalism	Teacher has little sense of ethics and professionalism, and contributes to practices that are self-serving or harmful to students. Teacher fails to comply with school and district regulations and timelines.	Teacher is honest and well-intentioned in serving students and contributing to child-centered decisions in the school Teacher complies minimally with school and district regulations, doing just enough to "get by."	Teacher displays a high level of ethics and professionalism in interactions with both students and the school community, and complies fully with school and district regulations.	Teacher assumes a leadership role in ensuring that school practices, decisions and procedures ensure that all the students' interests are addressed. Teacher displays the highest standards of ethical conduct.

***It is understood that the support of the building administrator is essential for the teacher to be successful**

****Job-embedded means during the school day - Participation in school or district events outside of the school day will not affect the teacher's summative rating.**

NOTE: The CPS Framework for Teaching has been adapted, with permission, from Charlotte Danielson's *Framework for Teaching*.

Appendix E: CPS Checklist

FORM 5A: CLASSROOM TEACHER VISITATION
(Required)

*This form is **required**. It should be used in conjunction with the "Post-Observation Framework Feedback Form" (Form 5B).*

Teacher's Name:_____ **Room**_____ **Date**_____

School_____ **Subject/Grade**_____

(Place a (√) or brief comment in the appropriate column.)

		Strength	Weakness	Does Not Apply
I.	**Instruction**			
	a) Provides written lesson plans and preparation in accordance with the objectives of the instructional program.	_____	_____	_____
	b) Establishes positive learning expectation standards for all students.	_____	_____	_____
	c) Periodically evaluates pupils' progress and keeps up-to-date records of pupils' achievements.	_____	_____	_____
	d) Applies contemporary principles of learning theory and teaching methodology.	_____	_____	_____
	e) Draws from the range of instruction materials available in the school.	_____	_____	_____
	f) Exhibits willingness to participate in the development and implementation of new ideas and teaching techniques.	_____	_____	_____
	g) Provides bulletin board and interest areas reflective of current student work.	_____	_____	_____
	h) Exhibits and applies knowledge of the curriculum content related to subject area and instructional level.	_____	_____	_____
	i) Shows evidence of student performance and progress.	_____	_____	_____
II.	**School Environment**			
	a) Establishes and maintains reasonable rules of conduct within the classroom consistent with the provisions of the Student Code of Conduct.	_____	_____	_____
	b) Maintains attendance books, lesson plan, seating chart(s) and grade book accurately.	_____	_____	_____
	c) Uses recommendations and suggestions form conference and special education staffings.	_____	_____	_____
	d) Encourages student growth in self discipline and positive self-concept.	_____	_____	_____
	e) Makes students aware of the teacher's objectives and expectations.	_____	_____	_____
	f) Practices fairness in teacher-pupil relationships.	_____	_____	_____
	g) Exhibits an understanding and respect for students as individuals.	_____	_____	_____
III.	**Professional and Personal Standards.**			
	a) Presents an appearance that does not adversely affect the students' ability to learn.	_____	_____	_____
	b) Demonstrates proper diction and grammatical usage when addressing students.	_____	_____	_____
	c) Uses sound and professional judgment.	_____	_____	_____
IV.	**Local School Unit Criteria**			
	a) CPS Framework for Teaching and related process (See attached).			
	b) _____			
	c) _____			

COMMENTS:_____
_____
_____