Abstract Title Page

Title:
Teacher Evaluation in Practice: Understanding Evaluator Reliability and Teacher Engagement in Chicago Public Schools

Authors and Affiliations:
Susan E. Sporte, University of Chicago Consortium on Chicago School Research
Jennie Y. Jiang, University of Chicago Consortium on Chicago School Research
Stuart Luppescu, University of Chicago Consortium on Chicago School Research

Abstract Body

Problem / Background / Context:
One of the most persistent and urgent problems facing education policymakers is the provision of highly effective teachers in all of our nation's classrooms. However, teacher evaluation policies across the nation have had well documented shortcomings (Hanushek and Rivkin, 2010). In particular, previous teacher evaluation systems produced the same general ratings for all teachers, providing little information on which teachers excelled or which needed improvement. They also failed to provide a way for teachers to receive feedback and rarely provided actionable information to teachers about how they could improve their practice.

Recent national policy, including the 2010 federal Race to the Top (RTTT) competition, has emphasized a dramatic overhaul of evaluation systems. State and local education agencies and districts responded by replacing traditional evaluation approaches with new systems that incorporate multiple methods of assessing teachers and formal processes to provide teachers with feedback. In recent years, districts across the nation have begun implementing these new evaluation systems.

Although these new evaluation systems incorporate student test score data to estimate a teacher's idiosyncratic contribution to student learning, a majority of teachers teach in grades or subjects in which they are not administered. As a result, classroom observation measures of teacher performance remain critically important components of teacher evaluation and typically comprise most of a teacher's summative evaluation score. Observations are also the key lever for providing teachers with timely and individualized feedback on their classroom practice. New evaluation systems have updated their observation process to include more frequent observations and formal structures for feedback.

Observations as measures of teacher performance rely heavily on evaluators who are highly trained and able to differentiate teacher performance using an observation rubric (Kane and Staiger, 2012). In addition, observations as key levers of instructional improvement rely heavily on teacher engagement in the process. Both ensuring the reliability of evaluators and engaging teachers in the process is vital to successful implementation.

Starting in 2012-13 we have worked in partnership with Chicago Public Schools (CPS) and the Chicago Teachers Union (CTU) to study implementation of Chicago's new teacher evaluation system. In this proposed presentation we share findings as well as experiences from our collaboration with CPS and the CTU as they seek to identify real-time strategies for improving implementation. We focus on findings from two key components for implementing a new evaluation system: the reliability of evaluators and teacher perceptions of their evaluators and the new process.

Purpose / Objective / Research Question / Focus of Research
We seek to share findings from the following research questions about Chicago's new teacher evaluation system:

- To what extent do systematic differences between evaluators affect teacher ratings?
- What are teacher perceptions of their evaluators? What are teacher perceptions of the new observation process?

In addition to these questions we will share how the district utilized findings to guide changes to the implementation of Chicago's new evaluation system.

Improvement Initiative / Intervention / Program / Practice:
CPS' new teacher evaluation system —Recognizing Educators Advancing Chicago's Students (REACH)— began district-wide implementation in the 2012-13 school year. REACH seeks to provide a measure of individual teacher effectiveness that can meet the district's dual needs of supporting instructional improvement and differentiating teacher performance. It incorporates teacher performance ratings based on multiple classroom observations together with student growth measured on two different types of assessments. The main components of REACH include multiple classroom observations using a modified version of the Charlotte Danielson Framework for Teaching, required feedback after each observation, and the inclusion of two different measures of student growth (see Appendix B Figure 1 and 2 for more information about REACH).

Another significant component of REACH is extensive training requirements for evaluators (currently only principals and assistant principals may act as evaluators). Prior to conducting any observations, administrators had to complete an online certification process that included video-based scoring practice and an assessment of their rating accuracy. Beyond this initial certification, the district required administrators to attend professional development sessions and participate in individualized coaching and calibration sessions throughout the year.

CPS and the CTU formed a joint committee to develop REACH. This joint committee continues to meet frequently to monitor implementation and make continuous improvements to the system.

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

Population / Participants / Subjects:
In school years 2012-13 and 2013-14, CPS employed approximately 23,000 teachers each year. Five hundred seventy eight schools are covered by this initiative, which have an enrollment of about 400,000 students each year (see Appendix B Table 2 for more details). Chicago students are likely to be from low-income families (87 percent), and 42 percent are African American and 44 percent are Latino.

Research Design:
To understand systematic differences in evaluators and their effect on teacher observation ratings, we focus on three aspects of evaluator behavior (1) evaluator severity and (2) evaluator internal consistency and (3) reliability using a generalizability theory approach. Differences in evaluator severity result in systematic differences in average ratings across evaluators — a more severe evaluator will award systematically lower ratings to an average teacher; a more lenient evaluator will give systematically higher ratings. Differences in evaluator internal consistency are systematic differences in the ratings patterns for all ratings of a particular evaluator. We analyzed all observation data looking for unusual patterns in the ratings. In addition, we are a conducting a generalizability study to quantify the amount of variance in ratings that is due to

differences in actual teacher performance compared to the amount of variance in ratings that is due to other factors.

Evaluator Severity
We use a randomly selected sample of 48 administrators across the district along with independent, joint, randomly assigned observations with 16 external observers to analyze severity. We use the Many-facet Rasch approach to model the probability of a teacher being given a certain rating on a component by an evaluator. Log-odds of the rating is modeled as a linear function of the measure of the performance of teacher *n*, β_n; the difficulty of component *i*, δ_i; and the severity of evaluator *j*, ν_j ; and the rating scale structure denoted by τ_x. The following equation shows the relationship among the rating and the three parameters plus the rating scale structure.

$$\pi_{nijx} = \frac{\exp \sum_{l=0}^{x} [\beta_n - (\delta_i + \upsilon_j + \tau_x)]}{\sum_{k=0}^{m_i} \exp \sum_{l=0}^{k} [\beta_n - (\delta_i + \upsilon_j + \tau_x)]}$$

Where π_{nijx} is the probability of teacher *n* receiving a rating of *x* from evaluator *j* on component *i* where teacher performance is β_n, rating scale structure is τ_x, component difficulty is δ_i and evaluator severity ν_j. These are accumulated across rating categories *l* from 0 to *x* in the numerator, and in the denominator across all *k* ratings categories and across the m_i category thresholds in component *i*.

The Many-facet Rasch model uses an iterative, modified Newton-Raphson estimation method and can estimate the three parameters independently of each other. This enables us to examine evaluator severity without the confounding effects of teacher performance or component difficulty.

Evaluator Internal Consistency We analyzed all observation data from 2012-13 for unusual patterns in the ratings. Similar to difficulty levels of test items, some components of the observation framework are more difficult to get higher ratings on than others. Using this component difficulty and all the ratings of all the observations of each evaluator, we can look at the patterns in evaluators' responses and identify evaluators who are not assigning ratings consistently.

There are two types of rating patterns that violate the expected: (1) erratic patterns are cases where evaluators give unexpectedly higher ratings to harder components and lower ratings to easier components; (2) muted patterns are cases where evaluators assign very similar ratings across all components regardless of component difficulty. Both patterns indicate evaluators are providing data that may not be reliable or useful (see Appendix B Figure 3 for a complete set of equations).

Generalizability Study
A fully crossed generalizability study is currently being conducted: teachers (*t*) x component (*c*) x evaluators (*r*) to analyze what proportion of the variance in ratings across teachers is measuring true teacher performance. In this approach, reliability is calculated by taking into account variability from components of the scoring rubric as well as variation among evaluators. Analysis will be complete this summer and included in our presentation.

Teacher and Administrator Surveys
Survey data includes survey items from CPS's annual My Voice, My School survey. This survey was administered to all teachers in March 2013 and had a response rate of 81 percent. All principals and assistant principals were surveyed in April/May 2013 and had a 57 percent response rate.

Data Collection and Analysis:
Data for this presentation includes CPS personnel and administrative data from the 2012-13 school year, observation ratings from external observers, and survey data from the 2012-13 and 2013-14 school year. Teacher and administrator personnel data includes individual-level data about tenure status, years of experience in the district, demographic information as well as evaluation data such as ratings and value-added scores (Please see Appendix B Table 2 for an overview of evaluation data)

Observation data from external observers consists of 123 independent joint observations conducted with CPS principals. External observers were randomly assigned to this joint observations and this data allowed us to analyze for evaluator severity.

Findings / Outcomes:
Findings from the end of the first year of REACH implementation include:

- In general administrators were able to apply the rating rubric in ways that were consistent with what we would have predicted given component difficulty. About 3 percent of administrators were identified as muted and about 5 percent of administrators were erratic.
- There is considerable variation in evaluator severity, but few are so extreme as to have a very substantial effect on teacher ratings. About 10 percent of our sample of evaluators is extremely severe or lenient. In general most administrators and external observers were rating within 0.5 points of the average (see appendix B Figure 4 for more details).
- In the first year of REACH implementation, teachers were generally positive about the accuracy of the ratings they received from evaluators and the observation process (see Appendix B Table 3 for details.

Conclusions:
We worked in partnership with CPS and the CTU to identify real-time strategies for challenges identified by our findings. Our conclusions on the systematic differences in evaluators led the district to target professional development for evaluators. In addition CPS is beginning develop confidence intervals for teacher observation ratings to account for measurement error in part due to the variability in scores assigned by evaluators. Our survey results are being used to gauge the quality of implementation in schools and target support and communication to teachers. As an independent research partner with both the district and the union, we have been able to provide findings and evidence that have allowed both sides to work together in improving implementation of this new policy.

Appendices

Appendix A. References

Hanushek, E.A., Rivkin, S.G. (2010). Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review, 100*(2), 4.

Kane, T.J., and Staiger, D.O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Appendix B. Tables and Figures

Figure 1: The Elements of REACH

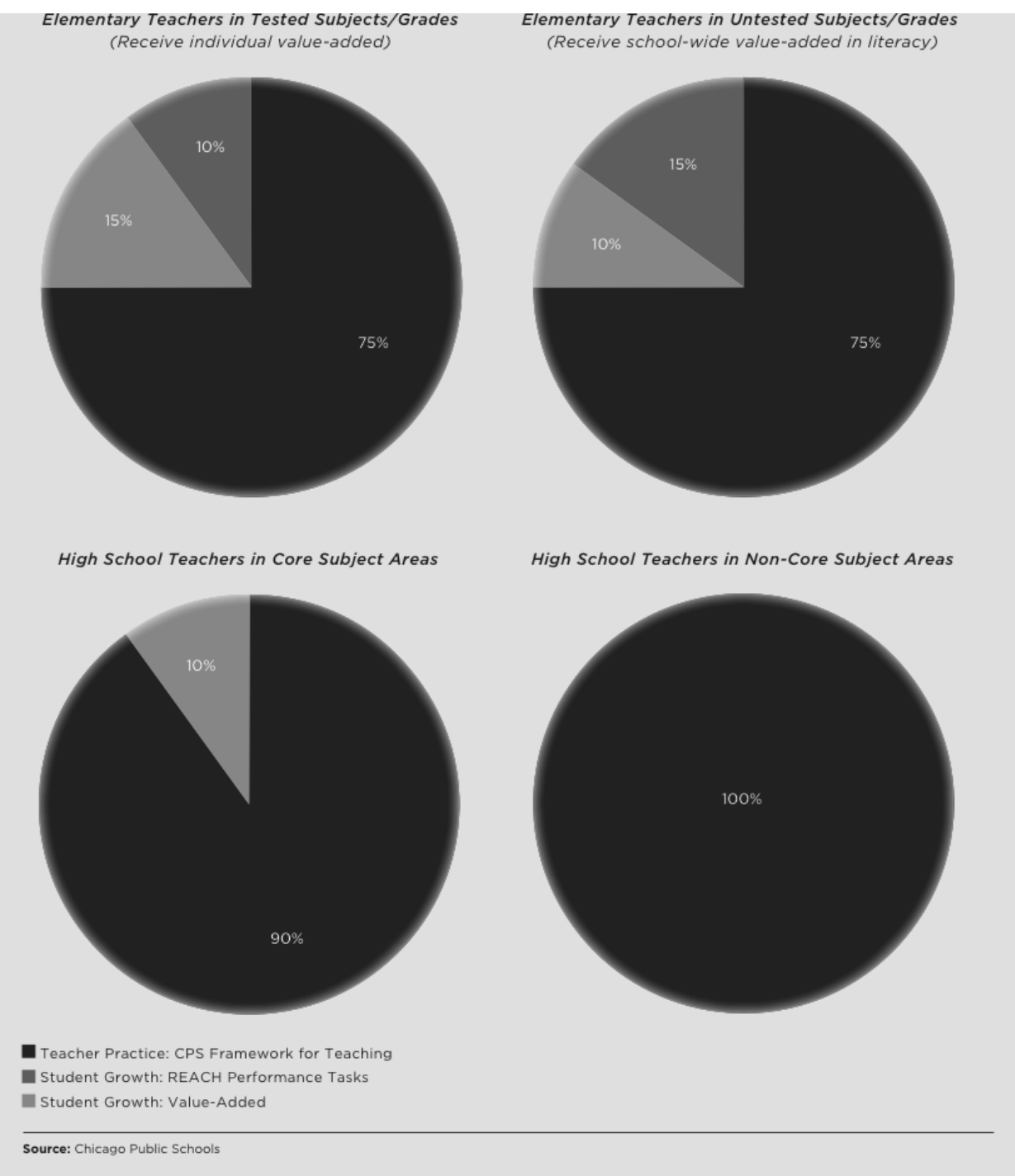


Figure 2: The CPS Framework for Teaching

The CPS Framework for Teaching

Adapted from the *Danielson Framework for Teaching* and Approved by Charlotte Danielson

Domain 1: Planning and Preparation

a. Demonstrating Knowledge of Content and Pedagogy
- Knowledge of Content Standards Within and Across Grade Levels
- Knowledge of Disciplinary Literacy
- Knowledge of Prerequisite Relationships
- Knowledge of Content-Related Pedagogy

b. Demonstrating Knowledge of Students
- Knowledge of Child and Adolescent Development
- Knowledge of the Learning Process
- Knowledge of Students' Skills, Knowledge, and Language Proficiency
- Knowledge of Students' Interests and Cultural Heritage
- Knowledge of Students' Special Needs and Appropriate Accommodations/Modifications

c. Selecting Instructional Outcomes
- Sequence and Alignment
- Clarity
- Balance

d. Designing Coherent Instruction
- Unit/Lesson Design that Incorporates Knowledge of Students and Student Needs
- Unit/Lesson Alignment of Standards-Based Objectives, Assessments, and Learning Tasks
- Use of a Variety of Complex Texts, Materials and Resources, including Technology
- Instructional Groups
- Access for Diverse Learners

e. Designing Student Assessment
- Congruence with Standards-Based Learning Objectives
- Levels of Performance and Standards
- Design of Formative Assessments
- Use for Planning

Domain 2: The Classroom Environment

a. Creating an Environment of Respect and Rapport
- Teacher Interaction with Students, including both Words and Actions
- Student Interactions with One Another, including both Words and Actions

b. Establishing a Culture for Learning
- Importance of Learning
- Expectations for Learning and Achievement
- Student Ownership of Learning

c. Managing Classroom Procedures
- Management of Instructional Groups
- Management of Transitions
- Management of Materials and Supplies
- Performance of Non-Instructional Duties
- Direction of Volunteers and Paraprofessionals

d. Managing Student Behavior
- Expectations and Norms
- Monitoring of Student Behavior
- Fostering Positive Student Behavior
- Response to Student Behavior

Domain 4: Professional Responsibilities

a. Reflecting on Teaching and Learning
- Effectiveness
- Use in Future Teaching

b. Maintaining Accurate Records
- Student Completion of Assignments
- Student Progress in Learning
- Non-Instructional Records

c. Communicating with Families
- Information and Updates about Grade Level Expectations and Student Progress
- Engagement of Families and Guardians as Partners in the Instructional Program
- Response to Families
- Cultural Appropriateness

d. Growing and Developing Professionally
- Enhancement of Content Knowledge and Pedagogical Skill
- Collaboration and Professional Inquiry to Advance Student Learning
- Participation in School Leadership Team and/or Teacher Teams
- Incorporation of Feedback

e. Demonstrating Professionalism
- Integrity and Ethical Conduct
- Commitment to College and Career Readiness
- Advocacy
- Decision-Making
- Compliance with School and District Regulations

Domain 3: Instruction

a. Communicating with Students
- Standards-Based Learning Objectives
- Directions for Activities
- Content Delivery and Clarity
- Use of Oral and Written Language

b. Using Questioning and Discussion Techniques
- Use of Low- and High-Level Questioning
- Discussion Techniques
- Student Participation and Explanation of Thinking

c. Engaging Students in Learning
- Standards-Based Objectives and Task Complexity
- Access to Suitable and Engaging Texts
- Structure, Pacing and Grouping

d. Using Assessment in Instruction
- Assessment Performance Levels
- Monitoring of Student Learning with Checks for Understanding
- Student Self-Assessment and Monitoring of Progress

e. Demonstrating Flexibility and Responsiveness
- Lesson Adjustment
- Response to Student Needs
- Persistence
- Intervention and Enrichment

2012

Figure 3: Internal Consistency Equations

$$\text{Infit (information weighted) mean square}: \mathrm{v_j} = \sum_{n=1}^{N} y_{nij}^2 / \sum_{n=1}^{N} W_{nij}, E(v_j) = 1$$

$$\text{Variance of } x_{nijx}: W_{nijx} = \sum_{k=0}^{m} (x - E_{nij})^2 \pi_{nijx}$$

$$\text{Expectation}: E_{nijx} = \sum_{k=0}^{m} x \pi_{nijx}$$

$$\text{Raw residual}: y_{nijx} = x - E_{nijx}$$

$$\text{Standardized residual}: z_{nij} = y_{nijx} / W_{nijx}^{1/2}$$

The fit statistic described above (called the "information weighted mean square") quantifies the degree to which the pattern of ratings fit the expected pattern. A value of greater than 1.5 for this statistic indicates that there was more than 50% greater variation in the ratings data than expected. This is the "erratic" pattern. A value of 0.5 or less indicates that there was less than 50% of the expected amount of variation in the ratings. This is the "muted" pattern.

Table 1: CPS Schools and Personnel

CPS School and Personnel Statistics (2012-13)

Schools*	578
Elementary Schools	472
High Schools	106
Non-Tenured Teachers	5,743
Tenured Teachers	15,109
Administrators**	1,195

Source: CPS Stats and Facts, Administrative records

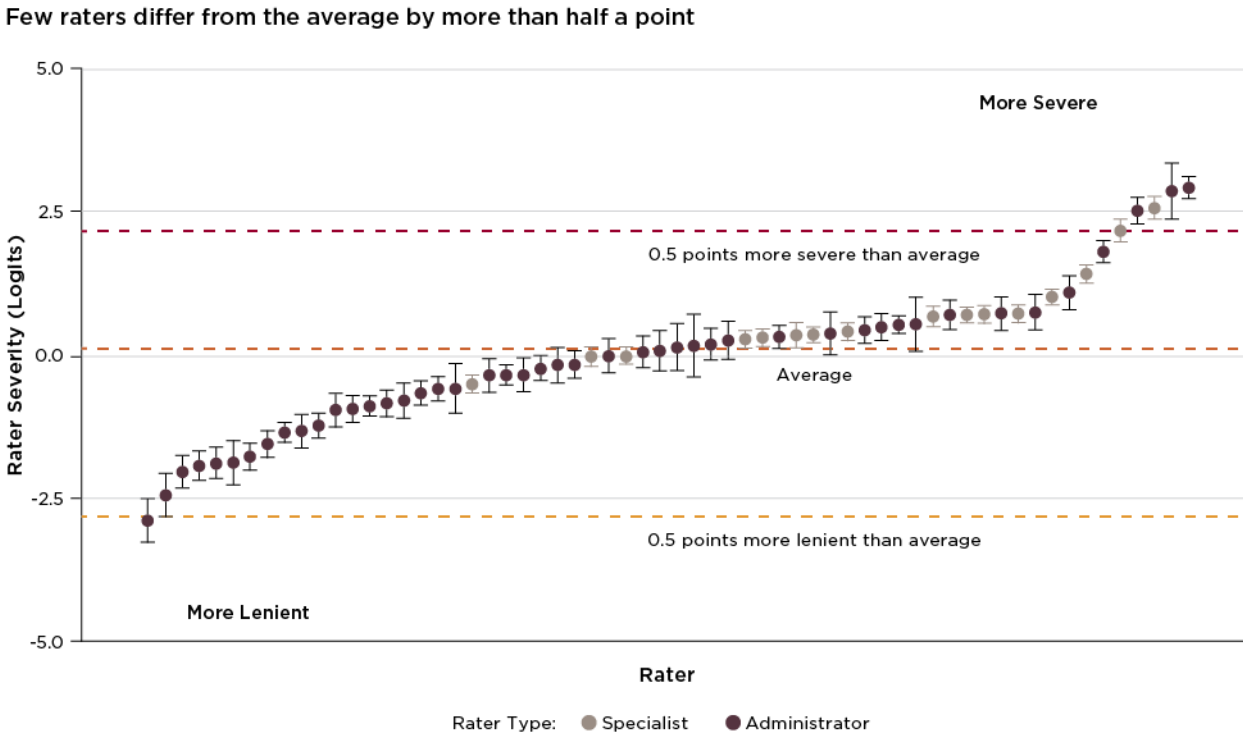
* Does not include charter or contract schools
** Only includes principals and assistant principals

Table 2: REACH Evaluation Data 2012-13

2012-13 REACH administrative data

	Elementary Non-Tenured	Elementary Tenured	High School Non-Tenured	High School Tenured
All Teachers	4,353	10,785	1,575	4,082
Summative Ratings	3,147	0	1,270	0
Observation Scores	4,326	10,050	1,573	3,975
Individual VA Scores	1,213	3,939	5	19
School-Wide VA Scores	3,095	6,719	0	0
Performance Task Scores	4,353	10,785	1,099	2,775

Figure 4: Evaluator severity



HOW TO READ FIGURE 4

Each dot on the figure represents an evaluator: purple dots are administrators, and gray dots are specialists. The lines above and below the dot are error bars, extending 1.96 standard errors above and below. Longer bars mean the estimate of the principal's severity is less precise; the true severity measure could fall anywhere within that bar. Error could come from two sources: (1) fewer classroom observation data from the evaluator and (2) how extreme the ratings are compared to average. The vertical axis is severity on the original log-odds units scale. The 0 point is the average severity of our study sample; higher indicates more severity while lower indicates more lenient. Evaluators at the top of the figure are more severe and evaluators at the bottom are more lenient. The three horizontal dashed lines indicate the average severity (in the middle), one half-point more severe than average, and one half-point more lenient than average.

Table 3: Teacher Perceptions of the REACH Evaluation

My Voice My School Teacher Survey
May 2013

Overall Evaluation Process

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

Evaluator

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

Feedback

	Somewhat Useful or Very Useful
How useful is your evaluator's feedback for your instruction?	67.03%

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