Study of Emerging Teacher Evaluation Systems

Study of Emerging Teacher Evaluation Systems

Prepared by:

Leslie M. Anderson
Alisha Butler
Andrea Palmiter
Erikson Arcaira

Policy Studies Associates

Prepared for:
Policy and Program Studies Service
Office of Planning, Evaluation, and Policy Development
U.S. Department of Education

November 2016

This report was produced under U.S. Department of Education Contract No. ED-CF0-10-A-0021-0002 with Policy Studies Associates. Elizabeth Eisner and Joanne Bogart served as the contracting officer's representatives. The views expressed herein do not necessarily represent the positions or policies of the Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service or enterprise mentioned in this publication is intended or should be inferred. For the reader's convenience, this publication contains information about and from outside organizations, including hyperlinks and URLs. Inclusion of such information does not constitute an endorsement by the Department.

U.S. Department of Education
John King
Secretary

Office of Planning, Evaluation and Policy Development
Amy McIntosh
Acting Assistant Secretary
Delegated Duties of Assistant Secretary

Policy and Program Studies Service
Jennifer Bell-Ellwanger
Director

November 2016

This report is available on the Department's website at
http://www2.ed.gov/about/offices/list/opepd/ppss/reports.html#tq.

Availability of Alternate Formats
Requests for documents in alternate formats such as Braille or large print should be submitted to the Alternate Format Center by calling 202–260–0852 or by contacting the 504 coordinator via email at om_eeos@ed.gov.

Notice to Limited English Proficient Persons
If you have difficulty understanding English you may request language assistance services for Department information that is available to the public. These language assistance services are available free of charge. If you need more information about interpretation or translation services, please call 1–800–USA–LEARN (1–800–872–5327) (TTY: 1–800–437–0833), or email us at: Ed.Language.Assistance@ed.gov. Or write to: U.S. Department of Education, Information Resource Center, LBJ Education Building, 400 Maryland Ave. SW, Washington, DC 20202.

Content Contact:
Joanne Bogart
Phone: 202–205–7855
Email: Joanne.Bogart@ed.gov

Acknowledgements

We wish to thank several individuals who contributed to the completion of this report. First, we are grateful for the guidance and support of the U.S. Department of Education. In particular, we thank Elizabeth Eisner and Joanne Bogart of the Policy and Program Studies Service for their careful guidance and support throughout the study.

We are particularly grateful to the many district administrators, principals, and teachers who participated in interviews and focus groups as part of the data collection for this study. Without their efforts, this report would not have been possible, and we deeply appreciate their assistance.

We also would like to acknowledge the thoughtful contributions of M. Bruce Haslam who successfully launched this complex study before retiring from Policy Studies Associates and then remained involved, generously sharing his wisdom and good humor throughout the study's data collection and reporting phases. In addition, we would also like to thank Meredith Ludwig and Ellen Behrstock Sherratt of the American Institutes for Research and Peter Youngs of the Curry School of Education at the University of Virginia for their valuable contributions to study design and data collection activities as well as for their thoughtful review of the draft report. Finally, we want to thank members of our technical working group who helped inform the design of the study, including Joanna Cannon, Laura Goe, Paul LeMahieu, Leigh McGuigan, Robert Meyer, Tony Milanowski, Pamela Moran, Susan Sclafani, Andy Sokatch, and Rob Weil. Although we appreciate the assistance and support of these individuals, any errors in judgment or fact are the responsibility of the authors.

Contents

Exhibits

Chapter I

Chapter II

Chapter III

Chapter IV

Chapter V

Chapter VI

Chapter VII

Appendix A

Appendix B

Appendix C

Executive Summary

Comprehensive teacher evaluation systems are a core element of current state and local strategies to support improved teaching and learning in elementary and secondary education in the United States. When fully implemented, these systems can inform a range of personnel decisions, including decisions pertaining to teacher support and professional development, career advancement and tenure, and compensation. Comprehensive teacher evaluation systems have the potential to contribute to improved teaching practice by providing information about a teacher's strengths and weaknesses. Until very recently, however, almost all teacher evaluation systems in the United States had very limited capacity to identify either effective or ineffective teachers (Weisberg et al., 2009).

Today, efforts are underway across the country to transform teacher evaluation into a useful tool for improving teaching and learning. These efforts are supported by state statutes as well as by improvements in state and local data systems and the development of new student assessments. In 2015, the National Council on Teacher Quality (NCTQ) reported that only five states "have no formal state policy requiring that teacher evaluations take objective measures of student achievement into account in evaluating teacher effectiveness" (NCTQ 2015, ii). In addition, state and local efforts to design and implement teacher evaluation systems have been supported by federal resources appropriated through several programs, including the State Fiscal Stabilization Fund (under the *American Recovery and Reinvestment Act*), Race to the Top (RTT), the Teacher Incentive Fund (TIF), and the Investing in Innovation (i3) Fund. The 2015 enactment of the *Every Student Succeeds Act* (*ESSA*) maintains a federal priority of ensuring that all students have access to high-quality educators. States that elect to use Title II, Part A funds to develop, improve, or provide assistance to local educational agencies (LEAs) to design and support the implementation of teacher, principal, or other school leader evaluation must describe in their state plans how they will use student growth and other measures of educator performance to provide clear, timely, and useful feedback to teachers, principals, or other school leaders.

This study provides descriptive information on the design and early implementation of teacher evaluation systems in eight local school districts. This exploratory study is intended to help other districts and states learn from the experiences of these eight districts to inform future research on the effects of teacher evaluation systems on teacher professional practice and student performance.

The study sample included four districts that were "fully implementing" their teacher evaluation systems at the time data were collected in 2012 and early 2013, and four districts that were considered to be "partially implementing" their systems. For the purposes of this study, "fully implementing districts" were defined as those that had used their teacher evaluation systems to evaluate every teacher in the district and tied teacher scores to specific results, including salary increases, bonuses, promotions, and other personnel decisions, while "partially implementing districts" were those that had designed and pilot tested their teacher evaluation system components but had not yet begun to use them to evaluate every teacher in the district and/or had not yet tied teacher scores to specific outcomes. The eight sample districts were:

Fully Implementing Districts
District of Columbia Public Schools (Washington, DC)
Hamilton County Department of Education (Chattanooga, TN)
Harrison School District 2 (Colorado Springs, CO)
Hillsborough County Public Schools (Tampa, FL)

Partially Implementing Districts
Austin Independent School District (Austin, TX)
Pittsburgh Public Schools (Pittsburgh, PA)
Plattsburgh City Public Schools (Plattsburgh, NY)
St. Mary's County Public Schools (Leonardtown, MD)

Highlights

- Teachers and central office staff generally agreed that the foremost goal of the teacher evaluation systems was to improve instruction.

- Teacher and principal input during the design and/or pilot test phase strongly influenced decisions regarding system modification in six districts, according to district administrators.

- Classroom observations varied in frequency, duration, and degree of formality in all eight districts In addition, principals reported challenges in finding time to conduct teacher observations.

- Six districts used multiple strategies for analyzing teacher impact on student performance, including individual and/or school-level value-added models (VAMs).

- Districts used teacher evaluation results for a range of purposes, including targeted professional development and support, career ladders and performance pay, and in some instances, redeployment or release of teachers identified as ineffective.

- The majority of districts created relatively simple, streamlined structures to administer their teacher evaluation systems.

- Teachers reported that that classroom observations and feedback helped them become better teachers.

Study Overview

This study sought answers to the following questions:

1. What key priorities and measures informed the design of new teacher evaluation systems?
2. What steps did the districts take prior to full implementation to test the system and prepare teachers and staff to implement it?
3. How did the districts structure and conduct the classroom observation component of their teacher evaluation systems?
4. How did the districts analyze student performance and other data to evaluate teacher performance?
5. How did the districts use, or plan to use, teacher evaluation results to make personnel decisions? To what extent were professional development and career advancement decisions tied to evaluation results?
6. What administrative structures and supports did districts use to support their new teacher evaluation systems?
7. What are the perceived early effects of the teacher evaluation systems on the professional practices of teachers, principals, and district administrators?

The Case Study Districts

Each teacher evaluation system selected for the study had to include the key features that all RTT grantees were required to include in their new teacher evaluation systems. Case study districts were selected because their teacher evaluation systems relied on multiple measures of teacher performance, including growth in student performance and observations of classroom practice. In addition, their systems: (1) distinguished among teachers at various levels of effectiveness; (2) included a formative component that provided timely feedback to teachers; and (3) used — or planned to use — results of the evaluation process to inform personnel decisions. Many contextual factors affected the design and implementation of the eight districts' teacher evaluation systems, including the following:

- Overall, the eight districts selected for the study had designed or implemented their evaluation systems within the last decade, with five of the eight districts commencing work within the last five years. Three districts, however, began work on their teacher evaluation systems as early as 2007 and 2009.

- Every district included in the study was ahead of statewide efforts to design teacher evaluation systems and, consequently, some eventually had to adjust their evaluation systems to comply with new state expectations and requirements. Indeed, five of the eight districts' evaluation system designs were influenced by state legislation, including legislation passed to support RTT applications.

- Seven of the eight districts' efforts to design their teacher evaluation systems grew from their participation in federal programs such as TIF, RTT, and i3. In addition, two districts also benefitted from Empowering Effective Teachers grants from the Bill and Melinda Gates Foundation.

- Seven of the eight districts' teachers were represented by unions affiliated with the American Federation of Teachers, the National Education Association, or both. Seven districts included union representatives in the evaluation design process.

In addition to classroom observations and growth in student performance, three of the eight districts included additional measures of teacher performance in their evaluations, including student surveys and professional expectations. Together, these measures were used to determine an overall rating of teacher performance. All eight districts in the study used four- or five-point scale rating systems to classify teacher performance, and six defined the ways in which they intended to use the evaluation results to make personnel decisions

Study Design

Each teacher evaluation system selected for the study had to include the key features required for all RTT grantees in their new teacher evaluation systems. Specifically, these systems: (1) relied on multiple measures of teacher performance, including gains in student achievement and observations of classroom practice, to assess individual teachers; (2) distinguished among teachers at various levels of proficiency and effectiveness; (3) included a formative component that provided timely feedback to teachers; and (4) used — or planned to use — results of the evaluation process to inform personnel decisions and decisions about teacher professional development and support. Sites were selected based on recommendations from members of the study's Technical Working Group and U.S. Department of

Education officials; searches of state education agency and school district websites; and conversations with district staff responsible for planning or managing design or implementation of new teacher evaluation systems.

Data collection involved site visits to the districts between spring 2012 and winter 2013. The site visit teams interviewed district staff who were responsible for and/or knowledgeable about the teacher evaluation system, as well as several key external stakeholders, including leaders and representatives of local teacher unions, principals, teachers (individually or in focus groups), instructional facilitators and coaches, peer observers, and administrators in six state education agencies (SEAs). In three of the four fully implementing districts included in the study (Harrison, Hamilton County, Hillsborough County), the site visit teams conducted teacher focus groups with approximately 120 teachers. The focus groups were intentionally limited to the fully implementing districts in an effort to capture teacher perceptions of the entire system, including its uses with respect to teacher compensation, recognition, and retention.[1] Finally, the study used extant data, including district reports and reports from external contractors and consultants, to verify data collected on-site about the structure of the teacher evaluation systems (including districts' teacher performance standards used to measure teacher professional practice), as well as to verify the number of classroom observations and types of student assessments districts used to evaluate teachers.

This study employed qualitative analytic procedures and adhered to a set of standards to limit bias and ensure reliable findings, including standards of evidence and triangulation of data. Site visitors reviewed and coded their interview notes and transcript data and compiled them into a standardized site visit report template used to capture verbatim quotes and summarize key findings that addressed the study questions. In addition, a data capture tool was used to categorize information collected on the components of each teacher evaluation system, including component weights, number of observations, and types of student performance data used to measure teacher performance.

Study Limitations

This study has three main limitations. First, it relies on data collected from a purposively selected group of eight districts, thus limiting the range of potential successes and challenges districts might experience in designing and implementing complex teacher evaluation systems. Second, statements regarding program processes, challenges, and successes represent the perspectives of the individuals making them and may not represent the full range of views among district and school staff. Nevertheless, key administrators in all eight study districts reviewed this report for its factual accuracy. Finally, because this study describes the design and implementation of eight districts' teacher evaluation systems at a particular point in time (from late 2012 to early 2013), the evaluation systems will likely change and evolve. Indeed, district administrators reported that the longer they implement, the better their insights into their system's design and, accordingly, they were regularly modifying and adapting those designs to better serve district, principal, and teacher needs. Another factor that is likely to influence these districts' designs is state teacher evaluation policy, which may change in response to the implementation of *ESSA* and of assessments aligned with college and career-ready standards. Despite these limitations, however, the report offers some important insights into the design and implementation processes related to comprehensive teacher evaluation systems and highlights some potential early outcomes.

[1] One of the four fully implementing districts requested that the study team not conduct teacher focus groups due to the district's concerns about the time burden the focus group interviews would place on teachers.

Key Findings

Designing an Evaluation System

Many choices and tasks go into the design of a comprehensive teacher evaluation system that relies on multiple measures of teacher performance to assess teacher effectiveness. Determining district purposes for developing a new teacher evaluation system was usually the first step in the design process for the districts in this study. Among the most central design tasks were designing observation rubrics to measure classroom practice and planning how to collect and analyze student achievement data to assess teacher impact on student performance. Other important steps in system design included involving stakeholder groups, particularly teacher unions, and assigning weights to each measure of teacher performance.

- **Respondents in all eight districts agreed that the foremost goal of their teacher evaluation systems was to improve instruction.** In each of the districts, central office staff, observers, and teachers alike expressed the view that the district's evaluation system was focused on improving student achievement through improved instruction. In four of the districts, some of the teachers interviewed said that they — or their fellow teachers — believed that the systems were also designed to identify and remove ineffective teachers.

- **Charlotte Danielson's *Framework for Teaching* (FFT) was the exclusive basis for the design of the observation rubrics in half of the districts included in the study.** The FFT divides the elements of teaching practice into 22 standards clustered into four domains: (1) planning and preparation, (2) classroom environment, (3) instruction, and (4) professional responsibilities. The other four districts developed their own rubrics to examine instructional practice but drew from existing frameworks as sources of reference, such as the FFT and Robert Marzano's *Classroom Instruction that Works.*

- **All eight districts used multiple assessments to assess teachers' influence on their students' performance.** In addition to the state test, districts used district-developed assessments, end-of-semester or end-of-course assessments, performance assessments, and standardized assessments, including DIBELS, PSAT, IB, and AP exam scores. Using multiple types of assessments was a common district strategy that addressed a three-fold purpose: (1) accommodated teachers of non-state tested grades and subjects; (2) offered all teachers other opportunities to demonstrate effectiveness through, for example, district curriculum-based assessments; and (3) guarded against one assessment being the primary determinant of teacher performance ratings.

- **To account for the many types of assessments and data sources used to measure teacher impact on student performance, seven of the eight districts assigned separate weights to each type of student assessment included in their evaluation system.** In addition, seven of the eight districts ensured that the same overall weight (i.e., either 40 or 50 percent of a teacher's score) was applied to student performance data for teachers of non-state tested grades and subjects as was applied to teachers of state-tested grades and subjects.

Early Implementation

The early implementation phase — pilot testing the system, introducing it to teachers, and training classroom observers — varied across the eight districts, yet may have been a critical juncture in determining the successful transition to a new, comprehensive teacher evaluation system.

- **Three of the eight districts did not pilot test their teacher evaluation systems before implementing them.** Administrators in these districts explained that holding teachers accountable as early as possible was important for achieving real, measurable change in teaching and learning.

- **Each of the eight districts included in the study introduced their teachers to the new evaluation system through one or more strategies, including distributing written materials, creating online resources, or using school-based teams of evaluation experts.** Teachers in three districts suggested that training should be ongoing because it sometimes took years to fully understand and respond to the complexities of their districts' teacher evaluation systems.

Conducting Classroom Observations

Districts' approaches to conducting classroom observations varied with respect to the types of observers used and the frequency, duration, and degree of formality of the observations.

- **The frequency of classroom observations varied widely across the eight districts included in the study, ranging from two observations for experienced teachers in one district to 18 observations for new teachers in another.** In some districts, these observations were preceded by a pre-observation conference between the observer and the teacher, and in all districts the observations were followed by feedback, often in the form of a post-observation conference.

- **Three districts used peer observers, in addition to principals, to conduct classroom observations. Administrators in these districts explained that peer observers were intended to offer an unbiased, objective opinion of a teacher's performance.** In Austin and Hillsborough County, the peer observers were teachers from within the district who chose to take a leave of absence from the classroom to observe and evaluate other teachers. The District of Columbia hired former teachers who had worked in an urban district for at least five years. District staff and teacher focus group participants in two districts expressed their view that the legitimacy of an observer — and the value of the feedback provided — was partly based on the strength of the match between the observer's background and experience and that of the teacher being observed.

Using Student Performance and Other Data

Districts used varied strategies in their evaluation systems to measure teacher impact on student performance.

- **Six of the eight districts used multiple methods for measuring teacher impact on student performance, including individual and/or school-level value-added models (VAMs).** VAMs are statistical models that attempt to explain teacher (or school) contributions to student academic growth over time by controlling for school- and student-level variables that may affect student learning.

- **Seven districts used student learning objectives (SLOs) among their methods for measuring teacher impact on student performance.** Some highlighted the challenges associated with SLOs including setting realistic and consistent goals for measuring student growth and ensuring that principals gave teachers fair and consistent advice on what the SLOs should be and how to measure them.

- **In four districts, some teachers who participated in focus groups or individual interviews criticized the use of student achievement data to evaluate their performance.** In particular, they expressed concern that the selected tests (especially the state assessments) do not fully capture their students' actual growth or are not aligned with the school curriculum.

Using Teacher Evaluation Results

Seven of the eight districts included in the study had defined — or begun to define — the results they would apply to teacher evaluation scores, including compensation, professional development, career advancement, tenure, and redeployment or release.

- **All eight districts included in the study used similar rating systems for classifying a teacher's overall performance.** Administrators in two districts reported that early outcomes data suggested that the new teacher evaluation systems were succeeding in rating teachers across a wider range of performance levels than the districts' previous evaluation systems.

- **Six districts used or planned to use teacher evaluation scores to redeploy or release low-performing teachers but narrowly defined the circumstances under which this could happen.** Three districts, for example, released only non-tenured or new teachers on the basis of their evaluation scores whereas the decision to release tenured or veteran teachers required more evidence. Two districts opted to use the observation data and feedback conferences to counsel out or redeploy their ineffective teachers rather than release them outright.

- **Three districts linked — or planned to link — teacher evaluation scores to a career ladder, which sometimes included performance pay.** Only one district indicated in its design documents that it planned to use evaluation results to redistribute high-quality staff to low-performing schools.

Administering Evaluation Systems

Despite the complexity of these evaluation systems that used multiple classroom observations and measures of student performance to assess teacher professional practice, the majority of the districts included in the study created streamlined, modest administrative structures to manage the systems.

- **Six of the eight districts created simple administrative structures for their evaluation** systems, with, on average, five staff administering their teacher evaluation systems.

- **Four districts worked with outside contractors to create data management systems for their teacher evaluation data**, and two districts worked with outside contractors to train their classroom observers.

- **Hiring peer observers was among the most expensive features of the teacher evaluation systems** in the three districts that took this approach.

Perceived System Effects

Respondents at all levels reported perceiving positive effects of their new teacher evaluation systems, even in the partially implementing districts.

- **Administrators, principals, and teachers in every district included in the study reported perceiving positive effects of their new teacher evaluation systems, notably that observations and feedback help teachers' improve their professional practice.** No district administrator, however, was ready to claim that the teacher evaluation system — still in the early stages of implementation — had improved student achievement.

- **Principals in each of six districts reported that the new teacher evaluation system had caused them to want to be better instructional leaders.**

- **Teachers who participated in focus groups or individual interviews in every district reported that they believed that the classroom observations and subsequent feedback had helped them become better teachers.** Teachers in three districts, however, questioned whether it was possible to demonstrate excellence on the full range of competencies included in their respective districts' observation rubrics.

Summary

Developing a teacher evaluation system that uses multiple measures to assess teacher performance is a complex process that the eight districts included in this study approached in myriad ways. They designed teacher evaluation systems that varied widely in terms of the number of annual classroom observations they required, the types of assessments they used — and the way they used them — to measure teacher impact on student performance, the weight they ascribed to each measure, and the personnel decisions they tied to teachers' final ratings. The eight districts — even those considered to have fully implemented their teacher evaluation systems — continued to modify and fine-tune their systems in response to teacher feedback, pilot test data, and early implementation experiences. Indeed, all were designing and redesigning their evaluation systems as they were implementing them. Telephone updates in fall 2014 confirmed that these districts were continuing to learn about the functionality, practicality, and effectiveness of their teacher evaluation systems and modifying and adapting their designs accordingly.

I. Introduction

Comprehensive teacher evaluation systems are a core element of current state and local strategies to support improved teaching and learning in elementary and secondary education in the United States. When fully implemented, these systems can inform a range of personnel decisions, including decisions pertaining to teacher support and professional development, career advancement and tenure, and compensation. Comprehensive teacher evaluation systems have the potential to contribute to improved teaching practice by providing information about a teacher's strengths and weaknesses. Until very recently, however, almost all teacher evaluation systems in the United States had very limited capacity to identify either effective or ineffective teachers (Weisberg et al., 2009).

Today, efforts are underway across the country to transform teacher evaluation into a useful tool for improving teaching and learning. These efforts are supported by state statutes as well as by improvements in state and local data systems and the development of new student assessments. In 2015, the National Council on Teacher Quality (NCTQ) reported that only five states "have no formal state policy requiring that teacher evaluations take objective measures of student achievement into account in evaluating teacher effectiveness" (NCTQ 2015, ii). In addition, state and local efforts to design and implement teacher evaluation systems have been supported by federal resources appropriated through several programs, including the State Fiscal Stabilization Fund (under the *American Recovery and Reinvestment Act*), Race to the Top (RTT), the Teacher Incentive Fund (TIF), and the Investing in Innovation (i3) Fund. The 2015 enactment of the *Every Student Succeeds Act* (*ESSA*) maintains a federal priority of ensuring that all students have access to high-quality educators. States that elect to use Title II, Part A funds to develop, improve, or provide assistance to local educational agencies (LEAs) to design and support the implementation of teacher, principal, or other school leader evaluation must describe in their state plans how they will use student growth and other measures of educator performance to provide clear, timely, and useful feedback to teachers, principals, or other school leaders.

This study provides descriptive information on the design and early implementation of teacher evaluation systems in eight local school districts. This exploratory study is intended to help other districts and states learn from the experiences of these eight districts as well as to inform future research on the effects of teacher evaluation systems on teacher professional practice and student performance. It is important to note, however, that these eight districts were designing and redesigning their evaluation systems as they were implementing them and, therefore, will likely continue to modify and adapt them as they gain better insights into ways these systems can better serve district, principal, and teacher needs.

The study sample included four districts that were "fully implementing" their teacher evaluation systems at the time data were collected in 2012 and early 2013, and four districts that were considered to be "partially implementing" their systems. "Fully implementing districts" were defined as those that had used their teacher evaluation systems to evaluate every teacher in the district and tied teacher scores to specific results, including salary increases, bonuses, promotions, and other personnel decisions while "partially implementing districts" were those that had designed and pilot tested their teacher evaluation system components but had not yet begun to use them to evaluate every teacher in the district and/or had not yet tied teacher scores to specific outcomes. The eight sample districts were:

Fully Implementing Districts
District of Columbia Public Schools (Washington, DC)
Hamilton County Department of Education (Chattanooga, TN)
Harrison School District 2 (Colorado Springs, CO)
Hillsborough County Public Schools (Tampa, FL)

Partially Implementing Districts
Austin Independent School District (Austin, TX)
Pittsburgh Public Schools (Pittsburgh, PA)
Plattsburgh City Public Schools (Plattsburgh, NY)
St. Mary's County Public Schools (Leonardtown, MD)

The following discussion includes a description of the study design and its limitations as well as an overview of the case study districts.

Study Design and Limitations

This study sought answers to the following questions:

1. What key priorities and measures informed the design of new teacher evaluation systems?
2. What steps did the districts take prior to full implementation to test the system and prepare teachers and staff to implement it?
3. How did the districts structure and conduct the classroom observation component of their teacher evaluation systems?
4. How did the districts analyze student performance and other data to evaluate teacher performance?
5. How did the districts use, or plan to use, teacher evaluation results to make personnel decisions? To what extent were professional development and career advancement decisions tied to evaluation results?
6. What administrative structures and supports did districts use to support their new teacher evaluation systems?
7. What are the perceived early effects of the teacher evaluation systems on the professional practices of teachers, principals and district administrators?

Site Selection

Each teacher evaluation system selected for the study had to include the key features that all RTT grantees were required to include in their new teacher evaluation systems. Specifically, these systems:

- Relied on multiple measures of teacher performance, including growth in student achievement and observations of classroom practice, to assess individual teachers
- Distinguished among teachers at various levels of proficiency and effectiveness
- Included a formative component that provided timely feedback to teachers
- Used — or planned to use — results of the evaluation process to inform decisions about teacher professional development and support.

Sites were selected based on: (a) recommendations from members of the study's Technical Working Group and U.S. Department of Education officials; (b) an extensive and iterative search of materials available on state education agency and school district websites; (c) a review of the literature on recent trends in teacher evaluation as well as case studies of individual teacher evaluation programs; (d) conversations with district staff responsible for planning or managing design or implementation of new teacher evaluation systems. In addition to the RTT-related criteria, the eight systems included in the study met the following additional criteria:

- Considered gains in student learning as measured by a value-added model or some other growth model as a significant factor in assessing teacher performance
- Served as the district's system of record for evaluating teachers[2]
- Included provisions for professional development and other kinds of support intended to foster improvements in teachers' professional practice

Overall, 15 sites were invited to participate in the study and seven declined the invitation, citing concerns about the burden the study would place on district and school staff, other planned or ongoing research in their district, or a shifting teacher evaluation system design.

Data Collection

Data collection involved two- or three-person site visits to the districts between spring 2012 and winter 2013. The study team interviewed district staff who were responsible for and/or knowledgeable about the teacher evaluation system, including peer observers in Austin, the District of Columbia, and Hillsborough County,[3] as well as several key external stakeholders, including leaders and representatives of local teacher unions, principals, teachers (individually or in small groups), instructional facilitators and coaches, peer observers, and administrators in six state education agencies (SEAs).[4] The teachers who participated in individual or focus group interviews were, in most cases, teachers who participated in designing their district's teacher evaluation system (Exhibit 1).

Exhibit 1
Number of interview respondents, by type

Type of respondent	Number of respondents
District administrators	63
State administrators	9
Superintendents or deputy superintendents	8
Teacher unions or association representatives	13
Teacher union or association presidents	3
Peer observers	9
Principals	40
Teachers (i.e., does not include focus group participants)	31
TOTAL Number of interview respondents	176

Exhibit reads: Across the eight districts, 63 district administrators participated in site visit interviews.

In three of the four fully implementing districts included in the study (Harrison, Hamilton County, and Hillsborough County), the study team conducted teacher focus groups. The focus groups were

[2] Note that this consideration eliminates districts that operate separate evaluation- and performance-based compensation systems and teacher evaluation systems that operate at the school level without a district infrastructure to support them.

[3] "Peer observer" is a generic term we used to identify all observers districts hired from outside the school or district to observe and evaluate teachers. In Hillsborough County, peer observers are called "peer evaluators," in the District of Columbia, they are called "Master Educators," and in Austin, they are called "peer observers."

[4] The District of Columbia's Office of the State Superintendent of Education was not included nor was the Texas SEA because, as per the study's decision rule, neither organization was involved in the design or approval of teacher evaluation systems.

intentionally limited to the fully implementing districts in an effort to capture teacher perceptions of the entire system, including its uses with respect to teacher compensation, recognition, and retention.[5] The team conducted seven focus group sessions in both Harrison (46 teachers) and Hillsborough County (43 teachers), and three focus group sessions in Hamilton County (31 teachers), for a total of 120 teachers. The teacher focus groups varied in size across the three districts and included all types of teachers (Exhibit 2).

Exhibit 2
Characteristics of teacher focus group participants

School level	Number of teachers
Elementary	50
Middle	34
K–8	2
High School	34
TOTAL by school level	120
Subject area	**Number of teachers**
General (elementary teachers)	27
English or reading	13
Math	10
Science	10
Social studies	11
Career/technology	3
Special education	25
English Language Development	9
Other (e.g., foreign language, art, music, health)	12
TOTAL by subject area	120

Exhibit reads: Across the eight districts, 50 elementary school teachers participated in focus group interviews.

The district administrator with whom the study team consulted to plan and organize the site visits — usually the administrator directing the implementation of the district's teacher evaluation system — selected the teachers who participated in the individual or focus group interviews. District selections were based on site visitor guidance that teacher focus group participants should include special education teachers, teachers of English learners (ELs), and teachers *from at least two different schools at each level*, including elementary, middle, and high schools. **District administrators did not randomly select the focus group participants and the focus group participants' views may not represent those of all teachers in their respective districts.**

Because the study had a dual purpose of both capturing descriptive information on system features and conducting exploratory data collection on the development and implementation of teacher evaluation systems, site visitors did not use interview protocols like surveys. Some questions were systematically asked of interviewees with the same titles or responsibilities across districts, but other questions were open-ended in order to capture the most salient information based on the particular respondent's perspective and based on the district context. Appendix D includes the questions that interviewers posed to at least one district administrator in every district and questions posed to all principals and teachers who participated in site visit interviews. In addition, the appendix includes the questions

[5] One of the four fully implementing districts requested that the study team not conduct teacher focus groups due to the district's concerns about the time burden the focus group interviews would place on teachers.

interviewers posed to all state administrators who participated in telephone interviews. Similarly, teacher focus group facilitators followed a standard focus group interview protocol, posing a standard set of questions to each focus group (see Appendix D). The focus group protocol, however, included guidance that encouraged the facilitator to probe on issues that arose within each group, which sometimes resulted in the discussion diverging into issues that went beyond the scope of the protocol.

Finally, the study used extant data, including district reports and reports from external contractors and consultants, (a) to verify data collected on-site regarding the structure of the teacher evaluation systems, including districts' teacher performance standards or frameworks used to measure teacher professional practice, and (b) to verify the number of classroom observations and types of student assessments districts used to evaluate teachers. Appendix E includes a complete list of all the extant documents referenced for this study.

Data Analysis

This study employed qualitative analytic procedures and it adhered to a set of standards to limit bias and ensure reliable findings, including standards of evidence and triangulation of data. Site visitors participated in debriefings immediately following the completion of the site visits in the spring and fall of 2012 and the winter of 2013. After the debriefing, site visitors reviewed and coded their interview notes and transcript data (based on audio-recordings) and compiled them into a standardized site visit report template used to capture verbatim quotes and summarize key findings that addressed the study questions.

The project team also developed a spreadsheet to capture uniform interview data collected on specific program features, including component weights, number of observations, types of staff conducting observations, types of student achievement data used to measure teacher impact on student performance, and personnel decisions ensuing from evaluation results. Lead analysts reviewed these reports and spreadsheets in detail, noting text that needed clarification or elaboration or data that were questionable (e.g., more classroom observations conducted for experienced teachers than new teachers).

Lead analysts coded the site visit reports against classification headings or categories of activity from the interview protocols. Each analyst reviewed the data across districts, identifying common topics of interest (such as district administrators working with external organizations to administer their teacher evaluation systems, or principals struggling to incorporate multiple classroom observations into their daily routines). Once analysts identified the broad topics, they coded and re-sorted the data by those topics. The goal of the analysis was to compare, contrast, and synthesize findings and propositions from the individual cases to triangulate data and summarize the data across the eight participating districts.

With respect to categorizing and sorting the data, it is important to note that the original study design called for selecting and sorting districts based on two large analytic categories: districts that had fully implemented their teacher evaluation systems, and districts that had partially implemented or were piloting their teacher evaluation systems. Preliminary analyses revealed, however, that the eight districts did not look very different based on their implementation progress. That is, with the exception of tying personnel decisions to teacher evaluation ratings, which none of the partially implementing districts had yet done, all eight districts had been through the process of designing their teacher evaluation systems, and all had begun implementing those designs. The district experiences were not substantively different based on where they stood with respect to their implementation progress. It is possible that district administrators in partially implementing districts will encounter capacity struggles as they move into full implementation, but for principals and teachers, the analysis did not suggest that they will confront new or different struggles in these districts. Instead, the teachers and principals in the partially implementing districts who were interviewed for this study had experienced, for the most part, the full evaluation process and could speak to its benefits and drawbacks. With full implementation in these districts, the capacity of observers might change and the quality of their observations and subsequent feedback might also change, but there are still ample lessons to learn from these partial implementation experiences.

Study Limitations

This study has three main limitations. First, it relies on data collected from a purposively selected group of eight districts, thus limiting the range of potential successes and challenges districts might experience in designing and implementing complex teacher evaluation systems. Second, statements regarding program processes, challenges, and successes represent the perspectives of the individuals making them and may not represent the full range of views among district and school staff. Nevertheless, key administrators in all eight study districts reviewed this report for its factual accuracy. Finally, because this study describes the design and implementation of eight districts' teacher evaluation systems at a particular point in time (from late 2012 to early 2013), the evaluation systems will likely change and evolve. Indeed, district administrators reported that the longer they implement, the better their insights into their system's design and, accordingly, they were regularly modifying and adapting those designs to better serve district, principal, and teacher needs. Another factor that is likely to influence these districts' designs is state teacher evaluation policy, which may change in response to the implementation of *ESSA* and of assessments aligned with college and career-ready standards. Despite these limitations, however, the report offers some important insights into the design and implementation processes related to comprehensive teacher evaluation systems and highlights some potential early outcomes.

Overview of the Case Study Districts

Each teacher evaluation system selected for the study had to include the key features that all RTT grantees were required to include in their new teacher evaluation systems. Each system relied on multiple measures of teacher performance, including growth in student performance and observations of classroom practice. In addition, each districts' system distinguished among teachers at various levels of effectiveness and used—or planned to use—evaluation results to inform personnel decisions (Exhibit 3). Many contextual factors, including the following, affected the design and implementation of the eight district's teacher evaluation systems:

- Overall, the eight districts selected for the study had designed or implemented their evaluation systems within the last decade, with five of the eight districts commencing work since 2009. Three districts, however, commenced work on their teacher evaluation systems as early as 2007 and 2009.

- Every district included in the study was ahead of statewide efforts to design teacher evaluation systems and, consequently, some eventually had to adjust their evaluation systems to comply with new state expectations and requirements. Indeed, five of the eight districts' evaluation system designs were influenced by state legislation, including legislation passed to support RTT applications.

- Seven of the eight districts' efforts to design their teacher evaluation systems grew from their participation in federal programs such as TIF, RTT, and i3. In addition, two districts also benefitted from Empowering Effective Teachers grants from the Bill and Melinda Gates Foundation.

- Seven of the eight districts' teachers were represented by unions affiliated with the American Federation of Teachers, the National Education Association, or both. Seven districts included union representatives in the evaluation design process.

In addition to classroom observations and growth in student achievement, three of the eight districts included additional measures of teacher performance in their evaluations, including student surveys and professional expectations. Together, these measures were used to determine an overall rating of teacher performance. All eight districts in the study used four- or five-point scale rating systems to classify teacher performance and six defined the ways in which they intended to use the evaluation results to make personnel decisions.

Exhibit 3
Summary characteristics of eight districts' teacher evaluation systems in 2012 and 2013

	Year system was piloted or implemented	Weight for classroom observations	Weight for student performance	Weight for other criteria	Rating levels	Uses of evaluation results
Austin Independent School District	2011–12 (Pilot)	40%	40%	20% (10% Student survey and 10% Professional expectations)	• Unsatisfactory • Developing • Effective • Highly effective	To be determined
District of Columbia Public Schools	2009–10	40% for teachers in state tested grades and subjects 75% for others	50% for teachers in state-tested grades and subjects 15% for others	10% (Commitment to school community; possible deductions in core professionalism)	• Ineffective • Minimally effective • Developing • Effective • Highly effective	• Career ladder • Compensation • Remediation • Professional development • Retention or release
Hamilton County Department of Education	2011–12	50%	50%	NA	• Does not meet standards • Improvement necessary • Effective • Highly effective	• Professional development • Retention • Redeployment or release
Harrison School District 2	2007–08	50%	50%	NA	• Unsatisfactory • Progressing • Proficient • Exemplary • Master	• Career ladder • Compensation • Professional development • Remediation • Release
Hillsborough County Public Schools	2010–11	60%	40%	NA	1–5 (5 was the highest score)[a]	• Compensation • Professional development • Redeployment or release
Pittsburgh Public Schools	2009–10 (Pilot)	50%	35%	15% (Student survey)	• Failing • Needs improvement • Proficient • Distinguished	• Compensation • Professional development • Promotion • Tenure
Plattsburgh City Public Schools	2010–11 (Pilot)	60%	40%	NA	• Ineffective • Developing • Effective • Highly effective	• Peer assistance plans • Retention or release
St. Mary's County Public Schools	2011–12 (Pilot)	50%	50%	NA	• Ineffective • Developing • Effective • Highly effective	• Professional development • Remediation • Retention or release

Exhibit reads: Austin piloted its teacher evaluation system in 2011-12 and assigned weights of 40 percent for classroom observations, 40 percent for student performance, 10 percent for student surveys, and 10 percent for professional expectations. The four-point rating levels included unsatisfactory, developing, effective, and highly effective. The district had not yet determined how it will use teacher evaluation results.

[a] For state reporting purposes, Hillsborough County converted its 1-5 rating scale to the state's rating system so that a score of 1=Unsatisfactory; 2=Needs improvement; 3=Effective; and 4 and 5=Highly effective.

Source: Extant data on system specifications (see Appendix D for complete list).

Capsule descriptions of the eight teacher evaluation systems included in the study appear below. In addition, and for further reference, diagrams depicting the structure of each district's teacher evaluation system are included in Appendix B.

Austin Independent School District, Austin, TX

In 2011–12, the Austin Independent School District began piloting a new teacher evaluation system. The pilot test of the new system built on REACH, Austin's performance-based compensation system, which began in 2007–08. The goal of the teacher evaluation system was to improve instructional practice. The system measured five domains of teacher performance: student growth (Domain 1); instructional practice (Domain 2); classroom climate (Domain 3); student perception surveys (Domain 4); and professional expectations (Domain 5). Trained evaluators, including principals and peer observers, observed teachers on a combined minimum of four occasions**.** At the time of data collection, Austin had not yet determined how the evaluation results would inform human resource decisions, as it had not completed pilot testing the system.

District of Columbia Public Schools, Washington, DC

Launched in 2009, the purpose of IMPACT was to identify strong teachers; establish clear, high expectations and establish a common language about effective teaching; provide regular feedback and support to all teachers to help them improve; recognize and retain top-performers; and remove teachers who are ineffective in their roles. IMPACT scores were based on four components, including: (1) classroom observations; (2) student performance (through value-added measures and/or student learning objectives); (3) commitment to school community, and (4) professionalism. A typical teacher had five formal, unannounced classroom observations based on the Teaching and Learning Framework (TLF) rubric. The TLF, which was based on sources such as Charlotte Danielson's *Framework for Teaching* and Robert Marzano's *Classroom Instruction that Works*, assessed performance based on nine domains or "Teaches." Principals conducted three observations and external evaluators conducted two. The District of Columbia partnered with Mathematica Policy Research to build and analyze a value-added model to assess teachers' impact on students who take the District of Columbia Comprehensive Assessment System (DC-CAS) test. The district used IMPACT results to determine teacher progression along the district's career ladder, which included increased professional responsibility and opportunities. The district also used IMPACT data to make decisions about compensation, remediation, improving practice, and dismissal.

Hamilton County Department of Education, Chattanooga, TN

In January 2010, Hamilton County Department of Education organized a committee of school principals, district officials, teachers, and union representatives to develop a plan to evaluate teacher performance. Hamilton County's evaluation system, designed to improve instructional practice through coaching, featured classroom observations and teacher and school value-added measures (i.e., the Tennessee Value Added Assessment System or TVAAS) developed by the state. The classroom observation rubric, Project COACH, was drawn from a teacher observation rubric developed by a New Leaders for New Schools' consultant, Kim Marshall. The rubric includes 40 components across six domains. Principals and assistant principals were responsible for conducting the 10-minute teacher observations. Teachers with a professional license were observed six times and teachers of all other license types (i.e., apprentice, transitional, etc.) were observed eight times during the school year. Throughout the year, teachers were encouraged to collect evidence demonstrating that they had met the performance level criteria associated with all six domains of professional practice. Results of the evaluation were used to plan and provide teacher professional development and make decisions related to teacher probation, remediation, release, tenure, and contract renewal.

Harrison School District 2, Colorado Springs, CO

In 2007, the then-superintendent of Harrison School District Two began developing a new teacher evaluation system to hold teachers accountable for high standards and provide them with feedback and incentives to improve. The system used classroom observation and student achievement data to determine teachers' ratings. The classroom observation rubric addressed seven standards of teaching informed by the district's curriculum map as well as frameworks developed by Charlotte Danielson and Robert Marzano. Student achievement data included the Colorado Student Assessment Program (CSAP),[6] curriculum-based measures or semester exams, quarterly district assessments, and teacher-determined student learning objectives. Educators were observed formally, by school administrators, at least once per year and informally, by either school administrators or an external evaluator, up to sixteen times per year for probationary teachers and up to eight times per year for experienced teachers. Performance ratings were used to guide decisions related to dismissal, remediation, compensation, and distinguished teacher review.

[6] In 2012, Colorado implemented the Transitional Colorado Assessment Program (TCAP) to reflect changes to the state-adopted academic content standards.

Hillsborough County Public Schools, Tampa, FL

In 2009, Hillsborough County Public Schools received a $100 million grant from the Bill and Melinda Gates Foundation, which it used, in part, to design a teacher evaluation system aimed at supporting teacher professional growth. In 2010–11, Hillsborough County launched its teacher evaluation system without a pilot test. System components included classroom observations and student achievement data, including the Florida Comprehensive Assessment Test (FCAT) and end-of-course assessments. The observation rubric was a modified version of Charlotte Danielson's *Framework for Teaching* and assessed 22 competencies and more than 70 required elements. Formal and informal (introduced in 2011–12) observations were conducted from four to eleven times a year by principals and peer observers. Hillsborough County varied observations based on previous evaluation ratings and experience. A value-added score was provided for each teacher based on analyses of a model developed by the Value-Added Research Center at the University of Wisconsin. Any teacher who received an "unsatisfactory" rating two years in a row was eligible for dismissal. These teachers received individualized support and had one school year to show improvement.

Pittsburgh Public Schools, Pittsburgh, PA

In partnership with the Pittsburgh Federation of Teachers (PFT), the Pittsburgh Public Schools developed the Research-Based Inclusive System (RISE) in 2009 to support the district's plan to improve teacher practice and recognize highly effective teachers. The Professional Growth System (PGS), which was part of RISE, featured three components: (1) classroom observations, conducted by school administrators or instructional teacher leaders, (2) teacher and school value-added measures, and (3) the TRIPOD student perception surveys, developed by Harvard University researcher, Ronald Ferguson. The classroom observation instrument, based on Charlotte Danielson's *Framework for Teaching*, addressed 24 dimensions across four domains of teaching. Novice teachers were observed at least eight times per year and experienced teachers were observed four times, although observations varied based on previous evaluation ratings and tenure status. Observations were conducted by principals, assistant principals, district administrators, or lead teachers. Student achievement data included the Pennsylvania System of School Assessment (PSSA), district-designed end-of-course assessments, and, where appropriate, the PSAT. Pittsburgh initiated a partnership with Mathematica Policy Research to develop and analyze teacher- and school-level VAMs. The results of the teacher evaluation system were intended to inform professional development planning and decisions on teacher retention, dismissal, promotion, and compensation.

Plattsburgh City School District, Plattsburgh, NY

In 2010, Plattsburgh City School District began developing a new teacher evaluation system as a member of a consortium convened by the American Federation of Teachers (AFT). Representatives from five districts in New York State convened regularly to design a new teacher evaluation system for districts that sought to improve professional practice by providing teachers with feedback on their performance. A 2010 federal i3 grant secured by the AFT supported this planning work. The teacher evaluation system included analyses of instructional practice and student performance. Plattsburgh used the Teacher Evaluation and Development (TED) framework from New York State United Teachers (NYSUT, the state AFT chapter) as a rubric for classroom observations. The framework, influenced by Charlotte Danielson's *Framework for Teaching*, assessed teacher performance using seven domains and 33 components. School administrators and external observers conducted one formal and one informal observation per teacher per year. Student performance was measured based on the state test and locally-assessed measures. Plattsburgh planned to incorporate student learning objectives as part of its teacher evaluation system. Results from the new teacher evaluation system were intended to help the district target professional development for teachers. Teachers who performed poorly on their evaluations were to be referred to the Peer Assistance and Review Panel for intensive supports and monitoring. Failure to improve could lead to termination.

St. Mary's County Public Schools, MD

Until 2010, teacher evaluation in St. Mary's County relied only on Charlotte Danielson's *Framework for Teaching*, which the district adopted in 2002 to evaluate instructional practice. Following direction from the state to comply with Maryland's RTT requirements for teacher evaluation, St. Mary's County began considering ways to incorporate student achievement data into its teacher evaluation system. Those interviewed for the study, including central office staff, principals, and teachers, described the purpose of the evaluation system as supporting professional development and improving instruction to increase student achievement. The evaluation consisted of two classroom observations a year for experienced teachers and four observations per year for new teachers; teachers were evaluated by school administrators on 21 components of the Danielson framework. St. Mary's County used several types of student achievement data to assess teacher performance, including summative assessments, formative assessments, performance assessments, and classroom performance. In 2011–12, five schools participated in the pilot test of a modified teacher evaluation system, and in 2012–13, the district expanded the evaluation system to include all schools for a "no fault" year. The 2013–14 school year was to be the first year of full implementation of the teacher evaluation system. Teacher evaluation scores took into account how "complexity factors" associated with a particular student or class, such as student mobility or family problems, could adversely influence student performance. The district expected to use the results of the teacher evaluation system primarily for identifying and addressing teacher professional development needs but would release ineffective novice teachers on the basis of an ineffective rating.

II. Designing an Evaluation System

Many choices and tasks go into the design of a complex teacher evaluation system that relies on multiple measures of teacher performance to assess teacher effectiveness. Determining district purposes for developing a new teacher evaluation system is usually the first step in the design process. Among the most central design tasks for districts in this study was designing observation rubrics to measure classroom practice and planning how to collect and analyze student achievement data to assess teacher impact on student performance. Other important steps in system design included involving stakeholder groups, particularly teacher unions, in the process, and assigning weights to each measure of teacher performance.

Key Findings

- Respondents in all eight districts agreed that the foremost goal of their teacher evaluation systems was to improve instruction. In each of the districts, central office staff, observers, and teachers alike expressed the view that the district's evaluation system was focused on improving student achievement through improved instruction. In four of the districts, some of the teachers interviewed said that they, or their fellow teachers, believed that the systems were also designed to identify and remove ineffective teachers.

- District administrators in all eight districts reported that principals and teachers made important contributions to the design of their teacher evaluation systems. Teachers, principals, district administrators, school board members, or representatives of local universities and teacher unions and associations were among the stakeholder groups the eight districts consulted when designing their teacher evaluation systems.

- The classroom observation rubrics used or developed in all eight study districts addressed common areas of teaching practice. The commonalities among districts' observation rubrics did not happen by chance, however. Charlotte Danielson's *Framework for Teaching* (FFT) was the exclusive basis for the design of the observation rubrics in half the districts included in the study. The other four districts developed their own rubrics to examine instructional practice, but drew from existing frameworks such as the FFT and Robert Marzano's *Classroom Instruction that Works* as sources of reference.

- All eight districts used multiple assessments to measure teachers' influence on their students' performance. In addition to the state test, districts used district-developed assessments, end-of-semester or end-of-course assessments, performance assessments, and standardized assessments, including DIBELS, PSAT, IB, and AP exam scores. Using multiple types of assessments was a common district strategy that addressed a three-fold purpose: (1) accommodating teachers of non-state tested grades and subjects; (2) offering all teachers other opportunities to demonstrate effectiveness through, for example, district curriculum-based assessments; and (3) guarding against one assessment being the primary determinant of teacher performance ratings.

- To account for the many types of assessments and data sources used to measure teacher impact on student performance, all eight districts assigned separate weights to each type of student assessment included in their evaluation systems. In addition, seven of the eight districts ensured that the overall weight (i.e., either 40 or 50 percent of a teacher’s score) applied to student performance data for teachers of non-state tested grades and subjects and teachers of state-tested grades and subjects was the same.

District Purposes for Developing New Evaluation Systems

Improving instruction was identified as the foremost goal of the evaluation systems.

In Hillsborough County, for example, central office staff, union leaders, and peer evaluators alike expressed the view that the district’s evaluation system was focused on improving student achievement through improved instruction. A union representative explained, “The whole idea was to have a system that helped everybody grow wherever they were in their career.” Two classroom observers echoed this statement, stating that the evaluation system was designed with teacher professional practice in mind.

In Hamilton County, too, respondents in different roles agreed that instructional improvement was the purpose. A district leader said that Project COACH was designed to “assist [teachers] in getting better, not necessarily to score people.” An education association official agreed “That’s our primary goal: better instruction, stronger instruction for our students.” The dialogue surrounding teacher evaluation began shortly after principals attended an annual district retreat in January 2010. The retreat featured a presentation by Kim Marshall, a former long-time principal and consultant for New Leaders for New Schools, on evaluating instruction. Principals and district staff explained in interviews that Marshall’s presentation on evaluating instruction sparked interest among principals to invest in teacher evaluation in order to improve instruction. As described later in this report, a design team comprising principals, teachers, central office staff, and the Hamilton County Education Association (HCES) developed the observation tool, which aimed to support teachers through targeted coaching.

Teachers in each of the eight districts — through individual and focus group interviews — described their perceptions regarding district purposes for developing new teacher evaluation systems. A sample of those views include the following:

> *So I think the purpose of any teacher evaluation system is two-fold: (1) ensure that there are excellent teachers in every single classroom, and (2) ensure that we have high expectations for what it means to be an excellent teacher. So [the teaching framework] serves as a guiding document for what teachers should focus on when they're in the classroom to maximize the time that they have with students [Teacher, District of Columbia].*
>
> *The way I see the [teacher evaluation system] is it really is more of a tool that will help teachers learn more about what they can improve on and what they're doing well at. And it includes a self-assessment [tool] as well as provides an opportunity for the principal or administrator to give you feedback. So it seems to be really much more of a learning tool than previous evaluations [Teacher, Hamilton County].*
>
> *I would say that [the goal] is to monitor the effectiveness of teachers, while at the same time, providing an avenue for offering direct feedback on where they can improve [Teacher, Plattsburgh].*

> *The new system gives the teacher an opportunity to be more reflective about our practice and craft, and it provides us with more common language when sitting with fellow teachers and administrators. It allows us to focus on our strengths and weaknesses. It makes the communication smoother [Teacher, Pittsburgh].*

While there was widespread agreement that teacher evaluation systems were intended to improve instruction, some teachers in four districts said that they, or their fellow teachers, believed that the systems were also designed to identify and remove ineffective teachers.

That is, some teachers interviewed in the District of Columbia, Harrison, Hillsborough County, and Plattsburgh said they believed that among the purposes of the new teacher evaluation systems was to provide objective data that districts could use to redeploy or release ineffective teachers. Union representatives in Plattsburgh said that they knew of teachers who viewed the new evaluation system as a "gotcha system" aimed at removing teachers. A teacher in another district explained that while she herself believed that the goal of the evaluation system was to help teachers improve practice, she recognized that her fellow teachers worried that the system's primary goal was to dismiss teachers. She explained, "I think transparency is key. [Our district] is having a really hard time shaking the [perception] that [the teacher evaluation system] is a tool to fire teachers."

Federal and other external funding sources helped seven of the eight districts lay the foundation for their teacher evaluation systems.

Seven of the eight district's efforts to design their teacher evaluation systems grew from their participation in federal programs such as the TIF, RTT, and the i3. Hillsborough County and Pittsburgh also benefitted from multi-million dollar Empowering Effective Teachers grants from the Bill and Melinda Gates Foundation. In addition to supporting the design and implementation of a teacher evaluation system, the Gates Foundation grant also supported the establishment of a new teacher induction program; enhancing professional development for teachers, a principal evaluation system; a revamped teacher compensation plan, and incentives for teachers to work with highest need students.

While the District of Columbia's teacher evaluation system benefitted from eventual support received from the federal RTT and TIF grants and from the Gates Foundation, the design of IMPACT began as an internal district initiative spearheaded by former chancellor, Michelle Rhee. That is, DC IMPACT was a district initiative rather than a district response to federal or foundation-funded programs.

Because every district included in the study was ahead of statewide efforts to design teacher evaluation systems, some districts eventually had to adjust their systems to comply with new state expectations and requirements.

As states competed for RTT awards, several passed legislation that established requirements and set expectations for teacher evaluation in their state. Six states relevant to this study and the District of Columbia were awarded RTT grants,[7] and districts in three affected states (Hamilton County, TN; Pittsburgh, PA; and St. Mary's County, MD) revised existing or adopted new evaluation components in response to requirements established by new legislation. Tennessee's First to the Top legislation, for example, required that teachers be evaluated annually using qualitative data, value-added student

[7] Race to the Top winners relevant to this study include: Tennessee (Round 1), the District of Columbia (Round 2), Florida (Round 2), Maryland, (Round 2), New York (Round 2), Colorado (Round 3), and Pennsylvania (Round 3).

growth, and a teacher-selected student achievement goal. To meet its state's requirements, Hamilton County adopted the Tennessee Value-Added Assessment System (TVAAS) to calculate student growth.

Conversely, Hillsborough County and Plattsburgh received approval from their states to use their existing evaluation systems without adaptation or revision to comply with state requirements. In Harrison, where Colorado had just passed legislation establishing requirements for teacher evaluation systems at the time of the site visit for this study, administrators did not expect that state requirements would affect the work they were doing because Colorado law provides for local control. As one administrator explained: "Provided you're within the guidelines of the law/expectations, you can develop any evaluation you want as long as it's reasonable and appropriate for teachers."

Pittsburgh similarly had developed elements of its teacher evaluation system ahead of statewide efforts. In June 2012, the Pennsylvania legislature passed Act 82 of 2012, which established requirements for teacher evaluation in the state. The bill specified that teacher ratings must be based 50 percent on teacher observations and 50 percent on student performance; districts must also include multiple measures of teacher performance to reach a single rating. At the time the bill was passed, Pittsburgh's evaluation design team had already developed a model to evaluate effectiveness that was mostly in-line with the state's guidelines. Pittsburgh submitted its model to the state education agency for approval for use beginning in 2013–14.

Finally, to comply with new state requirements, St. Mary's County agreed to add a student growth and classroom performance measure to its 10-year-old teacher evaluation system that had been based solely on classroom observations.

Involving Stakeholder Groups

Each of the eight districts' teacher evaluation system design teams were primarily staffed and led by district administrators. When designing their teacher evaluation systems, however, all eight districts consulted several stakeholder groups, such as teachers, principals, district administrators, school board members, local universities, and teacher unions and associations; parents, however, were not among them (Exhibit 4).

All eight districts involved some or all of their stakeholder groups directly in the design of their teacher evaluation systems.

District administrators in six of the eight districts described forging ongoing collaborations with many of their stakeholder groups when designing their teacher evaluation systems. Of its design process, a district representative from Pittsburgh said, "It was a very inclusive process. I would say that all of the key individuals were involved in planning [the evaluation system]."

Exhibit 4
Type of stakeholder groups involved in or consulted by the design teams for the teacher evaluation systems, by district

	District officials	Principals	Teachers	Teachers' union	Local school board	Other
Austin	✓	✓	✓	✓	✓	Austin Chamber of Commerce
District of Columbia	✓[a]	✓	✓			
Hamilton County	✓[a]	✓	✓	✓		
Harrison	✓[a]	✓	✓		✓	
Hillsborough County	✓[a]	✓	✓	✓	✓	Local university faculty; community members
Pittsburgh	✓	✓	✓	✓		
Plattsburgh	✓[a]	✓	✓	✓		Instructional facilitators/ coaches
St. Mary's County	✓	✓	✓	✓		Local university faculty and instructional facilitators/ coaches

Exhibit reads: Austin involved district officials, principals, teachers, the teacher union, the local school board, and the Austin Chamber of Commerce in the design work associated with their teacher evaluation system.

[a] Includes the district superintendent.

SOURCE: Site visit interviews; extant data on system specifications (see Appendix D for complete list).

Similarly, district administrators in St. Mary's County consistently described their design process as a collaborative one and did not believe that any groups or individuals were excluded. One central office administrator, for example, explained that the superintendent's vision for all projects is collaboration among the central office, principals, teachers, and education associations.

> *We build things with [stakeholders]. When we revise and renovate each year, we bring them to the table. They're pretty much at the table for everything. … I think that we've learned a really good lesson that could help a lot of others. If you can work with people, you don't have to work around people. It's been a rewarding experience. It's not always easy, but it's important.*

In Austin, a district administrator explained that the process created to design the teacher evaluation system was intentionally inclusive and collaborative and that no stakeholder group was left out:

> *We invited pretty much everybody to participate. So the invitation was on the table. We had the performance office, the research and evaluation office, the curriculum office, and [others]. I mean, we had pretty much all of the different offices that touch students in any way at the table, along with teachers and principals and [the teacher union], so I don't think there was anybody that was left out.*

In particular, Austin drew upon its strong collaborative relationship with the Austin Chamber of Commerce in an effort to learn about private sector practices for evaluating job performance. A district administrator explained the rationale behind their weekly meetings with Chamber representatives:

> *So the Chamber was important. We want their support, but their feedback is important too because they do a lot of things that we're thinking about. So, if we can use their best practices to inform what we're doing, we absolutely want to do that…We found that we can come to the table and explain things, and then there's sort of that give and take. We learn from them. They learn from us…It's a beneficial relationship to have*.

In St. Mary's County, district administrators included school leaders on the design committee who represented the variety of schools within the district, including large and small elementary, middle, and high schools, a Title I school, and a regional special education school. One principal explained the thinking this way: "A middle school, a high school, a large elementary school, a small elementary school. They really wanted to get a good representation that reflects the diversity of the schools in our system."

Hillsborough County's design team provided opportunities for teachers, principals, union representatives, and other stakeholders to make suggestions and present ideas, but acknowledged that the opinions of each stakeholder group did not carry equal weight. For example, the district engaged representatives from local colleges and universities, but one district administrator explained that their contributions did not carry as much weight as teachers and principals. As another district administrator in Hillsborough County explained, "I think we never could have done this without including teachers and [principals], but I just … don't necessarily believe that [parents, university faculty] are the best to design a program like this."

Principals and teachers made important contributions to the design of their teacher evaluation systems, according to district administrators in all eight districts.

All eight districts involved teachers in the design process, either by inviting teachers to join design committees or to provide feedback on the proposed design through focus groups and/or surveys. District staff in six districts (Austin, District of Columbia, Hamilton County, Hillsborough County, Pittsburgh, and St. Mary's County) suggested that teacher and principal input carried significant weight in the design of their respective evaluation systems. A district administrator in Austin, for example, said that she had deferred to teachers and principals over the opinions of all other stakeholders when designing the teacher evaluation system:

> *I still rely, quite honestly, on my principals and my teachers. I don't care to weigh heavier toward the union, nor toward the business people. When I have really outstanding teachers working on an issue with me, and outstanding principals, I'm going to go with them virtually every time.*

In St. Mary's County, school principals and teacher representatives from each of the pilot schools contributed to the design of the student growth model; teacher input brought a variety of "complex" issues to light and provided the group with items to consider. To identify teacher effectiveness measures for early childhood education, for example, one principal explained that the process of selecting growth measures was driven by teachers: "It is really teacher initiated — asking teachers and supervisors who know the content and know the assessments to come up with a model that really shows growth for those different roles." A second principal described focus groups that were convened to solicit teacher feedback as the design team developed the student performance component of the evaluation system:

> *At the various points, we would bring a group of teachers [to meetings]. I would either do volunteers or I'd selectively pick. I'd want to get one [music] teacher, one PE teacher,*

> *a sixth grade special education teacher, and the media specialist so I had a variety, depending on what feedback we wanted.*

Similarly, district staff in the District of Columbia, Hamilton County, and Hillsborough County sought feedback from teacher focus groups and a teacher advisory group. The design team developing the classroom observation rubric in Hamilton County, for example, solicited feedback on the rubric by surveying more than 2,000 teachers in the county. From the feedback, the design team decided to reduce the number of teaching standards from 60 to 40 as teachers reported that there were too many standards for which they should be held accountable, so the district collapsed some standards and deleted others. Hillsborough County and the District of Columbia included peer evaluators as classroom observers in response to teacher feedback — solicited by both districts — about principal bias. As is discussed later in this report, the peer observers were introduced, in part, to offer an unbiased, objective opinion of a teacher's performance that was unrelated to any personal connections to the school. In addition, however, the District of Columbia also introduced peer observers in response to teachers' interest in having an opportunity to receive content-specific feedback from outside experts.

The Role of the Teacher Unions

> **Administrators in six districts reported that their districts benefited from consulting their teacher unions or associations when designing their teacher evaluation systems.**

District administrators reported working closely with these groups to add or modify specific features of their evaluation systems' design. Austin, Hamilton County, Hillsborough County, Pittsburgh, Plattsburgh, and St. Mary's County involved their teacher unions in the evaluation system design process. According to one district representative in Hillsborough County, for example, representation from the teacher union gave teachers a voice in the design process and led to the decision to add peer observers to the evaluation system design: "The entire idea of peer [observers] was strictly from the union. We thought that it was the greatest thing since sliced bread, but it was their idea, their involvement." Similarly, a union representative (referred to as the Education Association) in Hamilton County reported that the union had worked "hand-in-glove" with the district to design the teacher evaluation system and that the close working relationship continues: "If I have a teacher who calls with a problem (i.e., he/she can't access last year's evaluation scores), I can just email [the district] and I get a rapid response to take care of that."

In Pittsburgh, collaboration was a consistent theme across interviews with union leaders, community members, central office staff, and teachers who participated in the design process. Indeed, several respondents across the aforementioned groups highlighted the collaborative nature of the design process as district leaders engaged union leaders and teachers at every step. One central office administrator, for example, described the teacher evaluation system as an example of a "productive collaboration" between the district and teacher union. Union leaders, in particular, were actively involved in designing the classroom observation instrument by recruiting schools to participate in the pilot and debriefing teachers on their experiences in using the instrument. Knowledge that the union was involved in the process encouraged teacher buy-in, as one teacher explained, "The fine details [of the evaluation system] are still being worked out, and it's been done hand-in-hand with the teacher union and the teachers. We really feel like everyone has had a stake and an opportunity to share their opinions."

In St. Mary's County, the local teacher union had begun to consider using student performance data in teacher evaluations to comply with the state's RTT plans and requirements, and a joint team of central office staff, principals, teachers, and union leaders played important roles in developing the student performance component of the teacher evaluation system. The design process began with a series of meetings during which each of the groups worked through the domain components to develop a complete vision of how it would work.

Overview of System Designs

The basic design of each of the eight districts' teacher evaluation systems was the same. Each system relied on multiple measures of teacher performance, including observations of classroom practice and measures of student performance, to assess individual teachers. The following describes the basic features of the classroom observation and student performance components of the eight districts' teacher evaluation systems.

Classroom Observation Rubrics

Designing a classroom observation rubric to measure teacher professional practice is a critical task in the development of a teacher evaluation system and one to which all eight districts devoted thought and time. In developing their classroom observation rubric, districts weighed the value of modifying or adapting an existing framework of teaching to assess teaching practice against developing a new framework tailored to their particular district. Ultimately, every district's rubric focused on several broad domains of professional practice, and seven of the eight districts broke those domains into multiple — and measurable — elements or standards of practice.

All eight districts used or developed observation rubrics that addressed common areas of teaching practice, including instruction and classroom environment.

For example, analyses across the eight observation rubrics revealed that all eight addressed instructional practice in some way. The District of Columbia's observation rubric, for example, defined instructional practice as a teacher's ability to "lead well-organized objective-driven lessons, explain content clearly; develop higher-level understanding through effective questioning; maximize instructional time; check for student understanding; respond to student understanding; develop high-level understanding; maximize instructional time; and build a supportive, learning-focused classroom." St. Mary's County's observation rubric defined instructional practice as a teacher who "communicates clearly and accurately; uses higher order questioning and discussion techniques; engages students in learning; uses assessments in instruction; and demonstrates flexibility and responsiveness" (Exhibit 5).

Similarly, all eight districts' observation rubrics measured classroom environment, which, among other things, included the extent to which teachers created environments of respect and rapport, established a culture for learning, and managed classroom procedures appropriately. In addition, six of the eight districts' observation rubrics addressed teacher professionalism (i.e., Austin and the District of Columbia measured teacher professionalism separately from classroom observations[8]), such as "participating in a

[8] In Austin and the District of Columbia, teacher professionalism is measured separately from classroom observation data. In the District of Columbia, for example, principals track teacher attendance, tardiness, compliance with school guidelines, and interactions with colleagues and twice a year, assess teachers' overall professionalism based on these measures. In addition,

professional community and growing and developing professionally (Pittsburgh) and “communicating with stakeholders” (Hillsborough County). (See Appendix C for samples of each of the eight districts’ classroom observation rubrics.)

teachers’ “Commitment to the School Community” (CSC) is also measured separately based on five competencies (DCPS, 2012, p. 46).

Exhibit 5
Types of classroom observation rubrics used to measure teacher professional practice, by district, in 2012 and 2013

District	Type of observation rubric	Domains of Professional Practice	Number of standards of practice
Austin	District-developed	1. Instructional practice 2. Classroom climate [a]	13
District of Columbia	District-developed *Teaching and Learning Framework*	1. Lead well-organized objective-driven lessons 2. Explain content clearly 3. Engage students at all learning levels 4. Provide students multiple ways to move toward mastery 5. Check for student understanding 6. Respond to student understanding 7. Develop high-level understanding 8. Maximize instructional time 9. Build a supportive, learning-focused classroom [b]	30 [c]
Hamilton County	Based on Kim Marshall's teaching framework with some modification	1. Planning and preparation 2. Classroom management 3. Delivery of instruction 4. Monitoring, assessment, and follow-up 5. Family and community 6. Professional responsibilities	40
Harrison	District-developed, but informed by Charlotte Danielson's *Framework for Teaching* (FFT) and Robert Marzano's *Classroom Instruction that Works*	1. Preparation for instruction 2. Use of data to inform instruction 3. Delivers quality instruction 4. Intervenes to meet diverse needs 5. Classroom environment 6. Leadership 7. Professionalism	27
Hillsborough County	*Framework for Teaching*	1. Planning and preparation 2. Classroom environment 3. Instruction 4. professional responsibilities	22
Pittsburgh	*Framework for Teaching* with some modification	1. Planning and preparation 2. Classroom environment 3. Instruction 4. Professional responsibilities	24
Plattsburgh	Developed through i3 state consortium of districts, but based on *Framework for Teaching*	1. Acquire knowledge of each student, demonstrate knowledge of student development, promote achievement for all students 2. Know content; plan instruction to ensure growth for all students 3. Implement instruction that engages and challenges all students to meet or exceed learning standards 4. Work with all students to create a dynamic learning environment that supports achievement and growth 5. Use multiple measures to asses and document student growth, evaluate instructional effectiveness, and modify instruction 6. Demonstrate professional responsibility and engage relevant stakeholders to maximize student growth, development, and learning 7. Set informed goal and strive for continuous professional growth	33
St. Mary's County	*Framework for Teaching*	1. Planning and preparation 2. Classroom environment 3. Instruction 4. Professional responsibilities.	21

Exhibit reads: Austin's classroom observation rubric was district-developed and included two domains of professional practice, including instructional practice and classroom climate. The rubric included 13 components within the two domains.

a: Principals or appraisers rate teachers' professionalism; the measure is not considered part of the formal classroom observation rubric.

b: Although the Teaching and Learning Framework (TLF) also included a planning component and a data-based decision making component, neither component was included in teachers' observation scores in 2012–13. In addition, the District of Columbia measures teacher professionalism and commitment to the school community separately, assigning these measures separate scores and weights as part of a teacher's overall evaluation score.

c: While the District of Columbia disaggregates its nine domains/standards or "Teaches" into 30 subparts or standards, they do not assign separate scores to each of these parts.

SOURCE: Site visit interviews; extant data on system specifications (see Appendix D for complete list).

Charlotte Danielson's Framework for Teaching (FFT) was the exclusive basis for the design of the observation rubrics in half of the districts included in the study.

The other four districts' rubric designs were influenced by Danielson and other existing observation rubrics, such as Robert Marzano's *Classroom Instruction that Works.* The observation rubrics developed in Hillsborough County, Pittsburgh, and St. Mary's County were informed by Charlotte Danielson's FFT. A fourth district, Plattsburgh, based much of its design on this framework.

Charlotte Danielson's *Framework for Teaching* defines "what teachers should know and be able to do in the exercise of their profession" (Danielson, 2013). Elements of teaching practice are divided into 22 standards clustered into four domains: (1) planning and preparation, (2) classroom environment, (3) instruction, and (4) professional responsibilities. Exhibit 6 provides an example of one domain and two of the five teaching standards within this domain. For each standard, the framework provides a rating scale, definitions for each rating level, and critical attributes of each standard. For example, for Domain 2, "Classroom Environment," Standard 2a, "Creating an environment of respect and rapport," a teacher given an "Unsatisfactory" rating does not call students by their names, while a teacher given a "Distinguished" rating has a personal rapport with each of his or her students and inquires about events in students' lives outside of school, such as extracurricular activities or hobbies.

Four districts developed their own rubrics to examine teacher professional practice, but drew heavily from existing frameworks as sources of reference.

Austin, the District of Columbia, Hamilton County, and Harrison did not adopt an existing framework but developed or adapted their own classroom observation rubrics. Austin's working group, for example, consulted the research literature as well as consulted system developers in Hillsborough County, the District of Columbia, and Pittsburgh. They also closely examined the FFT and the frameworks and rubrics included in the Bill and Melinda Gates Foundation-funded *Measures of Effective Teaching* study (MET, 2012). According to one district administrator, consulting a large number of sources helped the design team determine what practices would best align with the district's priorities:

> *I have a whole binder of [all the things we looked at]; we looked at everything that you can imagine. We pulled information from [districts'] websites; we looked at their rubrics, we looked at how they organized their systems. We actually called and talked to DC and had them talk to us about [peer observers] and what that process looks like. So we tried to reach out and get as much information that we could.*

Similarly, when developing its Teaching and Learning Framework (TLF), the District of Columbia's IMPACT design team consulted 20 existing frameworks, including the FFT*,* Robert Pianta's *Classroom Assessment Scoring System*, Teach for America's *Teaching as Leadership*, and Texas' TxBESS Framework. Hamilton County's evaluation design team drew primarily from the work of Kim Marshall, described earlier in the report, in developing its classroom observation rubric.[9] Finally, in Harrison, while the concept of a new teacher evaluation system came from the district's superintendent, the district's assistant superintendent led efforts to develop the classroom observation rubric based on best practices

[9] Marshall's observation rubric examines 60 standards of teaching across six domains of teachers' job performance, including: (1) planning and preparation; (2) classroom management; (3) delivery of instruction; (4) monitoring, assessment, and follow-up; (5) family and community outreach; and (6) professional responsibilities.

drawn from the FFT as well as from Robert Marzano's *Art and Science of Teaching Framework for Effective Instruction.*

Exhibit 6
Sample domain and standards of practice from Charlotte Danielson's *Framework for Teaching*: Domain 2a–b, the Classroom Environment

Domain 2: The Classroom Environment	Unsatisfactory	Basic	Proficient	Distinguished
2a: Creating an environment of respect and rapport	▪ *Teacher uses disrespectful talk towards students. Student body language indicates feelings of hurt or insecurity.* ▪ *Students use disrespectful talk towards one another with no response from the teacher.* ▪ *Teacher displays no familiarity with or caring about individual students' interests or personalities.*	▪ *The quality of interactions between teacher and students, or among students, is uneven, with occasional disrespect.* ▪ *Teacher attempts to respond to disrespectful behavior among students, with uneven results.* ▪ *Teacher attempts to make connections with individual students, but student reactions indicate that the efforts are not completely successful or are unusual.*	▪ *Talk between teacher and students and among students is uniformly respectful.* ▪ *Teacher responds to disrespectful behavior among students.* ▪ *Teacher makes superficial connections with individual students.*	In addition to the characteristics of "Proficient" ▪ *Teacher demonstrates knowledge and caring about individual students' lives beyond school.* ▪ *When necessary, students correct one another in their conduct towards classmates.* ▪ *There is no disrespectful behavior among students.* ▪ *The teacher's response to a student's incorrect response respects the student's dignity.*
2b: Establishing a culture for learning	▪ *The teacher conveys that the reasons for the work are external or trivializes the learning goals and assignments.* ▪ *The teacher conveys to at least some students that the work is too challenging for them.* ▪ *Students exhibit little or no pride in their work.* ▪ *Class time is devoted more to socializing than to learning.*	▪ *Teacher's energy for the work is neutral: indicating neither a high level of commitment nor "blowing it off."* ▪ *The teacher conveys high expectations for only some students.* ▪ *Students comply with the teacher's expectations for learning, but don't indicate commitment on their own initiative for the work.* ▪ *Many students indicate that they are looking for an "easy path."*	▪ *The teacher communicates the importance of learning, and that with hard work all students can be successful in it.* ▪ *The teacher demonstrates a high regard for student abilities.* ▪ *Teacher conveys an expectation of high levels of student effort.* ▪ *Students expend good effort to complete work of high quality.*	In addition to the characteristics of "Proficient": ▪ *The teacher communicates a genuine passion for the subject.* ▪ *Students indicate that they are not satisfied unless they have complete understanding.* ▪ *Student questions and comments indicate a desire to understand the content, rather than, for example, simply learning a procedure for getting the correct answer.* ▪ *Students recognize the efforts of their classmates.* ▪ *Students take initiative in improving the quality of their work.*

Exhibit reads: For Domain 2, Standard 2a: "*The Classroom Environment: Creating an environment of respect and rapport,*" teachers are rated "Unsatisfactory" if they use disrespectful talk towards students, if students use disrespectful talk towards one another with no response from the teachers, and/or if teachers display no familiarity with or caring about individual students' interests or personalities.

SOURCE: Danielson, C. (2013).

None of the eight districts tailored their observation rubrics to different grade-level or subject-area teachers.

The District of Columbia, for example, designed its TLF to be content neutral so that it could be applied to almost any teacher, regardless of their grade or subject. The TLF included nine domains/standards or "Teaches" that the design team believed represented high-quality instruction. In developing the nine

"Teaches," district administrators agreed that the practices should be familiar to all teachers. A district administrator stressed that the framework was not requiring any new practices, but was simply fundamental principles of good teaching:

> *We didn't see this [framework as setting] a brand new set of expectations for teaching. Philosophically, we just believe that there are some core tenets to good teaching. For example, you need to have a clear objective for what you're doing each day, whether it's kindergarten or AP physics. You want to push kids to think deeply, you want to probe for a deeper understanding.*

The District of Columbia, however, did adapt its TLF for Early Childhood Education (ECE) teachers (i.e., preschool, pre–Kindergarten, Kindergarten, and Special Education ECE teachers). That is, the nine "Teaches" were the same but were modified slightly to "better reflect best practices in early childhood settings" and to include "specific descriptors for effective group meetings and centers" (DCPS 2012). In addition, Special Education Autism teachers were evaluated based on the Autism Teaching Standards,[10] which were based on "Applied Behavior Analysis methodology" and "define excellence for autism teachers" (DCPS 2012).

Administrators in two districts described the importance of scrutinizing and testing the language of the observation rubric to ensure that it reflected clear, achievable, and measurable practices.

One principal in Hamilton County explained that the changes the district made to the language in its observation rubric were the result of piloting the instrument in the classrooms:

> *I was on the committee that created the rubric, but then as a principal, I'm sitting down to score them and I'm like, wait a minute, nobody could get a four on this. There was one standard for behavior that said something like, to be highly effective, 'Students would think it unthinkable to ever misbehave.' And I thought, wait a minute, I'm not scoring the teacher, I'm scoring what the students are thinking. We took that out.*

The design team in the District of Columbia described having gone through a deliberative and iterative process to design their observation framework, paying careful attention to the phrasing of each dimension to make sure it was clear and measuring the professional practices the district valued, as one district administrator explained:

> *We realized that words matter. We had words like 'rigorous' and 'compelling,' and 'invested in the learning,' but they were not really right. There were a lot of these weighted words that were vague, but you felt like they were [headed in] the right direction. The words were troublesome to people. We eventually cleared them all out and described more acutely what we meant.*

The administrator went on to describe how the design team used to debate the difference between 'a few' and 'some' and realized that individual definitions would always vary: "It's hard to say what your 'a

[10] Target and Track Learning Goals at Each Student's Level; Provide Frequent Opportunities to Practice and Demonstrate Skills; Promote Rigor and Improved Responding at Each Student's Level; Implement Instruction to Foster Development of Social and Communication Skills; Provide Instructive Feedback for Incorrect Responses and Adjust Instruction; Maximize Instructional Time through Organized Routines, Procedures, and Pacing; Reinforce Behaviors to Promote Engagement and Responding; Respond Consistently and Appropriately to Challenging Behaviors; Provide a Structured and Supportive Learning Environment

few' and my 'a few' and your 'some' and my 'some' are," he explained. Their solution was to assign values to the ratings, so that, for example, a teacher would receive a score of "4" if she had only one to three instances of misbehavior, a three if she had five to seven instances of misbehavior, and so forth. After some reflection, the district determined that it did not want classroom observers counting instances of certain behaviors, "So we moved away from that very specific kind of rubric."

Student Performance and Student Perception Data

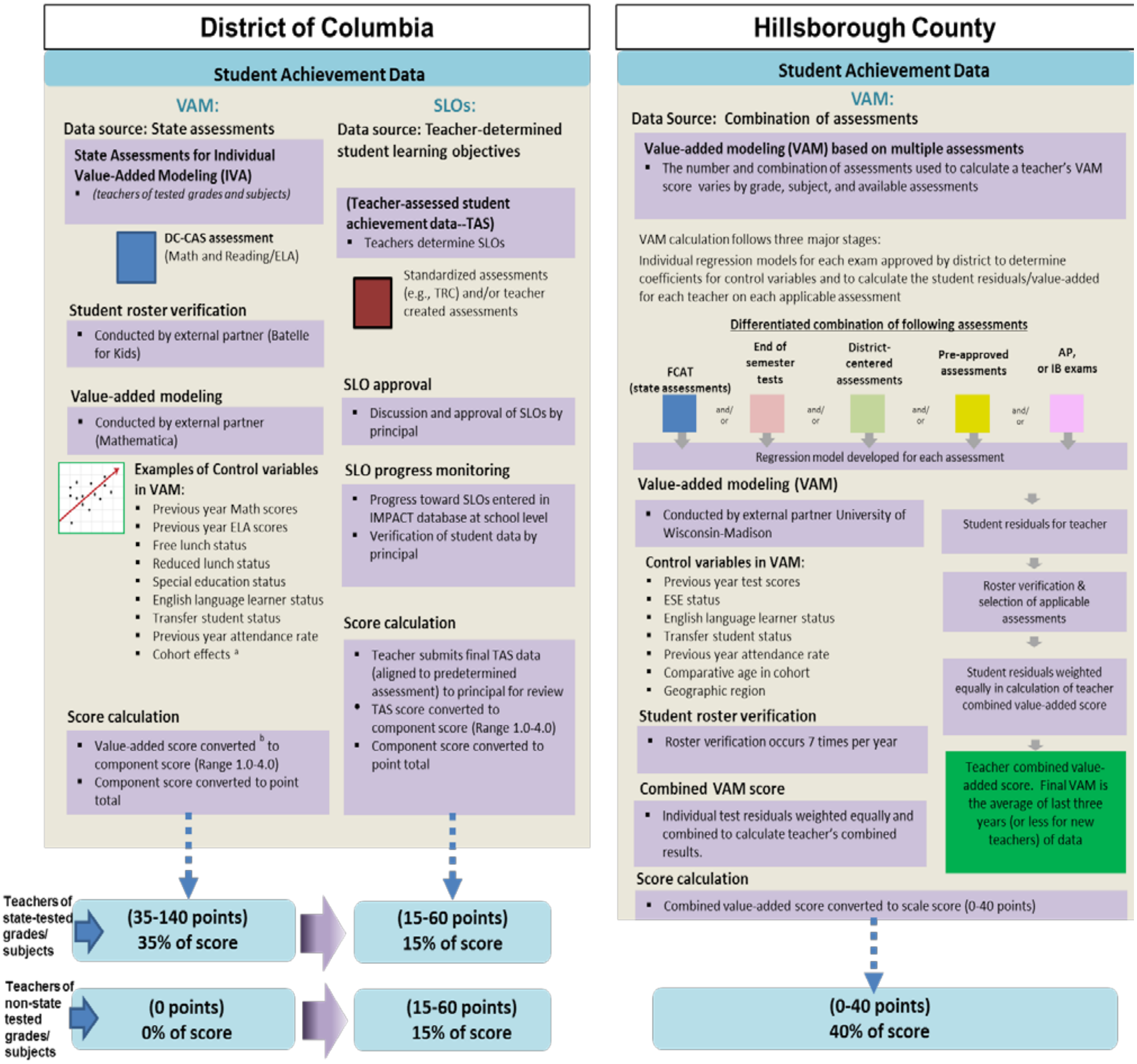
All eight districts used multiple assessments to assess teachers' influence on their students' performance.

A review of districts' evaluation system design and guidance documents revealed, and interviews with district administrators confirmed that, in addition to the state test, districts used district-developed assessments, end-of-semester or end-of-course assessments, performance assessments (e.g., art portfolios), and standardized assessments, including DIBELS, PSAT, IB, and AP exam scores. Using multiple types of assessments was a common district strategy that addressed a threefold purpose: (1) accommodated teachers of non-state tested grades and subjects; (2) offered all teachers other opportunities to demonstrate effectiveness through, for example, district curriculum-based assessments; and (3) guarded against one assessment being the primary determinant of teacher performance ratings.

The number of assessments districts used to evaluate teacher performance, nevertheless, varied widely. Exhibit 7 illustrates the variation in the number of assessments two districts used to measure teacher impact on student performance. The District of Columbia's system, for example, included only its standardized state test (the DC-CAS) and teacher-determined Student Learning Objectives (SLOs)[11] to measure teacher impact on student performance. Hillsborough County's system, by contrast, included state assessments, curriculum-based assessments, district assessments, and other external student assessments (AP and IB exams) to assess teacher impact on student performance. Complete illustrations of all eight districts' teacher evaluation system components can be found in Appendix B. The following discussion describes other types of student-level data districts used to assess teacher performance.

[11] The District of Columbia calls them "Teacher-Assessed Student Achievement Data" or "TAS."

Exhibit 7
Student achievement measures included in the District of Columbia and Hillsborough County teacher evaluation systems in 2012–13



District of Columbia

Student Achievement Data

VAM:

Data source: State assessments

State Assessments for Individual Value-Added Modeling (IVA)
- *(teachers of tested grades and subjects)*

DC-CAS assessment (Math and Reading/ELA)

Student roster verification
- Conducted by external partner (Batelle for Kids)

Value-added modeling
- Conducted by external partner (Mathematica)

Examples of Control variables in VAM:
- Previous year Math scores
- Previous year ELA scores
- Free lunch status
- Reduced lunch status
- Special education status
- English language learner status
- Transfer student status
- Previous year attendance rate
- Cohort effects [a]

Score calculation
- Value-added score converted [b] to component score (Range 1.0-4.0)
- Component score converted to point total

SLOs:

Data source: Teacher-determined student learning objectives

(Teacher-assessed student achievement data--TAS)
- Teachers determine SLOs

Standardized assessments (e.g., TRC) and/or teacher created assessments

SLO approval
- Discussion and approval of SLOs by principal

SLO progress monitoring
- Progress toward SLOs entered in IMPACT database at school level
- Verification of student data by principal

Score calculation
- Teacher submits final TAS data (aligned to predetermined assessment) to principal for review
- TAS score converted to component score (Range 1.0-4.0)
- Component score converted to point total

Teachers of state-tested grades/ subjects: (35-140 points) 35% of score; (15-60 points) 15% of score

Teachers of non-state tested grades/ subjects: (0 points) 0% of score; (15-60 points) 15% of score

Hillsborough County

Student Achievement Data

VAM:

Data Source: Combination of assessments

Value-added modeling (VAM) based on multiple assessments
- The number and combination of assessments used to calculate a teacher's VAM score varies by grade, subject, and available assessments

VAM calculation follows three major stages:

Individual regression models for each exam approved by district to determine coefficients for control variables and to calculate the student residuals/value-added for each teacher on each applicable assessment

Differentiated combination of following assessments

FCAT (state assessments) and/or End of semester tests and/or District-centered assessments and/or Pre-approved assessments and/or AP, or IB exams

Regression model developed for each assessment

Value-added modeling (VAM)
- Conducted by external partner University of Wisconsin-Madison

Control variables in VAM:
- Previous year test scores
- ESE status
- English language learner status
- Transfer student status
- Previous year attendance rate
- Comparative age in cohort
- Geographic region

Student roster verification
- Roster verification occurs 7 times per year

Combined VAM score
- Individual test residuals weighted equally and combined to calculate teacher's combined results.

Score calculation
- Combined value-added score converted to scale score (0-40 points)

Student residuals for teacher

Roster verification & selection of applicable assessments

Student residuals weighted equally in calculation of teacher combined value-added score

Teacher combined value-added score. Final VAM is the average of last three years (or less for new teachers) of data

(0-40 points) 40% of score

Exhibit reads: The District of Columbia used Value-Added Models (VAMs) and Student Learning Objectives (SLOs) to measure teacher impact on student performance.

[a] Cohort effects variables consist of 1) class's average test score from the previous year; 2) Percentage of the class eligible to receive free or reduced price lunch; and 3) extent of variation in the student's scores from the previous year.

[b] Regression models calculate the average likely scale score of students of a District of Columbia teacher serving a similar student population [by controlling for Value Added Model (VAM) variables]. The individual value-added or IVA score is the scale score difference between the average likely score and the average actual scores of the students taught by the teacher (e.g., average actual score of 64.0; average likely score of 60.0 = an IVA raw score of +4.0). The raw IVA score is then converted to an IMPACT score using Math and Reading conversion tables approved annually by the District of Columbia Office of State Superintendent of Education (e.g., Raw IVA math score of +4 = 3.3 component score).

SOURCE: Site visit interviews; extant data on system specifications (see Appendix D for complete list).

Student Learning Objectives

Seven of the eight study districts included student learning objectives (SLOs) among the outcome measures used to evaluate teacher impact on student performance.

SLOs are targets of student growth that teachers set at the start of the school year and strive to achieve by the end of the school year. Teachers set these targets based on a review of students' baseline skills and after consulting with administrators, usually the principal. For example, one Austin teacher's SLO established that by the end of the school year, 75 percent of her second-grade math class would score at 70 percent or higher on a test she designed measuring students' ability to use patterns to describe relationships and make predictions. In Austin, teachers identified and assessed two SLOs for all of their students each year.[12]

School-level Performance Data

Discussion was ongoing among administrators about whether and how to incorporate *school-level* performance data into districts' teacher evaluation systems.

While five of the eight districts (Austin, Hamilton, Harrison, Pittsburgh, and Plattsburgh) included or planned to include school-level performance in their teacher evaluation systems, they debated the extent to which teachers should be held accountable for school-level growth. One Austin district administrator reported that, although the school-level measure was designed to encourage teacher collaboration, many of the stakeholder groups, including Austin central office staff, principals, teachers, Education Austin,[13] and the Association of Texas Professional Educators, questioned whether it would. The District of Columbia, meanwhile, removed the school-level value-added score (SVA) from its teacher evaluation system. District administrators explained that a school-wide value-added score did not sufficiently differentiate among the performances of teachers within a school and created a disincentive for teachers to work in the lowest-performing schools. In addition, district administrators explained that the original purpose of the school-level VAM was to encourage school-based collaboration and a sense of shared commitment to student outcomes, which they later determined was a construct better encouraged through another component of the teacher evaluation system which measures teachers' "commitment to school community."

Student Perception Surveys

Student perception surveys were among the measures used to evaluate teacher impact on student performance in two districts. [14]

Austin's student course evaluation was designed to provide teachers with students' perceptions of their classroom climate and instructional practice. The survey asked students to rate their teachers on 10

[12] SLOs were not new in Austin; the district had been using SLOs in an earlier iteration of their teacher evaluation system.

[13] The teacher unions in Austin, Texas.

[14] Findings from the MET Project (2012b) showed that combining student survey data with classroom observation data improved correlations with student achievement data. The authors believed that the results of their study provided strong evidence for combining student survey data with structured observation data in "the grades and subjects where student achievement gains are not measured" (MET Project 2012b, p. 3). The MET Project authors noted that their research provided

domains identified as indicators of high-quality teaching, including student engagement, checking for students' understanding, and establishing rigorous academic expectations, to name a few. The survey was administered to all pre–K through 12th grade students. District staff in Austin piloted the survey in three schools in the spring of 2012 and added the survey to evaluations for all teachers in its 12 pilot schools the following school year. The results of the student survey accounted for 10 percent of a teacher's final evaluation score.

To capture student perceptions, Pittsburgh uses the Tripod Student Perception Survey, which was developed by Harvard University researcher, Ronald Ferguson. The survey asked students to rate their teachers on seven constructs, including care, control, clarify, challenge, captivate, confer, and consolidate (see Exhibit 8 for a sample item). The district used the surveys for three grade bands: K–2, 3–5, and 6–12. The district piloted the Tripod surveys in 2011–12 and planned to implement them as part of their teacher evaluation system in 2012–13. The district intended to weight teacher performance on the student survey at 15 percent of a teacher's final evaluation score.

Exhibit 8
Sample TRIPOD Survey Item (7C, Care)

Elementary School Version	Response Options for Grades 3–5	Secondary School Version	Response Options for Grades 6–12
I like the way my teacher treats me when I need help.	• No, never • Mostly not • Maybe, sometimes • Mostly yes • Yes, always	My teacher in this class makes me feel s/he really cares about me.	• Totally untrue • Mostly untrue • Somewhat • Mostly true • Totally true
My teacher is nice to me when I ask questions.		My teacher seems to know if something is bothering me.	
My teacher in this class makes me feel that s/he really cares about me.		My teacher really tries to understand how students feel about things.	
If I am sad or angry, my teacher helps me feel better.			
The teacher in this class encourages me to do my best.			
My teacher seems to know if something is bothering me.			
My teacher gives us time to explain our ideas.			

Exhibit reads: The TRIPOD survey gives students in grades three through five the option of responding "No, never," "Mostly not," "Maybe, sometimes," "Mostly yes," or "Yes, always" to the prompt: "I like the way my teacher treats me when I need help."

SOURCE: Measures of Effective Teaching Project (2012a).

Weighting System Components

Determining the relative weight or value assigned to each component of a teacher evaluation system has implications for how teachers are ultimately rated. Study findings suggest that the eight districts

support for the use of multiple measures in teacher evaluation (e.g., classroom observation instruments and student surveys), especially with teachers in non-tested grades and subjects. In their words, "We found that combining any of the observation instruments with the Tripod student survey data improved correlations with student achievement gains . . . in subjects such as social studies, history, and science, where the nature of effective instruction is plausibly similar to math and ELA, the combination of observations and student survey data could substitute for student achievement gains (2012b, pp. 57-58).

included in this study assigned similar weights to — or similarly valued — the various components of their teacher evaluation systems.

For four of the districts included in the study, state legislation strongly influenced district decisions related to how they weighted each evaluation component of their teacher evaluation systems.

Hamilton County's weighting system, for example, complied with Tennessee's *First to the Top* legislation, which specified that student achievement should account for 50 percent of a teacher's final rating. Within the student performance component, 35 percent of a teacher's score should be based on the TVAAS score (or school-level performance) and 15 percent on an additional measure of student performance. Similarly, Plattsburgh's weighting system complied with the state's policy on teacher evaluation in which 60 percent of a teacher's score is based on practice and 40 percent is based on student achievement.

Pittsburgh's teacher evaluation design team had already begun the process of determining the weights for its system components when the Pennsylvania legislature passed House Bill 1901 in 2012. The new state legislation specified that teacher evaluation ratings should be based 50 percent on teacher practice and 50 percent on student achievement, including (1) 15 percent on teacher-level data, (2) 15 percent on school-level data, and (3) 20 percent on teacher-determined data (i.e., SLOs). The Pittsburgh design team modeled a weighting approach based on the state guidelines but explored seven additional models to determine which achievement measures were most within a teacher's control. After testing each model with student achievement data, a review committee determined that 30 percent of the achievement component weight should be based on teacher-level scores from VAMs (i.e., individual VAMs) or SLOs, 15 percent on student surveys, and 5 percent on school-level growth on the state assessment. In the end, the design committee gave the school-level data less weight than prescribed by the state legislation. Central office administrators submitted this weighting model to the Pennsylvania Department of Education (PDOE), which approved the model for the 2013–14 school year.

The Maryland State Department of Education required districts to assign a 20 percent weight to student performance on the state test in order to be in alignment with what the state proposed in its RTT application. One central office administrator in St. Mary's County explained that the designers of St. Mary's County's teacher evaluation system did not want one assessment to be the primary determinant of teacher performance, and consequently, the district tried to convince the state to let it maintain a 10 percent weight for student performance on the state test. The state, however, ultimately denied the request.

In all eight districts, classroom observation and student performance data were weighted roughly the same, each accounting for approximately 40 to 50 percent of a teacher's overall evaluation score.

For classroom observation scores, the weights varied very little for six of the eight districts included in the study (Hamilton County, Harrison, Hillsborough County, Pittsburgh, Plattsburgh, St. Mary's County), where each assigned a weight of either 50 or 60 percent to teachers' classroom observation scores.[15] The remaining two districts, Austin and the District of Columbia,[16] weighted classroom observation

[15] Hamilton County did not determine the size of the weights for classroom observation scores or student performance data. Rather, the district assigned the weights based on the requirements of the Tennessee Department of Education.

[16] For teachers of non-state tested grades and subjects, classroom observation scores were assigned a weight of 75 percent.

scores at 40 percent of a teacher's overall evaluation score. However, these districts also assigned separate weights of 20 and 10 percent, respectively, to other teacher evaluation criteria, such as student surveys and professional expectations in Austin and teacher commitment to the school community in the District of Columbia (Exhibit 9).

For student performance data, the weights also varied very little for seven of the eight districts (Austin, District of Columbia, Hamilton County, Harrison, Hillsborough County, Plattsburgh, St. Mary's County). Each of the seven districts assigned weights of 40 or 50 percent to teachers' student performance data. While Pittsburgh's weight of 35 percent for student performance data was lower than the majority of districts included in the study, it also assigned a separate weight of 15 percent to student surveys. The District of Columbia assigned a total weight of 50 percent for student performance data for teachers of state-tested grades and subjects. For teachers of non-state tested grades and subjects, however, the weight was 15 percent.

To account for the many types of assessments and data sources used to measure teacher impact on student performance, seven of the eight districts assigned separate weights to each type of student assessment included in their evaluation system.[17]

For example, Austin assigned weights of 30 percent to SLOs[18] data and 10 percent to school-level student performance data. In addition, seven of the eight districts (all but the District of Columbia) ensured that the same overall values of the weights were applied to student performance data for teachers of non-state tested grades and subjects (non-TSTGS), and teachers of state-tested grades and subjects (TSTGS). For example, in Hamilton County, for TSTGS, a weight of 35 percent was applied to teacher-level state test results and 15 percent for SLOs. For non-TSTGS, a weight of 35 percent was assigned to school-level state test results and 15 percent for SLOs. In St. Mary's County, to offset the 10 percent weight assigned to state testing data for TSTGS , the district reduced the weight assigned to their SLOs by 10 points so that neither TSTGS nor non-TSTGS' final scores would be based on student performance data that was weighted more than 50 percent.

[17] Hillsborough County does not assign separate weights to each of the assessments used to measure teacher impact on student performance. Rather, each assessment is weighted equally in a teacher's final VAM score.

[18] Described in greater detail in Chapter IV.

Exhibit 9
Teacher evaluation system components and their assigned weights in 2012 and 2013, by district

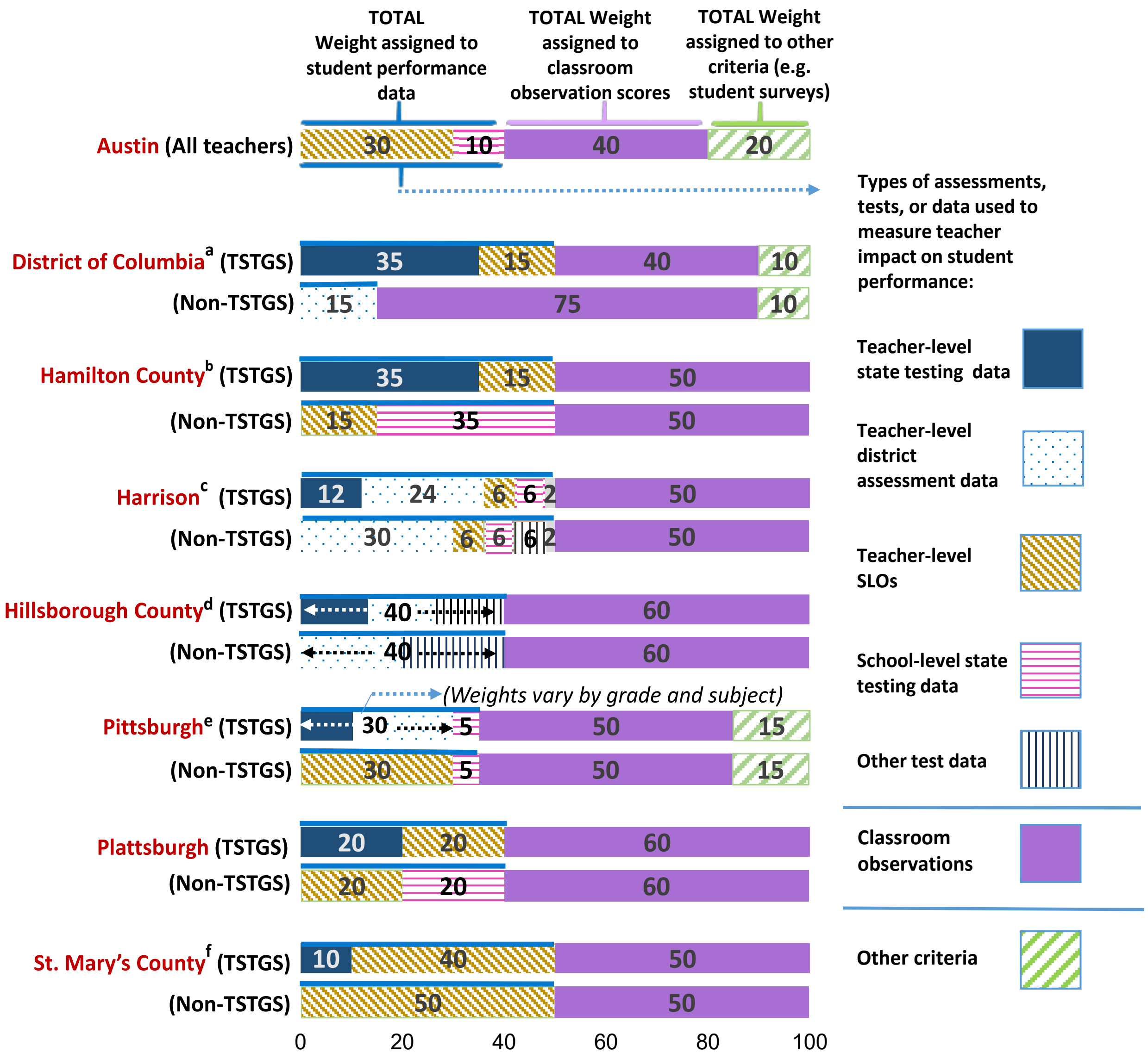


Exhibit reads: Austin assigned a weight of 40 percent to student performance data, 40 percent to classroom observation scores, and 20 percent to other teacher performance criteria.

*TSTGS=Teachers of state-tested grades and subjects.

[a] Classroom observation scores for special education and itinerant EL teachers were weighted at 65 and 85 percent, respectively.

[b] School-level assessment data could be the state assessment or student growth on an alternate assessment such as ratings of art portfolios

[c] Every teacher had a multi-part achievement template based on their grade, discipline, or specialty. Every teacher received two baseline points.

[d] The number of assessments used in a teacher's final VAM score varied by grade and subject, but each assessment was weighted equally.

[e] School-level test data could include other assessments such as end-of-course assessments and/or the PSAT.

[f] The district divided weights among four types of data, including formative district assessments, quarterly performance assessments by content area, student growth for a selected group of students, and classroom performance based on student grades.

SOURCE: Site visit interviews; extant data on system specifications (see Appendix D for complete list).

The District of Columbia was the only district among the eight included in the study that had assigned significantly divergent weights to classroom observation and student performance data for TSTGS (DC-CAS reading and math assessments administered in grades 4–8[19]) compared with teachers of non-TSTGS. That is, the weight assigned to classroom observation data for TSTGS was 40 percent compared with a weight of 75 percent applied to the classroom observation scores of non-TSTGS. The weights were similarly unequal for student performance data, with a weight of 50 percent assigned to student performance data for TSTGS compared with a weight of 15 percent for student performance data for non-TSTGS.

Administrators in the District of Columbia have recognized the concerns about the uneven weights and made some adjustments, including adding end-of-course exams and other outcome measures for the evaluations of non-TSTGS in the middle and high school grades. In addition, it reduced the weight assigned to scores based on the state test (from 50 to 35 percent of a teacher's IMPACT score) and added SLOs[20] to the performance measures used to evaluate teachers of state-tested grades and subjects. These changes were made in response to teacher concerns about the size of the weight assigned to their value-added scores in light of the fact that the scores were not made available until July, while their jobs and salaries hung in the balance:[21]

> *I think we are just uncompromising in our belief that how students perform — how they learn and grow — is the most important measure of a teacher's job [performance]. But we have been listening to teachers and are compelled by their concerns. While we still strongly believe in value-added, we realize that even our best teachers were feeling anxious about [value-added] since the data come back late and since the [individual value-added] IVA data do not capture all the learning that takes place over the course of the year. We decided to pull IVA back a bit and expand our teacher-assessed student achievement data (TAS) measure so that additional measures of achievement could be captured as well.*
>
> *[District Administrator, District of Columbia]*

The administrator also acknowledged that in the light of the fact that the IVA scores do not inform teacher practice by identifying ways in which teachers might improve instruction during the school year, reducing the weight of the score seemed a reasonable adjustment to make.

[19] Calculating an individual value-added (IVA) score requires student test scores from the prior year. Consequently, although third and tenth graders are administered the DC-CAS, their teachers are not included in the IVA analyses because neither third nor tenth graders have prior scores on which to measure achievement gains (neither second nor ninth grades are tested).

[20] We use the term "SLO" generically here. The District of Columbia's official term is "Teacher-assessed student achievement data" or TAS.

[21] An administrator explained that, for planning purposes, the district provides teachers with their potential score range in June, so that teachers know whether there is any possibility that they might be dismissed once their IVA scores are calculated. If their potential score does not fall within the dismissal range, they know that they are not at risk of termination. If their potential score does fall into the dismissal range, then teachers know to start looking for another job, and the school knows that it needs to look into finding a replacement.

Deliberations among Austin's stakeholder groups illustrated the challenges design teams confronted in trying to strike a balance between how to measure and weight the scores of teachers in the state-tested grades and subjects versus those of the teachers in the non-state tested grades and subjects.

The director of REACH explained that, although the working group agreed to include multiple measures of teacher performance in the evaluation system, it took time to settle on the appropriate weights to assign those measures. The working group began the process by looking into how other states weighed each component of their teacher evaluation system. For student achievement, Austin's working group struggled with issues of equity and fairness because not all teachers would have a value-added score: "When we talked about 20 percent [for value-added scores], there were a lot of people in the room who didn't want it to be 20 percent. They wanted it to be five percent." In the end, the working group settled on a 10 percent weight for the school-level VAM.

The full set of weights Austin established were the following:

– Individual SLO	20%
– Team SLO	10%
– School-wide student growth	10%
– Professional expectations	10%
– Student surveys	10%
– Classroom observations	40%

A central office administrator explained that the decision to assign the largest weight to individual SLOs was based on the belief that it was the area over which teachers exercised the most control:

> *[The debate] was partly about what a teacher has control over and what's the purpose of appraisal? Ultimately, we tried to balance the two, scoring the individual teacher versus using the school-level score as an opportunity to encourage the kind of behaviors [in schools] that we wanted to see.*

III. Early Implementation

The early implementation phase—pilot testing the system, introducing it to teachers, and training classroom observers — varied across the eight districts, yet may be a critical juncture in determining a district's successful transition to a new, comprehensive teacher evaluation system. For example, whether districts pilot tested their evaluation systems and assessed their validity may affect teachers' perceptions of the extent to which the system accurately assesses their professional practice. In addition, offering teachers information and training about the new evaluation system may help to facilitate the early implementation process by answering common questions, confronting common misconceptions, and encouraging participation.

This chapter begins by describing how districts tested their new evaluation systems and used the test results to modify or fine-tune their system designs. Several districts, in fact, made changes — sometimes significant ones — to their evaluation systems design based on the pilot test results. In addition, this chapter describes how the eight districts introduced their new evaluation systems to their teachers and how teachers, in turn, responded. Finally, the chapter addresses how districts trained classroom observers using the observation rubric in order to assure high-quality and internally consistent observations.

Key Findings

- Teacher and principal pilot test input carried significant weight in the design of four districts' evaluation systems.

- Three of the eight districts did not pilot test their teacher evaluation systems before implementing them. Administrators in these districts explained that holding teachers accountable as early as possible was important for achieving real, measurable change in teaching and learning.

- Each of the eight districts introduced their teachers to the new evaluation system through one or more strategies, including distributing written materials, creating online resources, or using school-based teams of evaluation systems experts. Teachers in three districts suggested that training should be ongoing because it sometimes took years to fully understand and respond to the complexities of their respective districts' teacher evaluation systems.

- All eight districts provided some type of observer training. Six of the eight districts used sample videos to help observers compare and contrast observer scores based on a common set of observed lessons. In an effort to ensure the highest quality training, three districts relied on the expertise of external organizations to train their observers rather than developing their own internal training capacity. Two districts conducted ongoing norming exercises to better ensure inter-rater reliability, and five districts trained observers on ways to provide constructive feedback to teachers.

Pilot Testing the System

Several studies of districts adopting teacher evaluation systems (Sartain, Stoelinga, and Brown 2011; the MET Project 2010; Milanowski 2004) showed that they tended to introduce their systems or individual components through pilot tests, allowing new practices to be introduced into relatively low-risk, low-stakes environments. Among the eight districts included in the study, five pilot tested their teacher evaluation systems prior to implementation. The following section describes the results of those pilot tests and their influence on system design, the rationale surrounding three districts' decisions not to pilot test their evaluation systems prior to implementation, and the varied efforts of districts to test the validity and reliability of their systems in evaluating teacher effectiveness.

Pilot Testing

All four of the partially implementing districts reported having piloted their teacher evaluation system and having made adjustments — sometimes significant ones — based on those results. In contrast, only one of the four fully implementing districts (Hamilton County) reported having piloted its teacher evaluation system prior to full implementation.

Teacher and principal pilot test input strongly influenced central office decisions regarding system modification in four districts, according to district administrators.

In Austin, Hamilton County, Pittsburgh, and St. Mary's County, teachers and principals played a significant role in system design and implementation. St. Mary's County, for example, piloted the teacher evaluation system in five schools; the principals and one teacher from each of the pilot schools were part of the district's planning committee for the roll-out of the teacher evaluation system. In particular, the planning committee worked on defining the student achievement component of the evaluation system. According to one district administrator, getting teacher buy-in and input early in the process and piloting the system in five schools were key steps to system design. For example, school principals and a teacher representative from all five pilot schools took active roles in developing the model for measuring teacher impact on student growth. The groups met at least once a month and used focus groups to talk with a variety of teachers in different roles about the types of data they would use to show the growth of their students. This proved to be an important step in the development process.

Austin's pilot test resulted in some modifications to its evaluation system, including removing student growth VAMs and adding student surveys, as described above. In Pittsburgh, during its pilot of the RISE classroom observation instrument, teams of Pittsburgh's central office staff and union representatives conducted site visits to 20 of its pilot schools to learn about implementation. During two and a half hour-long visits, teams interviewed principals and teachers about their experiences using the classroom observation instrument. The visits then continued with the pilot test of the student survey. As a union representative explained, "[The visits were] very encouraging because the teachers opened up. It was a good exchange and we could gather a lot of data." The design team later used the feedback from principals and teachers to make adjustments to the evaluation system, including adjusting administration procedures so that student surveys would be administered in two classes rather than one for each secondary teacher, thus capturing a wider, and theoretically better, representation of the many students who teachers work with at the secondary school level. In addition, the district agreed to lengthen the survey administration period to allow more time for students to complete the survey and to account for student absences.

In Hamilton County, five principals and central office administrators reported valuing the opportunity to pilot the classroom observation component, with four of these respondents underscoring the point that the pilot allowed principals and central office administrators to troubleshoot the system. As one district administrator explained, "In hindsight, [the pilot] was great for us because we actually saw the problems that we would encounter with such a large number of schools, such a large number of teachers being evaluated." For example, while scheduling short observations was easy for principals to manage, scheduling the feedback conferences was more difficult, so principals learned to create more structure in their schedules whereby they would allocate regular days and times to meet with teachers to provide observation feedback.

By piloting individual components of evaluation systems, central office staff in Pittsburgh and Hamilton County believed they helped principals and teachers become more comfortable with the evaluation process.

Central office staff in Pittsburgh rolled out individual system components, starting with the RISE classroom observation instrument during the 2009–10 school year; during the 2011–12 school year, the district piloted VAMs and the TRIPOD student surveys. At each stage, central office staff shared evaluation results with select staff (e.g., only teachers in subjects and grades for which VAM scores were calculated received their results; principals did not have access to individual teacher VAM data) but those results did not affect teachers' standing in the district. As one district administrator explained, this strategy allowed teachers to become more comfortable with the observation process and review their data before they were used in a formal evaluation; this strategy was particularly important in securing teacher buy-in: "With implementation, we allowed our teachers to sit with their data before we started to use the data in high-stakes evaluation. Some districts didn't do that; they didn't give teachers an opportunity to see what [the process and the data] look like."

Two respondents in Hamilton County, one central office administrator and one principal, explained that piloting the evaluation system resulted in less push back from stakeholders, including teachers. As the principal explained, "The best decision they made was to pilot [Project COACH] because it was received so much better. … When it came down to [fully implementing Project COACH] so many people were then comfortable with it."

Implementing without Piloting

Three of the eight districts did not pilot test their teacher evaluation systems before implementing them. Administrators in these districts explained that holding teachers accountable as early as possible was important for achieving real, measurable change in teaching and learning.

Staff in the District of Columbia, for example, explained the decision not to pilot IMPACT by noting that the district had spent two years on research and design activities and wanted to improve the quality of teaching quickly by establishing a common language for effective instruction and by improving the quality and quantity of the feedback delivered to teachers. The district believed that delaying IMPACT's implementation would also delay improvement in instruction. One district administrator explained the thinking behind the decision not to pilot test, contending that the best insights into system performance come from actual rather than simulated implementation:

> *I know some districts are taking 2–5 years to ramp up [their evaluation systems]. We ripped the band-aid off. You can't act on things until it becomes and feels real. I just think that if we had piloted [IMPACT] for a year, we would have lost a year for having actionable data for what's best for kids. [The data from a pilot test] also would have been noise. Our strategy has largely been one of 'Just do it.'*

Staff in Hillsborough County provided similar explanations for their decision not to pilot test. One district administrator, for example, reflected on the pilot testing in another district in another state and argued that teachers there would not fully engage in a teacher accountability system until it holds high-stakes consequences for them: "I'll be real interested in a couple of years when [that district] goes to a high-stakes evaluation system. … They are doing much more piloting over several years to get everyone ready for it, and I'm genuinely curious about when they finally do make it high stakes, whether teachers will take [the results] any better because it's been piloted slowly."

In the end, whether they pilot tested their teacher evaluation system or not, all eight districts were continuing to change, adjust, and fine-tune their systems in late 2012 and early 2013 in response to a variety of information, including teacher feedback, pilot test data, and early implementation experiences.

Validity and Reliability Testing

Only the District of Columbia and Hillsborough County had had the validity of their evaluation systems (i.e., the extent to which classroom observation scores are correlated with student growth measures) tested by outside experts. Pittsburgh had hired an external research organization to validate its teacher evaluation system but had not received the test results when the data for this study were collected. According to district administrators in Plattsburgh, limited district capacity meant that validity testing was unlikely in the near future; however, they suggested that the New York State Department would conduct validity testing once the new teacher evaluation systems had been fully operational for a few years.

> **District administrators in the District of Columbia and Hillsborough County reported that independent researchers had tested the validity of their teacher evaluation systems and determined that there was a firm and consistent relationship between measures of teacher practice and measures of student growth.**

As a Hillsborough County administrator explained, "We find that there is clearly a [link] between the written evaluations and the value-added scores [of teachers]. In general, [each measure is] identifying the same teachers as being effective."

> **None of the eight districts included in the study conducted a formal test of the inter-rater reliability of the observation instruments to determine whether it was possible to objectively and consistently score teacher practices using their respective observation rubrics.**

Nevertheless, five districts used observer trainings to ensure that evaluators used the observation rubric as designed and consistently and reliably scored observed teacher practices. We discuss this process, and other district approaches to observer training, in the next chapter.

Introducing the System to Teachers

A study of teacher evaluation in Chicago found that providing teacher orientation and training on the evaluation system may affect the success of that system's roll-out (Sartain, Stoelinga, and Brown 2011). The study showed that, compared with principals, teachers received less professional development and information on the evaluation purpose, process, and standards of practice on which they would be evaluated. Consequently, teachers were frustrated with the evaluation system; they did not trust in the process and, in some cases, resisted being evaluated. The following section describes the processes and products districts used to introduce teachers to their new evaluation systems and the extent to which teachers believed the training helped them understand and support their teacher evaluation systems.

Each of the eight districts introduced their teachers to the new evaluation system through one or more strategies, including distributing written materials, creating online resources, or using school-based teams of evaluation experts.

District administrators in all eight districts described providing teachers with written information about their evaluation system's purpose and structure, such as a handbook or answers to frequently asked questions (Exhibit 10). In addition, five districts (Austin, District of Columbia, Hamilton County, Hillsborough County, and Pittsburgh) offered online resources for teachers to learn more about the evaluation system, including webinars, videos of sample lessons with ratings, and web portals that allowed teachers to access relevant evaluation information and materials. Hamilton County, for example, offered numerous online resources for teachers through the district's *T-Eval* website, including answers to Frequently Asked Questions, help links, and sample evaluation reports. In addition, Hillsborough County's email account, "*greatteachers*," provided a place where teachers could submit questions or raise concerns about the evaluation system and to which district administrators could respond directly. Austin provided teachers with a training and evaluation resources website that only teachers could access.

Five districts (Austin, District of Columbia, Hamilton County, Harrison, St. Mary's County) armed their principals with informational materials and relied on them to introduce the evaluation system to their teachers, to varying success. The central office in Hamilton County, for example, provided principals with a PowerPoint presentation to offer teachers. In Austin, after the first pilot year, the district asked principals to show their teachers a training video created by the district to better explain the evaluation process. A district administrator explained that although the video answered many teachers' questions about the system, not all principals ensured that their teachers watched it. Peer observers in Austin, however, reported developing a website where resources were made available for teachers who have questions related to the observation rubric and what instructional strategies observers looked for related to a particular component of the rubric.

Exhibit 10
District strategies to introduce teachers to the new teacher evaluation system

District	Distribute handbooks, FAQs, other written materials	Have principals introduce teachers to the evaluation system	Create school-based teams of evaluation system experts	Provide online informational resources	Offer school-based trainings and professional development on the evaluation system
Austin	✓	✓		✓	✓
District of Columbia	✓	✓		✓	✓
Hamilton County	✓	✓		✓	✓
Harrison	✓	✓	✓		✓
Hillsborough County	✓			✓	✓
Pittsburgh	✓		✓	✓	✓
Plattsburgh	✓				
St. Mary's County	✓	✓			

Exhibit reads: To introduce teachers to the new teacher evaluation system, Austin distributed handbooks, FAQs, other written materials; had principals introduce teachers to the system; provided online informational resources; and offered teachers school-based trainings and professional development on the evaluation system.
SOURCE: Site visit interview data and other extant data (see Appendix D for complete list).

Two districts created teams of school-based evaluation system experts to promote teacher buy-in and provide a school-based resource for teachers with questions about the new evaluation system.[22] For example, Pittsburgh created teams within each school composed of the principal and four teachers. Each teacher on the team received training from an external provider (Batelle for Kids) and became a resident expert on a component of the evaluation system (i.e., observation, teacher VAMs, student surveys, and combined measures). Similarly, Harrison established teams of evaluation experts within the schools. These teams attended monthly, district-wide Effectiveness and Results (E&R) Focus Group meetings, during which they learned about new district developments in the teacher evaluation system and brought that information back to their school buildings. The teams also brought questions and concerns raised among staff at their schools to the attention of district administrators.

To ensure that teachers were familiar with the professional practices for which they would be held accountable, Hamilton County required teachers to perform a self-assessment using the district's new classroom observation rubric.

In addition to having principals introduce the system, Hamilton County required that all teachers rate themselves on all parts of the rubric. As one district official described it, the self-assessment compelled teachers to familiarize themselves with the new evaluation requirements:

[22] In the District of Columbia, one central office staff member observed that the way teachers in a school perceived the new evaluation system depended largely on the principals' attitude towards this system and how much the principals knew about it themselves. In 2012–13, having learned from the rollout of the IMPACT system, the District of Columbia modified its strategy for introducing its career ladder program, LIFT. Using a model similar to Pittsburgh's and Harrison's, each school assigned a designated LIFT ambassador, selected because of his or her support of the LIFT system. The ambassador was responsible for introducing and explaining the new system to other teachers in the school. One central office official said that it was important to have someone in each school who could serve as a positive voice about the program to help stave off negativity.

> *One thing that we did right off the bat was we required the teachers to do a self-assessment, which does not count as part of their score in any way. But what it does do is it ensures that they read the rubrics. Because in order to score yourself a one, two, three or four, you have to read what's there.*

In addition, Hamilton County introduced teachers to the COACH model through its New Teacher Network trainings. Although these sessions were geared towards new teachers, they were open to all teachers in the district.

The District of Columbia invested in communicating with teachers about their value-added scores by providing a separate, mandatory, two-hour training during the school day for teachers receiving individual value-added scores or "IVAs."

A district administrator described the training as a general overview of how the model worked, what variables the model controlled for, and why the district believed it was fair. She added that the training was necessary because, despite providing handbooks and online information, IVA was complicated and commonly misunderstood: "We had separate guidebooks explaining IVA but still there were misconceptions about how it worked. It's complicated."

In Hillsborough County, several respondents at the district and school levels said that the district's lack of a teacher orientation or training requirement led to ongoing confusion and misinformation among teachers about the evaluation system.

Initially, due to budget constraints, Hillsborough County did not offer comprehensive training to all staff on the new evaluation system. Instead, the district provided videos and other online resources to explain how the system worked, and central office staff made time to answer teacher questions. A district administrator acknowledged that, while unfortunate, the district's decision to forego a comprehensive teacher training on the evaluation system was based on practical considerations related to funding.

Since 2012–13, Hillsborough County teachers have received school-based training on the instructional component of the observation rubric (Domain 3), including observing a video of a sample lesson, rating the lesson, and then comparing and discussing the ratings with district trainers. A peer observer explained that the training was "very specific" as it related to what an observer will do in a classroom, including how observers will record the observation, as well as how they will prepare for and conduct the pre- and post-observation conferences. Peer observers participating in a focus group interview reported that the quality of the teacher training had improved in the district to the extent that it was having a noticeable impact on the quality of the feedback conversations observers were able to have with teachers. Peer observers also emphasized that providing teachers with explicit, hands-on training on the use of the observation rubric for documenting evidence is a critical step for any district implementing a similar teacher evaluation system.

Teachers in Austin, Harrison, and Hillsborough County suggested that training needed to be ongoing, not a one-time event.

The teachers explained that sometimes it took years for them to fully understand and respond to the complexities of their evaluation system.

Training Classroom Observers

Multiple studies examining district implementation of classroom observation as part of a teacher evaluation system stress the importance of observer training. Adequate training for all observers and support for implementing ongoing quality control measures to sustain inter-rater reliability are critical to maintaining evaluation quality and limiting bias ((Sartain, Stoelinga, and Brown 2011; Milanowski, 2004; MET Project 2012). The studies suggest various strategies for training observers including requiring certification, using videos of classroom instruction, and comparing notes with trained observers. Ultimately, by the end of the training, observers should be able to demonstrate an understanding of the observation instrument.

The following describes the ways in which the eight study districts prepared their observers to use the observation rubric to evaluate the professional practice of teachers.

All eight districts provided some type of observer training, although the scope, intensity, and rigor of the training varied substantially across the districts.

Initial observer trainings ranged from six hours for principals and other staff in Pittsburgh to five weeks for peer observers in the District of Columbia.[23] Hillsborough County, for example, offered one of the longest and most comprehensive observer trainings among the eight districts included in the study. The seven-day training included two days of work reviewing the observation rubric and watching and scoring videos of classroom instruction, two days observing and scoring directly in classrooms, and three days working individually with a trainer on the complete observation cycle. As one district administrator in Hillsborough County described it: “It's quite extensive. The last three days [of the training] [are] one-on-one with a trainer and [are] really powerful. They go out and they do a complete observation cycle twice in that one-on-one setting and the trainer is watching you and coaching you throughout that process.”

Five of the eight study districts included in the study (District of Columbia, Hamilton County, Hillsborough County, Pittsburgh, and Plattsburgh) required that all observers be certified before they were permitted to conduct teacher observations. To become certified, observers had to complete the district-sponsored observer training and pass a certification test demonstrating a thorough understanding of the rubric. St. Mary’s County did not provide any training on the rubric, as it had been using the same rubric for over 10 years prior to the implementation of the current system and did not believe that principals needed any additional training. However, to improve understanding of the rubric and consistency across observers, the district offered principals ongoing training on various components of the rubric at their monthly administrator meetings.

[23] Because principals were responsible for introducing IMPACT to the teachers in their schools, initial training for principals in the District of Columbia focused largely on learning the rubric in order to effectively communicate expectations to teachers. Training on using the rubric to conduct observations took place throughout the school year during the Principal Academy, their monthly, day-long professional development meetings. Principals watched videos of lessons during these training sessions and discussed the appropriate ratings. One principal noted that because the trainings on IMPACT were spread out over the course of the year, she had to turn to other resources, such as conversations with her colleagues, to learn how to effectively conduct classroom observations. The District of Columbia has since expanded the training and supports it provides for principals conducting observations.

Six of the eight districts used sample videos as part of observer training in order to compare and contrast observer scores — and supporting evidence — based on a common set of observed lessons.

That is, Austin, the District of Columbia, Hamilton County, Harrison, Pittsburgh, and Plattsburgh trained their observers by having them view and score sample videos of classroom lessons. For example, through a partnership with the Gates Foundation, the District of Columbia developed *Align*, an online certification and calibration program through which observers[24] rate a series of videos showing a range of practice for each of the Teach standards. *Align* made use of the District of Columbia's extensive video library of sample classroom instruction (described in Chapter VI) demonstrating each of the metrics on the rubric, at different levels. The system then recommended additional training videos for observers, if necessary, to support consistent application of the observation procedures. As one district administrator explained:

> *Basically, you go onto this platform, watch a video, and rate it. The system then tells you that all these videos are anchored against a set of scores that have been determined by our best raters. The system tells you whether you're on or off track. If you're off track, it will refer you to additional training videos. We now are at a point where we ensure that all of our observers reach a certain threshold of norming.*

Similarly, observers (principals) attending Pittsburgh's Level I Institute examined sample lessons, rated them, and then discussed their ratings. Central office staff explained that they introduced the institute in response to concerns that had been raised about observer bias. As one central office administrator explained, "We found that some principals either consistently rate teachers higher than the evidence warrants or they consistently rate them lower than they should." The district did not certify principals to conduct classroom observations until they were able to demonstrate proficiency in rating sample lessons. As of 2012–13, all principals in the district had gone through the first institute and had been certified to conduct observations. Three of the principals who participated in site visit interviews rated the observer training institute highly. As one principal said, "There's a lot of on-the-job training [provided to principals in this district]. I thought [this training] responded very well to the questions that my colleagues and I had." Another principal added, "I think the support [the district] provided has been excellent."

Finally, while Austin provided an online training resource with examples, videos, and resources linked to the rubric, the district struggled to determine the best way to train observers. On the one hand, watching videos did not provide a complete and authentic classroom experience. However, district administrators worried that having observers train by observing a lesson directly in the classroom could cause unnecessary stress and strain for teachers:

> *Some principals felt like the video norming was too inauthentic and they wanted to norm using teachers on their campus, which really made us uncomfortable because we didn't want to have teachers seeing us come in with administrators and comparing notes because that's not a way to build trust and have open conversation with them and so that did not happen either.*

[24] *Align* was originally used with peer observers only. In 2012–13, the District of Columbia began piloting the system with principals.

Four districts also relied on the expertise of external organizations to train their observers.

In its first year of implementation, for example, Hillsborough County contracted with the Cambridge Education Group to train principals and peer observers on the observation rubric because they needed the extra manpower and because they wanted to have experts conduct the training. A Hillsborough County district administrator explained that the district initially sought outside assistance because of the time required and need for expertise regarding the use of the rubric.

Hillsborough County, however, eventually took over the training of its observers and worked to fine-tune the process. Peer observers participating in a focus group interview reported that the training had indeed "gotten much better" over the years. The district paired new observers with expert or "buddy" observers for several weeks, during which time the new observer watched the experienced observer conduct observations and provided teacher feedback. Eventually, they switched roles and the buddy watched and critiqued the novice conducting classroom observations. One observer described the structure and benefits of the training: "I was with a buddy for several weeks and was able to observe cycle after cycle and really just watch [the observations] be done very well, which was the best help; the best way [to train]." The observers explained that their observation "buddies" stayed with them as long as they needed them.

In Plattsburgh, principals participated in a five-day training in Albany, organized by the i3 consortium and involving consultants with experience conducting trainings on the state's teacher evaluation system. Like other study districts, the training consisted of watching videos, documenting evidence of instructional practice based on those videos, coding the evidence, and rating the instruction based on the observation rubric. The consultants then evaluated the observation ratings. Observers learned how to document and score their observations to capture a record of a teacher's lesson. Each observer took a test at the end of the training to become a certified evaluator. To ensure consistency across observers, only those who passed the certification test at the end of training could conduct classroom observations.[25]

In support of Hamilton County, the state education agency contracted with the National Institute for Excellence in Teaching in summer 2011 to provide a four-day training for Hamilton County's observers (principals and assistant principals) that primarily focused on the classroom observation rubric. The training required evaluators to pass an observer rating exam for which they had to demonstrate their understanding of the distinction across differing levels of teacher performance by viewing and rating videos of teachers delivering lessons.

Pittsburgh also contracted with Battelle for Kids to assist district staff in organizing the observer trainings, including working with experts from the Danielson Group with planning, communications, the development of training materials, and organizing several day-long trainings and design sessions from 2009–10 through the 2012–13 school years. The majority of the observer trainings, however, were presented and led by district staff as well as experts from the Danielson Group.

[25] All of the principals in Plattsburgh City School District (five at the time of the site visit) passed the evaluator certification test the first time, and thus the district had not established a system of consequences for principals who do not pass the certification test at the time of the study. All other district and school-based staff who participated in the initial evaluator training passed the test the first time as well. Unlike principals, these staff members served as evaluators as a supplement to their regular duties and were not required to conduct observations as part of their job description and would simply not conduct observations had they failed the certification test.

In addition to video-based observer training, Harrison and Hillsborough County conducted ongoing norming exercises to guard against potential observer bias and strengthen inter-rater reliability.

Staff in Hillsborough County, for example, regularly calibrated principal, peer observer, and mentor observer[26] scores to ensure their consistency. According to one district administrator, the purpose of the calibration exercises was to ensure that observers collected consistent and accurate data to best help teachers improve practice: "We validate how closely aligned the [peer observers] and the principals are when they evaluate teachers because when they observe during the year, they're sharing the results of their observations so they can help the teacher." Efforts to calibrate observers were supported by consultants from Cambridge Education Group and were focused on the alignment of scores across all peer and mentor observers and principals. In interviews, principals said that they participated in calibration exercises approximately three to four times a year. One principal noted that principals lose points on their own evaluations if the ratings they give their teachers differ too much from peer and mentor observers who observe the same teachers, which holds principals accountable for consistent use of the rubric. Through an analysis of observer ratings across the district, Hillsborough County determined that principals, on average, scored teachers higher on their classroom observations than did the peer observers. Accordingly, and because peer observers and principals had heretofore trained separately, the district began calibration exercises to improve scoring consistency between the two groups.

In Harrison, calibration exercises consisted of district administrators (i.e., the Superintendent and assistant superintendents), together with principals, conducting a classroom observation, scoring the observation, and comparing their scores. Principals participated in calibration exercises approximately four times a year. According to one district administrator, the district was "constantly calibrating observations to make sure our notion of proficiency was the same." According to several high school teachers who participated in focus group interviews, as many as four evaluators would observe a teacher at one time for purposes of calibrating their scores.

To maximize the impact of classroom observations on teachers' professional practice, six districts trained observers on ways to provide constructive feedback to teachers.

Austin, the District of Columbia, Hamilton County, Hillsborough County, Pittsburgh, and Plattsburgh offered observers additional training to strengthen their skill at providing constructive — and sometimes difficult — feedback to teachers based on their classroom observation scores. Pittsburgh, for example, created an institute that offered observers strategies for providing helpful instructional feedback and support to teachers. According to a district administrator, the training helped evaluators know how to provide high-quality feedback to teachers:

> *We wanted the principals to not only use the observation tool in a consistent way; we also wanted to ensure that they could provide appropriate feedback to teachers based on how they were rating them. The [training] has a particular focus on subject matter; how leaders can use data from the observations to guide teachers with regard to curriculum and content. This can be*

[26] Hillsborough County uses three types of observers to conduct classroom observations. Principals conduct two to four observations per teacher, peer observers (teachers employed by the district who temporarily leave their classrooms to conduct classroom observations) conduct one to four observations of experienced teachers, and mentor observers conduct four observations of new, inexperienced teachers.

> *a challenge in high schools where the administrators have a background in just one or two subjects. Some observers can be intimidated by the idea of providing subject-specific feedback.*
>
> *[District Administrator, Pittsburgh]*

Similarly, principals in Plattsburgh received training on professional conversations and instructional coaching, both during their initial "boot camp" observer training and in follow-up consortium training sessions in Albany. Among other things, principals read the book, *Results Coaching* (Kee, Anderson, and Dearing, 2010), to help them learn ways to establish coaching relationships with teachers and provide feedback that maximizes the strengths of each individual teacher.

To help observers more effectively communicate to teachers the strengths and weaknesses of the lessons they observed, Hillsborough County contracted with Fierce Conversations[27] after the first year of system implementation. That is, according to district administrators, observers at times found themselves in confrontational conversations with teachers for which they felt unprepared. In addition, they often struggled to provide feedback that balanced praise with constructive criticism. For their part, teachers complained to the district that their observers had not always given them complete and honest feedback during their conferences and, as a result, teachers' understanding of their performance did not align with the ratings they received.

A peer observer in the District of Columbia reported that observers also struggled to have difficult conversations with teachers, and that training observers on how to provide effective post-observation feedback was an ongoing feature of their twice-monthly meetings with observers. He described the goal of the post-observation feedback training this way: "The ability to give bad news in a delicate way and push [teachers to do more] is a hard thing to do. It's not just that we give a score; we name things that have gone right and things that can be improved. You do feel like you've done a good job if you give a teacher a Level 2 score and they still [respect] you."

District respondents in Austin and Hamilton County described providing very specific feedback training. In Austin, for example, observers received training on how to remain objective and use questioning strategies, as opposed to saying "you should" or "why don't you do this instead?" In Hamilton County, a district administrator described conducting classroom walk-throughs with principals, watching them observe and provide feedback to teachers, and then giving the principals "feedback on their feedback." The administrator said that she observed every principal conduct an observation and provide teacher feedback. In addition, she reviewed the observation data and feedback: "Last year, I got to watch each one of my principals at least once do an observation with a teacher, bring the teacher in, and just hear the comments and watch it going on. And I'm taking notes, too. After the teacher leaves, then I give that principal feedback on [the feedback they gave the teacher]." She said that if a principal's feedback was largely indistinguishable from one teacher to the next, it meant that the principal was struggling with the task. In response, the administrator would offer the principal suggestions about strategies to look for in the observations and ways to provide better, more targeted feedback.

[27] Fierce Conversations is a professional development program developed by Fierce, Inc. that is designed to improve the quality of staff feedback in order to help organizations — including schools — establish and work toward common goals.

IV. Conducting Classroom Observations

Recent studies of teacher evaluation systems that include structured observations generally agree that multiple observations of individual teachers are needed to gain an accurate measure of teacher effectiveness since individual scores can vary considerably from one lesson to the next (Sartain, Stoelinga, and Brown2011; Milanowski, 2004; MET Project, 2012). These study authors also generally agreed that because of issues with inter-rater reliability, periodic observations conducted by impartial external evaluators should be part of the teacher evaluation system as a quality control check on regular internal observations.

While the research literature has reached consensus around the need for multiple observations and the inclusion of external evaluators to observe classroom instruction, these studies do not suggest an ideal number of observations, nor do they suggest who should conduct them or how they should be conducted.

Districts' approach to conducting classroom observations among the eight districts included in this study reflected the lack of consensus in the research literature. Indeed, the observations varied significantly with respect to the types of observers used and the frequency, duration, and degree of formality of the observations they conducted. In addition, principals reported challenges finding time to conduct the observations while maintaining their other responsibilities. All observations in every district, however, yielded at least some feedback to teachers that was intended to help them improve their instructional practice.

Key Findings

- The frequency of classroom observations varied widely across the eight districts included in the study, ranging from two observations for experienced teachers in one district to 18 observations for new teachers in another. In some districts, these observations were preceded by a pre-observation conference between the observer and the teacher, and in all districts the observations were followed by feedback, often in the form of a post-observation conference.

- Classroom observations represented a significant time commitment for principals in all eight districts; on average, principals interviewed for the study reported devoting approximately one-third of their time to this task. Implementation of the new teacher evaluation systems in the eight districts brought increased responsibilities for principals. While most districts attempted to offset these added responsibilities by offering principals additional supports and reducing their other administrative responsibilities, principals reported having to make many adjustments to their workday to accommodate the increased workload.

- Three districts used peer observers, in addition to principals, to conduct classroom observations in an effort to offer an unbiased, objective opinion of a teacher's performance. Each of the three districts hired experienced teachers to work full time as peer observers on behalf of the district.

- District staff and teacher focus group participants in two districts expressed their view that the legitimacy of an observer — and the value of the feedback provided — was partly based on the strength of the match between the evaluator's background and experience and that of the teacher being observed.

- No teacher who participated in focus groups or individual interviews, in any district, raised concerns about the legitimacy of the professional practices — as measured by the observation rubrics — to which they were being held accountable. Teachers in three districts, however, questioned whether it was possible to demonstrate excellence on the full range of competencies included in their respective districts' observation rubrics.

The Observation Cycle

All eight districts included in the study assessed instructional practice by having trained observers conduct a defined number of formal and informal observations of classroom lessons using the district-adopted classroom observation rubric. In some districts, these observations were preceded by a pre-observation conference between the observer and the teacher, and in all districts, the observations were followed by feedback, often in the form of a formal post-observation conference.

The frequency, duration, and degree of formality of the classroom observations varied widely across the eight districts included in the study.

The number of annual observations for a typical teacher ranged from two observations for experienced teachers in Plattsburgh and St. Mary's County to 18 observations for new teachers in Harrison (Exhibit 11). Two districts (District of Columbia, Hillsborough County) varied the number of observations for experienced teachers based on their previous year's performance. In the District of Columbia, for example, administrators based the number of observations an experienced teacher received in the spring on the teacher's rank or designation on the district's career ladder (i.e., "Expert," "Distinguished," or "Advanced") from the previous year and on their observation scores in the previous fall. Teachers who received an average observation score of 3.0 or higher (on a four-point scale) in the fall of the school year could waive additional observations in the spring.

Exhibit 11
Number and type of observations conducted for new and experienced teachers, by observer type, by district

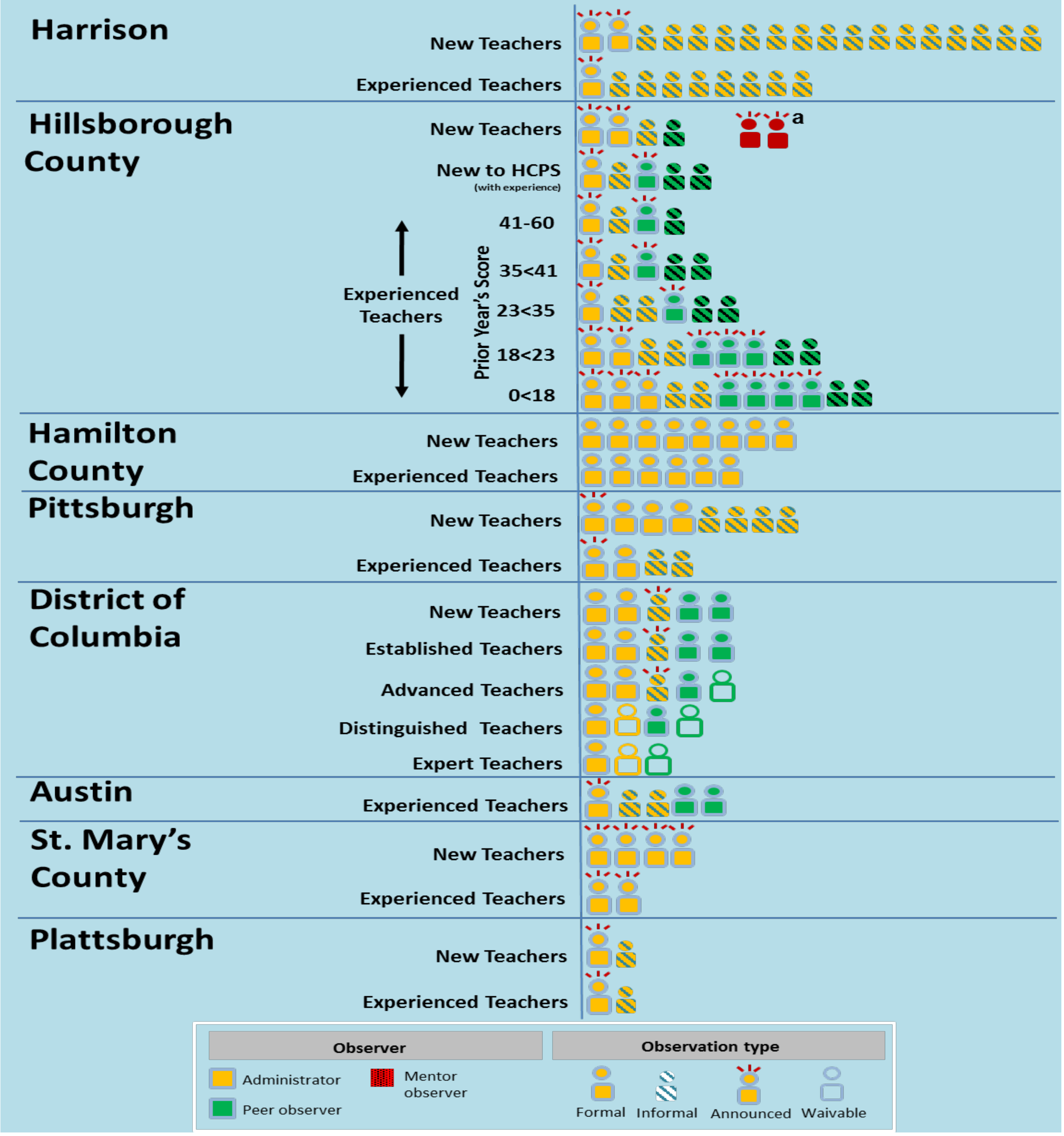


Exhibit reads: Harrison's teacher evaluation system required that each year, new teachers receive two formal, announced observations and 16 informal, unannounced observations from their principal.

[a] Hillsborough County's mentor observation ratings do NOT count toward new teachers' final evaluation scores (see Exhibit 12). NOTE: "New" and "Experienced" teachers are generic terms we used to identify teachers who have experience working in schools versus those who are relatively new to schools. However, many districts included in the study labeled these groups differently. Harrison, for example, referred to experienced teachers as "non-probationary" and to new teachers as "probationary" teachers. Hamilton County referred to these two groups as "professionally licensed" or "non-professionally licensed," respectively. Finally, St. Mary's County called new teachers "novices."

SOURCE: Site visit interviews; extant data on system specifications (see Appendix D for complete list).

Five of the seven districts whose evaluation systems included new teachers[28] (the District of Columbia, Hamilton County, Harrison, Pittsburgh, St. Mary's County) required that new teachers receive from two to nine more observations per year than experienced teachers. In the remaining two districts, Plattsburgh made no distinction in the number of observations required for new versus experienced teachers. Hillsborough County, however, limited the number of observations to which they held new teachers accountable (i.e., only four, compared with as many as 11 for experienced teachers whose evaluation score from the previous year was low) in order to allow these novices to develop their professional practice without fear of consequence (see Exhibit 12).

Exhibit 12
Evaluating new teachers: Hillsborough County's distinctive approach

Hillsborough County's first- and second-year teachers were assigned mentors who supply feedback and coaching support based on their regular observations of new teachers' classroom practices. Mentors provided new teachers 90 minutes of coaching and support once a week for first-year teachers and every two weeks for second-year teachers. The district characterized the mentor observations and coaching support as formative rather than summative assessments that were intended to help new teachers improve their practice. In order to maintain a positive coaching relationship, however, mentors did not conduct observations of their own mentees that counted toward a teacher's formal evaluation score. Instead, mentors swapped caseloads with other mentors to conduct formal observations for the EET evaluation. A district administrator stressed the importance of ensuring that mentors did not formally evaluate their mentees in order to preserve the honest, collaborative relationship that had been forged between mentor and mentee:

> *I was adamant that mentors would not evaluate their own caseload of teachers. I felt that it would compromise that relationship, [because] some teachers may not feel as comfortable sharing and opening up if they knew at the end of the day that you're going to be the one who evaluates me.*

The regular observations and feedback from the mentor gave new teachers continuous insight into how to improve their instruction without the pressure of those observations counting toward their evaluation score.

In response to staff feedback, two districts adjusted the number and type of required observations to lessen the observation burden on school staff and to accommodate the needs of new teachers.

After its pilot year, and in response to teacher feedback surveys, Hamilton County reduced the required number of observations for new teachers from 10 to eight in order to reduce some of the pressure on principals' time. In the District of Columbia, experienced teachers who consistently scored high ratings could choose to opt out of observations, eventually allowing teachers who had been rated highly effective for six consecutive years to require only one observation per year. In addition, and in response to concerns teachers raised about the fairness of new teachers receiving their first observation from a peer observer, the District of Columbia began requiring principals to conduct an initial, informal, and non-binding observation with new teachers in order to familiarize those teachers with the observation process. As a district administrator explained:

> *We made a guarantee to teachers that their first observation would be informal. [And] we're going to delay their [peer observer] observation such that they get some feedback before any high stakes are introduced. And this observation is a regular, full observation: same length, same*

[28] Austin did not include new teachers in its pilot of the teacher evaluation system.

> *written report afterwards, same debrief, it just doesn't count.* [This] *was a big change and something that was really popular even for teachers to whom it didn't apply.*

By including informal observations in the observation cycle, six districts reported that they were able to capture a more complete picture of teacher professional practice without the accompanying time burdens associated with full documentation and feedback.

All eight districts conducted formal observations of teachers (Exhibit 13) that typically: (1) were scheduled in advance, (2) covered the entire observation rubric, and (3) included pre- and post-observation conferences. In five of these districts (as well as in Hamilton County, which only includes formal observations), a formal observation was preceded by a pre-observation conference where the observer and teacher reviewed the teacher's lesson plan[29] and a post-observation conference to discuss what the observer saw during the lesson, review ratings, and provide specific strategies for improvement. Teachers may have the option of presenting additional evidence to demonstrate aspects of their work that may not have been evident in the observed lesson or that simply cannot be observed.[30]

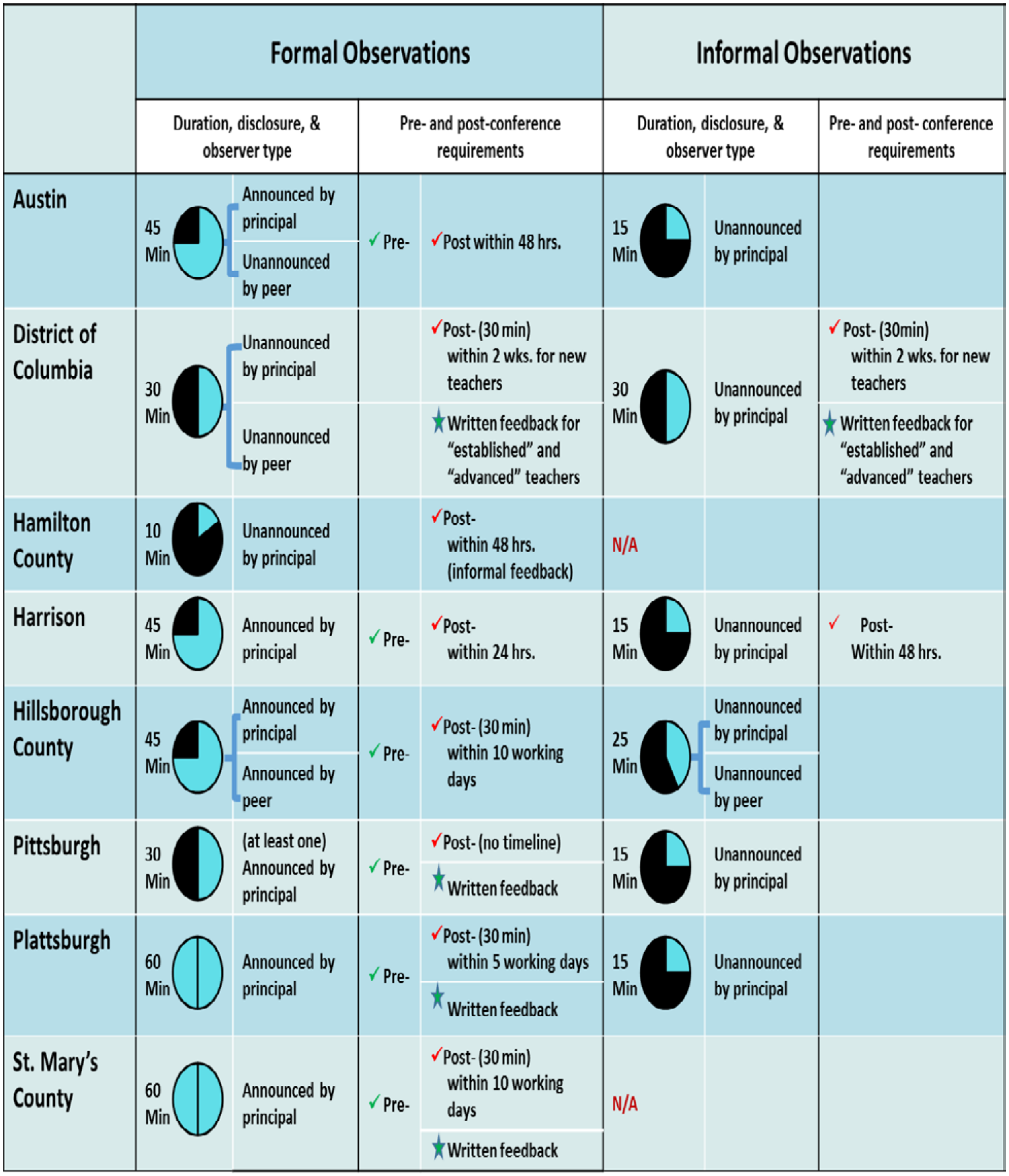
District administrators described an informal observation, by contrast, as a short, unannounced visit to a classroom where the observer could provide constructive feedback without the time burden of full documentation and conferencing associated with a formal observation. Among the six districts that conducted informal observations, Austin, the District of Columbia, Harrison, Hillsborough County, Pittsburgh, and Plattsburgh included informal observations — "walkthroughs" in Austin — in the observation cycle. Four counted those observations towards a teacher's overall evaluation score (Exhibit 13).[31] A principal in Hillsborough County said that these informal observations allowed him to better assess what was really happening in a teacher's classroom, in contrast to a "dog and pony show."

[29] In the District of Columbia, formal observations were unannounced in order to capture a teacher on a typical day; there was no pre-observation conference to review the lesson materials. Pittsburgh required that only one formal observation be announced and included a pre-observation conference; the remaining formal observation could be unannounced.

[30] For example, one teacher in Plattsburgh described bringing logs of phone calls and e-mails to parents to a post-observation conference to address the parent engagement component of the district's observation rubric. The ratings from these formal observations counted towards a teacher's overall evaluation score.

[31] The District of Columbia and Pittsburgh were the exception to this rule. In the District of Columbia, teachers (all but those rated "Distinguished" or "Expert") had one informal observation a year. The principal or other school administrator conducted the first observation and provided feedback and ratings. In Pittsburgh, tenured teachers had two informal observations and new teachers had four informal observations. Informal observations in both districts provided teachers with feedback, but did not count towards a teacher's overall evaluation score.

Exhibit 13
Duration and type of classroom observations and observation feedback, by district, in 2012 and 2013



	Formal Observations				Informal Observations		
	Duration, disclosure, & observer type		Pre- and post-conference requirements		Duration, disclosure, & observer type		Pre- and post- conference requirements
Austin	45 Min	Announced by principal Unannounced by peer	✓ Pre-	✓ Post within 48 hrs.	15 Min	Unannounced by principal	
District of Columbia	30 Min	Unannounced by principal Unannounced by peer		✓ Post- (30 min) within 2 wks. for new teachers ★ Written feedback for "established" and "advanced" teachers	30 Min	Unannounced by principal	✓ Post- (30min) within 2 wks. for new teachers ★ Written feedback for "established" and "advanced" teachers
Hamilton County	10 Min	Unannounced by principal		✓ Post- within 48 hrs. (informal feedback)	N/A		
Harrison	45 Min	Announced by principal	✓ Pre-	✓ Post- within 24 hrs.	15 Min	Unannounced by principal	✓ Post- Within 48 hrs.
Hillsborough County	45 Min	Announced by principal Announced by peer	✓ Pre-	✓ Post- (30 min) within 10 working days	25 Min	Unannounced by principal Unannounced by peer	
Pittsburgh	30 Min	(at least one) Announced by principal	✓ Pre-	✓ Post- (no timeline) ★ Written feedback	15 Min	Unannounced by principal	
Plattsburgh	60 Min	Announced by principal	✓ Pre-	✓ Post- (30 min) within 5 working days ★ Written feedback	15 Min	Unannounced by principal	
St. Mary's County	60 Min	Announced by principal	✓ Pre-	✓ Post- (30 min) within 10 working days ★ Written feedback	N/A		

Exhibit reads: Austin required that observers conduct pre-observation conferences and that they also conduct post-observation conferences within 48 hours of an observation. Informal observations were 15 minutes in duration and were unannounced by the principal.

SOURCE: Site visit interviews; extant data on system specifications (see Appendix D for complete list).

Although the intent of unannounced visits was to capture evidence of authentic practice, observers and teachers in five districts (Austin, the District of Columbia, Harrison, Hillsborough, Pittsburgh) reported that some teachers gave lessons that might gain favor with the observers rather than teach the actual lesson that was planned for that day. Observers in districts with unannounced visits were generally aware of this phenomenon. As one peer observer in Hillsborough County pointed out, evaluators structured the observation process so that they could talk to students to determine whether the instruction was unusual or seemed out of place. She noted that it was usually evident when a teacher made adjustments for the observation, saying, "The teacher may be able to fake things, but … it will come out because of the nature of the way that we talk [to students to learn about what they had been doing in the classroom prior to the observation] and things like that."

Finally, the duration of the observations ranged from 10 to 60 minutes for formal observations and, among the six districts that had them, 15 to 30 minutes for informal observations (Exhibit 13). The two districts with the fewest number of observations, St. Mary's County (four observations for new teachers) and Plattsburgh (two observations for new teachers) conducted the longest observations, at 60 minutes each. The remaining districts' observations, however, varied in length without any clear pattern relating to the number of observations administrators conducted.

The Role of Principals

> **The introduction of new teacher evaluation systems increased —sometimes considerably — the number of required teacher observations, which brought increased responsibilities for principals.**

Indeed, by increasing the frequency and duration of the observations, districts expanded the amount of time principals were required to devote to conducting the observations and to providing feedback to teachers. To offset these added responsibilities, all eight districts also trained assistant principals to conduct classroom observations. In addition, in four districts, principals could delegate responsibility for conducting classroom observations to central office staff (Hillsborough, Pittsburgh, Plattsburgh, and St. Mary's County) (Exhibit 14). In seven of the eight districts, observations conducted by assistant principals and central office staff counted towards a teacher's official rating, while in St. Mary's County, content area supervisors provided their observation comments to the principals, who were ultimately responsible for a teacher's score.

Exhibit 14
Type of staff conducting teacher observations, by district

	Principals	Assistant principals	District staff	Teacher leaders	Peer observers
Austin	✓	✓			✓
District of Columbia	✓	✓			✓
Hamilton County	✓	✓			
Harrison	✓	✓	✓		
Hillsborough County	✓	✓	✓[a]		✓
Pittsburgh	✓	✓	✓	✓	
Plattsburgh	✓	✓	✓	✓	
St. Mary's County	✓	✓	✓		

Exhibit reads: In Austin, principals, assistant principals, and peer observers conduct teacher observations.

[a] District supervisors conducted one observation for low-scoring (0–17.99 pts.), experienced teachers.
SOURCE: Site visit interviews; extant data on system specifications (see Appendix D for complete list).

In interviews, every principal respondent in each of the eight districts said that classroom observations represented a significant time commitment.

On average, across the eight study districts, the interviewed principals reported devoting approximately one-third of their time to classroom observations, but individual estimates ranged from five percent (District of Columbia) to 60 percent of their time (Harrison, St. Mary's County).

The average number of observations principals conducted per teacher, per year ranged from two for veteran teachers in Plattsburgh to 18 observations for new teachers in Harrison (Exhibit 15). Estimates of the number of principal hours required to evaluate a single teacher — including conducting classroom observations and providing feedback — ranged from one to 4.5 hours per teacher, per year (Exhibit 16) based on the authors' analyses of system specifications and documentation. When asked about the time burden, principals in Plattsburgh reported that it was excessive and expressed concerns about being able to keep up with the work over time. One principal estimated the time required to prepare for, conduct, and document one observation as taking about 15–20 hours per teacher.

Exhibit 15
Number of principal observations conducted per teacher, per year, by teacher type and by district, in 2012 and 2013

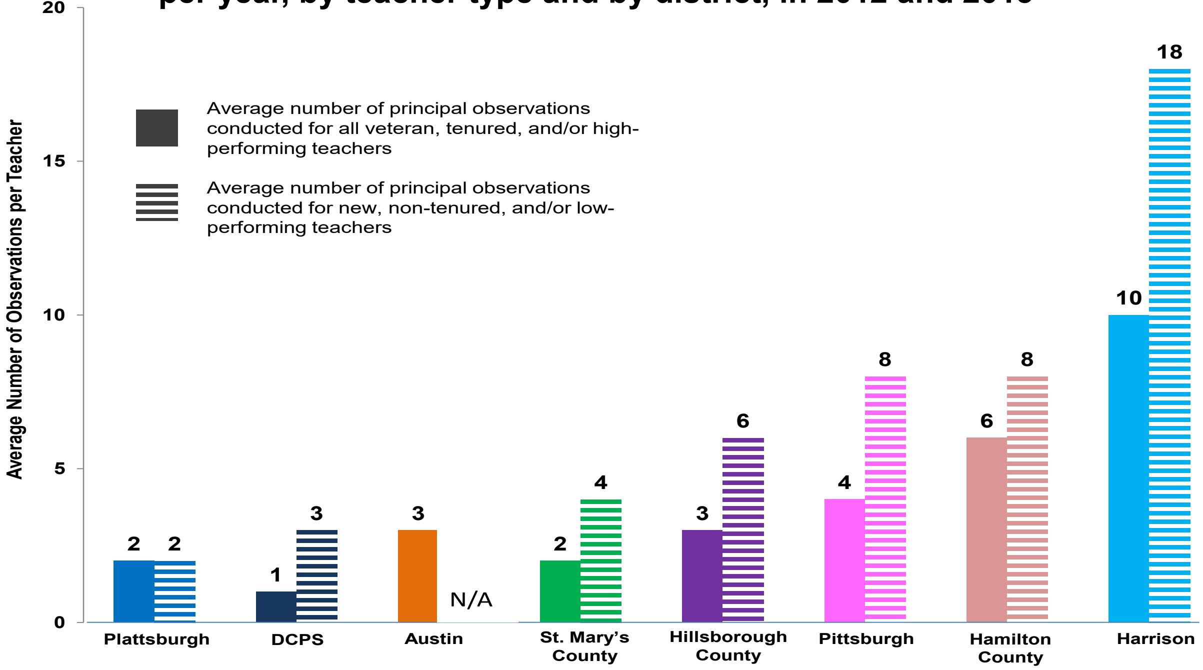


Exhibit reads: On average, principals in Plattsburgh conducted two observations per teacher, per year, based on the authors' analysis of system specifications and documentation.
SOURCE: Extant data on system specifications.

Exhibit 16
Number of principal hours devoted to conducting teacher observations and providing feedback, per teacher, per year, by district, in 2012 and 2013

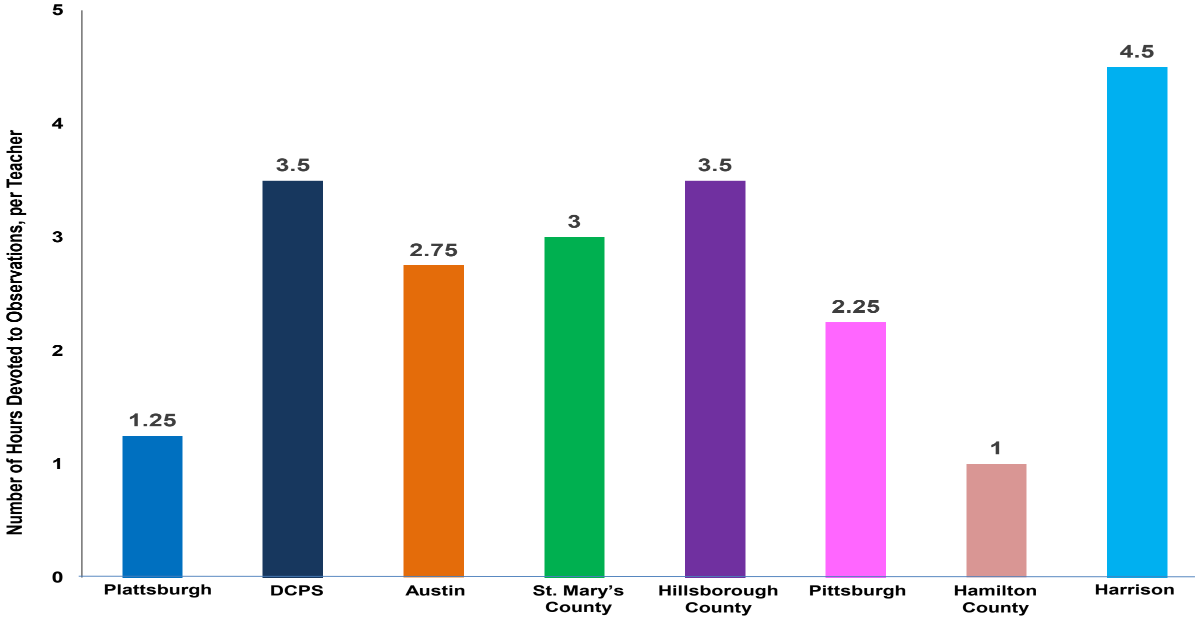


Exhibit reads: On average, principals in Hamilton County devoted one hour per teacher, per year, to conducting classroom observations, based on the authors' analysis of system specifications and documentation.
SOURCE: Extant data on system specifications.

One district signaled its long-term commitment to its new teacher evaluation system by making significant changes to the way it defined and supported the principal's role in the school.

In Harrison, the district leadership opted to hire more assistant principals, supervise custodians centrally, and do "everything we can to keep things off [principals'] plates," explained one district administrator.

The Role of Peer Observers

After the initial systems design phase had been completed, three districts began to use peer observers, in addition to principals, to conduct classroom observations.

Administrators in each of these districts reported using these peer observers to conduct teacher observations for two reasons: to offer an unbiased, objective opinion of a teacher's performance that was unrelated to any personal connections to a school; and to include the expertise of an observer who specializes in a particular subject with which the principal may not be familiar. As reported earlier, two of these districts (District of Columbia and Hillsborough County) began using peer observers in response to teacher feedback, which the districts collected after the initial design phase, but before implementation of their respective evaluation systems. In a third district (Austin), the evaluation design team, which believed strongly in using peer observers as a way to decrease the observation burdens on principals, integrated them into their evaluation system in the second year of the pilot. An Austin district administrator reported that the initial teacher response to peer observers was overwhelmingly positive, noting that teachers who experienced peer observations for the first time appeared to have really valued the feedback, which their principals did not have the time to provide.

Each of the three districts hired experienced teachers to work full time as peer observers on behalf of the district.

In Austin and Hillsborough County, the peer observers were teachers from within the district who chose to take a leave of absence from the classroom to observe and evaluate other teachers. The District of Columbia, conversely, hired former teachers who had worked in an urban district for at least five years. Of the district's 42 peer observers, a subset came from within the District of Columbia. Austin and the District of Columbia hired peer observers as permanent employees, whereas Hillsborough County offered only temporary positions, which teachers were permitted to hold for up to three years.

Peer observers interviewed in all three districts asserted that the most important function they performed — the value they brought to their districts' teacher evaluation systems — was providing targeted feedback and coaching to the teachers they observed. One peer observer in Austin explained, for example, that the external nature of the peer observers allowed them to provide feedback that was perhaps more constructive and less punitive than a principal could: "I think the difference too is that a lot of times, [principals will] conduct walkthroughs where they come through and say 'You don't have this, you didn't do this, this is wrong.' [As peer observers], we really try hard to balance areas for growth with what [teachers are] doing well and really try to focus a lot on the positive and acknowledge their efforts." A peer observer in Hillsborough County offered a similar perspective on the valuable role peer observers played in the district, explaining that teachers wanted and valued having someone provide them with feedback and follow-up support.

District staff and teacher focus group participants in two districts expressed their view that the legitimacy of an observer — and the value of the feedback provided — was partly based on the strength of the match between the observer's background and experience and that of the teacher being observed.

For example, in Hillsborough County, a district administrator said that although the district had a 91 percent "hit" rate of matching peer observers to teachers based on content-area certification, certification in a content area was not the same as experience in teaching the content. As a result, according to the district administrator — and corroborated by teachers participating in focus group interviews — observers were evaluating teachers in subjects or grades they had not taught. For example, peer observers with middle school experience observed high school teachers, or observers with special education certification but no classroom experience observed special education teachers. Some teachers said that even "matched" peer observers did not have a sufficient understanding of the classroom context with which to provide an accurate evaluation.

Although the district had worked to improve the match rate, teachers participating in focus group interviews raised concerns about matching not just on content area expertise but also on grade and school levels. As one teacher explained:

> *We have somebody at our school who is [observing] fourth and fifth grade, but he taught high school. He doesn't really have the background; it's not the same. If you're putting someone that's middle school into elementary schools, that's different. I mean, it really is different. They're coming from a whole different background and they're not going to see stuff the same way. And I think that comes into effect when it comes to your scores.*

A district administrator said that she was aware of and was working to solve the mismatch problem: "We spent a lot of time trying to [strengthen] our pool of observers — and manage the conflicting schedules of observers and teachers — so that we could match better. And we've increased [the match rate] every year. We're not perfect yet, but I have regular meetings with the union [representatives] from the schools where [the matches] are still a problem."

In the District of Columbia, teachers also raised the matching issue as a problem, but with regard to principals rather than peer observers. One district administrator reported that teachers wanted peer observers to conduct observations because their principals did not always have the background or expertise in the teachers' particular content area or grade level to provide content-specific feedback. Unlike principals, the peer observers in the District of Columbia were required to have certification and teaching experience in a particular subject area to observe a teacher in that subject.

Observation Feedback

The eight districts all used classroom observations [designed] to provide teachers with feedback to help them improve their instructional practices. The timeframe within which administrators were required to deliver their observation feedback to teachers was sometimes very specific. In addition, several districts required that administrators deliver their feedback to teachers during scheduled formal post-observation conferences.

In three of the eight districts, because observers had to provide their comments to teachers within 48 hours of the observation, feedback tended to be brief and informal, and principals found it challenging to provide high-quality feedback within the prescribed timeframe.

As shown in Exhibit 13, Austin, Hamilton County, and Harrison gave observers a narrow timeframe within which to provide observation feedback to teachers. One principal in Hamilton County described how he made it work in practice:

> *I have my iPad and I observe. Then I go out and I stand in the hall and on my notes page I've got [a record of past] observations. I just go to the very next one and I jot down a few notes of things that I saw and particularly if there's something that stuck out that I needed to mention, I make sure that I make a note of that, just to jog my memory. And then the feedback that they get from the observation is really about that conversation. It's not, 'You did this and this and this,' but it's we had a conversation about what [the students] understood.*
>
> *[Principal, Hamilton County]*

Although having short observation conferences allowed observers to focus on a few key areas relevant to the lesson, observers and teachers alike reported that this time-sensitive feedback structure restricted the observer from having a comprehensive conversation about the teacher's work. Another principal in Hamilton County said that she struggled to have short feedback conversations because she had so many thoughts to share:

> *The feedback conversation to me was one of the hardest pieces. Having the conversation is easy. What was difficult for me was not getting into a 30-minute conversation. That will kill this framework because there's not the time to do that. And that's what I ended up doing. I would end up with a teacher, we'd be having a good conversation, and we would just keep on. And it really does need to be refined and kept short and hit the high points.*
>
> *[Principal, Hamilton County]*

In the remaining five districts, post-observation conferences were formal meetings between the observer and the teacher.

In the District of Columbia, Hillsborough County, Pittsburgh, Plattsburgh, and St. Mary's County, observer-teacher conferences typically lasted 20 to 30 minutes (Exhibit 13). Observers described some discussions as focusing on ways to improve practice, while in others, the observer simply explained the teacher's score. In the District of Columbia, peer observers described trying to provide very concrete ideas about ways teachers could improve.

Having a longer feedback conference (i.e., as compared with informal observations) did not always mean that the conversation would—or should—cover all the issues that arose during an observation. Observers in Hillsborough County, for example, were trained to identify two to three strengths and two to three areas for improvement to discuss in each post-observation conference striving to keep the discussion balanced and focused. These observers did not assign ratings to teachers until after the conference, allowing information gleaned from the post-observation discussion to factor into the score. In some cases, however, observers did not identify more than the required three points of weakness even if they had deducted points in other areas, leaving teachers surprised and upset by their scores. This may have been due to time constraints or due to the desire to avoid confrontation. To address the latter issue, Hillsborough County contracted with *Fierce Conversations* (see Footnote 26), to help observers provide more direct, useful, and sometimes difficult or negative feedback to teachers during the post-observation conferences.

> **Observers in three districts said they believed that the face-to-face, post-observation feedback conferences with teachers contributed to improved teacher professional practice.**

Observers in Austin, the District of Columbia, and Hamilton County said that although the face-to-face conferences with teachers were difficult, they were an important part of teacher evaluation systems. Peer observers in the District of Columbia, for example, said that although the post-observation conversations were challenging and difficult, they could be the most rewarding, and insisted that they should be a non-negotiable in any teacher evaluation system: “We have some great conversations with lower-performing teachers [about ways to improve their practice], but it’s emotionally exhausting,” explained one observer. The other observer added: “I know it’s hard, but we have to [have those face-to-face, post-observation feedback conferences] for the teacher to know how he/she can improve.”

In Hamilton County, principals did not give a score at the end of each observation. Instead, they simply made notes about what they observed and provided a score at the end of the year. This, a district administrator explained, allowed conversations to focus on practice rather than on a score and directly addressed teachers’ areas of need.

Teacher Perspectives on Classroom Observations

Observation Rubrics

> **No teacher who participated in focus groups or individual interviews, in any district, raised concerns about the legitimacy of the professional practices — as measured by the observation rubrics — to which they were being held accountable.**

Site visitors asked every teacher participating in a focus group or in a personal interview about the observation rubric their district’s teacher evaluation system used to assess their instructional practice. No teacher raised any questions or concerns about the rubric on which classroom observations were based. One of the teachers participating in site visit interviews felt that the rubric simply evaluated teachers on the practices they were already using:

> *You've got a very well-developed rubric. It's not perfect, but it's a darn good attempt to look at all the relevant and important aspects of teaching children. [The district has] revised it every year so it means that they're constantly looking at trying to make it better. And I think over the years, they have taken teacher input into account as they've undertaken to revise and make it better.*

Another teacher said that she liked the rubric and that it was “easy to follow. There are examples of what it means to be proficient and distinguished in a given area.”

Teachers raised concerns in five districts, not about the professional practices the observation rubrics measured, but about applying the observation rubric to special education teachers.

Administrators and teachers in Austin, District of Columbia, Hamilton County, Harrison, and Hillsborough County all raised concerns about applying the observation rubric in special education classrooms. In the District of Columbia, for example, a principal explained that although the observation rubric was a “good tool” and she had used it with a “wide variety of teachers,” it could be challenging to apply the rubric to special education teachers working with significantly disabled students.

Similarly, a special education teacher in Austin believed that her district’s observation rubric was not appropriate for special education teachers. She explained that if a teacher’s students are not capable of self-monitoring, the teacher will never get a rating of “four.” In Hillsborough County, a teacher raised similar concerns, explaining that her special education students were never going to demonstrate higher-order thinking skills in the traditional way those skills are defined and measured on the observation rubric. A teacher in Harrison offered the following rationale: “I’m all for raising the bar and pushing the kids, but they’re special education for a reason. I’m supposed to grow my kids a year, but I have kids who are cognitively delayed, and I think the rubrics are tough for special education teachers [and students] to meet.”

Teachers in three districts questioned whether it was possible to demonstrate excellence on the full range of competencies included in their respective districts’ observation rubrics during the timeframes within which they were observed.

One Hillsborough County teacher said that with 24 standards or competencies on the rubric, the bar was set very high for teachers to demonstrate competency at the highest level:

> *There's 24 elements to this in this little snapshot of the world that we have to have. And so we have to prepare for this, because on a daily basis there's no possible way that any human can be up and alive for 50 solid minutes doing 24 things on a constant basis. And it was never built that way. …I’ve had experiences where I've sat with the [peer observer] and she's literally said,‘You did this, this, this, and this, but this one little tiny thing you didn't do. Seven things listed on exemplary and I don't get exemplary.’*

Similarly, a Plattsburgh teacher questioned whether it was reasonable to expect to see excellence on every domain of the rubric in a single observation:

> *Not all these [competencies] apply all the time. And you can't, as a teacher, be excellent at all these things at the same time, because you've got to prioritize. You've got to choose where you're going to spend your time. It's very overwhelming. I can't imagine being a new teacher and having to deal with this.*

Observation Scores

Despite district efforts to train observers, norm observations, and improve inter-rater reliability, teacher focus group participants in three districts felt that observation scores remained inconsistent across observers.

One teacher in Harrison said, “Some evaluators prefer certain teaching styles. I’ve had three folks come in to observe [a principal, a district administrator, and an instructional coach], and they gave me three different scores.”[32] Similarly, teachers in Hamilton County expressed concern about inconsistent ratings from observers. One teacher noted that she had had a positive experience with the evaluation system initially but that it was used less as a supportive tool than in previous years.

Similarly, despite Hillsborough County’s efforts to ensure consistent observations across principals and peer observers through the use of calibration exercises, teachers pointed to differences between the scores that come from principals and those from peers:

> *My [peer observer’s] scores have been very consistent and my principal scores have been very consistent, but there's a huge gap between the two. And my students are typically fairly homogenous from year to year compared to many other classrooms. So there are a lot of consistent variables [that make it difficult to understand the inconsistent ratings].*

Perhaps a result of the sometimes imperfect matching of teachers to observers, some teachers who participated in interviews and focus groups criticized peer observer scores — which tended to be lower than those from principals — for failing to reflect the school and classroom contexts within which teachers were working.

In the District of Columbia, for example, two teachers perceived that they had received poor ratings from their peer observers because the peer observers were unfamiliar with their school and the teachers and therefore did not understand the instructional strategies the teachers were using. Teachers in Hillsborough County expressed similar concerns, explaining that because peer observers only occasionally visited their schools, they had a finite number of opportunities to catch teachers at a time when they could be observed, whereas principals could “drop in at any time.” One Hillsborough County teacher said that it was difficult to trust and accept the feedback of someone who came in from the “outside”: “It's much easier to give and get feedback from someone whom you trust and you know. Having someone come in who you don't know and who knows nothing about you, personally, is not helpful to me.” A teacher in Austin said that she would have preferred an evaluation from a teacher from within her school; someone who had more of a sense of the context within which the teacher was working.

[32] District administrators and instructional coaches in Harrison observed teachers periodically for purposes of monitoring and calibrating principal observation scores. Only the classroom observations conducted by a principal counted toward teachers’ final evaluation scores.

Observation Feedback

Teachers in three districts who participated in interviews and focus groups reported that they most valued the observation feedback that was accompanied by clear, specific strategies for ways to improve their practice.

A teacher in Pittsburgh, for example, described having received feedback from her principal that was specific to each domain of the rubric and "easy to follow." A teacher in the District of Columbia described the observation feedback as helpful and accompanied by specific improvement strategies:

> *Honestly, even if I didn't agree with everything that was going on and everything I was evaluated on, every post-observation conference I've had has been really helpful, and they do bring concrete things for you to implement in your classroom. In my observation last year, the [peer observer] had noticed that I was struggling with one class in particular … and he gave me really good, really concrete things to try, and I did and they worked.*

In Harrison, a teacher described the feedback as focusing on only one area of practice at a time rather than on several areas all at once: "[The feedback] is always on one area of growth that I have to focus on, which makes it manageable."

Teachers in the District of Columbia and Hamilton County who participated in interviews and focus groups reported that the informal, verbal feedback that immediately followed the observations was more useful than the written feedback.

These teachers explained that the informal, verbal feedback was more useful because it came faster than the written feedback and, consequently, they could immediately apply it to their practice. A Hamilton County teacher said, for example, that she preferred verbal feedback to a formal written report because the written report comes much later in the process. She explained: "If it's written, and because it comes later, I can't even remember who it was or where it was when [the observation] happened most of the time." A teacher in the District of Columbia said that the advantage to receiving immediate feedback was that she could turn around and immediately implement or incorporate the verbal feedback into her professional practice.

Teachers in four districts who participated in interviews and focus groups reported that they did not receive verbal feedback from their observer or that the feedback was generic and unhelpful.

One teacher in Hamilton said that her principal simply did not complete the observation form: "The major complaint at our school has been that when the principal filled out the observation form, he just recorded what you did and did not give you any feedback, either positive or areas to be strengthened."

Among teachers who received feedback, some reported that the feedback failed to provide them with specific guidance about ways to improve. In Hillsborough County, for example, a teacher said that she practically begged her observer to help her develop strategies to improve, to no avail: "… if you don't give me the example then I cannot do anything. Rather than saying, 'This is wrong, this is wrong.' Tell me why this is 'Developing' and what I can do to make it better?'" Similarly, a teacher in Hamilton County offered suggestions for ways to make the feedback more effective: "Give me some constructive feedback, talk to me about it. Do you like what you see; do you not like what you see? What can I do better in a non-punitive way [that] I think would be real effective, personally?” Two teachers, one in the District of Columbia and one in Hamilton County, suggested that observers simply did not know that they were supposed to provide feedback on ways teachers might improve their practice. The Hamilton teacher suggested that the district should provide observers with a specific feedback template or format to follow as a way of promoting appropriate and constructive feedback:

> *I don't think the administrators know exactly what the feedback is supposed to look like, and so you get different types of feedback from different administrators. So, if there was a [feedback] format or template, then maybe we would all be getting the same information in the same way.*

V. Using Student Performance and Other Data

Despite widespread agreement and empirical support for the notion that teachers contribute significantly to student learning (MET Project 2010; Kane and Staiger 2008; Harris and McCaffrey 2010; Wright, Horn, and Sanders 1997), the eight districts included in this study generated varied and changing strategies for measuring those contributions, suggesting that the ways in which teachers contribute to student learning are still difficult to measure. Nevertheless, measuring teachers' contributions using standardized student assessments has become common practice. Still, some teachers expressed concern about using student assessments to measure their performance when those assessments do not align well with the curriculum they teach.

Key Findings

- Six of the eight districts used multiple approaches for measuring teacher impact on student performance, including individual and/or school-level value-added models (VAMs). VAMs are statistical models that attempt to explain teacher (or school) contributions to student academic growth over time by controlling for school- and student-level variables that may affect student learning. One district discontinued using individual VAMs to assess teacher impact on student performance because VAMs created too much teacher anxiety and its result at the individual teacher level showed little variation to provide useful information.

- Seven of the eight districts included in the study used student learning objectives (SLOs) among the methods used to measure teacher impact on student performance. Central office staff, principals, and teachers in three districts, however, highlighted challenges associated with SLOs. One challenge was setting realistic and consistent goals for student growth. Another was ensuring that principals, who exercised a great deal of discretion when advising teachers on their selection of SLOs, gave teachers fair and consistent advice.

- In four districts, some teachers who participated in focus groups or individual interviews criticized using student achievement data to evaluate their performance, expressing concern that the selected tests (especially the state assessments) would not fully capture their students' actual achievement or that the tests for which they were being held accountable did not align well with the school curriculum.

Methods Used to Analyze Teacher Impact on Student Performance

The following describes the methods districts' used to analyze teacher impact on student performance, including the use of complex statistical techniques, such as individual and school-level value-added models as well as simpler analytic strategies, such as gains in student performance or the extent to which teachers met their student learning objectives (SLOs).

Using Value-Added Models (VAMs)

Value-added models can be distinguished from two other models that are often used to assess school or teacher performance: *status models*, which provide data on student performance at a single point in time, or *growth models*, which document change in the scores of individual students or groups as they move from one grade to the next.

Individual and/or school-level value-added models (VAMs) were among the methods that six of the eight districts used for measuring teacher impact on student performance.

Austin, the District of Columbia, Hamilton County,[33] Hillsborough County, Pittsburgh, and Plattsburgh used value-added statistical models that attempted to explain teacher (or school) contributions to student achievement gains over time. To calculate teacher effects on student performance gains, models require at least two years of student test scores and usually control for other student and school characteristics. That is, VAMs are designed to take account of and control for differences among students with regard to their demographic and educational characteristics, their family background, and other factors that affect performance.

Each of these six districts' VAMs controlled for different school- and student-level factors that influenced student learning.

For example, Hamilton County, which used its state's value-added assessment system model (TVAAS), only controlled for students' prior academic achievement. By contrast, the District of Columbia, Hillsborough County, and Pittsburgh all controlled for various student- and school-level characteristics — in addition to prior academic achievement — when measuring teacher impact on student performance (Exhibit 17). Hillsborough County's model, for example, controlled for prior year student attendance rates, special education status, EL status, and other factors that are predicted to influence student achievement. Because the district is geographically diverse, it included teachers' geographic region in its VAMs.

[33] Hamilton County's decision to include VAMs for measuring teacher impact on student performance was in response to state requirements.

Exhibit 17
Control variables for value-added models used in the District of Columbia, Hillsborough County, and Pittsburgh[a]

District of Columbia	Hillsborough County	Pittsburgh
Assessments:		
▪ Previous year’s math and English/ language arts scores on the District of Columbia Comprehensive Assessment System (DC-CAS)	Previous year’s scores on: ▪ Florida Comprehensive Assessment Test (FCAT) ▪ End-of-semester tests ▪ Other district-designed assessments ▪ AP or IB exams ▪ PSAT scores (only as an independent or control variable; not an outcome or dependent variable)	Previous year’s scores on: ▪ Pennsylvania System of School Assessment (PSSA) ▪ Course-based assessments ▪ Terra Nova ▪ PSAT
Student characteristics:		
▪ Attendance rate (prior year) ▪ EL status ▪ Special education status ▪ Transfer student status ▪ Free/reduced-price lunch eligibility	▪ Attendance rate (prior year) ▪ EL status ▪ Special education status ▪ Transfer student status ▪ Geographic region ▪ Comparative age in cohort	▪ Attendance rate (prior year) ▪ EL status ▪ Special education status ▪ Mobility ▪ Free/reduced-price lunch eligibility ▪ Behind grade for age ▪ Gifted status ▪ Enrolled in specialized courses (AP) ▪ Age ▪ Gender ▪ Race/ethnicity ▪ Grade repeater ▪ Suspension rate (prior year) ▪ District membership ▪ Course pass rate (HS only)
Peer effects/characteristics:		
▪ Class’s average test scores from previous year ▪ Variation in class’s test scores from previous year ▪ Percent of class eligible to receive free/reduced-price lunch eligibility		▪ Gender ▪ Percent free/reduced price lunch eligibility ▪ EL status ▪ Gifted status ▪ Disability rate ▪ Attendance (prior year) ▪ Suspension (prior year) ▪ District membership (prior year) ▪ Average PSSA math and reading

Exhibit reads: The District of Columbia included the DC-CAS subject pretest in math and reading/language arts in its value-added model as well as student characteristics such as attendance, EL status, special education status, whether they had transferred from another school or district, and whether they were eligible for free or reduced-price school lunch.

[a] Hamilton County’s VAM only controlled for students’ prior academic achievement. Plattsburgh controlled for students’ special education status, EL status, and poverty.

SOURCE: Extant data on system specifications (see Appendix D for complete list).

In addition to controlling for many of the same student-level factors as Hillsborough County, the District of Columbia and Pittsburgh also controlled for classroom-level factors, or “peer effects.” The District of Columbia model, for example, controlled for the class’s average test scores and variation in the class’s test scores from the previous year, and the percent of the class that was eligible to receive free or reduced-price school lunch.

Administrators in five of the six districts reported relying on external organizations, including their SEAs, to calculate their VAM scores because the district lacked the technical capacity to conduct their own in-house analyses.

Among the districts using VAMs, three engaged external vendors to develop and compute VAM scores, including Mathematica Policy Research (District of Columbia and Pittsburgh) and the University of Wisconsin Value-Added Research Center (Hillsborough County). District administrators in all three districts explained that, despite the added expense, involving external vendors was a worthwhile investment for their respective districts. As one central office representative in Hillsborough County, for example, explained, “I think it’s well worth [the expense] for us because we can defend any ratings that we have because [the models] have all been well thought through.” Indeed, stakeholders in two districts believed that engaging external vendors validated the evaluation process. “We do not have the expertise to do what Mathematica does,” explained an administrator in the District of Columbia. Similarly, a district administrator in Pittsburgh stated, “[The calculations are] done by people with expertise in statistics and analysis, so there’s validity that comes with this.” Among the remaining districts, Hamilton County and Plattsburgh relied on their state education agencies to build and analyze the value-added models and provide these districts with individual value-added scores.

Only Austin conducted its own VAM analyses and, after the first year of implementation, eliminated individual, teacher-level VAMs to evaluate teacher performance.

Indeed, Austin opted to rely only on school-level VAMs and teacher-level SLOs to evaluate teacher performance after determining that the teacher-level VAMs created too much teacher anxiety and the individual teacher effects varied too little to warrant their inclusion. A district administrator explained that the decision to transition from individual teacher-level VAMs to SLOs was based primarily on three factors: 1) value-added scores did not appear to affect the teachers’ final evaluation scores, with SLOs serving as a strong predictor of value-added data; 2) the level of teacher anxiety provoked by teacher-level value-added modeling did not seem worth the benefits; and 3) value-added modeling added unhelpful complexity and time lags to the evaluation system. Ironically, for similar reasons, as noted earlier in the report, the District of Columbia opted to eliminate school-level VAMs because they did not sufficiently differentiate between the performance of teachers within a school and because district administrators believed that including the measure in teacher scores created a disincentive for teachers to work in the lowest-performing schools.

Finally, in an effort to make the complex analytic process more transparent, the District of Columbia provided its teachers with a non-technical, step-by-step description of the process used to calculate individual value-added scores and then how those raw scores were converted to IMPACT scores (Exhibit 18).

Exhibit 18
Information the District of Columbia Public Schools provided to teachers about VAM calculations

Step 1: Confirm student rosters.

Teachers indicate which subject(s) they taught, which students they taught within each subject, and what proportion of the time each student was actually instructed by the teacher compared to the total instructional time each student was assigned to that teacher. Allows teachers to note whether student mobility is a factor or whether students are pulled out of class on a regular basis for special education services or other reasons.

Step 2: Calculate average *likely* DC-CAS score for each teacher's students.

After the DC-CAS tests have been scored, statisticians calculate the average score that a teacher's students were *likely* to have achieved by analyzing the performance of all students in the District of Columbia (e.g., a student who received a score of 20 on last year's DC-CAS is likely to perform about as well as other students in the same grade who received a 20 last year).

Step 3: Calculate average *actual* DC-CAS score for each teacher's students.

Statisticians average the actual scores of all of the students in a teacher's class at the end of the year, with each student weighted according to the amount of instructional time spent in the teacher's class.

Step 4: Subtract the average likely score from the average actual score.

The difference between how students actually perform and how they were likely to perform is the teacher's "value-added." For example, the students in a class have an average *actual* score of 65, which exceeds the average *likely* score of 60 by 5 points. Thus, this teacher has a value-added score of +5 (65 – 60 = +5).

Step 5: Convert the raw value-added score into an IMPACT score.

The raw value-added score is converted to a 1 to 4 IMPACT scale score. In reading, a raw value-added score of 5.8 and above converts to an IMPACT score of 4.0; a raw score of 3.0 to 3.3 converts to an IMPACT score of 3.3. In math, a score of 7.7 and above converts to an IMPACT score of 4.0; a raw score of 4.0 to 4.4 converts to an IMPACT score of 3.3.

Exhibit reads: The first step the District of Columbia Public Schools used to calculate teachers' VAM scores was to confirm their student rosters.

SOURCE: The District of Columbia Public Schools, 2012b, pp 8–10).

Analyzing SLOs and Other Student Performance Data

Each of the seven districts that included SLOs among the outcome measures used to evaluate teacher impact on student performance adopted different ways of analyzing the SLO results.

In St. Mary's County, for example, principals conducted mid- and end-of-year conferences with teachers to review their progress on their SLOs. During the mid-year conference, teachers could discuss their students' performance and their approach to addressing the needs of specific students who might be struggling to meet performance targets identified in the teacher's SLOs. At the end-of-year conference, the teacher presented evidence of student learning as reflected in the SLOs.

Based on that review, the principal produced a written report that reflected "the quality of performance based on the evidence of student learning, teacher interventions and supports, as well as considerations of [particular student needs]" (Exhibit 19). In addition, the principal gave a teacher a score on a scale of 1–4 for each element of his or her SLOs, including the summative assessments (i.e., the state test, if it applied), formative assessments, performance assessments, student growth, and classroom performance.

In Austin, principals calculated teachers' SLO scores based on the percentage of students who met the teachers' SLO. For example, if 100 percent of a teacher's students achieved the SLO, then the teacher received the maximum possible number of points awarded for that SLO; however, if only 50 percent of a teacher's students achieved the SLO, then the teacher received only half the points.

In Hamilton County, teachers could select SLOs that "most closely aligned with their responsibilities," and SLO scores were awarded based on a range in the percent of students who met the teacher's SLO. For example, if 80–100 percent of a teacher's students achieved the SLO, then the teacher received the maximum possible number of points for that SLO (5 points); if 60–79.99 percent of a teacher's students achieved the SLO, the teacher received 4 points; for 40–59.99 percent of the students achieving the SLO, the teacher received 3 points, and so forth.

Exhibit 19
St. Mary's County SLO scoring rubric for formative and local assessments

	Ineffective (1 pt.)	**Developing (2 pts.)**	**Effective (3 pts.)**	**Highly Effective (4 pts.)**
Overall Performance	The attributed students do not meet performance goals for formative/local assessments	The attributed students meet performance goals for formative/local assessments in the aggregate, yet identified groups of students may not meet the identified goals	The attributed students meet performance goals for formative/local assessments in the aggregate and across the majority of identified groups of students	The attributed students meet or exceed performance goals for formative/local assessments in the aggregate and across all identified groups of students

Exhibit reads: A teacher's overall performance on their SLOs is rated as ineffective if their attributed students do not meet their performance goals on the formative/local assessments.
SOURCE: St. Mary's County Public Schools (2012).

Central office staff, principals, and teachers in three districts highlighted challenges associated with using SLOs.

One challenge was setting realistic and consistent goals for student growth. Another was ensuring that principals, who exercised a great deal of discretion when advising teachers on their selection of SLOs, gave teachers fair and consistent advice on the selection of assessments and learning objectives. A district administrator in Austin made it clear that SLOs were not necessarily reliable across teachers, in part because principals have authority to determine SLOs in consultation with teachers. She explained,

> *The SLOs are implemented differently on some campuses than on others. There's a lot of principal discretion in terms of imposing different rules about what is acceptable and what's not acceptable for an SLO. And we're just now starting to study those things. So for example, in order to meet your SLO, at least 75 percent of the students have to meet the objective. But some principals have upped the threshold to 80 percent of students have to meet the objective for those teachers to meet their SLO. Other schools have said not only must 80 percent of students*

> *achieve the objective, but 80 percent of students have to reach a score of 75 on the assessment regardless of where they started. So it imposes a different level of rigor to the SLO.*

In an attempt to address this problem, Austin created audit teams that reviewed and approved every SLO at the beginning of the school year to ensure comparability and rigor across teachers' SLOs and to ensure that the assessments used to measure the SLOs were high quality. According to a district administrator, the auditing process overturned some of the SLOs.
At the end of the year, the audit team reviewed a sample of SLOs to ensure that the results were properly documented.

Similarly, although the District of Columbia principals were expected to help teachers ensure that their SLOs were not too easy or too difficult, teachers who participated in site visit interviews in the district believed that SLOs were not consistent across and within schools. One teacher stated, "[W]hat [an SLO] looks like in [our school] doesn't look the same in all other schools. Therefore, as a teacher, I wonder. This school may have a very easy way to go with [SLOs] and that person may be [given a score of] 4.0, whereas my principal is very 'You must have this, you must have that' and I end up with a [score of] 1.5 or 1.6."

St. Mary's County permitted teachers to select students for whom they would be held accountable, following concerns raised by the education association that teachers would be held accountable for students that they did not teach as well as for circumstances, such as student absences, that were beyond a teacher's control.

To allay these fears, the St. Mary's County educator effectiveness committee decided that teachers would have "attributed groups" and set specific goals for those groups. For example, a teacher could identify six students who were three years below grade level and say that his/her goal was to get those students to one year below grade level by the end of the school year. Teachers, in fact, were not required to have SLOs for all of the students they served and could include students in more than one attributed group.

Instead of a statistical model, Harrison created a point system for each measure of student performance included in a teacher's evaluation.

Specifically, each teacher had his or her own evaluation template that consisted of some combination of the state test, district-designed end-of-course and quarterly assessments, school-level performance, and a teacher-defined SLO. For teachers of non-state-tested subjects and grades, other performance criteria were included in their evaluation templates, including student performance assessments and portfolios. Ultimately, teachers' student performance scores consisted of eight parts, each worth six points for a subtotal of 48 points (the district awarded every teacher 2 bonus points, for a total possible score of 50 points).

To assess teacher impact on student performance, Harrison administrators placed each student into one of four peer groups — "advanced," "proficient," "partially proficient," and "unsatisfactory" — based on their previous year's performance on the state reading exam or the district reading test.[34] The district

[34] Students in kindergarten through third grade were placed into one of three groups based on their performance on the DIBELS assessment.

then calculated the median assessment score for each peer group, and teachers received a score based on how their students performed relative to the median within their respective peer groups.

Teacher Perspectives on Using Student Achievement and Other Data to Evaluate Performance

In four districts, some teachers participating in focus groups or individual interviews criticized the use of student achievement data to evaluate their performance.

In Hamilton County, Harrison, Hillsborough County, and Pittsburgh,[35] some teachers expressed concern that the selected tests would not fully capture their students' actual achievement, while others believed that the student assessments for which they were being held accountable did not align well with the school curriculum they taught. As one teacher in Harrison explained, "It's hard for us to know whether we're not teaching correctly or there's a problem with the test. It's frustrating when you know your kids are proficient but it doesn't show on the district assessment." One special education teacher in Hillsborough County noted that the tests the district used for the teacher evaluation system did not demonstrate the variation in students who were performing significantly below grade level.

A teacher in Pittsburgh raised concerns about using student test scores to measure her performance when there were too many factors affecting student performance that she could not control: "I'm worried about using student test scores and things we can't control. No doctor would want to be measured by their patient survival rate. It's inconsistent. If you look at scores everywhere, economically disadvantaged students will struggle." A principal in Hamilton County noted that the value-added scores that the teachers in her building received did not align with her perception of their abilities as teachers: "… so you see a wonderful, wonderful teacher [whose] got a score of one or two on their TVAAS, and you wonder, 'How did that happen?'"

Teachers interviewed in Austin and Pittsburgh had mixed views on the validity and usefulness of student surveys.

In Austin, interviews with teachers found that some teachers appreciated receiving student input through surveys. Two teachers explained that they used survey results to make adjustments to their teaching style. One said, "I always look at the student surveys to know how I'm going to improve because I deal with [the students] more than anybody else." A teacher in Pittsburgh explained that she benefited — perhaps unduly — from the grade level she taught: "I'm in a good spot as a kindergarten teacher; the kids love me." Two other Pittsburgh teachers said they valued the student input: "… the student input is appreciated. [The survey] can help teachers improve their practice," explained one. The other teacher said that although she understood why some teachers had concerns about the validity of the survey, she believed that the data had potential value: "If most students answer a certain question a certain way, it's something to think about."

Teachers participating in individual interviews in both districts — including teacher leaders (Pittsburgh) as well as elementary and high school teachers — however, expressed concerns about administering the survey to younger students. In Austin, teachers explained that the wording of the questions posed

[35] We were unable to conduct teacher focus groups in the District of Columbia.

challenges. One teacher, for example, said that many questions were confusing to her five- and six-year-old students: “The questions [are phrased as] double negative[s]: ‘Does your teacher ever not allow you to not follow the rules?’ … The kids seemed just as confused as anything.” A Pittsburgh reading teacher had a similar opinion about the student surveys: “Some of the kids can’t read and don’t always get the question.”

Teachers also worried about the validity of the survey as students’ feelings toward a particular teacher may influence their responses. For example, Austin conducted teacher focus groups to assess the utility of the student surveys and found that teachers worried that students may “have it in for them” and respond to the survey as negatively as possible. A teacher in Pittsburgh explained, “I feel sorry for the middle school team because that’s when kids fight authority. I think [middle school teachers] might get slammed, especially if [they’re] diligent [teachers].”

VI. Using Teacher Evaluation Results

Until very recently, almost all teacher evaluation systems in the United States had very limited capacity to identify both effective teachers as well as those who needed help or who were ineffective. “The Widget Effect,” a report from The New Teacher Project (Weisberg et al. 2009), summed up the problem this way:

> *In practice, teacher evaluation systems devalue instructional effectiveness by generating performance information that reflects virtually no variation among teachers at all. This fundamental failing has a deeply insidious effect on teachers and schools by institutionalizing indifference when it comes to performance. As a result, important variations between teachers vanish. Excellence goes unrecognized, development is neglected, and poor performance goes unaddressed* (p. 10).

Consequently, these systems contributed little to efforts to improve education. They could not, for example, provide useful data to inform planning for teacher professional development and decisions about which teachers could lead improvement efforts, which teachers should be rewarded, or which teachers should be released.

This chapter describes districts’ teacher evaluation rating systems and how they used teacher evaluation results for a range of purposes, including targeted professional development and support, career ladders and performance pay, and in some instances, redeployment or release of teachers identified as ineffective.

Key Findings

- All eight districts included in the study used similar rating systems for classifying a teacher’s overall performance. According to administrators in two districts, early outcomes data suggested that the new teacher evaluation systems were succeeding in rating teachers across a wider range of performance levels than previous evaluation systems.

- Seven of the eight districts put in place specialized, targeted professional development and support for teachers who were rated as low-performing or at risk of termination. For teachers who were not low-performing, however, five of the eight study districts included in the study had not yet created clear linkages between district-sponsored professional development and teacher evaluation results. That is, unless rated as low-performing, teachers needed to rely on the advice and feedback of their observer — usually the principal — to know what professional development they needed and how to access it within the district. None of these districts had made explicit the availability of district-sponsored professional development to address a particular need as identified in a teacher’s evaluation.

- Six of the eight study districts had used or planned to use teacher evaluation scores to redeploy or release low-performing teachers but narrowly defined the circumstances under which this could happen. Three districts, for example, released only non-tenured or new teachers on the basis of their evaluation scores whereas releasing tenured or veteran teachers required more evidence. Two districts opted to use the observation data and feedback conferences to counsel out or redeploy their ineffective teachers rather than release them outright.

- Three districts linked — or planned to link — teacher evaluation scores to a career ladder, which sometimes included performance pay. None of the districts, however, reported using teacher evaluation system results to redistribute high-quality staff to low-performing schools.

Teacher Evaluation Rating Systems

All eight districts included in the study used similar four- or five-level rating scales for classifying a teacher's overall performance.

The highest rating for teachers was "highly effective" (Austin, District of Columbia, Hamilton County, and Plattsburgh), "exemplary" (Harrison), "effective" (St. Mary's County), "distinguished" (Pittsburgh) or "5" (Hillsborough County) (Exhibits 20 and 21). The lowest rating was "ineffective" (District of Columbia, Plattsburgh, and St. Mary's County), "unsatisfactory" (Austin, Harrison, and Pittsburgh), "does not meet standards" (Hamilton County), or a "1" (Hillsborough County).

Although the District of Columbia had originally used a four-level rating system, it added a fifth level, "Developing," to the mid-point of its rating scale in 2012–13. A district administrator explained that the range of scores for an "Effective" rating in the district had been too wide (250–350 points) and sent the wrong message to teachers at the lower end of the range who, in fact, needed to improve. Indeed, more than two-thirds of teachers in the district were scoring in the "Effective" range, but other data sources, including student achievement results, classroom observations, and feedback from evaluators suggested that "Effective" teachers were quite diverse in terms of their professional practice and skill. More importantly, however, was the fact that very high performers at the top of the "Effective" scale were not being sufficiently recognized nor were teachers at the very bottom of the scale receiving needed supports. A district administrator explained the rationale behind the decision to add the rating level: "Teachers were getting the message that they were effective [even though they] weren't actually meeting their principals' expectations [or] achieving the kind of student results we needed but thought that they were meeting the bar." A principal confirmed the perception that the score sent teachers the wrong message:

> *In my building alone, I have teachers who score a 349 and teachers who score a 260, and they're totally different types of teachers with respect to planning, looking at student data, how they work with families to increase student achievement, their everyday instructional practices, being prepared just to teach every day. They're totally different teachers and I feel as if those two folks shouldn't walk around both being effective, because one is not very effective where the other is a point away from highly effective. I was a huge advocate of [adding] another category [to the rating system].*
>
> *[Principal, District of Columbia]*

Exhibit 20
Variation in teacher evaluation system ratings and results, by fully implementing districts

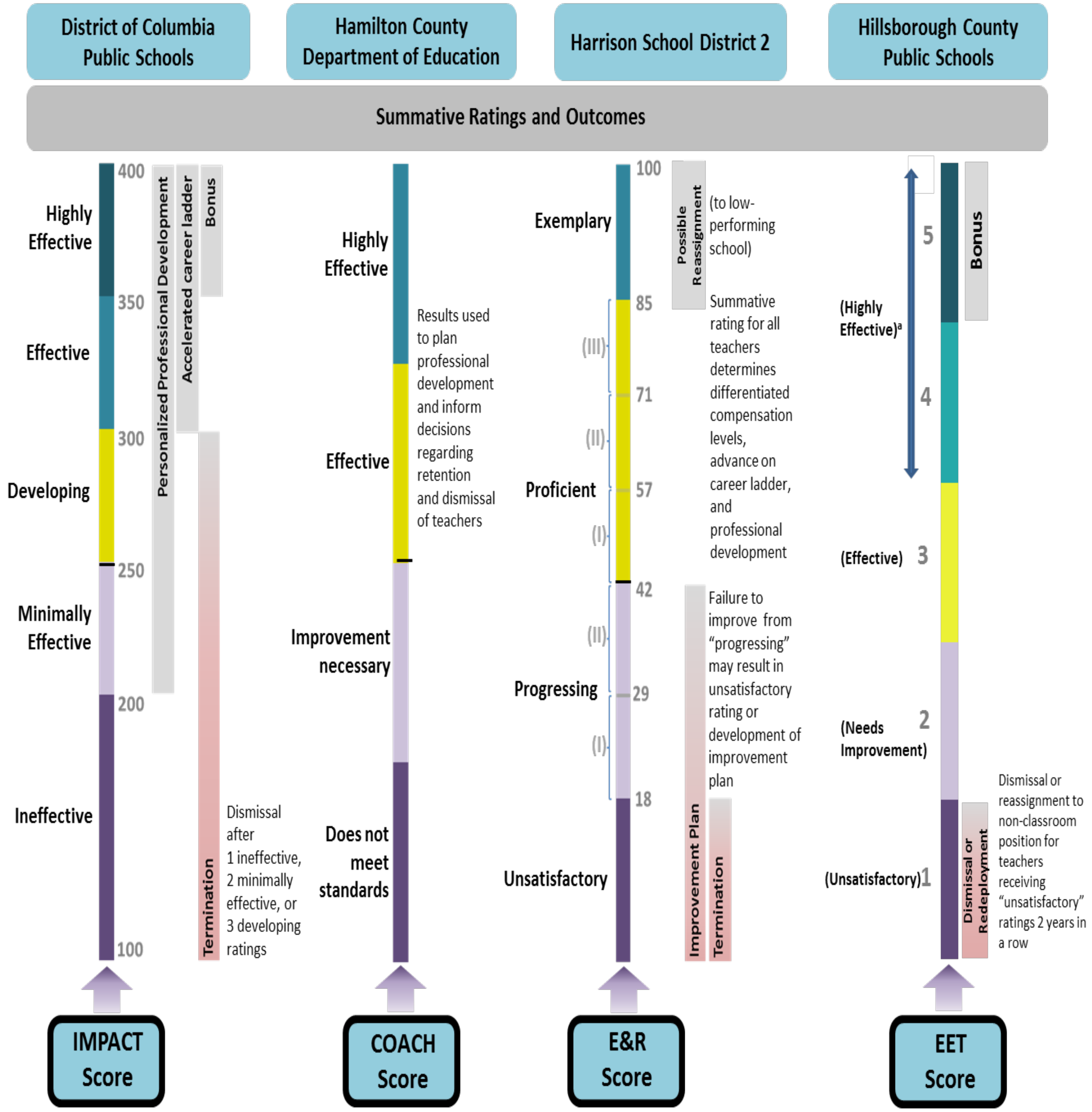
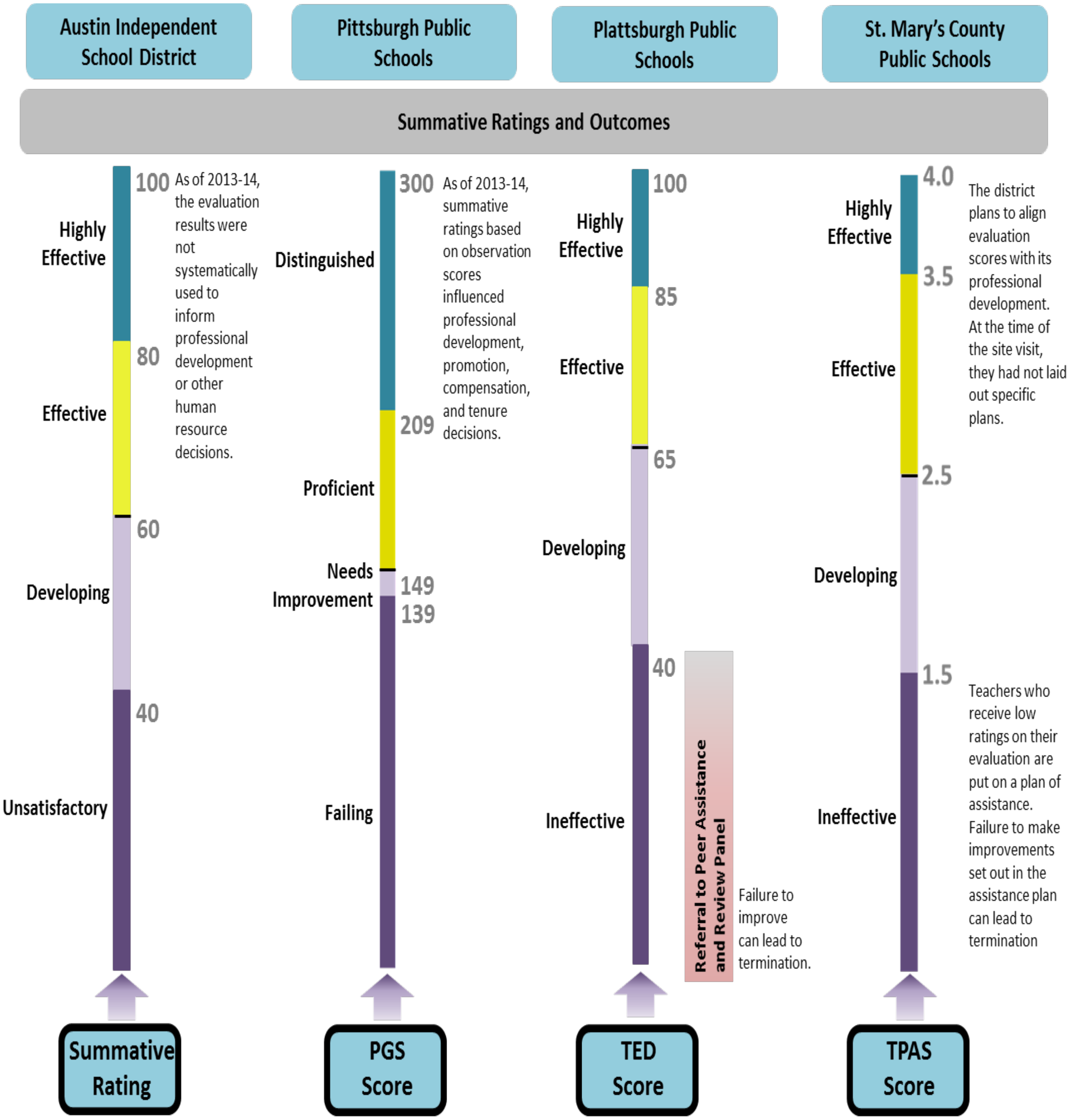

District of Columbia Public Schools
Hamilton County Department of Education
Harrison School District 2
Hillsborough County Public Schools
Summative Ratings and Outcomes
Highly Effective
Effective
Developing
Minimally Effective
Ineffective
400
350
300
250
200
100
Personalized Professional Development
Accelerated career ladder
Bonus
Termination
Dismissal after 1 ineffective, 2 minimally effective, or 3 developing ratings
IMPACT Score
Highly Effective
Effective
Improvement necessary
Does not meet standards
Results used to plan professional development and inform decisions regarding retention and dismissal of teachers
COACH Score
Exemplary
Proficient
Progressing
Unsatisfactory
100
85
71
57
42
29
18
(III)
(II)
(I)
(II)
(I)
Possible Reassignment
(to low-performing school)
Summative rating for all teachers determines differentiated compensation levels, advance on career ladder, and professional development
Failure to improve from "progressing" may result in unsatisfactory rating or development of improvement plan
Improvement Plan
Termination
E&R Score
5
4
3
2
1
(Highly Effective)a
(Effective)
(Needs Improvement)
(Unsatisfactory)
Bonus
Dismissal or Redeployment
Dismissal or reassignment to non-classroom position for teachers receiving "unsatisfactory" ratings 2 years in a row
EET Score

Exhibit reads: To receive a "Highly Effective" rating in the District of Columbia, teachers must earn an IMPACT score of between 350 and 400 points on the system's 400-pt. scale. Highly effective teachers receive a cash bonus, a base salary increase, and are placed on the district's accelerated career ladder program.

[a] Evaluation ratings are reported on a 1–5 scale at the district level ("1" is the lowest and "5" the highest overall evaluation score). For purposes of state reporting, Hillsborough County converts its 5-point scale to the state's 4-point scale by combining Level 4s and 5s into the "Highly Effective" group. Teachers scoring at Level 3 are assigned to the "Effective" group, Level 2s are "Needs Improvement," and Level 1s are "Unsatisfactory."

SOURCE: Site visit interviews; extant data on system specifications (see Appendix D for complete source list).

Exhibit 21
Variation in teacher evaluation system ratings and results, by partially implementing districts



As of 2013-14, the evaluation results were not systematically used to inform professional development or other human resource decisions.

As of 2013-14, summative ratings based on observation scores influenced professional development, promotion, compensation, and tenure decisions.

Failure to improve can lead to termination.

The district plans to align evaluation scores with its professional development. At the time of the site visit, they had not laid out specific plans.

Teachers who receive low ratings on their evaluation are put on a plan of assistance. Failure to make improvements set out in the assistance plan can lead to termination

Exhibit reads: To receive a "Highly Effective" rating in Austin, teachers must earn a summative rating score of 80 to 100 points on the system's 100-pt. scale. At the time of the study, the evaluation results were not systematically used to inform professional development or other human resource decisions.

SOURCE: Site visit interviews; extant data on system specifications (see Appendix D for complete source list).

District administrators in the District of Columbia and Pittsburgh—the two districts that had analyzed the distribution of evaluation scores over time—suggested that their new teacher evaluation systems were succeeding in rating teachers across a wider range of performance, from ineffective to highly effective.

A district administrator in Pittsburgh, for example, said that in 2008–09, 99 percent of Pittsburgh teachers were rated "Satisfactory" and 1 percent were rated "Unsatisfactory." Results in 2012–13, a no-stakes preview year that offered teachers a chance to understand their performance in the new system, showed 73 percent of teachers were rated "Proficient," 16 percent "Distinguished," 5 percent "Needs Improvement," and 9 percent "Failing." In the District of Columbia, because the district had already dismissed many of their ineffective teachers — and improved the practice of others — during IMPACT's first two years, the percent of teachers rated as "Ineffective" was small and the percent rated as "Effective" increased between 2010–11 and 2011–12 (Exhibit 22).

Exhibit 22
Percent of District of Columbia teachers scoring at each IMPACT rating level, in 2010–11 and 2011–12

	Percent of teachers in 2010–11	Percent of teachers in 2011–12
Ineffective	2	1
Minimally Effective	13	9
Effective	69	68
Highly Effective	16	21

Exhibit reads: Two percent of teachers in 2010–11 were rated "Ineffective" compared with one percent of teachers who received that rating in 2011–12.
SOURCE: DCPS, 2010–11; DCPS, 2011–12c.

A district administrator explained that the quality of instruction in 2012-13 was vastly different from that of 2007–08, when the process to develop a new teacher evaluation system began. "You walk into any one of my classrooms and teaching is happening across the board. That was not the case in 2007 when we [started]. When you look at, again, our ability to grow and develop teachers, I am confident that that is happening more now than ever before."

Redeployment or Release

Six of the eight study districts had used or planned to use teacher evaluation scores to redeploy or release low-performing teachers but narrowly defined the circumstances under which this could happen.

For example, Hamilton County, St. Mary's County, and Harrison would release only non-tenured or probationary teachers on the basis of an ineffective rating. In Hamilton County, two principals participating in site visit interviews explained that scoring "Ineffective" based on the observation data was sufficient evidence for them to make "non-renewal" decisions about teachers. One principal explained that because she did not have the value-added data available to her at the end of the school year[36] when she must make her non-renewal decisions, she was comfortable basing her decisions solely on the classroom observation data: "I wouldn't hire them back for a second year based on their observation scores [for which they received an 'Ineffective' rating]" (Exhibits 20 and 21).

Releasing tenured or experienced teachers (i.e., teachers who have been on the job for three or more years) in Hamilton and St. Mary's Counties was more complicated and required additional evidence beyond teacher evaluation scores. A principal in Hamilton County, for example, described having to produce an entire year of documentation to demonstrate that a teacher who is experienced who had been rated by the previous principal as "satisfactory" was, in fact, unsatisfactory, and should be dismissed.

In Harrison, non-probationary teachers receiving "Unsatisfactory" ratings could be recommended for release from the district. However, Colorado state statute provided protections for these teachers and placed the burden of proof on the district to release them through the teacher dismissal process, which included a hearing overseen by a hearing officer whom both the district and teacher have sanctioned. The process resembles a court proceeding, and the district and teacher are "almost always" represented by attorneys, according to a district administrator. The hearing officer issues a finding which the Board of Education may accept or reject. Colorado state statute requires that prior to dismissal proceedings, a teacher be given adequate time to improve classroom performance and/or student achievement, with support from the district. At a minimum, non-probationary teachers determined to be ineffective have two to three school years to show improvement. According to one district administrator, as of 2012, no non-probationary teacher had been deemed ineffective and assigned a formal improvement plan.

In the District of Columbia the consequences of the teacher evaluation system did not distinguish between new and experienced teachers and would release or "separate" from the district either after one poor rating. All teachers, in fact, were subject to the same consequences based on the results of their evaluation, regardless of experience. For example, teachers who received one "Ineffective" summary rating were dismissed at the end of the school year (as shown in Exhibit 20 above, this amounted to 2 percent of teachers in 2010–11, and 1 percent of teachers in 2011–12). Teachers receiving a "Minimally Effective" rating were given a chance to improve in the next school year but would be dismissed upon receiving a second "Minimally Effective," or lower, rating. Finally, the District of Columbia also dismissed teachers who received the mid-level, or "Developing," rating after they received three of them.[37]

Rather than release teachers outright, district administrators in Hamilton County and Hillsborough County described using the observation data and feedback conferences to counsel out or redeploy their low-performing teachers to other positions in the district.

In both districts, administrators explained that the observation data were extremely useful for purposes of identifying ineffective teachers. In Hamilton County, for example, a high-ranking district official said

[36] Because VAM scores stem from the state assessment results, which do not become available until the summer.

[37] A 2013 study of the District of Columbia's IMPACT system by the National Bureau of Economic Research (Dee and Wyckoff) found that dismissal threats increased the voluntary attrition rate among low-performing teachers, increased the retention rate among high-performing teachers, and improved the performance of teachers who had previously been deemed low-performing but who had remained in the district. In addition, the study found that the offer of financial incentives also improved the performance of high-performing teachers.

that 60 teachers had left their positions between January and May of 2012 due to low classroom observation scores. "I feel like we have done our job [well] if we lose low performers. We lost 60 teachers between January and May, and I feel like a good percentage of those people retired or resigned because they knew the evaluation was showing low performance." Similarly, and in an effort to avoid dismissing throngs of teachers, Hillsborough County counseled many of their lowest-scoring teachers out of their teaching positions. In addition, the district created additional pathways for their low performers, including allowing struggling teachers to transfer into assistant teaching positions. As one administrator explained: "… there were probably 50 to 60 who were counseled out or took different jobs, so we still achieved the purpose without having to fire someone. And then a large group just plain old improved, which is what we wanted in the first place." Nevertheless, in Hillsborough, any teacher who received an "unsatisfactory" rating two years in a row was eligible for release. These teachers received individualized support and had one school year to show improvement.

None of the eight study districts reported using teacher evaluation system results to redistribute high-quality staff to low-performing schools.

However, a review of Harrison's teacher evaluation system design documents indicated that the district intended to transfer highly rated teachers (at Proficient II or higher on the career ladder) to low-performing schools. In interviews, however, no district administrator in Harrison reported using evaluation system results for this purpose, and the design document suggested that this would only happen sparingly: "Proficient II and higher teachers should expect to be transferred to schools that required more skilled teachers. The minority of these higher level teachers will be transferred, as the District will not want to penalize a school that has worked hard to develop a highly effective team of teachers" (Harrison School District 2 2010; Section 14, p. 13). In the District of Columbia, teachers who taught in low-performing schools were awarded larger bonuses for receiving "highly effective" ratings than those who taught in higher performing schools, which arguably indirectly encourages the re-distribution of the teaching force.

Career Ladders

Three districts linked — or planned to link — teacher evaluation scores to a career ladder, which sometimes included performance pay.

Harrison, the District of Columbia, and Hillsborough County created career ladders that linked teacher evaluation scores to teacher opportunities to rise in stature and pay within their respective districts.

In site visit interviews, district officials said that Harrison's salary scale for beginning teachers was well above the salaries a highly rated teacher could command in most districts in the state. Also, according to district administrators, teachers in Harrison were able to progress through the teacher rating levels faster than a traditional teacher salary schedule would provide, generating a higher earning potential for a teacher sooner. Consequently, a teacher in Harrison had more earning power over the span of a teaching career than in other districts in the state. That is, while it would take a beginning teacher 12 years, on average, to earn $48,000 in most districts in the state of Colorado, a beginning teacher in Harrison rated as effective could move from Novice to Proficient I (earning a $48,000 annual salary) within four years (Exhibit 23).

Exhibit 23
Harrison School District 2 *Effectiveness and Results* career ladder salary scale

Teacher effectiveness levels	Salary scale	
1. Novice	$35,000	**Probationary Teachers**
2. Progressing I	$38,000	
3. Progressing II	$40,000–$44,000 [a]	
4. Proficient I	$48,000	**Non-probationary Teachers**
5. Proficient II	$54,000	
6. Proficient III	$60,000	
7. Exemplary I	$70,000	
8. Exemplary II	$80,000	
9. Master	$90,000	

Exhibit reads: In Harrison School District 2, probationary teachers rated as "Novice" received an annual salary of $35,000.

[a] $44,000 is reserved for teacher recruitment purposes.
SOURCE: Miles2011.

Offering higher salaries, Harrison district officials explained, offered the appropriate incentive to teachers to work hard and improve. To be affordable, however, the district designed the career ladder and evaluation system to ensure that not every teacher could achieve the highest level on the career ladder at the same time. To begin, the system did not permit teachers to advance more than one level on the career ladder in a single year, in part to create an incentive for teachers to remain in the district. Relatedly, beginning teachers were required to teach in the district a minimum of three years to achieve Proficient I status but risked termination if they failed to progress to at least Progressing II within their first three years in the district.

Each year, the district looked at historical achievement data to adjust student achievement cut scores for the teacher evaluation ratings. The target distribution identified about 40 percent of teachers who qualified for Novice, Progressing I, and Progressing II ratings. The balance of the target distribution identified those teachers who earned Proficient I or higher.

In the end, the Harrison career ladder system was designed in such a way that teachers' salaries would no longer be predictable but could fluctuate up or down based on their annual evaluation score.[38] As of 2011–12, approximately 15-20 percent (n=100-130) of the 650 classroom teachers on the Effectiveness and Results plan would be rated "Proficient II" or higher on the salary scale.[39]

[38] "Starting in the 2014–15 school year, a teacher may be moved to a lower level after two consecutive years of lower performance. The teacher will remain at the lower level for a least one year and will receive the salary commensurate with that level (except that the salary of a non-probationary teacher currently employed full time by the District may not be lower than his 2009–2010 salary)" (Miles 2011, pp. 36-37).

[39] http://www.chalkbeat.org/posts/co/2010/01/22/harrison-launches-new-merit-pay-plan/#.V8iU5Y-cFfw

The District of Columbia introduced its career ladder, LIFT, in 2012–13. The career ladder consisted of five teacher levels: (1) teacher, (2) Established teacher, (3) Advanced teacher, (4) Distinguished teacher, and (5) Expert teacher, and each was associated with increasing responsibilities and professional opportunities for teachers. Although teachers could steadily progress through the career ladder through years of experience, teachers earning highly effective ratings could accelerate their promotion. Teacher opportunities included leadership positions, fellowships, committee membership, and administrative positions. Teachers at higher levels on the career ladder could also receive fewer classroom observations each year. According to administrators, the LIFT career ladder was one of the more popular components of the IMPACT system, as it rewarded teachers and publicly recognized their success. One administrator said that she believed that the LIFT program "professionalized teachers" and provided them with the recognition and respect that, heretofore, was only achieved by moving out of the teaching profession and into administration: "I would likely have stayed in the classroom if the growth opportunities through LIFT had been available to me when I was a full-time teacher."

Since 2003, Hillsborough County has paid teacher bonuses, to which it allocates over $10 million in funding each year. To teachers, the opportunity to receive a cash bonus was among the most attractive features of the evaluation system. In considering the long-term implications of sustaining a performance pay system, however, one Hillsborough County community stakeholder cautioned districts to plan accordingly. She argued that a system that both supports and rewards continuous improvement faces the potential to financially cripple itself, ultimately, if every teacher performed at the highest standard of effectiveness. She argued that districts implementing performance pay must prepare for a full complement of highly effective teachers:

> *If this system works, then we will have more and more teachers demonstrating high effectiveness. And that's what you want. You don't want to ever cap excellence and ever say that just the top 10 percent or 25 percent will get to make a lot of money. That doesn't bring people into a field. And then, what is that saying? That the other 90 percent of the kids are not important? Tell that to their families. 'We're only going to give 10 percent of you [students] great teachers, and the rest of you are going to have this churn of teachers turning over because we're not going to pay them.' So if the system works, if it does what you want it to do, you should expect that 80 percent of your teachers are going to be highly effective, and you're going to have other people growing to that point while folks are leaving, retiring, or whatever.*
>
> *[Community Stakeholder, Hillsborough County]*

As a community stakeholder in another district explained, the system that linked high evaluation ratings with high salary increases or cash bonuses was financially viable because the district designed its evaluation system and hiring practices to ensure that it would never have a 100-percent exemplary teaching force:

> *I have no concern about most teachers eventually achieving at the upper levels of the pay scale — I know that's not going to happen. There will be people who come into this profession and decide that's not what they want to do. And we have seasoned teachers who are at progressing and not proficient. If it gets to the point where everybody gets to be a [highly effective] teacher, then we're not evaluating everybody appropriately.*

The Appeals Process

Three of the four fully implementing districts had implemented an appeals process as part of their teacher evaluation system design, but two of the districts limited the conditions warranting a legitimate appeal.

For example, the District of Columbia allowed an appeals process only for teachers facing possible dismissal or a hold on their salary step increase, and only in such cases in which an error occurred during the evaluation process, such as conducting an insufficient number of observations. The District of Columbia teachers could not contest scores that they believed did not accurately reflect their teaching.

In Hillsborough County, teachers could appeal a score based on procedural violations, such as a missed deadline by an observer or an incorrect data entry. In addition, Hillsborough County teachers could appeal their score only if principal scores and peer observer scores were substantially different. Like the District of Columbia, Hillsborough County teachers could not appeal their rating simply because they disagreed with it. One district administrator offered the following rationale for the limited appeals process in Hillsborough County: "… if you're a teacher, there were seven times during the year when you had a chance to make sure we were counting the right kids and counting the right tests. So in value-added, if you have a problem with your score, what you're really saying is you have a problem with the fact that you didn't correctly check the kids that counted."

Hamilton County simply adopted the state's appeals process, which was written into the state's new teacher evaluation law, and allowed teachers to file a grievance within 15 days of receiving their summative evaluation score. The appeals process included the following levels or stages:

Level 1: The teacher appeals their rating directly to the person who conducted their classroom observations (e.g., principal, assistant principal, district special education coordinator); and the evaluator either changes the rating, or keeps the rating and provides sufficient explanation so that the teacher withdraws the appeal.

Level 2: If the evaluator does not change the rating and the teacher is not satisfied with the evaluator's explanation, the teacher then appeals to the school district and the superintendent assigns a central office administrator to handle the appeal. The teacher can only appeal whether the evaluator followed the correct procedure in conducting the evaluation and a central office administrator must determine.

Among the three districts that had formal appeals processes, district administrators reported that the number of teachers appealing their evaluation scores was negligible.

To their great surprise, administrators in Hamilton County reported that only three teachers appealed their rating in the last year. As one administrator commented: "We kind of were geared up for the teacher union to try to force hundreds and hundreds of appeals to try to jam the system." In Harrison, final evaluation ratings were not subject to further review nor "grievable." However, if a teacher felt that the evaluation *process* was conducted inappropriately or that a principal was "being unfair," a teacher could file a grievance with the district's Human Resources Department. As of spring 2012, a district administrator said that they had received only one grievance in two years. In the District of Columbia, the number of appeals was somewhat higher. In 2010–11, for example, there were 252 appeals (or 7 percent of the 3,552 teachers in the district), and in 2011–12, there were 130 appeals (4 percent of teachers).

In four districts, teachers could informally appeal their scores during their feedback conferences with their observer.

In Austin, for example, peer observers used the post-observation conferences to afford teachers the opportunity to address any inaccuracies in the observation ratings or to point to whatever practice or competency the observer might have overlooked. If teachers were still dissatisfied, the district offered a formal process for teachers to dispute the results of a peer observation. In addition, for administrator observations, teachers could request a second observation by a different observer.

In Harrison, teachers in focus group interviews said that although their district had no formal appeals process, they believed that they could contest individual observations by talking directly to their observer, although some observers were more willing to change their scores than were others. Similarly, teachers in Plattsburgh, after their post-observation conference, could gather evidence to show success in areas in which they were considered weak. In fact, because principals did not finalize teachers' evaluation scores until the end of the year, teachers were permitted to collect samples of their work throughout the year in order to improve their score. Similarly, teachers in Pittsburgh could submit an addendum to the evaluation form within seven days of the observation.

Tying Evaluation Scores to Professional Development and Support

The extent to which teachers' learning and development needs shape school or district professional development services has implications for the extent to which evaluation systems achieve their stated purpose of improving instruction. As Isore (2009) observed in her OECD working paper: "Without a link to professional development opportunities, the evaluation process is not sufficient to improve teacher performance, and as a result, often becomes a meaningless exercise that encounters mistrust — or at best, apathy — on the part of teachers being evaluated" (p. 17).

As described in the following discussion, study findings suggest that seven of the eight districts put in place specialized, targeted professional development and support for teachers who were rated as low-performing or at risk of termination. For teachers who were not low-performing, however, five of the eight districts included in the study had not yet created clear linkages between district-sponsored professional development and teacher evaluation results. That is, unless rated as low-performing, teachers needed to rely on the advice and feedback of their observer — usually the principal — to know what professional development they needed and how to access it within the district. None of these districts explicitly communicated available district-sponsored professional development opportunities that would address the particular needs identified in teachers' evaluations.

For teachers who were rated as low-performing or who risked redeployment or release, seven districts put in place specialized, targeted professional development and support to help those teachers improve.

Hamilton County, for example, instituted a "Performance Improvement Plan" (PIP) for all teachers who, despite ongoing coaching and support, received a "does not meet standards" rating. The district required that the principal meet with the teachers at least monthly for a year to address their needs. One district administrator described the thinking behind the intervention strategy: “We wanted to be sure that we put everything in place so that we could be supportive of the teacher.” A principal described the intervention in some detail: “A performance improvement plan is very diagnostic and very specific and targets specific areas of need, and so you are being prescriptive almost to say ‘Look, you need to do this, this, this and this.’ If things don't improve after you do this, then we're going to move to an intensive assistance plan.” At the next level of support for tenured teachers who do not successfully complete a PIP, the district’s “Intensive Assistance Plan” (IAP) is, as a district administration described, “very, very intense and involving five or six employees from the central office who come in and work with that employee. And they're the ones who pretty much are conducting the observations on the teacher at that point.” Another administrator described the IAP as the last source of support offered to a low-performing teacher. If they fail to improve, they are terminated.

Harrison, Pittsburgh, Plattsburgh, and St. Mary’s County provided similarly intensive supports to teachers rated as low-performing. For example, Harrison placed teachers receiving an “Unsatisfactory” rating on a remediation plan requiring that they show improvement by the end of the year or risk losing their job. The district hired instructional coaches to assist teachers and created “observation classrooms” in which struggling teachers could sit behind a one-way mirror and observe and discuss with their instructional coach the best and the brightest teachers’ work. In a similar vein, Pittsburgh placed its low-rated teachers on an “Employee Improvement Plan” (EIP) intended to help them improve their instruction. Between 2011 and 2012, the district managed more than 200 improvement plans for marginally effective teachers. Finally, Plattsburgh and St. Mary’s County required that teachers with low scores participate in an assistance plan (the Peer Assistance and Review [PAR] in Plattsburgh) to receive support. In Plattsburgh, low-performing teachers were observed by and met with a teaching consultant regularly over the course of a school year. In both districts, failure to show improvement could lead to separation.

An administrator in Hillsborough County described a mixture of guidance and flexibility in the professional development that follows a low-rated evaluation. The district required teachers to develop an improvement plan with their principal, and teachers received advice from their principal and/or peer observer about the type of professional development they needed. However, a district administrator said that the district tried not to make the process too constricting, allowing for the fact that some teachers might choose different ways — different types of professional development — to solve their problems: “You know, they might not follow exactly the advice that the observer gave them, but at the same time they do something that improves them in that area.” Still, the district followed up to see if teachers had done something that helped them improve. Specifically, every principal received a district report three times a year that included a variety of teacher-level data, including a list of the professional development courses taken in the last 12 months. As of winter 2013, the district was planning to develop a professional development tracking system showing the link between the teachers who received poor evaluation scores in the area of, for example, questioning skills, and the teachers who had participated in a professional development training designed to develop teachers’ questioning skills.

For teachers who were not at risk of redeployment or release, the link between teacher evaluation results and available district-sponsored professional development was not explicit in five of the eight districts included in the study.

In fact, as of late 2012 or early 2013, none of these five districts’ teacher evaluation systems included a feedback process directly linking teacher evaluation results to available professional development resources within the district. Rather, these districts (Austin, Hamilton County, Pittsburgh, Plattsburgh, and St. Mary’s

County) relied primarily on principals[40] to identify and address teachers' professional development needs. In Pittsburgh, one principal explained that after discovering that she had given low ratings to many teachers in the area of "planning," she organized school-based professional development activities around planning: "So [the teacher evaluation system] has helped structure what teachers are getting in our own school-based professional development quite a bit." Another principal in the district believed that having principals define and address teachers' professional development needs made practical sense: "Because we're in classrooms so much, I'm able to see what the professional development needs are, and I'm able to see [whether] other teachers [within the school] can provide that professional development. So I've been able to create a total in-house professional development [system]."

Principals in Plattsburgh, however, raised concerns that might potentially apply to other districts that rely heavily or exclusively on principals to identify and address teachers' professional development needs. Plattsburgh principals questioned their ability based on the limited training they receive to diagnose or provide the high-quality professional development and support that teachers' need: "Because I can't be the go-to person for every single thing. I don't have that kind of expertise, and I don't feel that I should have to have expertise in every area [of instruction] …" A second principal echoed these concerns, commenting on how the expectation that principals provide professional development to help teachers improve could cause problems down the road:

> *To be honest with you, I don't think [the district] has the funding for [teacher professional development]. And they're looking to us, the building administrators, to become the professional developers, and it puts a huge strain on the administration. I mean, we're trying to approve a local assessment test, set up goals, do the observations, and now we're going to be the professional developers who help teachers who might be struggling with questioning techniques or in differentiating instruction.*

Administrators in Plattsburgh explained that the district had developed software to observe overall trends in teacher strengths and weakness but was working to determine how to align professional development with evaluation feedback.

Three districts worked to strengthen district-sponsored professional development to more directly and explicitly address teacher needs, as defined by their evaluation results.

For example, the District of Columbia restructured and more clearly defined the role of the instructional coach so that coaches could better respond to teacher professional development needs based on their IMPACT evaluations. That is, until 2012–13, the responsibilities of instructional coaches were ill-defined and principals were sometimes having them do myriad tasks unrelated to supporting the professional development needs of teachers, including substituting for absent teachers, doing lunch duty, administering assessments to students, and the like. The district restructured the coaching role so that coaches spent the majority of their time working with teachers on topics that teachers themselves had defined. In addition, the district produced a library of online lesson videos that teachers could watch featuring DCPS' highly effective teachers at various grade levels and subject areas demonstrating lessons for each of the districts nine Teach standards. In addition, coaches could use the video library to help support specific teacher needs (DCPS 2012).

[40] Peer observers also recommended professional development to teachers during the post-observation conference.

Austin planned to strengthen the linkages between professional development and teacher evaluation results by working with *truenorthlogic*[41] to design a human capital management system that integrated data on, among other things, payroll, professional development and teacher performance. Eventually, the system would permit users, including principals, assistant principals, and teachers, to track and link teacher evaluation ratings with professional development course information. In addition, the system would be used to track the individual components that go into overall evaluation scores. A district administrator explained the problem the district was trying to solve by creating the database:

> *The appraisal data were entered into an entirely different in-house system, and PD course scheduling and attendance were captured in yet a different software (Avatar). With* truenorthlogic*, campus administrators can enter their [classroom] observation data, track the status of [evaluations] for teachers, and recommend PD courses that are scheduled and in many cases offered within the same system. Teachers can sign up for courses, attend courses, and view their transcripts in relation to their evaluation data and principal recommendations for PD.*

Finally, Hillsborough County began to review how teachers had scored on each component of the observation rubric and added or revised district-sponsored professional development and support activities based on those results. As one district administrator explained:

> *Our teachers were scoring really well in [one area of the evaluation rubric]. Conversely, we weren't spending a lot of money and offering a lot of training in higher order questioning and student engagement … where teachers were consistently scoring [lower]. And so, at that point as a staff we decided to shift our focus and our financial support to [other areas of practice as defined by the observation rubric]. And I think for us, it gave us the data that we needed, because we were doing things, honestly, the way we'd always done them. … And we didn't realize we had those gaps [in support] until we had some more data to really show that we were not doing as much as we [thought] we were doing.*

The district still had work to do, however, to facilitate teacher access to professional development aligned with their evaluation scores. In 2012–13, Hillsborough County began developing a database linking the professional development that the district offered to the dimensions of professional practice included on the classroom observation rubric. Specifically, the district was re-writing the descriptions of district professional development to correspond to observation components and to make the database searchable by those components. One district administrator described the problem that the new database would hopefully correct:

> *Teachers are saying, 'I would love to be able to go through the online [professional development] registration system and search [observation rubric standard] 3A and see what's out there.' So we're in the process now of working with our vendor to make it easier. I mean, we try to embed it in course descriptions, but it's not easy to find. And we want to take some of the work off of our teachers to be able to do that. So I think that that's something that's in the works, and we're trying to find a way to work with that more.*

[41] *truenorthlogic* is a private, for-profit information technology company that designs software intended to support the "talent management" needs of K–12 organizations.

VII. Administering Evaluation Systems

Despite the complexity of these eight districts' teacher evaluation systems that use multiple classroom observations and measures of student performance to assess teacher professional practice, the majority of the districts created relatively simple, streamlined structures to administer their evaluation systems.

Key Findings

- Six of the eight districts created simple administrative structures for their evaluation systems, with, on average, five staff administering their teacher evaluation systems.
- Four districts worked with outside contractors to create data management systems for their teacher evaluation data, and two districts worked with outside contractors to train their classroom observers.
- Four districts that included state assessment data in their teacher evaluation systems regularly wrestled with the challenge of state test scores being unavailable to them until well after the school year ended. As a result, districts could not provide final evaluation scores for teachers until the summer or sometimes the fall.
- Hiring peer observers was among the most expensive features of the teacher evaluation systems in the three districts that took this approach.

Staffing the Evaluation System

Every district had its own distinctive approach to dividing design and implementation work among district staff and external organizations. Most districts' administrative structures were somewhat dependent on outside funding, however, and some of the structures were expected to change significantly once their funding ended.

Six of the eight districts assembled relatively simple, streamlined structures to administer their evaluation system.

Austin, Hamilton County, Hillsborough County, Pittsburgh, Plattsburgh, and St. Mary's County had fewer than 10 staff administering their teacher evaluation systems (Exhibit 24). Because four of these six districts were still piloting their evaluation systems at the time the data were collected for this study, it would be premature to draw conclusions about the size and type of staff needed to implement a comprehensive teacher evaluation system over time. Nevertheless, administrators in three of the four fully implementing districts (Hamilton County, Harrison, and Hillsborough County) reported hiring few staff to support their teacher evaluation systems. The work associated with designing and administering the teacher evaluation systems was absorbed into regular departmental routines and activities, with district staff supplemented, as necessary, by assistance from external organizations.

Exhibit 24
Number of administrative staff, peer observers, and external organizations employed to implement teacher evaluation systems, by district

District	Number of central office staff supporting system operations (in FTEs)	Number of peer observers	Contracts with external organizations
Austin (pilot)	5	15	▪ *truenorthlogic* (for 2013–14 — human capital management system for tracking and linking evaluation data with PD course information) ▪ SAS EVAAS (Education Value-Added Assessment System)(school-level VAM scores)
District of Columbia	13	42	▪ Mathematica Policy Research (validity testing and VAM scores) ▪ Batelle for Kids (roster validation) ▪ KSA Plus (end-of-year IMPACT reports for staff)
Hamilton County	3[a]	n/a	▪ Educator Software Solutions (data storage/warehousing)
Harrison	14[b]	n/a	▪ EduSoft (item banks for development of district assessments)
Hillsborough County	4	150 (Peer) 81 (Mentor)	▪ Value-Added Research Center, University of Wisconsin (validity testing and VAM scores) ▪ Cambridge Education Group (observation rubric; inter-rater reliability training) ▪ Fierce, Inc. (observer feedback training) ▪ Lawson Talent Management (data storage/warehousing)
Pittsburgh (pilot)	9	n/a	▪ Mathematica Policy Research (validity testing, VAM scores, technical assistance) ▪ Cambridge Education Group (administration of the Tripod Survey) ▪ Danielson Group ▪ Battelle for Kids
Plattsburgh (pilot)	4[a]	n/a	
St. Mary's County (pilot)	6[b]	n/a	▪ Performance Matters (data storage/warehousing)

Exhibit reads: Austin used five central office staff (in FTEs), 15 peer observers, and contracted with or planned to contract with two external organizations (*truenorthlogic* and SAS EVAAS) to support teacher evaluation system operations.

[a] State education agency conducts VAM analyses on behalf of the district.

[b] District does not conduct VAM analyses as part of its teacher evaluation system.

SOURCE: Site visit interviews; extant data on system specifications (see Appendix D for complete list).

The District of Columbia developed a relatively large and complex administrative structure for its teacher evaluation system while Hillsborough County used a simpler structure.

In the District of Columbia, a staff of 13 full-time employees managed the teacher evaluation system (Exhibit 25). Among its many administrative responsibilities, the IMPACT team was responsible for managing the peer observers; developing and managing the data system that housed the student achievement, observation, and other data used to evaluate teachers; overseeing the appeals process; running the teacher compensation, bonus, and career ladder systems; managing the *Align* team (described earlier); and coordinating the external organizations providing administrative support, including the value-added analyses.

Hillsborough County, by contrast, assembled a four-person team of existing central office staff to manage the day-to-day tasks associated with design and early implementation of the teacher evaluation system. Each team member, all former principals, specialized in a key area for the development of the system: technology, human resources, professional development, or assessment. Once the early implementation tasks were completed, however, and rather than create a new and separate office and infrastructure for managing the teacher evaluation system, the four-person team dispersed within the district office, transitioning to district departments relevant to their respective areas of expertise. For example, the staff person who initially managed the mentor evaluators for new teachers became the Director of Professional Development for the district, and the staff person who helped design and implement the teacher evaluation system became the Director of Human Resources. Thus, the district had no department dedicated specifically to managing the teacher evaluation system. A Hillsborough district administrator noted the importance of creating a design team from within the school district rather than hiring staff from the outside who would be unfamiliar with the way the district operated:

> *All of us [on the design team] had observed teachers and conferenced with teachers about their practice, so we had a good understanding of what that process actually entailed, how time consuming it was. We understand the scheduling difficulties, and that helps us in a couple of ways. You know, when peers, mentors, or principals have difficulty with observations I think that we have a better sense of being able to judge whether what they're saying is valid because we have all been in their position and know whether it's as big a problem as they think it is.*

Exhibit 25
Operation of the District of Columbia's IMPACT system

The work associated with the IMPACT system was housed in the District of Columbia Public Schools' Office of Human Capital. The Chief of Human Capital and his Deputy oversaw the IMPACT system. Within the Office of Human Capital, the Division of Teacher Effectiveness managed the IMPACT design process while the Division of IMPACT managed the day-to-day IMPACT operations.

The Division of Teacher Effectiveness led the initial IMPACT design process and was responsible for ongoing monitoring and review to strengthen and refine the system. For example, using evaluation data, research on best practices, and feedback collected through teacher and principal focus groups, the Division developed two updated versions of the evaluation system since its introduction in 2009: IMPACT 2.0 in 2010 and IMPACT 3.0 in 2012. The division was also working towards using IMPACT data to inform decisions about teacher recruitment and retention. For example, the district hoped to use IMPACT data to determine which local teacher preparation programs were producing the district's strongest teachers in various subject areas. This division also designed the district's LIFT career ladder program that tied IMPACT evaluation scores to leadership and professional growth opportunities for their effective teachers.

The second division, Division of IMPACT, managed IMPACT operations and included nine staff who were responsible for the following tasks:

- Managed 42 master educators, including matching them to teachers for observations, monitoring their progress, and providing them with on-going training and support.
- Oversaw analyses of the student achievement data used for the value-added analyses and the analyses of teacher-assessed student achievement (TAS) goals. Also, worked with the Office of Data and Strategy, which provided the student test scores and demographic data that the division then organized and transferred to its external contractors to analyze.
- Worked with Mathematica Policy Research to calculate the individual value-added scores and Battelle for Kids to manage the roster confirmation process that ensured teachers were held accountable for the appropriate students.
- Confirmed the caseloads of special education teachers. This process was similar to the roster confirmation process managed by Battelle for Kids and allowed teachers to confirm their student caseloads.
- Maintained and continually updated the IMPACT database used to store teacher evaluation ratings based on student performance, classroom observations, core professionalism, and commitment to school community data. One of the IMPACT operations coordinators managed and analyzed all the teacher evaluation data stored in the database while coordinating with external database consultant, Quickbase, to maintain and modify the database according to evolving needs.
- Produced end-of-year reports that provided teachers with their overall evaluation results, as well as their individual ratings on each of the IMPACT components. These reports were tailored to reflect the specific evaluation components included in a teacher's evaluation and, consequently, varied by grade and subject area.
- Managed the IMPACT appeals process for teachers facing termination as a result of their IMPACT rating.
- Managed the IMPACT-plus performance-based compensation system for highly effective teachers and the LIFT career ladder, which provided expanded professional opportunities to teachers as they performed well on their IMPACT evaluations.
- Provided ongoing training and support to teachers and principals as well as trained new teachers and principals on IMPACT. In addition, facilitated the sharing of best practices for principal observations and the TAS goal-setting process. Worked with the Office of Teaching and Learning to address teacher professional development needs based on IMPACT data.
- Responded to IMPACT-related questions and concerns from teachers, principals, and other district staff. The division also tracked decision rules made about specific cases — such as how to evaluate teachers who missed an observation because they were on leave — to ensure consistent application of evaluation policies.
- Supervise the *Align* team, a group of eight temporary contractors working in-house to design the district's observer training system.

SOURCE: Site visit interviews; extant data on system specifications (see Appendix D for complete list).

Contracting with External Organizations

Seven of the eight districts (Austin, District of Columbia, Hamilton County, Harrison, Hillsborough County, Pittsburgh, St. Mary's County) contracted with external organizations to support both the design and operation of their teacher evaluation systems. The strategy of outsourcing evaluation work was central to Hillsborough County's system administration plans; other districts used outside contractors to a lesser extent, as described below.

Four districts worked with outside contractors to create data management systems for their teacher evaluation data and two districts worked with outside contractors to train their classroom observers.

St. Mary's County worked with Performance Matters, a firm that had already been helping the district track its student-level achievement data, to expand the database to include classroom observation data (Exhibit 24). Building on the existing database, with which teachers were already familiar, made for a smooth transition to the new database. As noted in the previous chapter, Austin created a data management system that served a slightly different purpose. Working with *truenorthlogic*, Austin planned to design a human capital management system by integrating data on payroll, professional development, and teacher performance. Hamilton County and Hillsborough County also contracted with outside organizations to build their data management systems and to store the tremendous volume of observation and student achievement data collected for their teacher evaluation systems.

As mentioned earlier in the report, Hillsborough County and Pittsburgh used outside contractors to train their classroom observers. Hillsborough County worked with the Cambridge Education Group to conduct its initial large-scale observer training on the observation rubric. Although the district later took over training its new observers and conducting follow-up trainings for returning observers, it continued to contract with Cambridge to calibrate observers and assure inter-rater reliability. In addition, and in response to complaints from teachers that observers did not provide thorough feedback, Hillsborough County contracted with Fierce, Inc. to provide its *Fierce Conversations* training to improve the quality of feedback conferences. In addition, Pittsburgh contracted with Battelle for Kids to assist district staff and teacher leaders with planning, communications, and development of training materials.

Three districts turned to outside entities for other types of administrative support.

For example, in addition to outsourcing its value-added analyses, the District of Columbia turned to Battelle for Kids to conduct its roster confirmation process. The district initially managed roster confirmation in-house, but because it was a time-consuming and labor-intensive process, began contracting with an outside organization to manage the work. One district official noted that using an outside contractor for this task saved the district money, describing the contract as "money well-spent." Harrison accessed the item banks available through EduSoft to assist them in the development of the over 165 quarterly assessments and semester exams — 25 percent of which were revised or replaced every year — that contributed to teachers' performance evaluation scores. Finally, Pittsburgh worked with Cambridge Education Group to administer their Tripod Student Survey.

Managing Evaluation Data Systems

Every district included in the study had to expand their internal capacity to track, collect, and analyze the student performance data used to evaluate teachers. The inclusion of student achievement data in each district's teacher evaluation system required administrative oversight.

Five districts identified roster confirmation as a critical step in accurately measuring a teacher's impact on student growth.

In the District of Columbia, Harrison, Hillsborough County, Pittsburgh, and St. Mary's County, teachers were required to confirm that the students for whom the value-added or other student growth measure scores were calculated were the students they taught that school year. Hillsborough County, for example, had its teachers verify their student rosters seven times a year to ensure that the students included in the teachers' value-added calculations were students they actually taught. However, several teachers who participated in focus group interviews noted that the system counted students for whom the teachers did not believe they should be held responsible:

> *Five kids upgraded from seventh grade to my class last Thursday. They will count against me tomorrow in the [writing] FCAT and not one of them can write an essay. That is not fair.*

In the District of Columbia, teachers could report to the district any notable information about students in their class that may affect their performance. For example, a teacher could identify students who had transferred in or out of the class during the school year. Teachers also noted if a student was regularly pulled out of class for special education or other services; such students were weighted into the teacher's value-added score based on how much time they spent with the teacher.

Teachers at every school level in Harrison reported that the task of finding errors in the rosters was difficult and time-consuming because the rosters they received from the district were frequently inaccurate and because teachers could not request a change to the roster unless the change would affect their evaluation results. Teachers participating in focus group discussions reported that it was not uncommon to devote more than 10 hours on roster validation during the two-week window that the district made available for teacher verification of the district's data. As one teacher explained: "Every time there's an assessment, we need to spend hours checking if the kids are on our roster. The roster is always wrong, and you cannot change the roster unless it affects your results, so you have to check both the roster and whether your results are affected … You have to be an accountant."

St. Mary's County took a different approach to ensure that it held teachers accountable for the correct students included in a teacher's SLOs. Teachers selected "attributed students" for each of their SLOs. For example, a teacher might set one progress goal for all ELs in the class and another goal for special education students. Teachers could include a student in more than one goal, and they were not required to set performance goals for all the students they taught. Teachers tailored these goals further through the consideration of "complexity factors," wherein teachers could note specific personal or academic issues that a student faced that could influence his or her progress. This system allowed teachers to set relevant goals for their students, but as district administrators and principals explained, also allowed for the possibility of manipulating targets. Despite the opportunity to set the bar low, however, the principals participating in site visit interviews noted that teachers generally tended to set their goals too high.

In addition to the assessment data, Pittsburgh teachers had to verify the accuracy of the student rosters to ensure that the appropriate students were surveyed.

The district administered student surveys to samples of students, and these students were often not evenly distributed among teachers. One central office administrator explained that they discovered that they needed to extend the window for when the surveys were administered and to think carefully about how to sample one or two classes for each secondary teacher. “Last year,” he explained, “some teachers had three or four of their classes complete the Tripod surveys, and other teachers did not have any students complete the surveys.” In Austin, where student surveys were also administered, district administrators did not use a similar roster confirmation process for teachers.

Administrators in four districts described regularly wrestling with the challenge of state test scores being unavailable to them until well after the school year ended.

As a result, districts could not provide final evaluation scores for teachers until the summer or sometimes even the fall. In the District of Columbia, Hillsborough County, Plattsburgh, and St. Mary’s County, teachers with borderline scores who were potentially subject to termination did not know until the next school year (or late summer in the District of Columbia) whether they would have a job in these districts, and, consequently, districts did not know what their hiring needs would be.

Managing Peer Observers

The three districts that used peer observers reported that “staffing up” in this way was among the most expensive features of their teacher evaluation systems.

In the 2012–13 school year, the District of Columbia employed 42 peer observers at salaries of approximately $90,000 each. Peer observers were required to have at least five years of urban teaching experience and were selected for their ability to effectively communicate with teachers about their instructional and classroom management practices. In Hillsborough, teachers were recruited from within the district to serve as full-time peer observers and were paid a $5,000 stipend in addition to their regular salary and benefits. During their one- to three-year term as peer observers, these teachers were on temporary leave from their regular teaching assignment. The cost to the district was the $5,000 stipend, plus the cost in salary and benefits of replacing these teachers in the classroom. In Austin, the 15 full-time peer observers received salaries and benefits totaling approximately $500,000 a year.

In addition to paying their salaries and benefits and matching observers to teachers (described in Chapter 4), districts also devoted staff time to training and monitoring peer observers. In the District of Columbia, for example, a district administrator described monitoring peer observers’ performance to ensure that they conducted their observations in a timely manner and that they produced strong observation feedback reports. In particular, district staff monitored peer observers’ observation logs, working to ensure that they distributed their observations relatively evenly across teachers within each evaluation cycle. A district administrator explained that if peer observers waited until the end of each cycle to complete their work, their observations would be rushed and, perhaps, not as robust as observations conducted earlier in the cycle. As the administrator explained, “We wanted every teacher to have the same evaluation experience.” She also explained that district staff monitored peer observers’ observation feedback reports for purposes of quality control: “Writing [a feedback report] is a good interpretation of the rubric. If they’re putting the evidence in the wrong [domain], that shows a misunderstanding. We have found in the last year that our issue is not norming [peer observers] around interpretation of the rubric but on the collection of evidence [which is sometimes uneven among observers].”

Administrators in all three districts also reported devoting a significant amount of time to recruiting and hiring their peer observers. For example, in addition to written applications, Hillsborough County required its job candidates to participate in a panel interview and to conduct a mock classroom observation. One district administrator described the interview process: “We always pick a video of a young teacher who has great rapport [with her students] but whose instruction is all over the place. [The candidates] come in and role play. ‘Tell me a specific lesson that you taught?’ ‘How did you differentiate for ELL students?’ ‘How do you interact with parents?’” An administrator explained that hiring peer observers was difficult and time-consuming because while the district could teach job candidates to “say the hard things,” they needed candidates who already knew how to connect with and support teachers.

In Austin, a district administrator described the comprehensiveness of their peer observer application process. That is, in addition to the standard district job application, applicants for peer observer positions were required to submit their credentials, evidence of relevant experience, and professional recommendations from administrators and colleagues. Three reviewers used a rubric to score applicants’ applications and recommendations. Those receiving the highest scores were invited to participate in a personal interview as well as to complete a writing exercise that required answering a list of nine questions in 30 minutes. District administrators then scored each candidate’s responses, and the top scorers were offered the peer observer position.

VIII. Perceived System Effects

Districts’ investments in system design and implementation have begun to bear fruit, even among the partially implementing districts. Respondents at all levels reported perceiving positive effects of their new teacher evaluation systems, and while no respondent was ready to claim that these systems — still in the early stages of implementation — had improved student achievement, district administrators, principals, and teachers alike cited early effects and outcomes, particularly related to improved teacher professional practice.

Key Findings

- Principals in six districts reported that the teacher evaluation system had caused them to want to be better instructional leaders in their schools.

- Teachers who participated in focus groups or individual interviews in every district included in the study reported that they believed that the classroom observations and subsequent feedback had helped them become better teachers. Teachers in three districts, however, questioned whether it was possible to demonstrate excellence on the full range of competencies included in their respective districts’ observation rubrics.

Perceived Effects on Districts

None of the eight districts reported reassigning its teachers to schools based on evaluation ratings but said that some teachers sought transfers out of lower performing schools or into higher performing schools.

According to district administrators in the District of Columbia and Hillsborough County, some teachers transferred schools in response to the advent of the teacher evaluation system. Teachers might transfer to lower performing schools because they wanted the opportunity to earn a larger bonus (a noted, but extremely rare occurrence in the District of Columbia) or to higher performing, less challenging schools (Hillsborough County) because they believed that with fewer disciplinary challenges they would score higher on the evaluation rubric.

Administrators in two districts reported that the results of their teacher evaluation systems had caused them to change — or to consider changing — the criteria they used to recruit and hire teachers and principals.

The District of Columbia, for example, was considering using its IMPACT data in the long term to inform its teacher recruitment strategies, identifying the strengths and weaknesses of teachers prepared in particular programs or institutions and recruiting selectively from those training providers.

The District of Columbia also had developed a research partnership with the University of Virginia to analyze the IMPACT data to determine the system’s effects on the district. Early findings suggested that the district was differentially retaining staff. Specifically, the district was retaining their highest performing teachers and not retaining their lowest performers.[42]

An administrator in Hamilton County explained that the district had changed the way it hired principals because implementation of the new teacher evaluation system had revealed the need for high-quality principals who had the skills required to provide teachers with effective feedback. As she explained: "We've got to have someone instructionally sound that can give [teachers] that specific and meaningful feedback. I think Project COACH has had a great bearing on instruction in those schools where I think the instructional leader is aware of how [the evaluation system] can be used and what it can be used for."

Perceived Effects on Principals

At least two principals in each of six districts reported that the teacher evaluation systems had caused them to want to be better instructional leaders.

In Plattsburgh, for example, a principal described reading books on coaching to better prepare for post-observation feedback sessions with his teachers. A principal in Hamilton County described how Project COACH helped her to change her approach to instructional leadership:

> *[Project COACH] has allowed me as a principal to do the things I always thought I should do, like be in classrooms but also give feedback all the time. I have to do a certain number of observations on every teacher a certain number of times and [as a result], I have so many more conversations with teachers. I'm coaching them so much more. I'm so much more focused on instruction. I know what's going on so I know — I'm learning from the good teachers, which is a huge part. It's completely changed the way that we do our jobs, I think, because instead of doing all this huge amount of paperwork and scripting everything, you're [instead] just sitting there watching instruction and focusing on how you improve that instruction, how you help that teacher.*

Two other principals in Hamilton County described using the evaluation data to identify the professional development needs for their staff. One principal explained that she looks at the summative evaluation data for the entire school, which then helps her "structure what teachers are getting in our own school-based professional development quite a bit." The other principal reported developing a better insight into teacher professional development needs as a result of the classroom observations she conducted. In addition, she was able to identify teachers who were capable of developing and assisting other teachers within the school.

In Austin, principals who participated in interviews said that peer observers' work was causing them to raise their game and improve their performance with teachers. That is, the principals said they were hearing that their teachers highly valued the feedback that they had received from the peer observers. A principal said that she knew her teachers thought that the feedback from the peer observers was the best feedback they had ever received. She explained her response:

[42] A 2013 study of the District of Columbia's IMPACT system by the National Bureau of Economic Research (Dee and Wyckoff) found that dismissal threats increased the voluntary attrition rate among low-performing teachers, increased the retention rate among high-performing teachers, and improved the performance of teachers who had previously been deemed low-performing but who had remained in the district. In addition, the study found that the offer of financial incentives also improved the performance of high-performing teachers.

> *I want to find a way to give them that [kind of feedback] and to [increase my] time in the classrooms and have those follow-up conferences that I don't always make time for. I observe [teachers] all the time, I'm in their classrooms all the time, but I don't necessarily sit down and give them that kind of feedback after the fact.*

Principals in each of six districts reported having had to make many adjustments to their workday in order to conduct the number of teacher observations that their teacher evaluation systems required.

At least one principal interviewed in each of the six districts reported that the time burdens that their district's new evaluation system imposed represented a significant addition to their workload.

When asked about the ways in which they were adjusting to the demands of the new teacher evaluation system, one principal in Pittsburgh stated, "We're asking others to help take on some other responsibilities because of the time needed for [teacher observations]." A principal in Hamilton County reported that the number of observations she had to conduct and the system's 48-hour feedback requirement caused her to rethink her approach to time management:

> *Because you really have to have a schedule and you have to be able to manage your time to be able to give that feedback within 48 hours. And of course, we're pulled from our buildings for things, so it's been a struggle to just make sure that I'm getting the number of observations we need and then providing the feedback in a timely manner according to the district guidelines.*

Principals in the District of Columbia and Hillsborough County reported that working on teacher evaluations, particularly preparing the lengthy feedback reports, had begun to creep into their personal time, requiring them to work on evaluation reports after school hours. One principal in the District of Columbia described a typical scenario among principals: "… it gets to the point where it's like you're up 72 hours straight [preparing feedback reports] or you have pockets of principals together at Starbucks trying to keep each other awake to type the reports." A principal in Hillsborough County described the work as never-ending: "It could take me up to two hours to write a [observation feedback] report. I take my laptop wherever I go and I call it 'have rubric, will travel.' And I'm always working on at least one report."

Because the classroom observation component of St. Mary's County's evaluation system had been in place for 10 years, the principals did not report any additional job responsibilities or time pressures resulting from the observation component of the teacher evaluation system. Similarly, one principal in Austin reported that the only additional time burden from that of the previous system was in meeting with teachers to create SLOs; the observation portion, with just two teacher observations per year, did not take more time than had been required for the previous evaluation system.

Perceived Effects on Teachers

> **Teachers who participated in focus groups or individual interviews in every district included in the study reported that they believed that the classroom observations and subsequent feedback had helped them become better teachers.**

In Harrison, for example, there was near unanimous agreement among focus group participants that the evaluation system, particularly the frequent classroom observations, had made them better at their craft. A sentiment repeated by several teachers was: "I could go teach anywhere in the U.S. because of how they prepared me [in Harrison]." Teachers in every district in the study said that they believed they had grown professionally due to the quality of the feedback they received from their classroom observers. The following are a few of the related focus group and interview comments:

> *I do think that thinking about these elements [of the observation rubric] has improved my teaching.*
>
> *[District of Columbia]*

> *I can tell you for me, having been through the process, [the teacher evaluation system] completely changed the way that I taught, the way that I viewed my classroom, the way that I reflected, and the way that I saw my students. [It] totally changed the way that I looked at everything and I was able to do that because of the support that I had from both of my [peer observer] in the last two years.*
>
> *[Hillsborough County]*

In addition, teachers reported a number of ways in which their district's teacher evaluation system had caused them to reflect upon their practice and take more time to reflect upon their students' progress. The following quotes reflect some of what teachers reported in focus group and individual interviews:

> *I really do think the reflection is huge. It has made me a better teacher just in reflecting on how I could have done [something] differently. And I have to say that my peer [observer] has given me some good insights, but that reflection is huge. I think that that really* has *stepped up my game. Whenever I'm doing something now, afterwards I go, 'Okay, well, how can I make that better the next time I do it?' [It made a] big difference. I think I wouldn't have probably done that without the evaluation system.*
>
> *[Hillsborough County]*

> *It really challenged me as I thought about IMPACT, how to ask higher-order thinking questions that stimulate higher-order thinking skills. I think I would do it naturally, and I hope that would be a part of my process.*
>
> *[District of Columbia]*

> *It's given me a chance to reflect on my teaching practice and look at the components of what we're expected to complete to become more empowering and effective teachers.*
>
> *[Pittsburgh]*

At least one principal interviewed in each of the eight districts pointed to ways in which the classroom observations and subsequent feedback had helped teachers reflect on and improve their practice.

In particular, these principals explained that their respective observation rubrics provided a structure for good conversations about teaching and helped them and their teachers together identify areas for professional development. As one Hamilton County principal explained:

> *The level of reflection of teachers is much higher because of the unannounced aspect of [the observations]. I'm not just giving them a list of stuff to do. Directing somebody to do something is very different from helping them think. I see it as just building the capacity of our school system. As we get better and better at [evaluating teachers], teachers are becoming more and more reflective, and they're not going to need as much support because they will be their own support.*

The principals participating in site visit interviews agreed that teachers had gotten better at teaching. One principal in the District of Columbia said that his teachers had taken their previous year's IMPACT scores to heart and it made a difference: "The IMPACT scores are totally different, and their practice is totally different and how they plan and think about teaching is totally different [then it had been before]".

Despite early signs of positive effects on teachers, respondents in three districts warned that high-stakes teacher evaluation systems may have unintended consequences for teachers and the teaching culture.

In one district, for example, a teacher union representative explained that the evaluation system had turned teachers inward and made them more competitive. In another district, several teachers described how teachers had begun to compete with one another: "We used to collaborate across the district but now there's no time. We are too swamped with creating and grading the student assessments; a lot of resources are going to that." Along the same lines another teacher stated, "The concerns in our county were that we worked so hard to become collaborative with [grade-level] teams and common planning times ... but there was some fear that that was going to start breaking apart because our evaluations are based on assessments of students that teachers often share." The worry, he explained, was that teachers would no longer be willing to share lesson plans or think together about ways to address student needs because collaborating would benefit teachers with whom they are competing for their evaluation scores.

IX. Summary

Developing a teacher evaluation system that uses multiple measures to assess teacher performance is a complex process that the districts included in this study approached in myriad ways. They designed teacher evaluation systems that varied widely in the number of annual classroom observations required, the types of assessments used — and the way they used them — to measure teacher impact on student performance, the weight assigned to each measure used to rate teacher performance, and the consequences tied to teacher performance. These eight districts, taken together, offer no single model or standard of practice that other districts might replicate. Indeed, these eight districts — even those considered to have fully implemented their teacher evaluation systems — continued to change, adjust, and fine-tune their systems in response to teacher feedback, pilot test data, and early implementation experiences. They were designing and redesigning their evaluation systems as they were implementing them. Some of those system corrections or changes, as noted in the report, included the following:

- The number and types of required classroom observations changed in three districts. Two districts made adjustments in response to staff feedback, attempting to lessen the burden on school staff and to accommodate the needs of new teachers by reducing the number of required classroom observations. Another district added an initial, informal, and non-binding principal observation of new teachers in order to familiarize these teachers with the observation process.

- Three districts decided — after their teacher evaluation system had been designed — to use external peer observers to evaluate teachers. In two districts, teachers requested that their districts add the external peer observers to their respective evaluation systems, believing that such observers would offer an unbiased, objective opinion of a teacher's performance that was unrelated to any personal connections to the school. In the case of one of these two districts, teachers also wanted peer observers so that they would have an opportunity to receive content-specific feedback from an expert. The third district added external peer observers in response to pilot test data as a way to decrease the observation burdens on principals.

- One district opted to reduce the weight assigned to teacher-level VAMs from 50 to 35 percent and to add student learning objectives (SLOs) to the student performance measures used to evaluate teachers of state-tested grades and subjects. These changes were made in response to teacher concerns about VAM results not being made available until July, while teachers' jobs and salaries hung in the balance (i.e., because they carried such a large weight, VAMs were often the determining factor in district decisions regarding teacher retention or release). A district administrator explained that reducing the teacher-level VAM weight seemed reasonable because VAM scores were not available during the school year and therefore did not identify ways in which teachers might improve instruction.

- Two districts opted to stop using teacher-level or school-level VAMs. In one district, administrators explained that eliminating teacher-level VAMs was based on pilot test results that they believed suggested that teacher-level VAMs did not vary enough to affect or change teachers' final evaluation scores. In the second district, administrators said that they stopped using school-level VAMs because they perceived that they created a disincentive for teachers to work in the lowest performing schools. In addition, district administrators explained that the original purpose of the school-level VAM was to encourage school-based collaboration and a sense of shared commitment to student outcomes, which they later determined was a construct better encouraged through another component of the teacher evaluation system.

In addition to the above changes, telephone updates with district administrators in fall 2014 revealed that these systems continued to change. District administrators reported that the longer their districts implemented the teacher evaluation systems, the more they learned about ways to improve their systems' functionality, practicality, and effectiveness. For example:

- One district began having classroom observations address just a few portions of the teaching rubric in each observation cycle, recognizing that it was not possible for observers to give every domain of the rubric equal attention in the span of a single observation. In addition, this district used evaluation results to identify high-performing teachers to work with their lower performing colleagues.

- Another district adapted the way it interpreted the results of student performance, recognizing that some teachers, by virtue of the particular subject area and ability group they taught (e.g., teachers of an introductory-level foreign language class where student progress, starting from a baseline of around zero, can be assumed), will show stronger student growth than others.

- A third district added more feedback loops to the teacher evaluation process so that the district could work with instructional staff to make more timely adaptations and changes to the evaluation system.

Despite all the changes, this study offers a set of findings about teacher evaluation systems that may inform and support future efforts to design comprehensive teacher evaluation systems. Those findings include the following:

- Respondents in all eight districts agreed that improving instruction was the foremost goal of their teacher evaluation systems.

- Teacher and principal input during the design and/or pilot test phase strongly influenced decisions regarding system modification in six districts, according to district administrators.

- The classroom observation rubrics used or developed in all eight districts addressed similar areas of teaching practice, including instructional practice and classroom environment. Indeed, Charlotte Danielson's *Framework for Teaching* (FFT) was the exclusive basis for the design of the observation rubrics used in half the districts included in the study. The other four districts developed their own rubrics to examine instructional practice but drew from existing frameworks as sources of reference, such as the FFT and Robert Marzano's *Classroom Instruction that Works*.

- Student performance on the state test was but one of several assessments used to calculate a teacher's effect on student performance in all eight districts included in the study.

- All eight districts used multiple assessments to assess teachers' influence on their students' performance. Using multiple types of assessments addressed a three-fold purpose: (1) accommodating teachers of non-state tested grades and subjects; (2) offering all teachers other opportunities to demonstrate effectiveness through, for example, district curriculum-based assessments; and (3) guarding against one assessment being the primary determinant of teacher performance ratings.

- Seven of the eight districts assigned separate weights to each type of student assessment included in their evaluation systems. In addition, seven of the eight districts ensured that the overall weights applied to student performance data for teachers of non-state tested grades and subjects and teachers of state-tested grades and subjects were the same.

- The implementation of these comprehensive teacher evaluation systems brought increased responsibilities for principals. While most districts attempted to offset these added principal responsibilities by training assistant principals and district content specialists to conduct classroom observations, principals reported having to make many adjustments to their workday to accommodate the increased workload.

- Six districts used or planned to use teacher evaluation scores to redeploy or release low-performing teachers, but they narrowly defined the circumstances under which this could happen. Three districts, for example, released only non-tenured or new teachers on the basis of their evaluation scores whereas releasing tenured or veteran teachers required more evidence. In addition, two districts opted to use the observation data and feedback conferences to counsel out or redeploy their ineffective teachers rather than release them outright.

- Seven of the eight districts put in place specialized, targeted professional development and support for teachers who were rated as low-performing or at risk of termination. For teachers who were not at risk of termination, however, five of the eight study districts had not yet created clear linkages between district-sponsored professional development and teacher evaluation results. In all five districts, principals determined what professional development teachers needed.

- The majority of the districts included in the study created relatively simple, streamlined structures to administer their teacher evaluation system. However, every district had to develop and expand its internal capacity to track, collect, and analyze the student performance data used to evaluate teachers. Five districts, for example, identified roster confirmation as a critical and time-consuming step in accurately measuring a teacher's impact on student growth.

- The three districts that used peer observers reported that "staffing up" in this way was among the most expensive features of their teacher evaluation systems. In addition to paying their salaries and benefits, districts also devoted staff time to recruiting and managing the peer observers.

- Some districts discovered that their evaluation systems had unintended consequences. For example, some teachers sought transfers to avoid working in a low-performing school where a school-level score might negatively affect their evaluation results. Teachers and community stakeholders in another district raised concerns that the teacher evaluation system might be causing high rates of turnover among experienced teachers.

- Principals in five districts reported that the teacher evaluation system had caused them to want to be better instructional leaders in their schools. Indeed, the intensive time commitment appeared to have caused principals to change their professional practice and sometimes rethink their role as their school's instructional leader.

- Teachers who participated in focus groups or individual interviews in every district reported that they believed that the classroom observations and subsequent feedback had helped them become better teachers. Teachers in three districts, however, questioned whether it was possible to demonstrate excellence on the full range of competencies included in their respective districts' observation rubrics.

References

Danielson, C. (2013). *The framework for teaching evaluation instrument.* Princeton, NJ: The Danielson Group.

Dee, T. and Wyckoff, J. (2013). *Incentives, selection, and teacher performance: Evidence from IMPACT.* National Bureau of Economic Research Working Paper Series. Cambridge, MA: National Bureau of Economic Research.

District of Columbia Public Schools (2012a). *IMPACT: The District of Columbia Public Schools Effectiveness Assessment System for School-Based Personnel.* Washington, DC: Author.

District of Columbia Public Schools (2012b). *The District of Columbia Public Schools Guide to Value-Added.* Washington, DC: Author.

District of Columbia Public Schools (2012c). 2011–12 *IMPACT Results*. Washington, DC: Author.

District of Columbia Public Schools (2011). *2010–11 IMPACT Results.* Washington, DC: Author.

Doherty, K. M. and Jacobs, S. (2013). *State of the States: Connect the Dots. Using evaluations of teacher effectiveness to inform policy and practice.* Washington, DC: National Council on Teacher Quality.

Harris, D.N., and McCaffrey, D. (2010). Valued-added: Assessing teachers' contributions to student achievement. In M. Kennedy (Ed.), *Teacher assessment and teacher quality: A handbook.* San Francisco: Jossey-Bass.

Harrison School District Two (2010). *Professional Educator Evaluation System: Rules and Procedures.* Colorado Springs, CO: Author.

Isore, M. (2009). *Teacher evaluation: current practices in OECD countries and a literature review, OECD education working paper No. 23*. Paris, France: Organization for Economic and Community Development.

Kane, T.J., and Staiger, D.O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. NBER Working Paper No. 14607.

Kee, K.M., Anderson, K.A., and Dearing, V. S. (2010). *RESULTS Coaching: The New Essential for School Leaders.* Thousand Oaks, CA: Corwin Press.

Martirano, M. J. (2012). *St. Mary's County Public Schools framework for teaching: Teacher evaluation program materials.* St. Mary's County Public Schools: Leonardtown, MD.

Measures of Effective Teaching Project (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project.* Seattle, WA: Bill and Melinda Gates Foundation.

Measures of Effective Teaching (MET) Project (2012a). *Asking students about teaching: Student perception surveys and their implementation.* Seattle, WA: Bill and Melinda Gates Foundation.

Measures of Effective Teaching (MET) Project (2012b). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill and Melinda Gates Foundation.

Milanowski, A.T. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33–53.

Miles, M.F. (Jan. 2011). *Teacher Compensation Based on Effectiveness: The Harrison Pay-for-Performance Plan.* Harrison County, CO.

National Council on Teacher Quality (2015). *State of the States 2015: Evaluating Teaching, Leading and Learning.* Washington, DC: NCTQ.

Sartain, L., Stoelinga, S.R., and Brown, E.R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Chicago: Consortium on Chicago School Research.

Weisberg, D., Sexton, S., Mulhern, J., and Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* New York, NY: The New Teacher Project.

Wright, S.P., Horn, S.P., and Sanders, W.L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57-67.

Appendix A: Number and Type of Classroom Observations, by District

Exhibit A1
Number of formal and informal observations conducted for experienced teachers, by district

District	Number of observations conducted for experienced teachers							
	TOTAL number of observations	Formal observations				Informal/ walk-throughs		
		Number						
District	TOTAL number of observations	Principal	Peer observer	Duration	A/UN[a]	Number	Duration	A/UN[a]
Austin	5	1[b]	2	45 min.	P: A PO: UN	2 (P)	15 min.	UN
District of Columbia	1–5[c]	1–2[b]	0–2	30 min.	UN	0–1 (P)	30 min.	UN
Hamilton County[d]	6	6	__	10 min.	UN			__
Harrison	9	1		45 min.	A	8	15 min.	UN
Hillsborough County	4–11[c]	1–2[b]	1–4	45-50 min.	A	2–4	25 min.	UN
Pittsburgh	4	2	__	30 min.	A or UN[e]	2	15 min.	UN
Plattsburgh	2	1	__	60 min.	A	1	15 min.	UN
St. Mary's County	2	2	__	60 min.	A		__	

[a] Announced (A) or unannounced (UN) classroom observations.

[b] Observer could be an AP

[c] Number of observations varies by past rating.

[d] Hamilton County categorizes teachers into licensing categories. "Experienced teachers" are categorized as "Professionally Licensed" teachers and new or novice teachers are categorized as "Non-Professionally Licensed" teachers.

[e] At least one formal observation in Pittsburgh must be announced.

Exhibit A2
Number of formal and informal observations conducted for new teachers, by district

	Number of observations conducted for new, non-tenured, or probationary teachers								
			Formal observations				Informal/ Walk-throughs		
			Number						
District	TOTAL number of observations	Type of teacher	Principal	Peer Observer	Duration	A/UN[a]	Number	Duration	A/UN*
Austin[b]	—		—				—		
District of Columbia	5[c]	1st year	2	2	30 min.	UN	1	30 min.	UN
Hamilton County	8	Non-professionally licensed	8	__	10 min.	UN	__		
Harrison	18	Probationary	2	__	45 min.	A	16	15 min.	UN
Hillsborough County	4[c]	1st and 2nd year	2	2^^	45–50 min.	A	2 (P and PO)	25 min.	UN
Pittsburgh	8	Non-tenured	4	__	30 min.	A and UN^	4	15 min.	UN
Plattsburgh	2	Non-tenured	1	__	60 min.	A	1	15 min.	UN
St. Mary's County	4	Non-tenured	4	__	60 min.	A	__		

[a] Announced (A) or unannounced (UN) classroom observations.

[b] Austin does not include new teachers in its current teacher evaluation system.

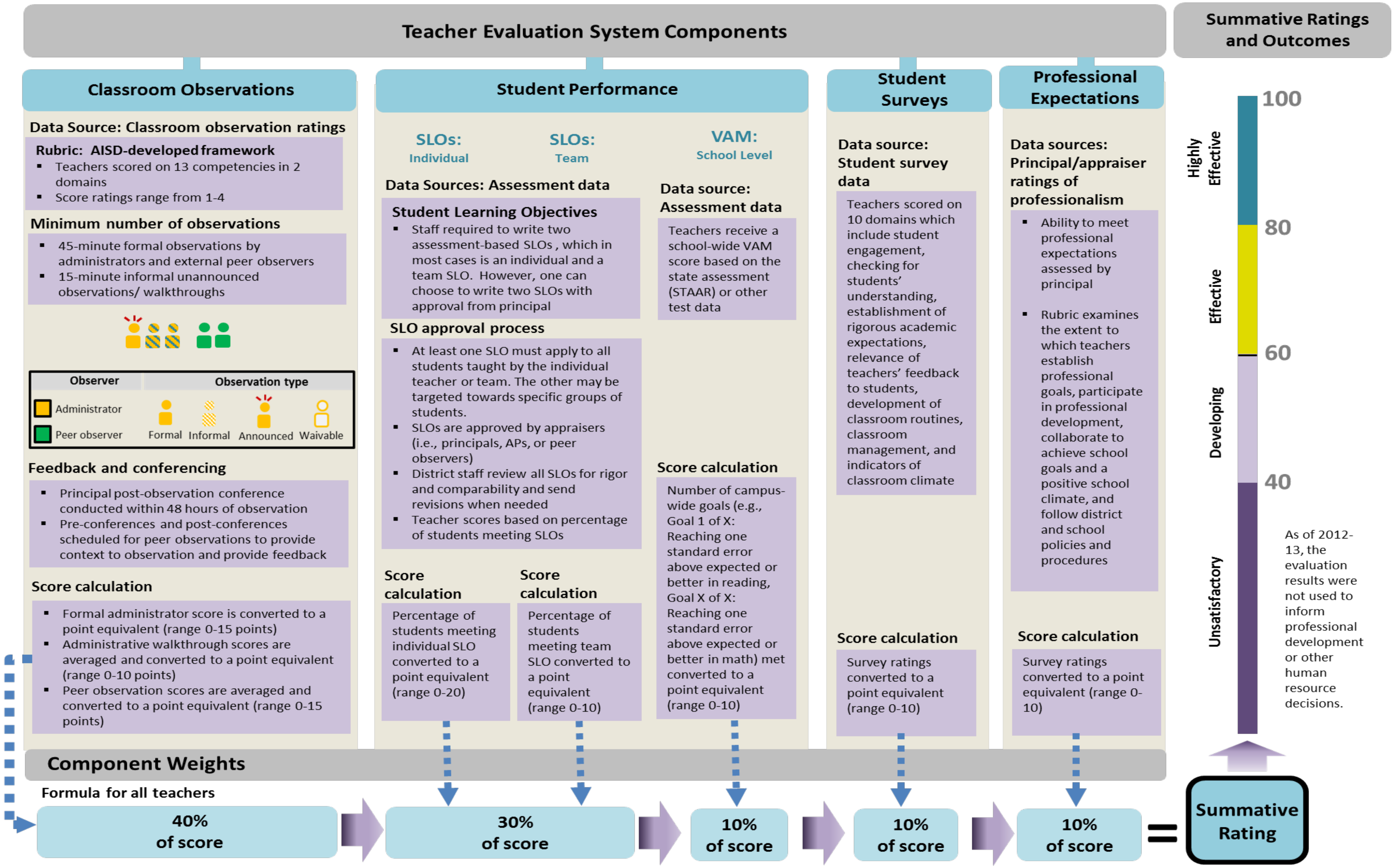
[c] Number of observations varies by past rating.

^ At least one formal observation in Pittsburgh must be announced.

^^ Conducted by mentor observers who work exclusively with new, inexperienced, non-tenured teachers.

Appendix B: Teacher Evaluation System Diagrams, by District

Exhibit B1
Austin Independent School District: Job Appraisal System



Teacher Evaluation System Components

Classroom Observations

Data Source: Classroom observation ratings

Rubric: AISD-developed framework

- Teachers scored on 13 competencies in 2 domains
- Score ratings range from 1-4

Minimum number of observations

- 45-minute formal observations by administrators and external peer observers
- 15-minute informal unannounced observations/ walkthroughs

Observer	Observation type
🟧 Administrator 🟩 Peer observer	Formal · Informal · Announced · Waivable

Feedback and conferencing

- Principal post-observation conference conducted within 48 hours of observation
- Pre-conferences and post-conferences scheduled for peer observations to provide context to observation and provide feedback

Score calculation

- Formal administrator score is converted to a point equivalent (range 0-15 points)
- Administrative walkthrough scores are averaged and converted to a point equivalent (range 0-10 points)
- Peer observation scores are averaged and converted to a point equivalent (range 0-15 points)

Student Performance

SLOs: Individual

SLOs: Team

VAM: School Level

Data Sources: Assessment data

Student Learning Objectives

- Staff required to write two assessment-based SLOs , which in most cases is an individual and a team SLO. However, one can choose to write two SLOs with approval from principal

SLO approval process

- At least one SLO must apply to all students taught by the individual teacher or team. The other may be targeted towards specific groups of students.
- SLOs are approved by appraisers (i.e., principals, APs, or peer observers)
- District staff review all SLOs for rigor and comparability and send revisions when needed
- Teacher scores based on percentage of students meeting SLOs

Score calculation

Percentage of students meeting individual SLO converted to a point equivalent (range 0-20)

Score calculation

Percentage of students meeting team SLO converted to a point equivalent (range 0-10)

Data source: Assessment data

Teachers receive a school-wide VAM score based on the state assessment (STAAR) or other test data

Score calculation

Number of campus-wide goals (e.g., Goal 1 of X: Reaching one standard error above expected or better in reading, Goal X of X: Reaching one standard error above expected or better in math) met converted to a point equivalent (range 0-10)

Student Surveys

Data source: Student survey data

Teachers scored on 10 domains which include student engagement, checking for students' understanding, establishment of rigorous academic expectations, relevance of teachers' feedback to students, development of classroom routines, classroom management, and indicators of classroom climate

Score calculation

Survey ratings converted to a point equivalent (range 0-10)

Professional Expectations

Data sources: Principal/appraiser ratings of professionalism

- Ability to meet professional expectations assessed by principal
- Rubric examines the extent to which teachers establish professional goals, participate in professional development, collaborate to achieve school goals and a positive school climate, and follow district and school policies and procedures

Score calculation

Survey ratings converted to a point equivalent (range 0-10)

Component Weights

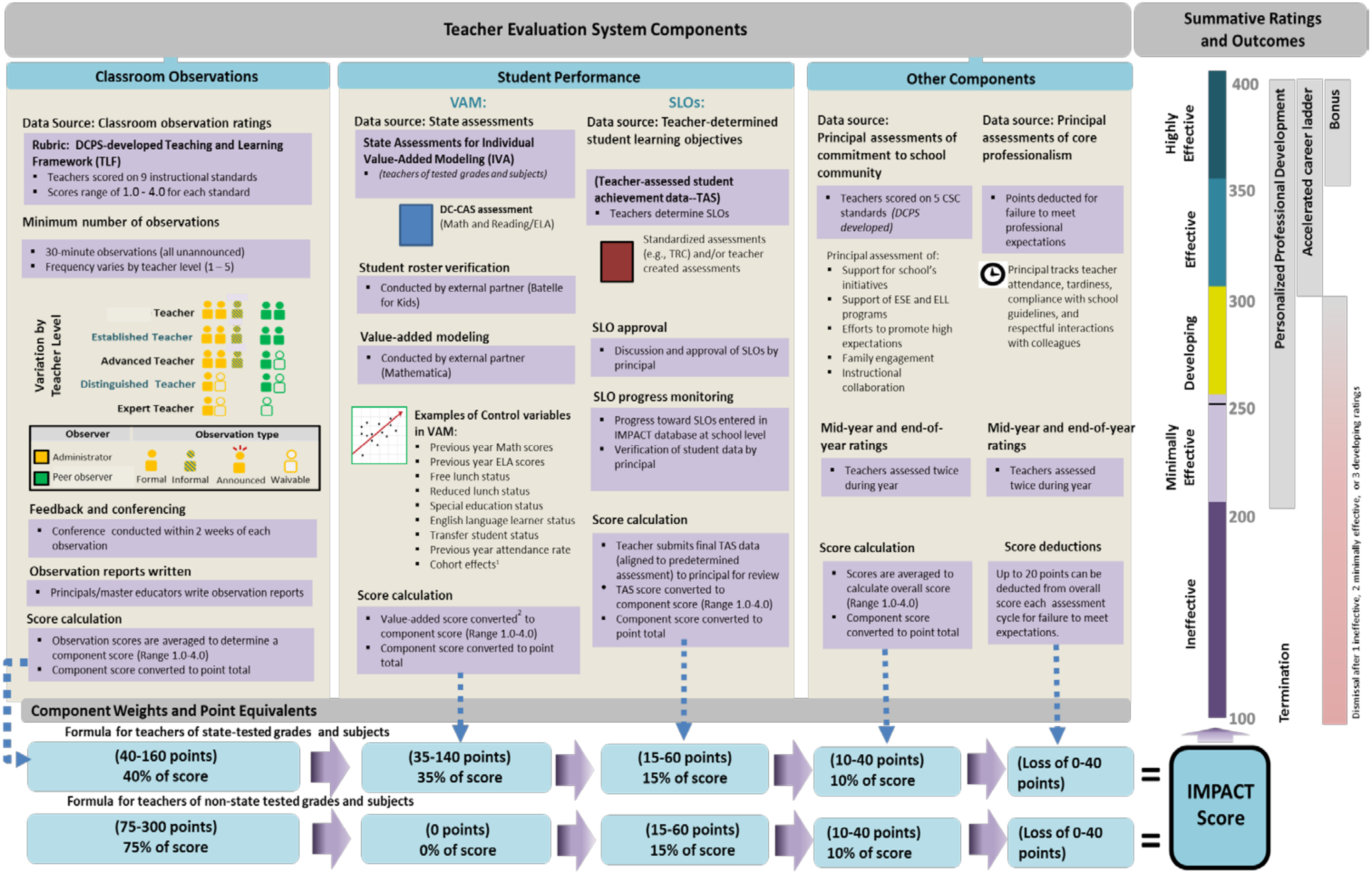
Formula for all teachers

40% of score + 30% of score + 10% of score + 10% of score + 10% of score = Summative Rating

Summative Ratings and Outcomes

As of 2012-13, the evaluation results were not used to inform professional development or other human resource decisions.

Exhibit B2
District of Columbia Public Schools: DC IMPACT



Teacher Evaluation System Components

Classroom Observations

Data Source: Classroom observation ratings

Rubric: DCPS-developed Teaching and Learning Framework (TLF)
- Teachers scored on 9 instructional standards
- Scores range of 1.0 - 4.0 for each standard

Minimum number of observations
- 30-minute observations (all unannounced)
- Frequency varies by teacher level (1 – 5)

Variation by Teacher Level

	Administrator	Peer observer
Teacher	● ● ▨	● ●
Established Teacher	● ● ▨	● ●
Advanced Teacher	● ● ▨	● ○
Distinguished Teacher	● ○	● ○
Expert Teacher	● ○	○

Observer	Observation type
🟨 Administrator 🟩 Peer observer	● Formal ▨ Informal ● Announced ○ Waivable

Feedback and conferencing
- Conference conducted within 2 weeks of each observation

Observation reports written
- Principals/master educators write observation reports

Score calculation
- Observation scores are averaged to determine a component score (Range 1.0-4.0)
- Component score converted to point total

Student Performance

VAM:

Data source: State assessments

State Assessments for Individual Value-Added Modeling (IVA)
- *(teachers of tested grades and subjects)*

DC-CAS assessment (Math and Reading/ELA)

Student roster verification
- Conducted by external partner (Batelle for Kids)

Value-added modeling
- Conducted by external partner (Mathematica)

Examples of Control variables in VAM:
- Previous year Math scores
- Previous year ELA scores
- Free lunch status
- Reduced lunch status
- Special education status
- English language learner status
- Transfer student status
- Previous year attendance rate
- Cohort effects[1]

Score calculation
- Value-added score converted[2] to component score (Range 1.0-4.0)
- Component score converted to point total

SLOs:

Data source: Teacher-determined student learning objectives

(Teacher-assessed student achievement data--TAS)
- Teachers determine SLOs

Standardized assessments (e.g., TRC) and/or teacher created assessments

SLO approval
- Discussion and approval of SLOs by principal

SLO progress monitoring
- Progress toward SLOs entered in IMPACT database at school level
- Verification of student data by principal

Score calculation
- Teacher submits final TAS data (aligned to predetermined assessment) to principal for review
- TAS score converted to component score (Range 1.0-4.0)
- Component score converted to point total

Other Components

Data source: Principal assessments of commitment to school community
- Teachers scored on 5 CSC standards *(DCPS developed)*

Principal assessment of:
- Support for school's initiatives
- Support of ESE and ELL programs
- Efforts to promote high expectations
- Family engagement
- Instructional collaboration

Mid-year and end-of-year ratings
- Teachers assessed twice during year

Score calculation
- Scores are averaged to calculate overall score (Range 1.0-4.0)
- Component score converted to point total

Data source: Principal assessments of core professionalism
- Points deducted for failure to meet professional expectations

Principal tracks teacher attendance, tardiness, compliance with school guidelines, and respectful interactions with colleagues

Mid-year and end-of-year ratings
- Teachers assessed twice during year

Score deductions

Up to 20 points can be deducted from overall score each assessment cycle for failure to meet expectations.

Component Weights and Point Equivalents

Formula for teachers of state-tested grades and subjects

(40-160 points) 40% of score → (35-140 points) 35% of score → (15-60 points) 15% of score → (10-40 points) 10% of score → (Loss of 0-40 points) = **IMPACT Score**

Formula for teachers of non-state tested grades and subjects

(75-300 points) 75% of score → (0 points) 0% of score → (15-60 points) 15% of score → (10-40 points) 10% of score → (Loss of 0-40 points) = **IMPACT Score**

Summative Ratings and Outcomes

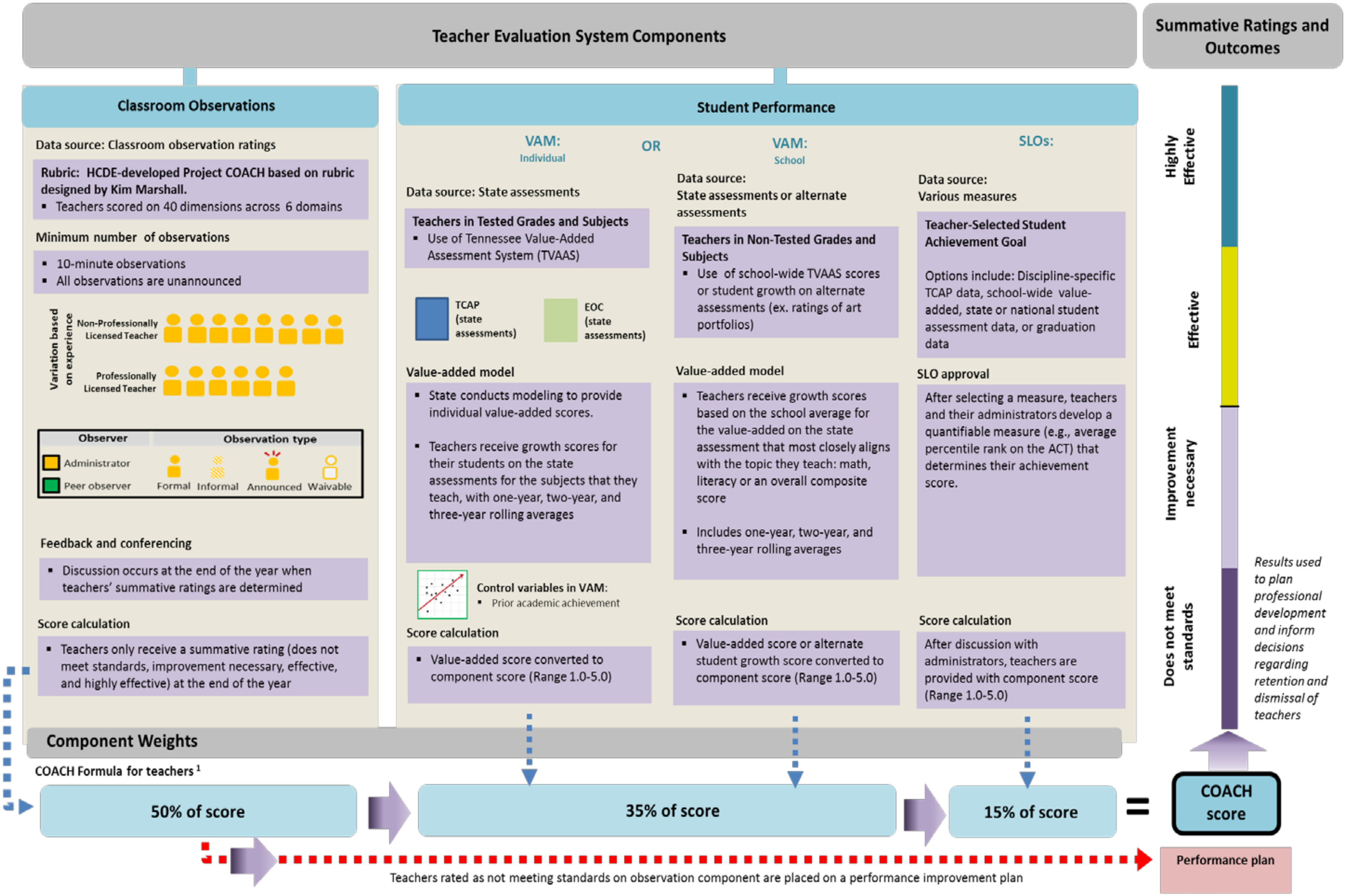
Rating	Points
Highly Effective	350–400
Effective	300–350
Developing	250–300
Minimally Effective	200–250
Ineffective	100–200

400
350
300
250
200
100
Personalized Professional Development
Accelerated career ladder
Bonus
Termination
Dismissal after 1 ineffective, 2 minimally effective, or 3 developing ratings

[1] Cohort effects variables consists of a) class's average test score from the previous year; b) Percentage of the class eligible to receive free or reduced price lunch; and c) extent of variation in the student's scores from the previous year.

[2] Regression models calculate the average likely scale score of students of a DCPS teacher serving a similar student population (by controlling for VAM variables). The individual value-added or IVA score is the scale score difference between the average likely score and the average actual scores of the students taught

Exhibit B3
Hamilton County Department of Education: Project COACH



Teacher Evaluation System Components

Summative Ratings and Outcomes

Classroom Observations

Data source: Classroom observation ratings

Rubric: HCDE-developed Project COACH based on rubric designed by Kim Marshall.
- Teachers scored on 40 dimensions across 6 domains

Minimum number of observations
- 10-minute observations
- All observations are unannounced

Variation based on experience
Non-Professionally Licensed Teacher
Professionally Licensed Teacher

Observer	Observation type
🟧 Administrator 🟩 Peer observer	Formal · Informal · Announced · Waivable

Feedback and conferencing
- Discussion occurs at the end of the year when teachers' summative ratings are determined

Score calculation
- Teachers only receive a summative rating (does not meet standards, improvement necessary, effective, and highly effective) at the end of the year

Student Performance

VAM: Individual

OR

VAM: School

SLOs:

Data source: State assessments

Teachers in Tested Grades and Subjects
- Use of Tennessee Value-Added Assessment System (TVAAS)

Value-added model
- State conducts modeling to provide individual value-added scores.
- Teachers receive growth scores for their students on the state assessments for the subjects that they teach, with one-year, two-year, and three-year rolling averages

Control variables in VAM:
- Prior academic achievement

Score calculation
- Value-added score converted to component score (Range 1.0-5.0)

Data source:
State assessments or alternate assessments

Teachers in Non-Tested Grades and Subjects
- Use of school-wide TVAAS scores or student growth on alternate assessments (ex. ratings of art portfolios)

Value-added model
- Teachers receive growth scores based on the school average for the value-added on the state assessment that most closely aligns with the topic they teach: math, literacy or an overall composite score
- Includes one-year, two-year, and three-year rolling averages

Score calculation
- Value-added score or alternate student growth score converted to component score (Range 1.0-5.0)

Data source:
Various measures

Teacher-Selected Student Achievement Goal

Options include: Discipline-specific TCAP data, school-wide value-added, state or national student assessment data, or graduation data

SLO approval

After selecting a measure, teachers and their administrators develop a quantifiable measure (e.g., average percentile rank on the ACT) that determines their achievement score.

Score calculation

After discussion with administrators, teachers are provided with component score (Range 1.0-5.0)

Results used to plan professional development and inform decisions regarding retention and dismissal of teachers

Component Weights

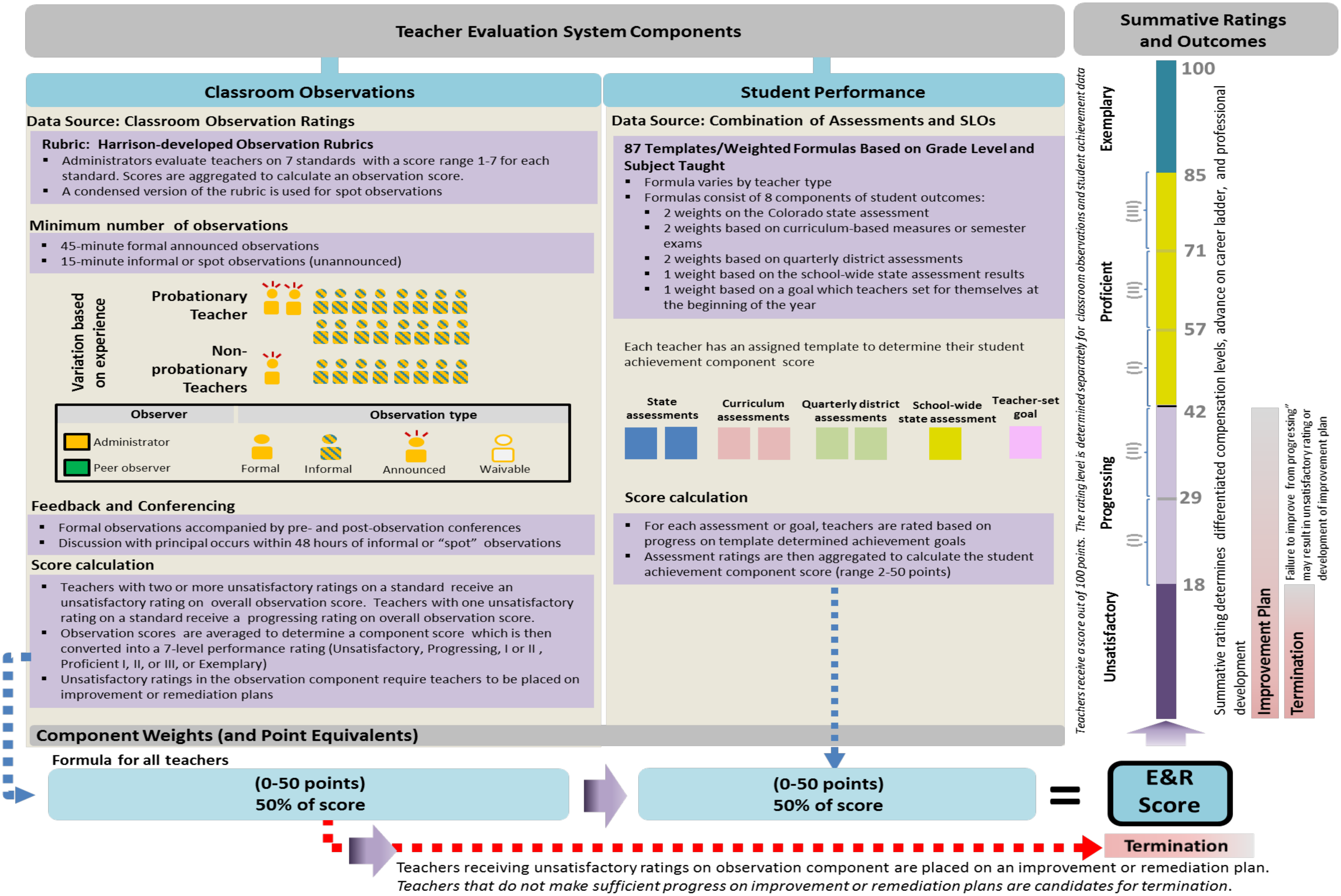
COACH Formula for teachers [1]

50% of score → 35% of score → 15% of score = COACH score

Performance plan

Teachers rated as not meeting standards on observation component are placed on a performance improvement plan

Exhibit B4
Harrison School District 2: Effectiveness and Results (E&R) System



Teacher Evaluation System Components

Classroom Observations

Data Source: Classroom Observation Ratings

Rubric: Harrison-developed Observation Rubrics

- Administrators evaluate teachers on 7 standards with a score range 1-7 for each standard. Scores are aggregated to calculate an observation score.
- A condensed version of the rubric is used for spot observations

Minimum number of observations

- 45-minute formal announced observations
- 15-minute informal or spot observations (unannounced)

Feedback and Conferencing

- Formal observations accompanied by pre- and post-observation conferences
- Discussion with principal occurs within 48 hours of informal or "spot" observations

Score calculation

- Teachers with two or more unsatisfactory ratings on a standard receive an unsatisfactory rating on overall observation score. Teachers with one unsatisfactory rating on a standard receive a progressing rating on overall observation score.
- Observation scores are averaged to determine a component score which is then converted into a 7-level performance rating (Unsatisfactory, Progressing, I or II , Proficient I, II, or III, or Exemplary)
- Unsatisfactory ratings in the observation component require teachers to be placed on improvement or remediation plans

Student Performance

Data Source: Combination of Assessments and SLOs

87 Templates/Weighted Formulas Based on Grade Level and Subject Taught

- Formula varies by teacher type
- Formulas consist of 8 components of student outcomes:
  - 2 weights on the Colorado state assessment
  - 2 weights based on curriculum-based measures or semester exams
  - 2 weights based on quarterly district assessments
  - 1 weight based on the school-wide state assessment results
  - 1 weight based on a goal which teachers set for themselves at the beginning of the year

Each teacher has an assigned template to determine their student achievement component score

Score calculation

- For each assessment or goal, teachers are rated based on progress on template determined achievement goals
- Assessment ratings are then aggregated to calculate the student achievement component score (range 2-50 points)

Summative Ratings and Outcomes

Teachers receive a score out of 100 points. The rating level is determined separately for classroom observations and student achievement data

Summative rating determines differentiated compensation levels, advance on career ladder, and professional development

Improvement Plan

Termination

Failure to improve from progressing" may result in unsatisfactory rating or development of improvement plan

Component Weights (and Point Equivalents)

Formula for all teachers

(0-50 points) 50% of score + (0-50 points) 50% of score = E&R Score

Termination

Teachers receiving unsatisfactory ratings on observation component are placed on an improvement or remediation plan. *Teachers that do not make sufficient progress on improvement or remediation plans are candidates for termination.*

Exhibit B5
Hillsborough County Public Schools: Empowering Effective Teachers (EET) System

Teacher Evaluation System Components

Classroom Observations

Data source: Administrator observation ratings

Rubric: Modified Charlotte Danielson framework

- Principals/administrators observe 17 of 22 components in Domains 1-3 of rubric. Separate from observations, rate 5 components under Domain 4 (professional responsibilities).

Minimum number of observations

- 45-50-minute formal and informal observations (announced)

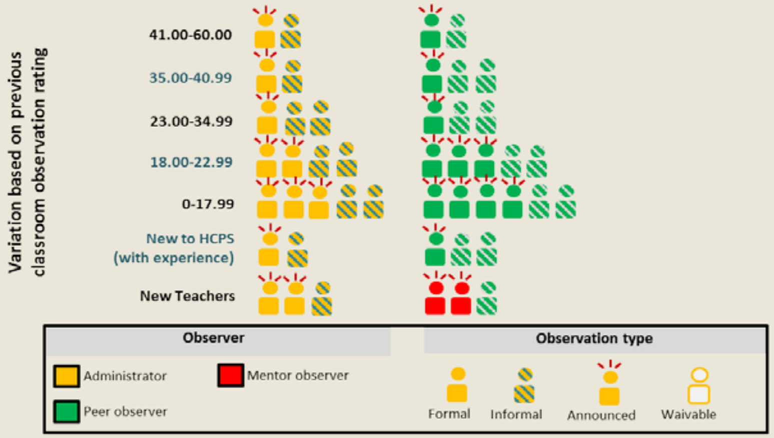
Data Source: Peer/Mentor Observation ratings

Rubric: Modified Charlotte Danielson framework

- Peers/mentors observe 17 of 22 components of the evaluation rubric

Minimum number of observations

- 45-50-minute formal and informal observations (announced)



Feedback and conferencing

- 30-minute conference conducted within 1 week of each formal observation

Score calculation

- Teachers received a score from 0-3 for each domain component. Component scores are aggregated. Administrator and peer observer review scores and determine a holistic evaluation rating.

Feedback and conferencing

- 30-minute conference conducted within 1 week of each formal observation

Score calculation

- Teachers received a score from 0-3 for each domain component. Component scores are aggregated. Administrator and peer observer review scores and determine a holistic evaluation rating.

Student Performance

VAM:

Data Source: Combination of assessments

Value-added modeling (VAM) based on multiple assessments

- The number and combination of assessments used to calculate a teacher's VAM score varies by grade, subject, and available assessments

VAM calculation follows three major stages:

Individual regression models for each exam approved by district to determine coefficients for control variables and to calculate the student residuals/value-added for each teacher on each applicable assessment

Differentiated combination of following assessments

FCAT (state assessments) and/or **End of semester tests** and/or **District-centered assessments** and/or **Pre-approved assessments** and/or **AP, or IB exams**

Regression model developed for each assessment

Value-added modeling (VAM)

- Conducted by external partner University of Wisconsin-Madison

Control variables in VAM:

- Previous year test scores
- ESE status
- English language learner status
- Transfer student status
- Previous year attendance rate
- Comparative age in cohort
- Geographic region

Student roster verification

- Roster verification occurs 7 times per year

Combined VAM score

- Individual test residuals weighted equally and combined to calculate teacher's combined results.

Score calculation

- Combined value-added score converted to scale score (0-40 points)

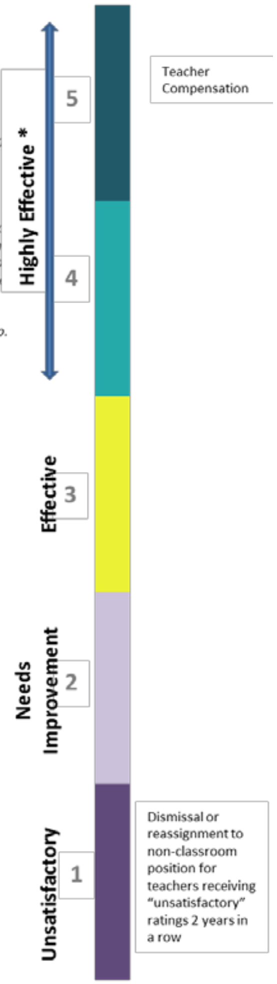
Student residuals for teacher

Roster verification & selection of applicable assessments

Student residuals weighted equally in calculation of teacher combined value-added score

Teacher combined value-added score. Final VAM is the average of last three years (or less for new teachers) of data

Summative Ratings and Outcomes



*Evaluation ratings are reported on a 1-5 scale at the district level and are converted to a 4-point scale to comply with the state system by combining Level 4s and 5s into the Highly Effective group.

Teacher Compensation

Dismissal or reassignment to non-classroom position for teachers receiving "unsatisfactory" ratings 2 years in a row

Component Weights (and Point Equivalents)

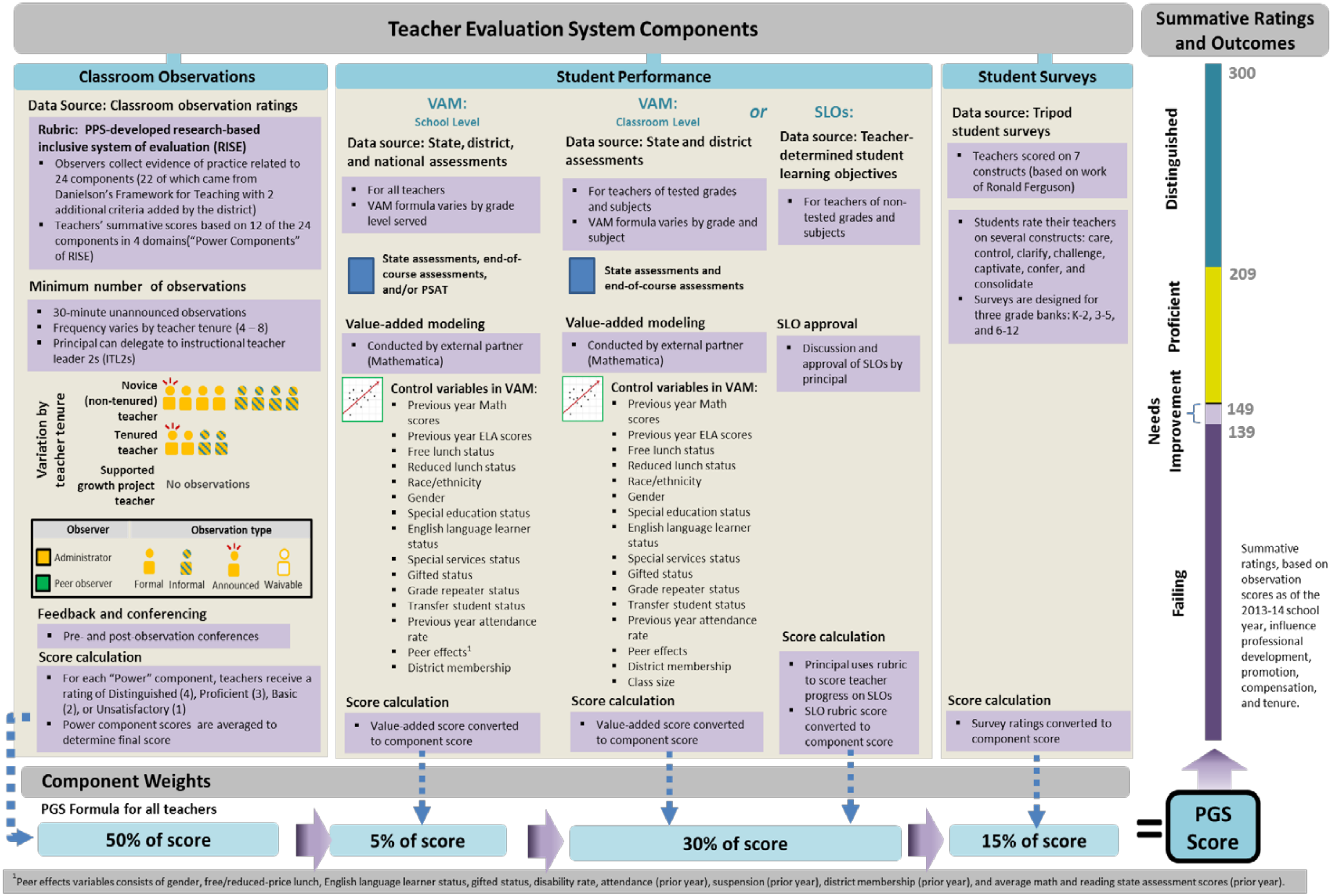
(0-30 points)
35.1% of score

(0-30 points)
24.9% of score

(0-40 points)
40% of score

EET Score

Exhibit B6
Pittsburgh Public Schools: Professional Growth System



Teacher Evaluation System Components

Classroom Observations

Data Source: Classroom observation ratings

Rubric: PPS-developed research-based inclusive system of evaluation (RISE)
- Observers collect evidence of practice related to 24 components (22 of which came from Danielson's Framework for Teaching with 2 additional criteria added by the district)
- Teachers' summative scores based on 12 of the 24 components in 4 domains("Power Components" of RISE)

Minimum number of observations
- 30-minute unannounced observations
- Frequency varies by teacher tenure (4 – 8)
- Principal can delegate to instructional teacher leader 2s (ITL2s)

Variation by teacher tenure

Novice (non-tenured) teacher

Tenured teacher

Supported growth project teacher — No observations

Observer	Observation type
Administrator Peer observer	Formal Informal Announced Waivable

Feedback and conferencing
- Pre- and post-observation conferences

Score calculation
- For each "Power" component, teachers receive a rating of Distinguished (4), Proficient (3), Basic (2), or Unsatisfactory (1)
- Power component scores are averaged to determine final score

Student Performance

VAM: School Level

Data source: State, district, and national assessments
- For all teachers
- VAM formula varies by grade level served

State assessments, end-of-course assessments, and/or PSAT

Value-added modeling
- Conducted by external partner (Mathematica)

Control variables in VAM:
- Previous year Math scores
- Previous year ELA scores
- Free lunch status
- Reduced lunch status
- Race/ethnicity
- Gender
- Special education status
- English language learner status
- Special services status
- Gifted status
- Grade repeater status
- Transfer student status
- Previous year attendance rate
- Peer effects[1]
- District membership

Score calculation
- Value-added score converted to component score

VAM: Classroom Level

Data source: State and district assessments
- For teachers of tested grades and subjects
- VAM formula varies by grade and subject

State assessments and end-of-course assessments

Value-added modeling
- Conducted by external partner (Mathematica)

Control variables in VAM:
- Previous year Math scores
- Previous year ELA scores
- Free lunch status
- Reduced lunch status
- Race/ethnicity
- Gender
- Special education status
- English language learner status
- Special services status
- Gifted status
- Grade repeater status
- Transfer student status
- Previous year attendance rate
- Peer effects
- District membership
- Class size

Score calculation
- Value-added score converted to component score

or

SLOs:

Data source: Teacher-determined student learning objectives
- For teachers of non-tested grades and subjects

SLO approval
- Discussion and approval of SLOs by principal

Score calculation
- Principal uses rubric to score teacher progress on SLOs
- SLO rubric score converted to component score

Student Surveys

Data source: Tripod student surveys
- Teachers scored on 7 constructs (based on work of Ronald Ferguson)

- Students rate their teachers on several constructs: care, control, clarify, challenge, captivate, confer, and consolidate
- Surveys are designed for three grade banks: K-2, 3-5, and 6-12

Score calculation
- Survey ratings converted to component score

Summative Ratings and Outcomes

Summative ratings, based on observation scores as of the 2013-14 school year, influence professional development, promotion, compensation, and tenure.

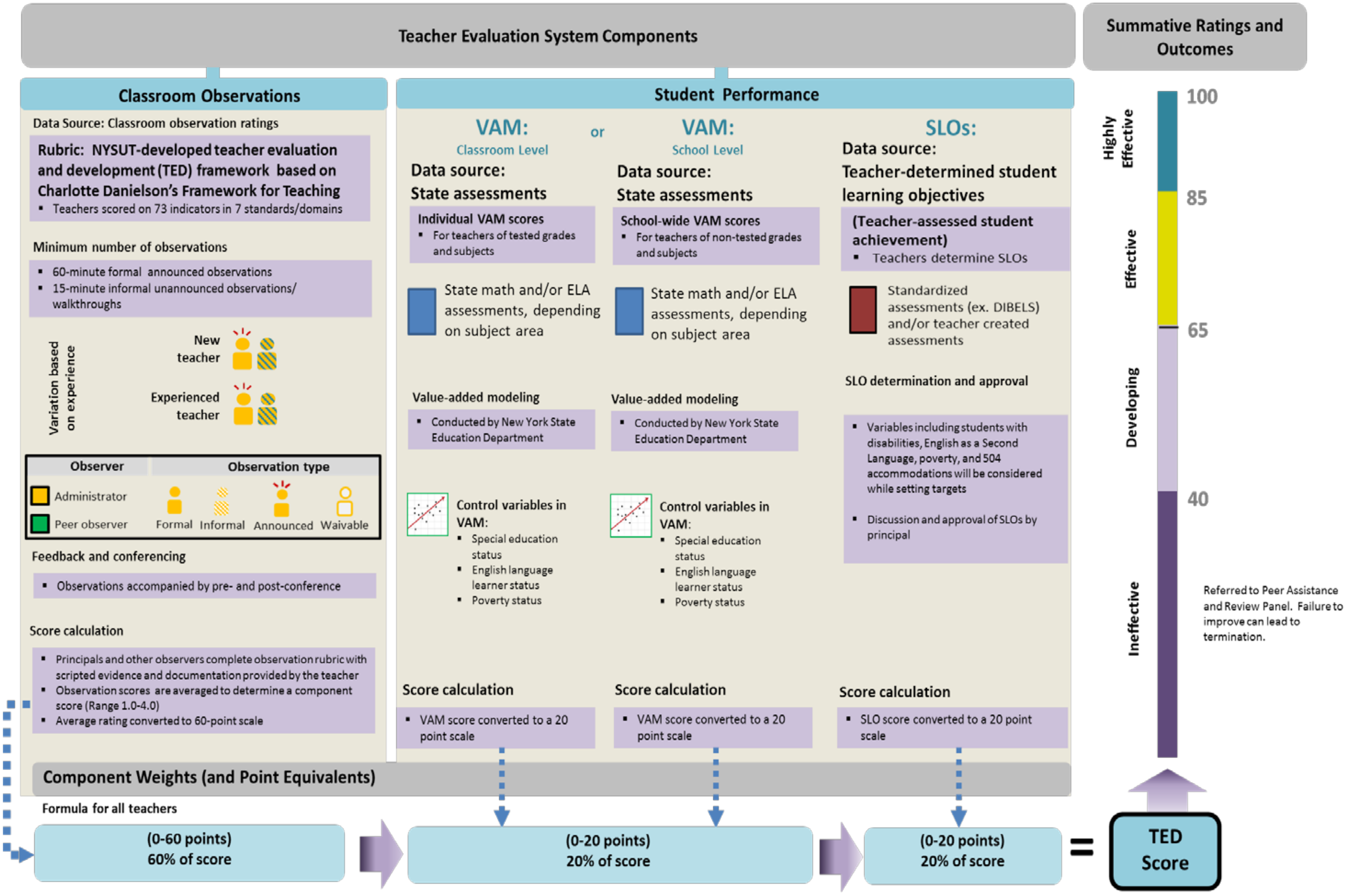
Component Weights

PGS Formula for all teachers

50% of score + 5% of score + 30% of score + 15% of score = PGS Score

[1] Peer effects variables consists of gender, free/reduced-price lunch, English language learner status, gifted status, disability rate, attendance (prior year), suspension (prior year), district membership (prior year), and average math and reading state assessment scores (prior year).

Exhibit B7
Plattsburgh Public Schools: Teacher Evaluation and Development (TED)



Teacher Evaluation System Components

Summative Ratings and Outcomes

Classroom Observations

Data Source: Classroom observation ratings

Rubric: NYSUT-developed teacher evaluation and development (TED) framework based on Charlotte Danielson's Framework for Teaching

- Teachers scored on 73 indicators in 7 standards/domains

Minimum number of observations

- 60-minute formal announced observations
- 15-minute informal unannounced observations/ walkthroughs

Variation based on experience

New teacher

Experienced teacher

Observer	Observation type
🟨 Administrator	Formal
🟩 Peer observer	Informal
	Announced
	Waivable

Feedback and conferencing

- Observations accompanied by pre- and post-conference

Score calculation

- Principals and other observers complete observation rubric with scripted evidence and documentation provided by the teacher
- Observation scores are averaged to determine a component score (Range 1.0-4.0)
- Average rating converted to 60-point scale

Student Performance

VAM: Classroom Level

or

Data source: State assessments

Individual VAM scores

- For teachers of tested grades and subjects

State math and/or ELA assessments, depending on subject area

Value-added modeling

- Conducted by New York State Education Department

Control variables in VAM:

- Special education status
- English language learner status
- Poverty status

Score calculation

- VAM score converted to a 20 point scale

VAM: School Level

Data source: State assessments

School-wide VAM scores

- For teachers of non-tested grades and subjects

State math and/or ELA assessments, depending on subject area

Value-added modeling

- Conducted by New York State Education Department

Control variables in VAM:

- Special education status
- English language learner status
- Poverty status

Score calculation

- VAM score converted to a 20 point scale

SLOs:

Data source: Teacher-determined student learning objectives

(Teacher-assessed student achievement)

- Teachers determine SLOs

Standardized assessments (ex. DIBELS) and/or teacher created assessments

SLO determination and approval

- Variables including students with disabilities, English as a Second Language, poverty, and 504 accommodations will be considered while setting targets
- Discussion and approval of SLOs by principal

Score calculation

- SLO score converted to a 20 point scale

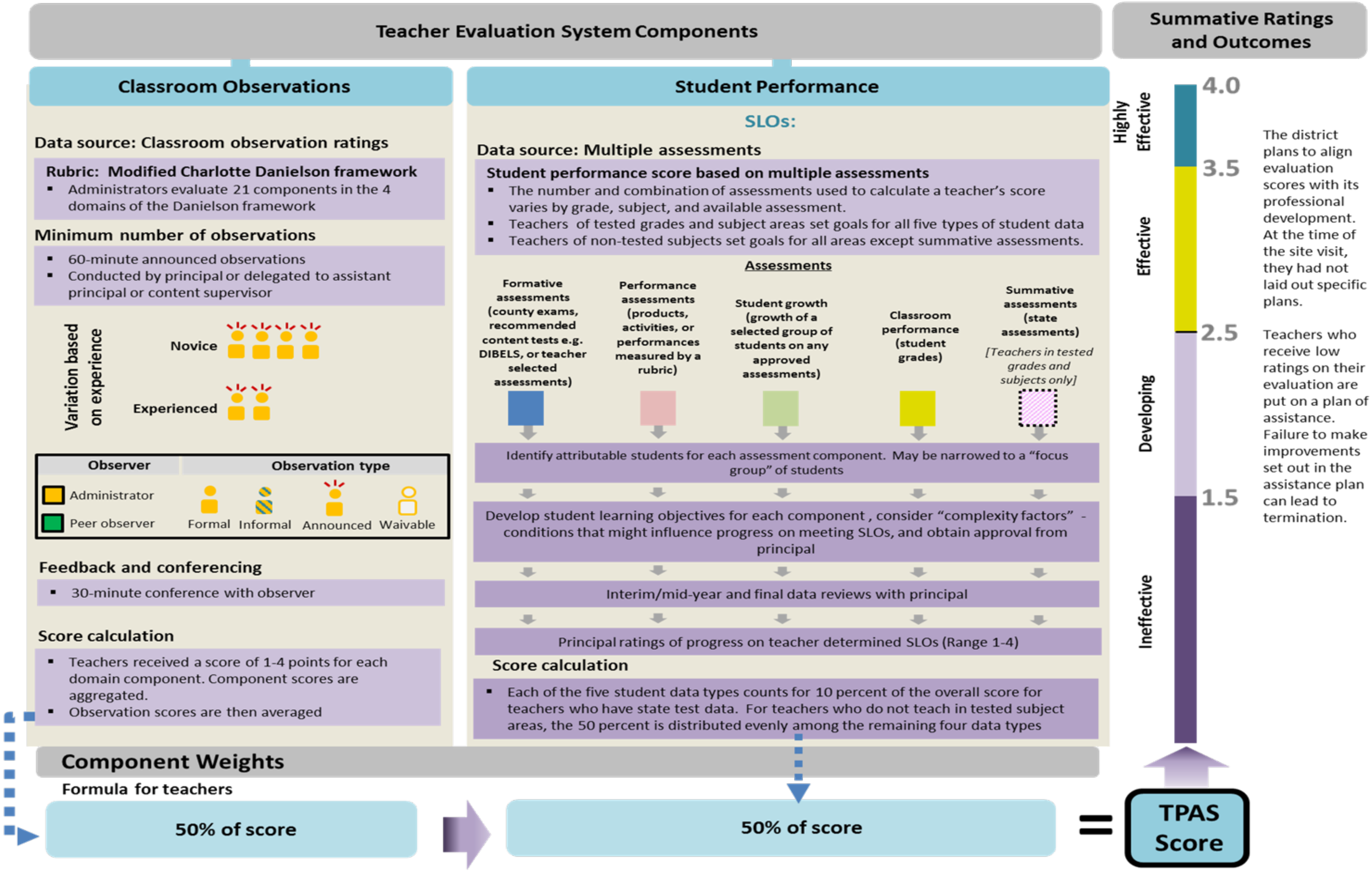
Component Weights (and Point Equivalents)

Formula for all teachers

(0-60 points) 60% of score → (0-20 points) 20% of score → (0-20 points) 20% of score = TED Score

Referred to Peer Assistance and Review Panel. Failure to improve can lead to termination.

Exhibit B8
St. Mary's County Public Schools: Teacher Performance Assessment System (TPAS)



Teacher Evaluation System Components

Classroom Observations

Data source: Classroom observation ratings

Rubric: Modified Charlotte Danielson framework
- Administrators evaluate 21 components in the 4 domains of the Danielson framework

Minimum number of observations
- 60-minute announced observations
- Conducted by principal or delegated to assistant principal or content supervisor

Observer	Observation type
🟨 Administrator 🟩 Peer observer	Formal · Informal · Announced · Waivable

Feedback and conferencing
- 30-minute conference with observer

Score calculation
- Teachers received a score of 1-4 points for each domain component. Component scores are aggregated.
- Observation scores are then averaged

Student Performance

SLOs:

Data source: Multiple assessments

Student performance score based on multiple assessments
- The number and combination of assessments used to calculate a teacher's score varies by grade, subject, and available assessment.
- Teachers of tested grades and subject areas set goals for all five types of student data
- Teachers of non-tested subjects set goals for all areas except summative assessments.

Assessments

- Formative assessments (county exams, recommended content tests e.g. DIBELS, or teacher selected assessments)
- Performance assessments (products, activities, or performances measured by a rubric)
- Student growth (growth of a selected group of students on any approved assessments)
- Classroom performance (student grades)
- Summative assessments (state assessments) *[Teachers in tested grades and subjects only]*

Identify attributable students for each assessment component. May be narrowed to a "focus group" of students

Develop student learning objectives for each component , consider "complexity factors" - conditions that might influence progress on meeting SLOs, and obtain approval from principal

Interim/mid-year and final data reviews with principal

Principal ratings of progress on teacher determined SLOs (Range 1-4)

Score calculation
- Each of the five student data types counts for 10 percent of the overall score for teachers who have state test data. For teachers who do not teach in tested subject areas, the 50 percent is distributed evenly among the remaining four data types

Component Weights

Formula for teachers

50% of score → 50% of score = **TPAS Score**

Summative Ratings and Outcomes

The district plans to align evaluation scores with its professional development. At the time of the site visit, they had not laid out specific plans.

Teachers who receive low ratings on their evaluation are put on a plan of assistance. Failure to make improvements set out in the assistance plan can lead to termination.

Appendix C: Excerpts from Classroom Observation Rubrics, by District

Exhibit C1: Austin Independent School District

Classroom Climate

	Performance Rating			
	Level 1	Level 2	Level 3	Level 4
Sets and implements classroom routines and procedures that support student learning	– Does not design or implement consistent classroom routines and procedures that run smoothly – Does not use instructional time so that students are engaged from the beginning of class to the end of class – Fosters limited student independence or no independence through inefficient classroom routines and procedures	– Designs and implements classroom routines and procedures but does not implement them consistently or teach them to students – Uses instructional time so that some students are engaged from the beginning of class to the end of class – Fosters some student independence through some shared responsibilities for classroom routines and procedures	– Effectively designs and implements consistent classroom routines and procedures that run smoothly – Effectively uses instructional time so that students are engaged from the beginning of class to the end of class – Fosters student independence through shared responsibilities for classroom routines and procedures	– Effectively designs and implements consistent classroom routines and procedures that incorporate student responsibility and run smoothly – Students assume responsibility for utilizing instructional time from the beginning of class to the end of class – Students assume responsibility for routines and procedures and carry them out in an efficient manner with little or no direction from the teacher
Establishes and maintains standards for student behavior	– Does not clearly communicate high student behavioral expectations – Does not reinforce appropriate behavior as needed – Does not encourage or reinforce positive behavior – Does not address off-task or inappropriate behavior efficiently – Off-task or inappropriate behavior interferes with student learning	– Communicates some student behavioral expectations – Reinforces some appropriate behavior – Inconsistently follows through on consequences – Encourages and reinforces positive behavior inconsistently – Addresses off-task or inappropriate behavior inconsistently – Off-task or inappropriate behavior does not interfere with student learning some of the time	– Clearly communicates high student behavioral expectations – Reinforces appropriate behavior as needed – Consistently follows through on consequences – Encourages and reinforces positive behavior – Addresses off-task or inappropriate behavior efficiently – Off-task or inappropriate behavior does not interfere with student learning	– Students demonstrate high behavioral expectations through their actions and require little redirection from the teacher – Students hold each other accountable for appropriate behavior – Consistently follows through on consequences – Students encourage and reinforce positive behavior – Addresses off-task or inappropriate behavior efficiently – Off-task or inappropriate behavior does not interfere with student learning or does not occur
Creates a safe and secure classroom environment that is organized and engages students	– Classroom is not a safe environment. Learning is accessible to some students – Class arrangement is not conducive to learning and does not change as needed for lessons – Classroom environment does not display student work and exemplars Students do not have access to appropriate resources and technology – Students are not invested in their work and do not value their academic success – Students are not able to take risks and challenge themselves	– Classroom is a safe environment. Learning is accessible to most students. – Class arrangement is conducive to learning but does not change as needed for lessons – Classroom environment displays some student work or exemplars – Students have access to some resources and technology – Students are invested in some of their work and sometimes show that they value their academic success – Some students are able to take risks and challenge themselves	– Classroom is a safe environment. – Learning is accessible to all students. – Class arrangement is conducive to learning and changes as needed for lessons – Classroom environment displays student work and exemplars – Students have access to appropriate resources and technology – Students are invested in their work and value their academic success – Students are able to take risks and challenge themselves	– Classroom is a safe environment – Learning is accessible to all students – Class arrangement is a resource that is conducive to individual and group learning and students are able to contribute to the changing design of the environment – Classroom environment displays student work and exemplars – Students have access to appropriate resources and technology – Students are invested in their work and value their academic success as shown through their ownership of classroom routines and behaviors – Students openly take risks and challenge themselves during class

SOURCE: Austin Independent School District. *Teacher Evaluation System: AISD REACH 2011–2012.* Austin, TX: Author.

Exhibit C1 (cont.): Austin Independent School District

Classroom Climate

	Performance Rating			
	Level 1	**Level 2**	**Level 3**	**Level 4**
Establishes a climate that promotes fairness, respect, and diversity	– Students do not actively listen or respond positively to each other and the teacher – Teacher does not communicate or model expectations for respect of student differences – Teacher does not have a positive rapport with students and does not ensures that all students contribute and their opinions' are valued – Teacher does not celebrate student accomplishments	– Students listen occasionally and respond to each other and the teacher – Teacher communicates and models expectations for respect of student differences some of the time – Teacher has a rapport with students and ensures that some students contribute and their opinions' are valued – Teacher celebrates some student accomplishments	– Students actively listen and respond positively to each other and the teacher – Teacher communicates and models expectations for respect of student differences – Teacher has a positive rapport with students and ensures that all students contribute and their opinions' are valued – Teacher celebrates student accomplishments	– Students take initiative socially and participate in creating a climate of respect – Students demonstrate respect for student differences and encourage positive peer interactions – Teacher develops a positive, caring rapport with students and ensures that all students contribute and their opinions' are valued – Students celebrate each other's accomplishments
Provides responsive communication to parents throughout the year	– Does not communicate with parents/guardians regarding performance, behavior, and school activities – Does not promptly respond to parents/guardians within 1–2 school days – Does not celebrate with parents/guardians academic and social successes – Does not maintain a communication log – Does not engage parents in students' academic success	– Communicates infrequently with parents/guardians regarding performance, behavior, and school activities – Responds to parents/guardians – Celebrates with some parents/guardians some academic and social successes – Maintains a sparse communication log – Engages some parents in students' academic success	– Regularly communicates with parents/guardians regarding performance, behavior, and school activities – Promptly responds to parents/ guardians within 1–2 school days – Celebrates with parents/guardians academic and social successes – Maintains a communication log – Engages parents in students' academic success	– Regularly communicates with parents/guardians regarding performance, behavior, and school activities and that communication results in changes in student behavior – Promptly responds to parents/guardians within 1–2 school days – Regularly celebrates with parents/guardians academic and social successes – Maintains a thorough and precise communication log – Engages parents in students' academic success

Exhibit C2: District of Columbia Public Schools

Teach 1: Lead well-organized, objective-driven lessons

	Performance Rating			
	Level 1 (Ineffective)	**Level 2 (Minimally Effective)**	**Level 3 (Effective)**	**Level 4 (Highly Effective)**
Teach 1: Lead well-organized, objective-driven lessons	The lesson is generally disorganized: Parts of the lesson have no connection to each other, most parts of the lesson are not aligned to the objective, or most parts of the lesson do not significantly move students toward mastery of the objective.	The lesson is somewhat organized: Some parts of the lesson are not closely connected to each other or aligned to the objective, or some parts do not significantly move students toward mastery of the objective.	The lesson is well-organized: All parts of the lesson are connected to each other and aligned to the objective, and each part significantly moves students toward mastery of the objective.	The lesson is well-organized: All parts of the lesson are connected to each other and aligned to the objective, and each part significantly moves students toward mastery of the objective.
	The objective of the lesson is not clear to students, or does not convey what students are learning or what they will be able to do as a result of the lesson. For example, students might be unclear or confused about what they are learning and doing, or the objective stated or posted might not connect to the lesson taught.	The objective of the lesson is clear to some students and conveys what students are learning and what they will be able to do as a result of the lesson, but it is not clear to others. For example, the teacher might state the objective, but students' comments, actions, or work products suggest that not all students understand what they are learning or what they will be able to do as a result of the lesson.	The objective of the lesson is clear to students and conveys what students are learning and what they will be able to do as a result of the lesson. For example, students might demonstrate through their comments, actions, or work products that they understand what they are learning and what they will be able to do as a result of the lesson.	The objective of the lesson is clear to students and conveys what students are learning and what they will be able to do as a result of the lesson. Students also can authentically explain what they are learning and doing, beyond simply repeating the stated or posted objective.
	Students do not understand the importance of the objective.	Students do not fully understand the importance of the objective. For example, the teacher might explain the importance of the objective to students in a way that is too general, such that the explanation is not entirely effective in building students' understanding.	Students understand the importance of the objective. For example, the teacher might effectively explain how the objective fits into the broader unit or course goals or how the objective connects to the unit's essential questions or structure; or students might demonstrate through their comments, actions, or work products that they understand the importance of what they are learning and doing.	Students understand the importance of the objective. Students also can authentically explain why what they are learning and doing is important, beyond simply repeating the teacher's explanation.

SOURCE: District of Columbia Public Schools (2012). IMPACT: The District of Columbia Public Schools Effectiveness Assessment System for School-Based Personnel 2012–2013, Group 1: General Education Teachers with Individual Value-Added Student Achievement Data. Washington, DC: Author.

Exhibit C3: Hamilton County Public Schools

Classroom Management

	Performance Rating			
	Does Not Meet Standards	**Improvement Necessary**	**Effective**	**Highly Effective**
Expectations	Comes up with ad hoc rules and punishments as events unfold during the year.	Announces and posts classroom rules and punishments.	Clearly communicates and consistently enforces high standards for student behavior.	Is direct, specific, consistent, and tenacious in communicating and enforcing very high expectations.
Relationships	Is sometimes unfair and disrespectful to the class; plays favorites.	Is fair and respectful toward most students and builds positive relationships with some.	Is fair and respectful toward students and builds positive relationships.	Shows warmth, caring, respect, and fairness for all students and builds strong relationships.
Respect	Is not respected by students and the classroom is frequently chaotic and sometimes dangerous.	Wins the respect of some students but there are regular disruptions in the classroom.	Wins almost all students' respect and refuses to tolerate disruption.	Wins all students' respect and creates a climate in which disruption of learning is unthinkable.
Social-emotional	Publicly berates "bad" students, blaming them for their poor behavior.	Often lectures students on the need for good behavior, and makes an example of "bad" students.	Fosters positive interactions among students and teaches useful social skills.	Implements a program that successfully develops positive interactions and social-emotional skills.
Routines	Does not teach routines and is constantly nagging, threatening, and punishing students.	Tries to train students in class routines but many of the routines are not maintained.	Teaches routines and has students maintain them all year.	Successfully inculcates class routines up front so that students maintain them throughout the year.
Responsibility	Is unsuccessful in fostering self-discipline in students; they are dependent on the teacher to behave.	Tries to get students to be responsible for their actions, but many lack self-discipline.	Develops students' self-discipline and teaches them to take responsibility for their own actions.	Gets all students to be self-disciplined, take responsibility for their actions, and have a strong sense of efficacy.
Repertoire	Has few discipline skills and constantly struggles to get students' attention.	Has a limited disciplinary repertoire and some students are not paying attention.	Has a repertoire of discipline "moves" and can capture and maintain students' attention.	Has a highly effective discipline repertoire and can capture and hold students' attention any time.
Efficiency	Loses a great deal of instructional time because of confusion, interruptions, and ragged transitions.	Sometimes loses teaching time due to lack of clarity, interruptions, and inefficient transitions.	Maximizes academic learning time through coherence, lesson momentum, and smooth transitions.	Skillfully uses coherence, momentum, and transitions so that every minute of classroom time produces learning.
Prevention	Is unsuccessful at spotting and preventing discipline problems, and they frequently escalate.	Tries to prevent discipline problems but sometimes little things escalate into big problems.	Has a confident, dynamic presence and nips most discipline problems in the bud.	Is alert, poised, dynamic, and self-assured and nips virtually all discipline problems in the bud.
Incentives	Gives out extrinsic rewards (e.g., free time) without using them as a lever to improve behavior.	Uses extrinsic rewards in an attempt to get students to cooperate and comply.	Uses incentives wisely to encourage and reinforce student cooperation.	Gets students to buy into a highly effective system of incentives linked to intrinsic rewards.

SOURCE: Tennessee Department of Education. Tennessee Instructional Leadership Standards (TILS) Evaluation Rubric. Nashville, TN: Author.

Exhibit C4: Harrison School District 2

Preparation for Instruction

		Performance Rating			
		Unsatisfactory	**Progressing I Progressing II**	**Proficient I Proficient II**	**Proficient III Exemplary**
1a: Establish a culture of high expectations for learning and achievement	***Expectations and Inclusion***	Teaching practices maintain the status quo and do not contribute to the building culture of high expectations for students.	Acts in ways that demonstrate support of the building culture as one of inclusion and high expectations for most students.	Teaching practices reinforce and strengthen the building culture as one of inclusion and high expectations for *all* students.	Initiates and engages in problem solving to advance the culture of the building as one of inclusion and high expectations for *all* students.
	Culture of Excellence	The culture in the classroom reinforces low-level learning expectations and/or plans to meet even minimal student achievement goals are not clear.	The classroom culture supports student improvement efforts suitable for most students and the teacher outlines the steps to meet student achievement goals.	Establishes a culture in the classroom that challenges *all* students to continuously improve. Develops a plan to measure progress toward meeting challenging student achievement goals.	Creates a culture of excellence in the classroom that focuses on stretching student achievement for *all* student groups. Differentiated plans to meet rigorous student achievement goals are developed and there is a system in place to continuously measure progress toward goal attainment.
	Communicating Expectations	There is little to no evidence that achievement expectations have been communicated to students in advance and/or achievement goals are low.	Achievement expectations are not communicated well to students and/or the achievement goals are not high enough for some students.	Achievement expectations are communicated to students and the teacher provides example of how students can meet challenging achievement goals.	Achievement expectations are communicated in advance and if asked, the student is able to articulate what the goals are. It is evident that students know where they are in relation to the goals.
1b: Use district-adopted curriculum maps and content knowledge to design coherent lessons	***Curriculum and assessment alignment***	Lesson plans, when available, do not align with the district adopted curriculum maps and/or district assessments.	Lesson plans are partially aligned to the district adopted curriculum maps and district assessments.	Lesson plans closely align to the district adopted curriculum maps and district assessments.	Lesson plans are based on a thorough understanding of how to “unpack” the district adopted curriculum maps and alignment of district assessments.
	Content knowledge	Lesson plans reference outdated content knowledge. Information presented in class contains content errors.	Lesson plans are based on a general understanding of content knowledge. While information presented in class is accurate, it may not reflect the most current knowledge of the discipline.	Lessons plans are based on solid content knowledge. Information presented in class is accurate and current.	Lesson plans are based on extensive content knowledge. Information presented in class is accurate, current and consistent with well-established concepts or sound practices of the discipline.
	Lesson and/or unit design	Lesson and unit planning is inadequate. Learning activities do not follow an organized progression and time allocations are unrealistic.	Lesson plans or units are based on activities or resources, rather than focused on objectives. Progression and pacing of learning activities is sporadic, thus, time allocations are not always reasonable.	Lesson or unit is planned in detail around clearly defined lesson objectives. Progression and pacing of the planned learning time (instructional strategies, student activities, use of resources, assessment tasks) is constant, with reasonable time allocations.	Lesson or unit is precisely planned with explicit attention to detail leading to the demonstration of learning of the lesson objectives. The progression and pacing of planned learning time (instructional strategies, accessing materials, use of resources, student activities, and assessment tasks) is highly coherent.

SOURCE: Miles, M. (January 2011). Teacher Compensation Based on Effectiveness: The Harrison Pay-for-Performance Plan. Colorado Springs, CO: Author.

Exhibit C4 (cont.): Harrison School District 2

Preparation for Instruction

		Performance Rating						
		Unsatisfactory	**Progressing I**	**Progressing II**	**Proficient I**	**Proficient II**	**Proficient III**	**Exemplary**
1.c Post aligned lesson objectives and plan for demonstrations of learning	***Posts Lesson Objectives***	Lesson objectives cannot be found or the teacher simply posts a list of activities.	Posts lesson objectives that reference grade-level and/or course content. The objectives are used to re-focus student's attention to task.		Posts lesson objectives that align to a grade level or course essential content or skills. The lesson objectives effectively focus student attention at the beginning of the lesson and they are used to refocus student's attention to task.		Posts lesson objectives that align to cross disciplinary grade-level course essential content and skills. The objectives serve to effectively focus student's attention to learning targets throughout the lesson.	
	Student Understanding of Lesson Objectives	Lesson objectives are not known to students and students do not know what they are expected to know and be able to do.	Although students are aware of where to find the posted lesson objectives, they rely on teacher direction to focus them on what they are expected to know and be able to do.		Lesson objectives are written in student-friendly language and students understand what they are expected to know and be able to do by the end of each lesson.		Students have been well prepared to know that the lesson objective and the demonstration of learning provide direction for them in understanding exactly what they are expected to know and be able to do. This clarity promotes both autonomy and independence in accomplishment of student tasks.	
	Plans for DOL's	The Demonstration of Learning (DOL) is not developed in advance of instruction and/or not aligned with the lesson objective.	The Demonstration of Learning (DOL) is minimally developed and/or may be loosely connected to the lesson objective.		The Demonstration of Learning (DOL) is developed in advance of instruction and is aligned with the lesson objective.		The Demonstrations of Learning (DOL) are designed in advance of instruction, tie closely with the lesson objective and provide multiple ways for students to demonstrate what they have learned.	

Exhibit C5: Hillsborough County Public Schools

Classroom Environment

	Performance Rating			
	Requires Action (0 points)	**Progressing** (1 point)	**Accomplished** (2 points)	**Exemplary** (3 points)
2a: Creating an Environment of Respect and Rapport	Classroom interactions, between the teacher and students and/or among students, are negative, inappropriate, or insensitive to students' cultural backgrounds and are characterized by sarcasm, put-downs, or conflict.	Classroom interactions, between the teacher and students and among students, are generally appropriate and free from conflict, but may be characterized by occasional behaviors and/or language that compromise the promotion of learning.	Classroom interactions between the 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 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 importance and relevancy of learning goals of the lesson, low expectations for student achievement, little or no student pride in work and no evidence that students believe that they can succeed if they work hard.	The teacher's attempt to create a culture for learning is partially successful, with moderate teacher commitment to the importance and relevancy of learning goals of the lesson, some evidence that students are committed to success beyond completion of assignments, modest expectations for student achievement, and little student pride in work.	The classroom culture is characterized by high expectations for most students, the belief that students can succeed if they work hard, and genuine commitment to the subject by both teacher and students, with students demonstrating pride in their work.	High levels of student energy and teacher passion for the subject create a culture of learning in which everyone shares a belief in the importance of the subject and the belief that students can succeed if they work hard. All students hold themselves to high standards of performance — for example, by initiating improvements
2c: Managing Classroom Procedures	Much instructional time is lost because of inefficient classroom routines and procedures for transitions, handling of supplies, and performance of non-instructional duties.	Some instructional time is lost because classroom routines and procedures for transitions, handling of supplies, and performance of non-instructional duties are only partially effective.	Little instructional time is lost because of classroom routines and procedures for transitions, handling of supplies, and performance of non-instructional duties, which occur smoothly. Students willingly assist with procedures when asked.	Students contribute without prompting 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 does not reinforce positive behavior. The teacher does not address off-task, inappropriate, or challenging behavior efficiently. Inappropriate and off-task student behavior has significant negative impact on the learning of students in the class.	It appears that the teacher has made an effort to establish standards of conduct for students and tries to monitor student behavior and respond to student misbehavior, but these efforts are not always successful. The teacher reinforces positive behavior. The teacher addresses some off-task, inappropriate, or challenging behavior efficiently. Inappropriate and off-task student behavior has some negative impact on the learning of students in the class.	Standards of conduct appear to be clear to students, and the teacher monitors student behavior against those standards. The teacher's response to student misbehavior is appropriate and respectful to students. The teacher strategically reinforces positive behavior. The teacher addresses most off-task, inappropriate, or challenging behavior efficiently. Inappropriate and off-task student behavior has little negative impact on the learning of students in the class.	Standards of conduct are clear, with evidence of student participation in setting them. The teacher's monitoring of student behavior is subtle and preventive, and responses to student misbehavior are sensitive to individual student needs. Students actively monitor the standards of behavior. The teacher strategically reinforces positive behavior AND there is significant evidence that students reinforce positive classroom culture. The teacher addresses almost all off-task, inappropriate, or challenging behavior efficiently. Inappropriate and off-task behavior has no negative impact on student learning.
2e: Organizing Physical Space	The physical environment is unsafe, or many students don't have access to learning. Alignment between the physical arrangement and the lesson activities is poor.	The classroom is safe, and essential learning is accessible to most students. The teacher may attempt to modify the physical arrangement to suit learning activities with partial success.	The classroom is safe, and learning is accessible to all students; the teacher ensures that the physical arrangement supports the learning activities. The teacher makes effective use of physical resources.	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. The teacher uses technology skillfully, as appropriate to the lesson.

SOURCE: Hillsborough County Public Schools (July 2012). Classroom Teacher Evaluation Instrument. Tampa, FL: Author.

Exhibit C6: Pittsburgh Public Schools

Professional Responsibilities

	Performance Rating			
	Unsatisfactory	**Basic**	**Proficient**	**Distinguished**
4a: **Reflecting on teaching and student learning**	The teacher provides an inaccurate and/or provides no assessment of the lesson's effectiveness, and the degree to which outcomes were met based on evidence of student learning. The teacher offers no suggestions for lesson improvement. In reflecting on practices, the teacher does not indicate that it is important to reach all students.	The teacher provides an inconsistent assessment of the lesson's effectiveness, and the degree to which outcomes were met based on evidence of student learning. The teacher offers general suggestions for lesson improvement. In reflecting on practice, the teacher indicates the desire to reach all students, but does not suggest specific strategies to do so.	The teacher provides an accurate assessment of the lesson's effectiveness, and the degree to which outcomes were met based on evidence of student learning. The teacher offers specific suggestions for lesson improvement. In reflecting on practice, the teacher cites multiple approaches undertaken to reach students having difficulty.	The teacher provides a thoughtful and accurate assessment of a lesson's effectiveness, and the degree to which outcomes were met based on evidence of student learning. The teacher offers many and specific suggestions for lesson improvement and can discuss the probable success of different course of action. In reflecting on practice, the teacher can cite others in the school and beyond who s/he has contacted for assistance in reaching some students.
4b: **System for managing students' data**	The teacher's systems for maintaining both instructional and/or non-instructional records are non-existent or in disarray, resulting in errors and confusion. Data is not shared with students.	The teacher's systems for maintaining both instructional and non-instructional records are rudimentary. Some data is shared with students.	The teacher's systems for maintaining both instructional and non-instructional records are accurate and efficient. Most data is shared with students.	The teacher's systems for maintaining both instructional and non-instructional records are accurate, efficient and effective. Data is shared with students and students understand the importance of the data and contribute to the maintenance of the data.
4c: **Communicating with families**	Rarely responds to parent/ guardian inquires and rarely shares information with them about their student and/or instructional program. Contacts with families are culturally insensitive or inappropriate with content. Does not initiate positive communication with families.	Responds to parent/guardian inquires and shares information with them about their student and/or instructional program. Contacts are culturally appropriate and acceptable content. Occasionally initiates positive communication with families.	Initiates contact with parents/ guardian and responds to inquiries from parents/guardian informing them about their student and/or instructional program. Contacts are culturally appropriate and acceptable in content. Regularly initiates positive communication with families.	Regularly communicates with parent/guardian to inform them about their student and/or instructional program and allow students to have a role in the communication. Contacts are culturally appropriate and acceptable in content. Consistently and regularly initiates positive communication with families that results in a two-way dialogue between home and school.
4d: **Participating in a professional community**	Professional relationships are negative or self-serving. Avoids participation in the professional learning community and does not participate in the culture of inquiry.	Professional relationships with other educators are cordial/appropriate. Participates in the professional learning community and in the culture of inquiry when asked or invited.	Professional relationships are positive and characterized by mutual support and cooperation. Actively participates in the professional learning community.	Professional relationships are positive and characterized by mutual support and cooperation where the teacher makes substantial contributions to the development of the community. Assumes leadership in the professional learning community and promotes a culture of inquiry.
4e: **Growing and developing professionally**	Participation in professional development activities is rare and the teacher makes no effort to share knowledge with others or to assume professional responsibilities. Feedback on performance is not accepted and/or used to improve performance.	Participation and contributions to professional development are limited. Feedback on performance is accepted but only some of the feedback is used to improve performance.	Actively participates in and/or pursues building, district and other professional development opportunities, and shares new information with colleagues. Feedback on performance is welcomed and used to improve performance.	Consistently contributes to the professional learning community at the building and/or district. Feedback on performance is solicited and successfully used to improve performance.
4f: **Showing professionalism**	Interactions and practice are characterized by a lack of honesty, integrity, and awareness of student needs. Decisions are self-serving, and/or do not comply with district initiatives. Inappropriate dress.	Interactions and practice are characterized by honest, but inconsistent attempts to serve students. Decision-making is based on limited information, and/or minimal compliance with district initiatives. Somewhat appropriate dress.	Interactions and practice are characterized by honesty, integrity, confidentially and/or assurance that student needs are consistently met. Decision-making contributes to a culture of continuous improvement in district initiatives. Appropriate dress.	Professional interactions and practice display the highest standards of honesty, integrity, confidentially. Decision-making is focused upon student learning, reflects and assumption of a leadership role with colleagues, and an abundance of information used in challenging negative attitudes/ practices, in encouraging a culture of continuous improvement in district initiatives. Appropriate dress.

SOURCE: Pittsburgh Public Schools. RISE Teacher Evaluation Process 2011–2012. Pittsburgh, PA: Author.

Exhibit C7: Plattsburgh City Public Schools
NYSUT's Teacher Practice Rubric Aligned with New York State Teaching Standards

Knowledge of Students and Student Learning

		Performance Rating			
		Ineffective	**Developing**	**Effective**	**Highly Effective**
Teachers demonstrate knowledge of child and adolescent development, including students' cognitive, language, social, emotional, and physical developmental levels.	**Describes developmental characteristics of students**	Teacher is unable to describe orally or in writing the developmental characteristics of the age group.	Teacher describes orally and in writing some knowledge of the developmental characteristics of the age group.	Teacher describes orally and in writing an accurate knowledge of the typical developmental characteristics of the age group, as well as exceptions to the general patterns.	In addition to accurate knowledge of the typical developmental characteristics of the age group, and exceptions to the general patterns, teacher describes orally and in writing the extent to which individual students follow the general patterns and how 21st Century Skills fit into this knowledge base.
	Creates developmentally appropriate lessons	Teacher does not create lessons that are developmentally appropriate or that address individual student learning needs.	Teacher creates lesson plans that are generally appropriate to the developmental needs of students and meet the student learning needs of groups of students.	Teacher creates lesson plans that are appropriate to the developmental needs of students and meet the student learning differences and needs of groups of students.	Teacher creates lesson plans that are appropriate to the developmental needs of students and meet the student learning differences and needs of each individual student.
Teachers demonstrate current, research-based knowledge of learning and language acquisition theories and processes.	**Uses strategies to support learning and language acquisition**	Teacher designs lessons with few strategies that support student learning and language acquisition needs. Teacher does not adjust instruction.	Teacher designs lessons to include some instructional strategies that support the learning and language acquisition needs of some students. Teacher is able to adjust instruction by implementing one or two additional strategies.	Teacher designs lessons to include several instructional strategies that support the learning and language acquisition needs of most students. Teacher is able to adjust instruction by adapting and/or adding strategies to meet the needs of specific students.	Teacher designs lessons to include several instructional strategies that support the learning and language acquisition needs of each student. Teacher is able to adjust instruction by adapting and/or adding strategies to meet the needs of specific students. Students suggest specific strategies that help them achieve the outcomes of the lesson and teacher supports the students' suggestions.
	Uses current research	Teacher is unable to cite current research to explain instructional decisions.	Teacher cites limited or dated research to explain instructional decisions.	Teacher cites current research to explain instructional decisions.	Teacher cites current research to explain instructional decisions and seeks out additional research to inform practice.
Teachers demonstrate knowledge of and are responsive to diverse learning needs, strengths, interests, and experiences of all students.	**Meets diverse learning needs of each student**	Teacher does not vary or modify instruction to meet diverse learning needs of students.	Teacher varies or modifies instruction to meet the diverse learning needs of some students.	Teacher varies or modifies instruction to meet the diverse learning needs of most students.	Teacher varies or modifies instruction to meet the diverse learning needs of each student. Students suggest ways in which the lesson might be modified to advance their own learning and teacher acknowledges the suggestion.
	Plans for student strengths, interests, and experiences	Teacher does not plan instruction to address the strengths, interests, and experiences of students.	Teacher plans instruction to address the strengths, interests, and experiences of some students.	Teacher plans instruction to address the strengths, interests, and experiences of most students.	Teacher plans instruction to address the strengths, interests, and experiences of each student and is able to adapt the lesson as needed.

SOURCE: New York State Education Department (February 2012). Guidance on New York State's Annual Professional Performance Review Law and Regulations. Albany, NY: Author.

Exhibit C7 (cont.): Plattsburgh City Public Schools
NYSUT's Teacher Practice Rubric Aligned with New York State Teaching Standards

Knowledge of Students and Student Learning

		Performance Rating			
		Ineffective	**Developing**	**Effective**	**Highly Effective**
Teachers acquire knowledge of individual students from students, families, guardians, and/or caregivers to enhance student learning.	**Communicates with parents, guardians, and/or caregivers.**	Teacher does not communicate directly with student's parents, guardians, and/or caregivers to enhance student learning and/or does not accommodate the communication needs of the family.	Teacher occasionally communicates directly with student's parents, guardians, and/or caregivers to enhance student learning. Communication is occasionally modified to meet the needs of the family.	Teacher regularly communicates directly with student's parents, guardians, and/or caregivers to enhance student learning. Communication is frequent and uses multiple modes of contact to accommodate the needs of the family.	Teacher communicates directly with student's parents, guardians, and/or caregivers to enhance student learning. Multiple modes of contact are used to accommodate the needs of the family. Students and parents/guardians initiate communication.
Teachers demonstrate knowledge of and are responsive to the economic, social, cultural, linguistic, family, and community factors that influence their students' learning.	**Incorporates the knowledge of school community and environmental factors**	Teacher does not incorporate knowledge and understanding of the school community when designing or implementing instruction.	Teacher incorporates general knowledge of the school community when planning and implementing instruction.	Teacher incorporates detailed and specific knowledge of the school community when planning and implementing instruction reflecting a deep understanding of the school community.	Teacher incorporates detailed and specific knowledge of the school community when planning and implementing instruction, reflecting a deep understanding of the school community. Teacher continuously seeks additional information to impact instruction.
	Incorporates multiple perspectives	Teacher does not consider students' personal and family experiences when discussing content.	Teacher considers students' personal and family experiences when discussing content by incorporating more than one perspective.	Teacher considers students' personal and family experiences when discussing content by incorporating multiple perspectives.	Teacher considers students' personal and family experiences when discussing content by incorporating multiple perspectives. Students are supported by the teacher to share their personal perspective as it relates to the content.
Teachers demonstrate knowledge and understanding of technological and information literacy and how they affect student learning.	**Understands technological literacy**	Teacher does not use available technological tools or a variety of communication strategies to engage students or assist them in becoming critical users of quality information. Teacher is unaware of 21st Century Skills.	Teacher uses available technological tools and communication strategies to engage some students and/or to assist them in becoming critical users of quality information. Teacher's knowledge of 21st Century Skills is rudimentary.	Teacher uses available technological tools and communication strategies to engage most students, and to assist them in becoming critical users of quality information. Teacher's knowledge of 21st Century Skills is current and embedded in the communication strategies.	Teacher uses available technological tools and communication strategies to engage each student. Teacher's knowledge of 21st Century Skills is current and embedded in the communication strategies. Students contribute to the variety of technological strategies used to engage them in their own learning and become critical users of quality information.

Exhibit C8: St. Mary's County Public Schools

Instruction					
		Performance Rating			
		Ineffective	**Developing**	**Effective**	**Highly Effective**
Communicates Clearly and Accurately	**Expectations for Learning**	Teacher's purpose in a lesson or unit is unclear to students.	Teacher attempts to explain the instructional purpose, with limited success.	Teacher's purpose for the lesson or unit is clear, including where it is situated within the unit or long range plan.	Teacher's purpose of the lesson or unit clear, including where it is situated within unit or long range plan, linking that purpose to student expectations and relevance.
	Directions and Procedures	Teacher's directions and procedures are lacking or are confusing to students.	Teacher's directions and procedures are clarified when directions are confusing or contain too much or too little detail.	Teacher frequently uses both auditory and visual cues as well as modeling when appropriate to ensure that directions and procedures are clear to students. Techniques are in place to check for understanding.	Teacher consistently uses both auditory and visual cues as well as modeling when appropriate to ensure that directions and procedures are clear to students. Teacher anticipates possible student(s) misunderstanding, checks for understanding, and plans are evident for re-teaching.
	Explanations of Content	Teacher's explanation of the content is unclear, confusing or inappropriate.	Teacher's explanation makes limited connections with students' knowledge and experience.	Teacher's explanation of content is appropriate and connects with students' knowledge and experience.	Teacher's explanation of content is appropriate and connects with students' knowledge and experience. Students contribute to explaining concepts to their peers.
	Oral and Written Language	Teacher's spoken language is inaudible, or written language is illegible. Spoken or written language contains grammar and syntax errors. Vocabulary is inappropriate, vague, or used incorrectly, leaving students confused.	Teacher's spoken language is audible, and written language is legible. Both are used correctly. In some situations, the vocabulary is limited or not appropriate to students' ages or backgrounds.	Teacher's spoken and written language is clear and correct. Vocabulary is appropriate to students' age and interests.	Teacher's spoken and written language is clear, correct, and expressive, with well-chosen vocabulary that enriches the lesson.
Uses Higher Order Questioning and Discussion Techniques	**Quality Of Questions**	Teacher's questions are virtually all of low cognitive challenge (e.g., recall or yes/no) and/or low quality (poorly worded or incomprehensible).	Teacher's questions are virtually all of low cognitive challenge (e.g., recall or yes/no). Teacher introduces some variety into the level of questioning. Few questions invite an extended response.	Teacher uses a variety of questions that invite an extended response and require higher level thinking skills. Adequate response time is available. Students often formulate thoughtful responses.	Teacher consistently and skillfully uses questioning techniques that elicit thoughtful response for effective decision-making or problem solving. Adequate response time is provided.
	Discussion Techniques	Interaction between teacher and students is predominantly recitation style, with teacher mediating all questions and answers.	Teacher makes some attempt to engage students in teacher-directed discussions through follow up questions with inconsistent results.	Classroom interaction represents true discussion with teacher stepping to the side when appropriate. Both teacher and students use follow up questions and other techniques.	Teacher creates environment in which students assume considerable responsibility for the success of discussions, initiating and expanding upon topics appropriately.
	Student Participation	Teacher actively engages only a few students in discussions and learning activities. Participation reflects gender, cultural, ability, seating, or other patterns of which the teacher is unaware.	Teacher attempts to actively engage students in discussions and learning activities. Teacher displays some limited recognition of gender, cultural, ability, seating, or other patterns of participation and some attempts are made to address them.	Teacher actively engages all students in discussions and learning activities. Participation generally reflects equity in gender, cultural, ability, seating, or other patterns of responses and involvement.	Teacher actively engages all students in discussions and learning activities. Participation reflects equity in gender, cultural, ability, seating, or other patterns of responses and involvement. Teacher creates an environment in which the perspectives of all students are sought and in which all students are drawn into the discussion and learning activities.

SOURCE: Martirano, M. (2012). St. Mary's County Public Schools Teacher Performance Assessment System. Leonardtown, MD: St. Mary's County Public Schools.

Exhibit C8 (cont.): St. Mary’s County Public Schools

Instruction					
		Performance Rating			
		Ineffective	Developing	Effective	Highly Effective
Engages Students in Learning	**Presentation of Content**	Methods (examples, analogies, graphic representations) used to represent or relay content are limited, inappropriate and/or unclear.	Methods (examples, analogies, graphic representations) used to represent or relay content are limited and/or inconsistent in quality.	Methods (examples, analogies, graphic representations) used to represent or relay content are varied and link well with students’ knowledge and experience, promoting understanding and retention.	Methods (examples, analogies, graphic representations) used to represent content are rich, varied, and link well with students’ knowledge and experience, promoting understanding and retention. Students contribute to representations of content.
	Activities and Assignments	Activities and assignments are rarely appropriate for students in terms of their age or backgrounds. Students are not cognitively engaged.	Some activities and assignments are appropriate and cognitively engage students.	Most activities and assignments are appropriate to students. Almost all students are cognitively engaged.	All students are cognitively engaged in the activities and assignments in their exploration of content. Students initiate or adapt activities and projects to enhance understandings.
	Flexible Grouping of Students for Differentiated Instruction	Instructional groups do not support students’ instructional goals. There is no differentiation of instruction.	Instructional groups support students and are moderately successful in advancing the instructional goals of a lesson. There is some differentiation in the instructional strategies.	Instructional groups are productive and fully appropriate to the students or to the instructional goals of a lesson. There is a clear differentiation of instruction for the various groups.	Instructional groups are productive and fully support instructional outcomes, goals from county and state guidelines and standards for the lesson. Teacher provides for flexible grouping to accommodate student differences in readiness, interest, and learning needs.
	Instructional Resources	The use of instructional materials and resources (including appropriate and available instructional technologies) are absent, unsuitable to the instructional outcomes and standards, or do not engage students cognitively or in active learning.	The use of instructional materials and resources (including appropriate and available instructional technologies) are partially suitable to the instructional outcomes, and moderately engages students cognitively.	The use of instructional materials and resources (including appropriate and available instructional technologies) are suitable to the instructional outcomes and engage students in higher level thinking skills.	The use of instructional materials and resources (including appropriate and available instructional technologies) are suitable to the instructional outcomes, and engage students in higher level thinking and active learning. Students initiate the choice, exploration, adaptation, or creation of materials to enhance their own purposes.
	Lesson/Unit Structure	The lesson, unit/theme has no clearly defined structure, or the pacing of the lesson, unit/theme is inappropriate.	The lesson, unit/theme has a recognizable structure although it is not uniformly maintained throughout the lesson. Pacing is inconsistent or inappropriate.	The lesson, unit/theme has a clearly defined structure around which problem-solving activities are organized. An opening, closure, and appropriate transitions are included. Pacing is generally appropriate.	The lesson, unit/theme structure is highly coherent. Appropriate time is provided for comprehending, applying, analyzing, synthesizing, and evaluating, as well as for reflection and closure. Transitions are clear, and pacing is appropriate for all students.
Uses Assessments in Instruction	**Assessment Criteria**	Teacher does not articulate the criteria and performance standards by which students’ work will be evaluated or the criteria is unclear.	Teacher articulates some of the criteria and performance standards by which students’ work will be evaluated, but some are unclear.	Teacher clearly articulates the criteria and performance standards by which students’ work will be evaluated.	Teacher clearly articulates the criteria and performance standards by which students’ work will be evaluated, and students are involved in the review of the criteria.
	Monitoring of Student Learning	Teacher does not monitor student learning.	Teacher monitors the progress of the class as a whole, but elicits little or no diagnostic assessment information for individual students.	Teacher monitors the progress of groups of students relative to the learning objectives, making use of the diagnostic assessments to elicit information.	Teacher actively and systematically elicits diagnostic assessments from individual students regarding their understanding and monitors the progress of individual students.
	Utilizes a Variety of Assessment Instruments	Teacher does not use appropriate assessment instruments.	Teacher utilizes few assessment instruments.	Teacher utilizes a variety of assessment instruments (e.g., formal, informal, traditional, and non-traditional methods).	Teacher utilizes a wide variety of assessment instruments (e.g., formal, informal, traditional, and non-traditional methods) and develops assessment instruments, as appropriate. Differentiation is based on both informal and formal assessment results.

Exhibit C8 (cont.): St. Mary's County Public Schools

Instruction					
		Performance Rating			
		Ineffective	**Developing**	**Effective**	**Highly Effective**
Uses Assessments in Instruction	**Feedback to Students**	Feedback is either missing, of overall poor quality, or untimely.	Feedback is inconsistent in quality. Some elements of high quality (i.e., it is timely, specific to the learning outcomes, and clear) are present; others are not.	Feedback is consistently of high quality (i.e., it is timely, specific to the learning outcomes, and clear). Teacher usually helps students see how they can use feedback in their learning.	Feedback is consistently of high quality (i.e., it is timely, specific to the learning outcomes, and clear), Teacher consistently assists and guides students to use feedback to enhance and demonstrate advanced levels of learning.
	Interprets Results and Uses Data to Make Decisions	Instructional decisions are not based upon assessment data.	There is limited use of assessment data when making instructional decisions.	Instructional decisions are consistently based upon analysis of assessment data.	Instructional decisions are consistently based upon analysis of assessment data. Individual student results.
Demonstrates Flexibility and Responsiveness	**Lesson/Unit Adjustment**	Teacher adheres rigidly to instructional plans, even when an adjustment will clearly improve a lesson.	Teacher makes conscious attempts to adjust lessons to meet student needs, with mixed results.	Teacher consciously makes minor adjustments to lessons and instructional plans to meet student needs, and such adjustments occur smoothly.	Teacher skillfully makes adjustment to lesson/units to meet student needs.
	Response to Students	Teacher does not respond to students' questions, interests, or level of proficiency.	Teacher attempts to accommodate students' questions, interests, or level of proficiency, but frequently loses the instructional focus.	Teacher successfully accommodates students' questions, interests, or level of proficiency on a regular basis while accomplishing instructional goals.	Teacher seizes opportunities to enhance learning, building on spontaneous events.
	Persistence	When a student has difficulty learning, the teacher either gives up or blames the student, does not persist with the students, blames external factors, or makes excuses about the environment for the student's lack of success.	Teacher accepts responsibility for the success of all students and is developing a repertoire of instructional strategies.	Teacher persists in seeking approaches for students who need help, and possesses a good repertoire of strategies.	Teacher persists in seeking effective approaches for students who need help, using an extensive repertoire of strategies and soliciting additional resources from the school.

Appendix D: Interview Questions

District Administrators

Site visitors posed the following questions to at least one administrator interviewed in every district included in the study:

- What is the purpose of the teacher evaluation system in this district?

- How does the teacher evaluation system consider student learning gains, instructional practices, and other measures in assessing performance? Does the district administer additional standardized tests for the purpose of measuring teacher effects on student achievement for the teacher evaluation system, or is the district using existing tests (with no additional test administrations)? If additional tests are administered only for the purposes of the teacher evaluation, please explain which tests are administered, to whom, and how frequently those tests are administered during the school year.

- How was the teacher evaluation system developed and what stakeholders were included in the process? What groups (e.g., teachers, union leaders, building administrators, district staff/administrators, school board members, parents, business representatives) were involved in planning the new teacher evaluation system? In your opinion, were there any groups or individuals who were not involved who should have been involved?

- How, if at all, is the system tailored for teachers serving in different roles (e.g., special education, ELL, etc.)?

- Once the teacher evaluation system was developed, how many weeks, months, or years were required to fully implement the system? Was the system pilot tested? Why/why not? If pilot tested, how long did the pilot testing take to complete?

- How, if at all, has the district tested the validity of the teacher evaluation system? That is, to what extent has the system's capacity to validly measure teacher performance been tested? In addition, to what extent has the district tested the consistency of the ratings that principals and other observers (internal and/or external) assign teachers based on their observations of teachers' classroom practices? Are observers shared across schools? How was the number of required classroom observations determined for purposes of evaluating teacher effectiveness? What are the challenges associated with scheduling observations?

- Who is responsible for collecting information on a teacher's classroom practice and other performance indicators? How are these people prepared for their responsibilities? Are training materials provided to observers, such as handbooks or other reference materials? Do the reference materials include examples of the types of practices observers should look for? Are observers certified as having received the required training? What is required for certification (e.g., participation in specific observer training, written exam, inter-rater reliability testing, etc.)? Are observer scores reviewed for within-school and across-school irregularities/inconsistencies? What happens when scoring inconsistencies/ irregularities are identified for an individual observer? In your opinion, has the preparation been adequate? If not, what were the gaps in the preparation?

- To what extent are professional development and other kinds of assistance available to help teachers address areas of concern that may be identified through the evaluation system?

- How, if at all, has the teacher evaluation system affected teacher practice?

- To what extent have any changes (i.e., organizational, policy, programmatic, personnel, etc.) occurred in the school district that are attributable to the teacher evaluation system?

- What resources are required to develop and maintain the teacher evaluation system? Does the district have any contracts with outside organizations for analysis of teacher evaluation data or for assistance with the administration of the evaluation system? If so, what are the purposes of these contracts and what are the annual costs of these contracts? What district-level staff are involved in the operation and administration of the teacher evaluation system? What are their responsibilities? Did the district create any new positions that are devoted largely or entirely for the new teacher evaluation system? If so, what are these positions? Were these positions added to the total number of positions in the district, or were existing positions converted into these positions?

- What type of staff are directly involved in a teacher's annual evaluation in the current system and in the previous system? Please estimate the amount of time (in hours or minutes) that each of these staff spends on one typical teacher's evaluation each year. To what extent do estimates vary for elementary versus secondary school teachers?

- How and to what extent, if at all, does the district use the results of teacher evaluation to make changes in the deployment of teachers across the district or for other purposes?

- What advice would [an administrator] offer a district trying to develop a teacher evaluation system?

Principals and Teachers

Site visitors posed the following questions to each principal and teacher interviewed in every district included in the study:

- What is the purpose of the teacher evaluation system in this district?

- How does the teacher evaluation system assess instructional practice, teacher planning and preparation, and/or other dimensions of professional practice? What are the questions and concerns, if any, about this component of the teacher evaluation system?

- Who is responsible for collecting information on a teacher's classroom practice and other performance indicators? To your knowledge, how are these people prepared for their responsibilities? In your opinion, has the preparation been adequate? If not, what were the gaps in the preparation?

- How does the teacher evaluation system use gains in student achievement as a factor in rating teacher performance? What are the questions and concerns, if any, about this component of the teacher evaluation system?

- In your opinion, is the overall design of the teacher evaluation system appropriate for all types of teachers (e.g., regular education teachers, special education teachers, teachers of EL learners, teachers in core academic subjects, teachers in other subject areas, elementary school teachers, middle school teachers, high school teachers)?

- What options, if any, do teachers have to appeal ratings of their effectiveness?

- In a normal week, approximately how much of your time is devoted to responsibilities associated with the teacher evaluation system and process?

- What changes and accommodations, if any, have you made in other areas of your job to be able to devote time to responsibilities associated with the teacher evaluation system?

- What challenges/problems, if any, have you encountered in using the results of the teacher evaluation system to inform decisions about the following:

  - Teacher compensation
  - Retention or release of individual teachers
  - Promotion of teachers to positions of increased authority or status
  - Teacher compensation
  - Planning and providing professional development
  - Reporting to parents or other stakeholders on teacher quality in your school

- Based on your experience and information from teachers, what, if anything, should the district have done differently regarding the design and implementation of its new teacher evaluation system?

- What other changes in the district and/or your school organization and culture, if any, have you observed that you believe to be attributable to the teacher evaluation system?

Teacher Focus Groups

Interviewers posed the following questions during focus group discussions with teachers in each of the three districts in which focus groups were conducted:

- In your opinion, what is the goal and purpose of the district's teacher evaluation system?
- Is the teacher evaluation system based on a specific framework or model of good teaching? If so, how appropriate is the framework for assignments like yours?
- How does the teacher evaluation system consider student learning gains in assessing your performance? Do you have any questions or concerns about this part of the process as it applies to your work as a teacher?
- How does the teacher evaluation system assess your instructional practices? In what other areas does the system assess your performance? Do you have questions or concerns about these parts of the process?
- Did you receive feedback from classroom observations or other parts of the evaluation process? If so, when do you receive the feedback? How do you receive it and to what extent is the feedback timely and useful? What questions and concerns do you have about this part of the process?
- Are professional development and other kinds of assistance readily available to help teachers address weaknesses or problems that may be identified in the evaluation process? If so, briefly describe the kinds of support that are available. In your opinion, how useful are they to teachers in your area?
- Are you teaching any differently as a result of going through the evaluation process?
- In your opinion, have there been changes in your school that are attributable to the new teacher evaluation system? If so, please briefly describe the changes.
- Overall, do you consider the teacher evaluation system to be fair and helpful to teachers and to the students? Why or why not?
- Looking at your experiences, what changes or improvements would you suggest?

State Administrators

Interviewers posed the following questions to each state administrator interviewed:

- What is the purpose of the teacher evaluation system in this district?

- What statutes and/or regulations, if any, are in place to guide and support the design of local teacher evaluation systems? Based on your knowledge of the statutes and regulations, what components of teacher evaluation systems are required of all systems and where are there areas of flexibility? In your opinion, what factors and/or issues led to the state statutes and/or regulations on teacher evaluation in this state?

- How, if at all, has the state tested the validity of the teacher evaluation system? That is, to what extent has the system's capacity to validly measure teacher performance been tested?

- In your opinion, is the overall design of the teacher evaluation system appropriate for all types of teachers? [If not] For which groups of teachers is the system less appropriate? What are the specific reasons why it is less appropriate? To your knowledge, are there any efforts underway to address this issue(s)?

- Does the state education agency require school districts to submit plans for new teacher evaluation systems? [If so] What are district plans expected to include (e.g., an implementation timeline, strategies for communicating about key components of the new teacher evaluation system, plans for developing guides, rubrics, and other artifacts)? How does the state review and provide feedback on the plans?

- Based on your experience and information available from districts, what are the key lessons learned from local implementation of the state model/requirements for teacher evaluation systems?

- Looking ahead, do you anticipate changes in the state model/requirements/ for local teacher evaluation systems? What specific changes do you anticipate? When will the changes be made?

Appendix E: Reference List of Extant Documents, by District

Austin

Austin Independent School District Department of Program Evaluation (April 2010). *AISD REACH Year 2 Evaluation Report II, 2008–2009.* Austin, TX: Author.

Austin Independent School District Department of Program Evaluation (February 2009). *Strategic Compensation Initiative REACH Pilot 2007–2008 Evaluation Report.* Austin, TX: Author.

Austin Independent School District Department of Research and Evaluation (Fall 2011). *AISD REACH Program Update 2010–2011: Participant Feedback.* Austin, TX: Author.

Austin Independent School District Department of Research and Evaluation (Fall 2011). *AISD REACH Program Update 2010–2011: Texas Assessment of Knowledge and Skills Growth and Student Learning Objectives.* Austin, TX: Author.

Austin Independent School District Department of Research and Evaluation (Fall 2011). *AISD REACH Program Update 2010–2011: Professional Development Units.* Austin, TX: Author.

Austin Independent School District. *Teacher Evaluation System: AISD REACH 2011–2012.* Austin, TX: Author.

Burns, S., Gardner, C., Meeuwsen, J. (August 2009). *An Interim Evaluation of Teacher and Principal Experiences during the Pilot Phase of AISD REACH.* Nashville, TN: Vanderbilt Peabody College.

District of Columbia

District of Columbia Public Schools (2012). *IMPACT: The District of Columbia Public Schools Effectiveness Assessment System for School-Based Personnel 2012–2013, Group 2: Grades 1–12 General Education Teachers without Individual Value-Added Student Achievement Data.* Washington, DC: Author.

District of Columbia Public Schools (2012). *LIFT: Leadership Initiative for Teachers 2012–2013.* Washington, DC: Author.

District of Columbia Public Schools (2012). *IMPACT: The District of Columbia Public Schools Effectiveness Assessment System for School-Based Personnel 2012–2013, Group 1: General Education Teachers with Individual Value-Added Student Achievement Data.* Washington, DC: Author.

District of Columbia Public Schools. (2012). *Key Changes to IMPACT for 2012–2013.* Washington, DC: Author.

Hamilton County

State Collaborative on Reforming Education. *Supporting effective teaching in Tennessee: Listening and gathering feedback on Tennessee's teacher evaluations.* Nashville, TN: Author.

Tennessee Department of Education (2012). *Teacher Evaluation in Tennessee: A Report on Year 1 Implementation.* Nashville, TN: Author.

Tennessee Department of Education. *Tennessee Instructional Leadership Standards (TILS) Evaluation Rubric.* Nashville, TN: Author.

Tennessee Department of Education (2011). *Overview of School-Level TVAAS Composites (Non-tested educators).* Nashville, TN: Author.

Tennessee Department of Education (2011). *Achievement Measures Process Summary*. Nashville, TN: Author.

Harrison

Colorado State Council for Educator Effectiveness (April 2011). *Report and Recommendations. Submitted to the Colorado State Board of Education.*

Harrison School District 2 (2010). *Professional Education Evaluation System Rules and Procedures.* Colorado Springs, CO: Author.

Miles, M. (January 2011). *Teacher Compensation Based on Effectiveness: The Harrison Pay-for-Performance Plan.* Colorado Springs, CO: Author.

Hillsborough County

Hillsborough County Public Schools (January 2013). *Empowering Effective Teachers Teacher/Admin Newsletter.* Tampa, FL: Author.

Hillsborough County Public Schools (July 2012). *Classroom Teacher Evaluation Instrument.* Tampa, FL: Author.

Hillsborough County Public Schools. *An Explanation of the Informal Observation Cycle.* Tampa, FL: Author.

Hillsborough County Public Schools. *Classroom Observation Summary.* Tampa, FL: Author.

Hillsborough County Public Schools. *Classroom Teacher Required Observations 2012–2013.* Tampa, FL: Author.

Hillsborough County Public Schools. *Conference Discussion Guide.* Tampa, FL: Author.

Hillsborough County Public Schools. *Peer Evaluation Protocols.* Tampa, FL: Author.

Hillsborough County Public Schools. *Pre-Observation Conference Questions.* Tampa, FL: Author.

Pittsburgh

Battelle for Kids (2011). *Value-Added 101: Introduction to Value-Added Measures.* Columbus, OH: Author.

Battelle for Kids (2011). *Value-Added 201 Part A: Deepening Your Understanding*. Columbus, OH: Author.

Battelle for Kids (2012). *Value-Added 301: Understanding Teacher Value-Added Reports.* Columbus, OH: Author.

Halloran, P. (Spring 2010). *RISE-ing to the occasion*. The Pittsburgh Educator.

Johnson, M., Lipscomb, S., Gill, B., Booker, K, Bruch, J. (February 2012). *Value-Added Models for the Pittsburgh Public Schools.* Cambridge, MA: Mathematica Policy Research.

Pittsburg Public Schools (August 2011). *Pittsburgh RISE: Research-based, Inclusive System of Evaluation, Pittsburgh Standards of Effective Teaching.* Pittsburgh, PA: Author.

Pittsburg Public Schools (July 2009). *Empowering Effective Teachers in the Pittsburgh Public Schools.* Pittsburgh, PA: Author.

Pittsburg Public Schools. *Building school leadership that drives student achievement.* Pittsburgh, PA: Author.

Pittsburgh Federation of Teachers. *Teachers/Professionals Tentative Collective Bargaining Agreement between the Pittsburgh Federation of Teachers and the Pittsburgh Board of Public Education. July 1, 2010 through June 30, 2015.* Pittsburgh, PA: Author.

Pittsburgh Public School and Pittsburgh Federation of Teachers. *Domain 5: Teaching and Professional Excellence.* Pittsburgh, PA: Author.

Pittsburgh Public School and Pittsburgh Federation of Teachers. *The Empowering Effective Teachers Plan: Introduction to the Research-based Inclusive System of Evaluation.* Pittsburgh, PA: Author.

Pittsburgh Public School (August 2013). *Update on Effective Teaching: Report to Board of Directors Education Committee.* Pittsburgh, PA: Author.

Pittsburgh Public Schools. *RISE Teacher Evaluation Process 2011–2012.* Pittsburgh, PA: Author.

Scott, A. and Correnti, R. (August 2013). *Pittsburgh's new teacher improvement system: helping teachers help students learn.* Pittsburgh, PA: A+ Schools.

Plattsburgh

New York State Education Department (February 2012). *Guidance on New York State's Annual Professional Performance Review Law and Regulations.* Albany, NY: Author.

New York State Education Department (March 2012). *Guidance on the New York State District Wide Growth Goal-Setting Process: Student Learning Objectives.* Albany, NY: Author.

New York State Union of Teachers. *Integrated Teacher Evaluation and Development System: A Plan for Teacher Evaluation and Development.* Albany, NY: Author.

New York State Union of Teachers (August 2012). *NYSUT's Teacher Practice Rubric Aligned with the New York State Teaching Standards.* Albany, NY: Author.

New York State Union of Teachers (July 2011). *NYSUT's Teacher Practice Rubric Aligned with the New York State Teaching Standards.* Albany, NY: Author.

New York State Union of Teachers. *Teacher Evaluation and Development Implementation Guide.* Albany, NY: Author.

Plattsburgh City School District (January 2013). *Annual Professional Performance Review Plan 2012–2013.* Retrieved from http://usny.nysed.gov/rttt/teachers-leaders/plans/docs/plattsburgh-appr-plan.pdf on June 10, 2013.

St. Mary's County

Grasmick, N and Weller, E. (June 2011). *Maryland Council for Educator Effectiveness Initial Recommendations: Statewide Educator Evaluation System.* Submitted to Governor Martin O'Malley, the Maryland General Assembly, and the Maryland State Board of Education.

Martirano, M. (2012). *St. Mary's County Public Schools Teacher Performance Assessment System.* Leonardtown, MD: St. Mary's County Public Schools.

Martirano, M., Dudderar, L., Maher, J. Smith, S. and Greely, R. (April 2012). *Creating a Balanced Evaluation System: Teacher Performance Assessment System.* Leonardtown, MD: St. Mary's County Public Schools.

Sadusky, B, and Weller, E. (June 2012). *Second interim report of the Maryland Council for Educator Effectiveness.* Submitted to Governor Martin O'Malley, the Maryland General Assembly, and the Maryland State Board of Education.

DEPARTMENT OF EDUCATION

UNITED STATES OF AMERICA

The Department of Education's mission is to promote
student achievement and preparation for global
competitiveness by fostering educational excellence and
ensuring equal access.
www.ed.gov