New Jersey Teacher Evaluation, RU-GSE External Assessment, Year 1 Report

By William A. Firestone, Cynthia L. Blitz, Drew H. Gitomer
Dessi Kirova, Anton Shcherbakov, and Timothy I. Nordon

January, 2013

Rutgers University-Graduate School of Education
10 Seminary Place
New Brunswick, NJ 08901

Table of Contents

Executive Summary

After the New Jersey Educator Effectiveness Task Force published its plan to revise teacher evaluation in the state, the New Jersey Department of Education (NJ DOE) began a pilot teacher evaluation program to help develop more effective teacher evaluation practices. NJDOE asked the Rutgers Graduate School of Education (RU GSE) to conduct an independent assessment to assess the first year of the pilot. This assessment asked three questions:

1. What was the status of implementation of the classroom observation portion of the teacher evaluation pilot program? [1]

2. What were teachers' and administrators' perceptions of teacher evaluation and the classroom observation rubrics being implemented?

3. What barriers and facilitators affected how the program was initially implemented?

This document briefly summarizes the methodology, findings, and conclusions that are described in more depth in the full report.

Methodology

Ten districts received notification of awards to begin their pilot teacher evaluation work in August, 2011. These districts varied substantially in size, location within the state, and socio-economic status. The evaluation began in January, 2012 and uses surveys of teachers and administrators, site visits to six districts, and an analysis of the districts' own teacher observation data to describe first-year status.

Teachers in all pilot districts were included in a March, 2012 survey. With the assistance of district leaders and NJDOE, a 59% response rate was achieved. Ultimately, 2,495 teachers in all grades completed the survey. The survey asked teachers to report on their experiences and perceptions of the teacher observation process during the spring of the first year.

Administrators were surveyed twice, once in spring and then in late summer. The response rate for the first survey was 60% (164 responses). The summer survey had a 54% response rate with 154 completed surveys. These surveys asked administrators about the number and quality of observations completed, in addition to the administrators' experiences and perceptions.

Site visits were made to six of the ten districts that varied in size, socioeconomic status, and location within the state and the teacher observation rubrics they used. In general, the assessment team interviewed the superintendent, the director of the teacher evaluation project, people who led professional development and data analysis, the teachers' association representative, and two

[1] The focus was on classroom observation because districts put the most time into that portion of the evaluation system and because student growth data did not become available to districts until January 2013.

principals. Two teacher focus groups were also held: one with teachers of tested subjects and one with teachers in untested subjects.[2] These interviews and focus groups covered many of the same topics as the surveys, but they provided an opportunity to learn about the history and process of pilot implementation in each district, participants' perceptions in their own words, and factors that were not anticipated when the study was designed.

Over the summer of 2012, districts provided computerized files of their observation data to NJDOE, which shared this information with the assessment team in a confidential form with personal identifiers removed. Although the quality of the data and amount of missing data varied, these files typically included information on the date of each observation, an identifier for the observer, the grade and/or subject of the teacher, and the scores recorded from the observation.

This study should help New Jersey educators and policy makers prepare for the broad implementation of teacher evaluation in the state, but the study has some limitations. First, while diverse, the districts sampled are not representative. Second, many issues that arose during the first year may not occur later because of what the pilot districts, NJ DOE, and the providers of observation services learned from the pilot itself.

Findings

The first year of the pilot program was a learning year for the districts and for the NJDOE. After receiving notification that they had won their pilot grants just before the school year began, districts had to work quickly to organize. Each district had to set up what are now called District Evaluation Advisory Committees (DEAC). They also needed to select a teacher observation rubric, from a list of teacher evaluation frameworks listed in the Notice of Grant Opportunity (NGO) (New Jersey Educator Effectiveness Task Force, 2011a) or to adopt any observation instrument that was research based and consistent with criteria established in the NGO. Districts were responsible for training both teachers and observers on how to use the new evaluation rubric and the data management tool through which observation information would be collected, stored, analyzed, and reported. In most cases, this work was largely done by December 2011, allowing observations to begin by January, 2012.

Given the challenges of the late startup and the need to learn new procedures and observation criteria, it is not surprising that while a substantial number of observations were completed, some districts struggled in completing the required number of observations. Depending on the district[3], between 60% and 91% of the teachers were observed using the new evaluation rubric.

[2] Tested subjects were supposed to be those where growth data would eventually be available, and untested subjects were all others. Since only 4th through 8th grade teachers would have growth data, this was a small group of teachers. However, since that distinction was not well understood, this group often included teachers of core subjects even if they would not actually receive growth data.

[3] Since the state statistics counted nurses and counselors as teachers and they were not observed, it is possible that these percentages underestimate the number of actual teacher observations conducted.

The assessment team looked at three other characteristics of the observation data. Although multiple observations were not required in the first year of the pilot, this kind of information is important for assessment of the quality of observation data generated by the districts: consistency among multiple observations is an adequate way to assess the reliability and stability of observation scores over time. In seven districts, most teachers were only observed once. Teachers were observed two to three times each in two other districts. The number of double observations--i.e., times when two observers saw the same lesson--was also indicated. The survey data suggests that very few double observations were conducted and that the numbers varied across districts. Overall, the distribution of observation scores in most districts tended to be somewhat higher on the observation scales than is the case when the evaluation rubrics are used by highly trained, experienced observers in research studies. However, this issue occurs more in some districts than in others. These data suggest that while New Jersey observers may need more training to become calibrated with expert observers nationally, some districts are already making better progress than others.

Perceptions

The survey data provided information on three aspects of district staff's perceptions of the teacher evaluation rubrics. The first is their view of its quality. Generally, teachers and administrators used the same criteria for assessing these rubrics. They looked for evaluation rubrics that were accurate, fair, provided useful feedback to help teachers improve their practice, and that could be used for personnel decisions. However, administrators generally had a more positive view of these evaluation rubrics than did teachers. For instance, 74% of administrators agreed that the evaluation rubrics assessed teachers accurately, as did 32% of the teachers. Similarly, 75% of administrators agreed that the rubrics generated information that provided useful individual feedback or guidance for professional development, as did 53% of teachers. These differences may not be surprising given that teachers are being evaluated and administrators are not, but the patterns persist across many items. On the other hand, districts differ in how their programs are perceived. For instance the percent of administrators who think the teacher evaluation rubrics are accurate ranges from 38% to 100% and the percent of teachers who agree on this point ranges from 20% to 52%, indicating that districts have quite different views of their teacher evaluation rubrics.

When we examined how district staff perceived the adequacy of administrators as observers, we found similar patterns except that more teachers acknowledge the quality of their evaluators than of the evaluation rubrics. For instance 94% of administrators believe they have the knowledge and competencies to appraise teachers as do 54% of teachers. This is an item where there is notable variation across districts. Administrators were more positive about the overall effects of the teacher evaluation rubrics than were teachers although they did not rate the broad effects as strongly as the quality of the rubrics or their ability to assess. When asked about the effects of the evaluation rubric on "your professional development", "collaboration with others", and "your school", 77 to 79% of administrators reported positive effects. On the same items, 32 to 42% of the surveyed teachers reported positive effects.

Barriers and Facilitators

The assessment team explored four possible factors that might affect implementation of the teacher evaluation rubric: time, training, the data management tool, and resistance. The significant time concern was that 90% of the administrators surveyed reported that they were spending more time conducting observations and entering observation data than they had previously. While there is always room to question whether individuals exaggerate the time demands when they are asked to do something new, there are two reasons to give these findings some credence. First, in interviews administrators described new time demands that were now being made of them, including doing more pre-observation conferences with teachers, doing longer post-observation conferences, and providing more detailed records of observations. Second, teachers in several schools reported things that were not getting done by administrators once the observations started that had been getting done before, including attending to discipline issues. In fact, teachers were notably sympathetic to the time pressures on administrators. Some portion of time demands came from learning to use the new rubrics, a one-time issue that would end once administrators understood the new rubrics. However, administrators will most likely have to do substantially more observations in subsequent years than they did in the first. Still, districts where the superintendent made completing teacher observations a priority managed to conduct more observations than did other districts.

Administrators reported that the training they received over the year was very useful in helping them understand the evaluation rubrics and what they were supposed to do. Teachers were less convinced. One important reason is that administrators received substantially more training on the new rubrics than teachers--four times as many teachers as administrators reported receiving less than eight hours of training on the rubrics--largely because it was administrators who would conduct the observations. As a result, administrators knew better what to expect of the new evaluation rubrics. Two promising practices helped educators benefit from the training provided. First, some districts provided extra training to a cadre of teachers who helped their peers learn and understand the system better. Second, after the initial training, which was sometimes rushed and excessively didactic, administrators especially benefited from opportunities to collectively review real observation data. These concrete experiences through regular meetings and walk-through like events allowed administrators to process what they learned in ways that most teachers did not have.

All of the teacher evaluation rubrics are accompanied by computerized data management tools. These often include a tablet-based element for recording observation data in the classroom, but they always include some means to record data, generate observation reports, share them with teachers, store the data in a central storage facility, and run analyses to identify patterns. Administrators in particular report that these tools are essential to the operation of the teaching practice evaluation instruments. These tools created their own learning issues that initially slowed recording, led to lost reports, and the like.Most of these problems were resolved by the end of the first year although most administrators reported more facility at inputting data and using evaluation rubrics to provide feedback to teachers than in analyzing data.

Finally, in light of the differences between administrators' and teachers' assessments of the evaluation rubrics and some reports that teachers were more guarded in their discussions with administrators than in the past, the assessment team explored possible sources of teacher resistance

to the new evaluation rubrics. Three distinct themes emerged from the qualitative data. First, many teachers described the evaluation rubrics as subjective. In particular, some described significant discrepancies between observations that they or their colleagues had received at different times. Second, in some districts, teachers were warned that the scale for assessing teachers would change from the past leading to fewer distinguished evaluation ratings. To many teachers, this anticipated change seemed like an arbitrary quota. Finally, teachers described "inappropriate" observation criteria, especially those that were good practice but that would not be expected in every lesson or for teachers of every subject area. They feared they would be graded down without those criteria. Although teachers had reservations about the observation process, there were comments to the effect that the quality of conversations about teaching and learning were getting better in the pilot districts as a result of the teacher observation programs. These observations were echoed even more strongly by administrators.

Summary

In sum, the pilot districts have accomplished a great deal in the first year. All observation systems were up and running, and districts have moved to refining them in the second year. The fact that every one of them succeeded in implementing an observation system is a significant achievement. Because the first year (2011-2012) of the Teacher Evaluation pilot was primarily focused on teacher observation, this report reflects only that particular implementation aspect.

At the same time, the evaluation has identified issues to address in the future. One is the time issue. The need to ensure that every teacher receives the required number of observations and that districts complete enough dual observations to assess agreement among raters may put a significant burden on administrators who already report substantial strains in completing their full range of tasks. Another is the challenge of generating valid, reliable observations. The assessment team cannot fully assess the magnitude of this issue as not enough data is yet available, but indications are that more work will be needed. Finally, we note the need for more and better communication with teachers so they better understand their teacher evaluation systems. We also note that some districts provide models of especially helpful ways to provide that communication through the use of teacher leaders and through embedded, concrete training using real observation data. As the program moves forward, a useful strategy will be to take advantage of the variation among districts in how they implement their teacher evaluation rubrics.

Finally, the team notes two issues that would benefit from future research. First, it will be useful to better understand variation among districts. Do the differences noted stem from differences among evaluation rubrics, in which case, it may be useful to try to identify one or more that are most promising? Do they stem from resource differences that districts have to allocate to evaluation (e.g., proportionately larger number of administrators)? Finally, do differences result from district implementation strategies such that districts that have done better could provide advice to those where indicators suggest that more progress is possible? The other issue is the quality of observation data. New Jersey needs better information on inter-rater reliability, stability of ratings, and the distribution of ratings within and across districts and evaluation rubrics to better understand how well the system is working and what improvements are needed.

Introduction

In March 2011, the New Jersey Educator Effectiveness Task Force put forth a plan for teacher evaluation and recommended a new State Educator Evaluation System. This evaluation system aims to be a major tool for improving student achievement and promoting equity in New Jersey by providing a firmer basis for awarding teacher tenure and setting compensation levels (New Jersey Educator Effectiveness Task Force, 2011a). The New Jersey teacher evaluation effort is part of a much broader national effort. More than a decade of federal and state-level legislation mandating teacher assessment and accountability policies, including Race to the Top competition requirements, provides the backdrop for the current debate over teacher quality and preparation. While there is broad agreement that high quality, effective teaching contributes to student learning, assessing teachers' effectiveness and the quality of their teaching has proven to be very challenging. Researchers, politicians, and practitioners are grappling with fundamental questions about what exactly constitutes effective teaching, which aspects of teaching are most likely to improve student learning, how to best measure teachers' effectiveness, and how assessment data should be used and for what purposes. Not surprisingly, for most of the recent past, teacher evaluation has not been terribly systematic. The use of student data to assess teachers only began to be seriously considered in the late 1990s (Wright & Sanders, 1997). In the past, personnel decisions were often linked to formal credentials—degrees, credits, and years of experience—because people had little faith in the objectivity of most observations of teachers (Podgursky & Springer, 2007).

During the 2011–2012 academic year, the New Jersey Department of Education (NJ DOE) launched the New Jersey Teacher Evaluation pilot program to help further the State Educator Evaluation System. This program was part of NJ DOE's commitment to "elevating the teaching profession, recognizing classroom excellence, and providing support to educators needing assistance." (New Jersey Department of Education, 2012, p. 5). The 2011-2012 Teacher Evaluation pilot program has two primary elements - measures of student growth and observation of classroom instruction. All school districts in New Jersey will be expected to implement the new educator evaluation program beginning in the 2013–2014 academic year.

To support this effort, the NJ DOE contracted with the Rutgers University Graduate School of Education (RU GSE) to conduct an external assessment of the Teacher Evaluation pilot program. The scope of this assessment included reporting on the implementation of new evaluation practices, documenting participants' perceptions of the pilot programs, and identifying factors that influenced the implementation process. The RU GSE assessment team collected data from various sources including administrator and teacher surveys, site visit interviews and focus group data, documents and artifacts collected from school districts, and teacher observation data, to help understand the implementation of the 2011-2012 Teacher Evaluation pilot program. This report summarizes the findings of the first-year evaluation of the pilot.

The RU GSE evaluation research of the pilot focused on three major questions:

1. What was the status of classroom observation implementation, including the selection of particular observation protocols, training, administration, data management, and evaluations of teachers?

2. What were the perceptions of district teachers and administrators with respect to the new teacher evaluation and more specifically, to the new evaluation rubric? To what extent were perceptions consistent across teachers and administrators?

3. What barriers and facilitators affected the successful implementation of the program?

The answers to each of these questions are discussed in the following sections. A critical issue that could not be examined was the relationship between measures of teacher effectiveness based on classroom observation and student achievement because student growth scores were not available.

It is important to recognize that the initial pilot is a critical step in developing an operational evaluation system in New Jersey. Due to the small size of the pilot and since much of the system continues to evolve, these findings should not be viewed as representative of all that is likely to be observed about teacher evaluation as the system becomes operational statewide in the 2013–2014 academic year.

Methodology

This assessment began when Rutgers Graduate School of Education and the New Jersey Department of Education finalized the Memorandum of Understanding on January 26, 2012. While the assessment team lacked direct access to the activities and decisions during the initial stages of the pilot or its development, the RU GSE team worked closely with the NJ DOE leadership to develop the assessment strategy within the overall scope of the project and to gain access to important sources of information once the external evaluation began.

Participating Districts

Ten school districts were selected by the NJDOE to participate in the first year of the pilot Teacher Evaluation, and they began their work in September 2011. Table 1 provides information on each district's location, district factor group, and student enrollment. The districts range in enrollment size and geographic distribution. Although we cannot consider the districts a representative sample of all districts in the state, they do represent a diverse sample of districts in the state. Three districts are from northern New Jersey, three are from southern New Jersey, and four are from central New Jersey. In addition, there is a mixture of low and high socioeconomic school districts represented in the pilot sample as indicated by the distribution from A to GH district factor grouping (DFG) in Table 1. The three lowest DFG categories make up 31% of the state population, and they make up 30% of the pilot districts. However, the middle three DFG categories make up 46% of the state, and represent 70% of the sample districts. Unfortunately, there are no districts from the two highest categories (I and J), which make up 24% of the state population. In terms of enrollment, we captured a variety of districts ranging from very small and suburban (e.g., Alexandria) to rural and mid-sized (e.g., Pemberton) to very large and urban (e.g., Elizabeth).

Table 1 - *Participating School Districts in Pilot Teacher Evaluation*

District	County	Location	2000 DFG[4]	Enrollment
Alexandria	Hunterdon	Central	GH	588
Bergenfield	Bergen	Northern	FG	3329
Elizabeth	Union	Northern	A	21734
Monroe	Middlesex	Central	FG	5493
Ocean City	Cape May	Southern	DE	2046
Pemberton	Burlington	Central	B	4863
Red Bank	Monmouth	Central	CD	993
Secaucus	Hudson	Northern	DE	2146
West Deptford	Gloucester	Southern	DE	3165
Woodstown-Pilesgrove	Salem	Southern	FG	1663

Primary Sources of Information for the External Assessment

Data came from site visits, surveys, and the records of the actual teacher observations conducted.

Site visits. Site visits allowed for collection of interview and focus group data from key stakeholders in piloting districts, including both teachers and administrators.

Sample of school districts for site visits. Our assessment team conducted site visits in six of the ten participating school districts. Budgetary limitations did not allow visits to all ten districts. In selecting districts to visit, we attempted to secure adequate representation of districts from different regions (northern, central, and southern New Jersey), as well as those of different student enrollment sizes and DFG.

Targeted participants in piloting districts. We developed a list of key individuals in pilot districts who were directly involved with the implementation of the pilot Teacher Evaluation program. The list included the superintendent, the key instructional leader in charge of curriculum and instruction, the district Teacher Evaluation project director, the director of professional development, the director of student data, the president of the local teachers' association, and two principals (elementary and secondary). We worked with the pilot Teacher Evaluation program project directors to schedule site

[4] The NJDOE introduced the District Factor Grouping (DFG) system in 1975. This system provides a means of ranking school districts in NJ by their socioeconomic status with A districts being the poorest in the state and IJ districts being the wealthiest (NJDOE web site, http://www.state.nj.us/education/finance/sf/dfgdesc.shtml).

visits and for all but one district, we met with all individuals on the developed list. For one district, because of scheduling constraints, several administrators and the teachers' association president were unavailable. Interview times ranged from 20–60 minutes. In addition, we conducted two focus groups with teachers of tested subjects and teachers of untested subjects. Focus groups consisted of 4–10 teachers and lasted from 30–60 minutes.

Interview and focus group topics. Members of our assessment team conducted all interviews and facilitated all focus groups, according to the site visit guide protocols (please refer to Appendix A). These protocols guided each meeting; however, the interviewer/facilitator prompted the interviewees for further detail, elaborations and clarifications as needed. The interview protocols focused on the following topics, organized around the two types of data used to evaluate teachers:

1. The Teacher Evaluation Rubric

    a) Choice of framework for teacher evaluation
    b) Training on the new framework
    c) Collection of teacher observation data
    d) Quality of observations
    e) Impact of the pilot program on professional development, professional collaboration, and school culture

2. Student Assessment Data

    a) Tested subjects
    b) Untested subjects

The focus group protocol focused on the following items:

1. Teacher Observations

    a) Compare and contrast current observations with past
    b) Sources of knowledge about teacher observation data
    c) Teacher evaluation rubric and planning, supervision, and professional development
    d) Expertise of observers
    e) District collaboration

3. Student Assessment Data

    a) Tested subjects
    b) Untested subjects

Surveys. Our assessment team designed three online surveys—one targeted toward all teachers participating in the pilot Teacher Evaluation program and the other two toward all administrators in year 1 pilot districts. Survey questions for teachers and administrators overlapped a great deal, although certain questions were specific to each group. Preliminary work included reviewing prior

state evaluations of other teacher evaluation rubrics as well as testing individual survey questions and the entire surveys. Once we finalized the questions and response categories, surveys were created in Qualtrics, an online survey tool. Both the administrator and the teacher surveys were approved by Rutgers University's Institutional Research Board, ensuring that data collection efforts complied with the strict federal and University requirements for the protection of human subjects. The three surveys were successfully administered as a result of the ongoing efforts of our assessment team in cooperation with district administration and NJ DOE staff.

Administrator surveys. We worked carefully with the pilot Teacher Evaluation project directors in all ten pilot districts to identify the correct respondents (also known as sampling frame) for both administrator surveys. The response rate for the first survey was 60 % with 154 responses. The response rate for the second survey was 54% with 134 responses. The surveys asked administrators about the numbers and quality of the observations they completed, in addition to their experiences and perceptions with the implementation of the new teaching practice evaluation instruments in their districts.

Sample for administrator survey. For the administrator survey sample, we were able to compare administrators' self-reported demographic information to the demographic data provided to us by the NJ DOE regarding the level of education and school district positions of various administrators. The data aligns well with the state data on district staff with significant (>20%) discrepancies in three districts where response rates were low or the accuracy of state data was not clear. For example, in one district where only 11 of the 22 administrators responded to our survey, the self-reported percentage (37.5%) of administrators with a Master's degree conflicted with the state report (82%). In another district, we had an 80% response rate, but more respondents (3 out of 4) reported having Master's degrees than state data suggested (2 out of 4). Discrepancies may be linked to outdated state data or faulty self-reports. In terms of sampling of positions, we had significant success. Overall, we were able to survey 8 out of 10 superintendents and at least half of the principals in 9 out of 10 districts. Further, there was similar representation for other positions. In general, we are confident that we captured a representative sample of the population of the ten districts.

Administrator survey topics. We generated questions for the administrator surveys in order to reflect the main components of the implementation process: we asked about the choice of teacher evaluation framework, experiences with training, the number and quality of observations, and perceptions of the quality, fairness, usefulness, and ease of use of the selected teacher evaluation framework. Please refer to Appendix B for the second administrator survey.

Teacher survey. For the teacher survey, the overall response rate was 59%, with 2,495 responses. Only one district had a response rate of 36%, while all of the others had a greater than 50% response rate. While there is some uncertainty whether there is any systematic bias in the survey results because of non-respondents, we believe that we have captured teachers' general attitudes and perceptions of the pilot Teacher Evaluation program.

Sample for teacher survey. With our teacher survey, we compared the self-reported educational level and school level taught to the state's demographic data. Overall, we were able to secure a representative sample of the ten districts. There were no significant discrepancies (>20%) in terms of

educational attainment, but three districts had significant differences in school level taught. For example, in one district that had a 36% response rate, 37% of our sample reported being at the middle school level, while state data reported that only 14% of teachers in the district taught at the middle school level. Thus, middle school teachers may have been slightly over-represented in our sample of this district. However, closer examination of our data also revealed that some teachers reported teaching at both the elementary and middle school levels, which may have contributed to apparent over-representation in our sample because the state data only allowed teachers to affiliate with one school level.

Teacher survey topics. Our assessment team organized questions in the teacher survey to parallel the topics covered by the administrator surveys so that we could compare the information that is common to both populations. The questions followed the main components of the pilot implementation: we asked about the selected teacher evaluation framework, about training on the district-selected framework for teacher evaluation, and about their experiences with the pilot Teacher Evaluation program in general. Additionally, we asked participants about their perception of the quality, fairness, usefulness, and ease of use of the selected teacher evaluation framework. The teacher survey appears in Appendix C.

Observation Data. During the pilot year, all pilot districts were expected to use the district's selected teacher evaluation framework to review every teacher, using the following procedures described in the Notice of Grant Opportunity (NGO) (New Jersey Educator Effectiveness Task Force, 2011a):

> a. For non-tenured teachers, conduct a minimum of three formal observations (i.e., with pre- and post-conference input and feedback) for one instructional period or a minimum of 40 minutes;
>
> b. For tenured teachers, conduct a minimum of two formal observations[5] (i.e., with pre- and post-conference input and feedback) for one instructional period or a minimum of 40 minutes;
>
> c. Conduct a minimum of two informal observations[6] (i.e., without pre- and post-conferences) with feedback;
>
> d. Prepare one summative evaluation that results in a mutually-developed teacher professional development plan;

[5] The formal observation process includes a pre- and post-observation conference. The pre-observation conference with the teacher must be held prior to observing the teacher for the purpose of discussing the lesson plan and intended outcomes of the lesson. The results of the observation are discussed at the post-observation conference and provided to the teacher in a written observation report. (New Jersey Educator Effectiveness Task Force, 2011a, p. 33).

[6] The informal observation process can be accomplished through a number of methods including short classroom visits for a specific purpose, power walk-throughs, or a review of artifacts of teaching. In the informal observation process, it is not necessary to have a specific pre-observation conference. The results of the observation are discussed with the teacher in a written observation report with specific feedback (New Jersey Educator Effectiveness Task Force, 2011a, p. 33).

e. At least once per year, conduct teacher self-assessments of their own practice and compare with the evaluator's assessments to calibrate teachers' personal vision of effective practice;

f. Promote an environment for supportive and accurate feedback on teacher practice; and,

g. Provide teachers with professional learning experiences to support improvement in teacher practice.

Each district was expected to provide to the NJ DOE a record of all classroom observations that were conducted as part of the pilot. These data were given to the state and then shared with the RU GSE assessment team. Generally, but not always, the districts provided data that included the teacher observed (a confidential, unique identifier), the grade and subject taught, the date of the observation, the observer (a confidential, unique identifier), and a summary score for the observation. Table 2 presents the percentage of districts that provided these data.

Table 2 - *Percentage of Pilot School Districts Providing Various Types of data to the NJ DOE*

Type of Data Sent to the NJ DOE	Percentage of Districts Providing Data
Any Observation Data	90%
Date of Observation	40%
Observer Identifier	80%
Grade or Subject Taught	70%

Other documents and artifacts collected from pilots. Our assessment team considered documents and artifacts, especially materials from training, an important data source. We expected to gain understanding of the districts' local pilot implementation based on the information about the district-selected framework for teacher evaluation, training on the framework, timeline of progress made, as well as on the specifics of the district-developed assessments to be used to measure student growth. We asked pilot districts to provide the following list of documents and artifacts:

i) Proposal for pilot

ii) Progress report

iii) Documentation on the provider rubric (including manuals, forms, training materials, and district teacher observation "tools"

iv) Training materials (including locally-developed forms and locally-developed rules and agreements)

v) Tools and documents related to student growth scores (tested subjects)

vi) Tools and documents related to student growth scores (untested subjects)

Although some districts were able to provide all of the requested documents, the majority of districts were only able to share certain ones. Some of the more problematic documents were those directly related to teacher observation, stemming from concerns about intellectual property, privacy, and confidentiality. Other documents, such as those related to measurement of student growth, were unavailable.

Data Analysis

Analysis of collected data began early as results from the first administrator survey became available and continued as other data sources provided more information about the ongoing implementation of the Teacher Evaluation pilot program. The mix of various data sources – interviews with participants, focus group data, survey data, observation data and artifacts of the teacher evaluation process - allowed us to triangulate many of the results derived from these separate analyses and put together a clearer picture of the pilot Teacher Evaluation program implementation.

Interview and focus group data. Our assessment team recorded, transcribed, and uploaded interviews and focus group sessions into Dedoose, a qualitative and mixed-methods software program. We developed a multilevel coding scheme that, in addition to Dedoose descriptors, allowed us to interpret the qualitative data and identify emerging themes. Themes from the interviews were matched with the goals of the evaluation research and used to elaborate and provide greater detail to the results of the survey instruments.

Survey data. Data from both administrator surveys and from the teacher survey were exported from Qualtrics (the online instrument used for all survey administration) and were analyzed using IBM's SPSS Statistics software.

The above-described methodology allowed our assessment team to address the questions that guided the pilot Teacher Evaluation program in the 2011–2012 academic year. The study findings are presented in the following three main sections:

1. Implementation of the teacher evaluation systems: the rubric chosen, the quantity of observations completed, and the basis for assessing the quality of the teaching practice evaluation instruments.

2. Perceptions of the teacher observation rubric: how teachers and administrators assessed the adequacy of the teaching practice evaluation instruments, the extent to which teachers and administrators reached the same conclusion, and how far they had moved toward actually using the rubrics for improving teaching given those perceptions.

3. Barriers and facilitators of implementation: the time it takes to conduct observations, the training provided, the new data management tools, and how potential resistance to change influences the extent and quality of implementation.

The report ends by identifying three challenges to teacher evaluation that New Jersey will face as all districts are required to implement the new teacher evaluation requirements. In addition, this last section identifies future research needs.

Implementation

This section describes issues of implementation of the evaluation rubric in each district, specifically with respect to classroom observations. This process began when districts were notified that they had been awarded Teacher Evaluation pilot grants in August 2011. At this point, pilots had to form an internal district advisory committee consisting of teachers, administrators and community members. Only then could they select a teacher evaluation rubric and associated data management system. After the districts had selected their teacher evaluation rubrics, they still had to train both teachers and administrators on the new teaching practice evaluation instruments. These steps meant that most districts could not begin using the new observation models until January 2012. Moreover, some of the variability among districts regarding the observation data is due to the differences in the chosen teacher evaluation framework.

The required number of observations by the NGO (New Jersey Educator Effectiveness Task Force, 2011a) differed depending on tenure status: a minimum of three formal observations for non-tenured teachers and a minimum of two formal observations for tenured teachers in addition to two informal observations (for both tenured and non-tenured teachers). As a requirement, this represents a significant increase from past years and leads to three broad questions:

1. What proportion of eligible teachers was included in the classroom observation component of the Teacher Evaluation pilot?

2. How many times over the course of the school year were teachers observed?

3. To what extent were formal and informal observations used to satisfy evaluation requirements?

Finally, an underlying premise of educator evaluation is that the rubric ought to be able to differentiate among teachers in terms of their instructional quality. Further, the respective protocols should identify strengths and weaknesses of teachers in ways that reflect the state of practice. Traditional evaluation practice typically has not differentiated teachers and instead has considered almost all teachers as being very strong (Weisberg, Sexton, Mulhern, & Keeling, 2009). We ask the following questions:

1. To what extent is there evidence that the rubric differentiates among teachers within the system?

2. To what extent is there evidence that the protocols are being used as they were designed and to identify strengths and weaknesses of teachers?

Protocol Adoption

Each district had latitude to select among a set of state-approved protocols. Districts also had the option of selecting a data management system. Often the data management system and teaching practice evaluation instrument were bundled together. Table 3 details the observation protocol and data management system for each of the participating pilot districts.

Table 3 – *Districts' Selected Teacher Evaluation Framework & Data Management System*

District	Teacher Evaluation Framework	Data Management System
Alexandria	James Stronge	Oasys
Bergenfield	Danielson	Teachscape
Elizabeth	Danielson	iObservation
Monroe	Marzano	iObservation
Ocean City	Danielson	iObservation
Pemberton	Danielson	Teachscape
Red Bank	Danielson	Teachscape
Secaucus	Danielson	Teachscape
West Deptford	McREL	McREL
Woodstown-Pilesgrove	McREL	McREL

Each of the protocols address multiple facets of practice and guides observers in making a rating on each of a set of dimensions identified in the protocol. The protocols do vary in their particulars, although recent research has demonstrated strong correlations among protocols when judging the same lessons (Bill and Melinda Gates Foundation, 2012). A brief summary of the teacher evaluation frameworks used by pilot districts follows.

Purpose. All four teacher evaluation frameworks used during the first year of the pilot (Danielson, Marzano, McREL and James Stronge) have a dual purpose – one aspect is the use of the system for improvement of teacher accountability through teacher evaluation, and the second aspect is the frameworks' aim to improve teaching quality by offering targeted professional development for individual teachers, based on their performance. Whether both aspects were equally addressed in the first year of the pilot is part of our discussion of the data findings.

Research-based. All four teacher evaluation frameworks are developed to reflect research-based standards of teaching quality. The McREL evaluation instrument and accompanying process is based on elements of a 21st century education and a set of rigorous research-based standards (www.mcrel.org), as is the Danielson model, which is aligned to the Interstate Teacher Assessment and Support Consortium (INTASC) standards (Danielson, 2011). Marzano (2011) claims to have evidence of a causal link between the teaching characteristics observed in his model and increased student achievement. The James Stronge framework is based on seven practice-tested teacher performance standards (Stronge, 2006).

Framework Content. As can be seen in Table 4, all four teacher evaluation frameworks consist of multiple domains/standards designed to address all aspects of teaching.

Table 4 – *Danielson, Marzano, McREL and James Stronge Frameworks for Teacher Evaluation*

Danielson	Marzano	McREL	James Stronge
Domain 1. Planning and Preparation	Domain 1. Classroom Strategies and Behaviors	Standard 1: Teachers demonstrate leadership.	Standard 1: Professional Knowledge
Domain 2. Classroom Environment	Domain 2. Planning and Preparing	Standard 2: Teachers establish a respectful environment for a diverse population of students.	Standard 2: Instructional Planning
Domain 3. Instruction	Domain 3. Reflecting on Teaching	Standard 3: Teachers know the content they teach.	Standard 3: Instructional Delivery
Domain 4. Professional Responsibilities	Domain 4. Collegiality and Professionalism	Standard 4: Teachers facilitate learning for their students.	Standard 4: Assessment of/for Learning
		Standard 5: Teachers reflect on their practice.	Standard 5: Learning Environment
			Standard 6: Professionalism
			Standard 7: Student Progress

Observations

The information discussed below is based both on district observation data (shared with RU GSE by the NJ DOE) and selected responses from the survey.

What proportion of teachers was included in the classroom observation component of the evaluation pilot? Most, but not all, teachers in each district were evaluated under the new system. The proportion of teachers who were actually observed ranged from 60%–91%, with most districts having participation rates between 74% and 84% of eligible teachers. However, these percentages are estimates; they were calculated by dividing the number of teachers observed by the number of teachers in the district according to a NJ DOE dataset, which may have overestimated the population of teachers in the district. For example, close examination of NJ DOE's dataset revealed that individuals such as guidance counselors, school nurses, and other individuals were occasionally classified as teachers. These classifications may have artificially inflated the population of teachers in the district. Thus, the percentage of teachers observed may somewhat underestimate the proportion of teachers actually evaluated under the new rubric.

Table 5 – *Implementation of Various Components of Observation Rubric by District*

District[7]	Percent of Teachers Observed	Observations per Teacher (Median)	Raters per Teacher (Median)	All Dimensions Rated
1	74%	1	1	YES
3	84%	1	1	YES [8]
4	74%	1	1	NO
5	83%	1	1	NO
6	64%	1	1	NO
7	80%	2 or 3[9]	2	YES
8	60%	1	1	YES
9	86%	1	1	YES
10	91%	2 or 3[6]	2 or 3[6]	YES

How many times over the course of the school year were teachers observed? Table 5 displays the median number of observations carried out for each teacher. For seven of the nine districts, the median number of observations was one. Two out of nine districts were able to fulfill the extensive requirements of the evaluation pilot program.

It is important to recognize that districts were not asked to conduct observations which included multiple observers in the first year of the pilot, although doing so will be a requirement in the second and subsequent years. Most observations were completed in the pilot year by a single observer, though some districts did conduct observations with multiple observers. This precluded the assessment team from collecting relevant information about the reliability of scoring classroom observations, an essential component in monitoring the quality of observation data. The second year of the pilot includes the requirement that a portion of observations must be conducted using two observers.

To what extent were formal and informal observations used to satisfy evaluation requirements? The observation data records provided by districts were not uniformly clear on the nature of each observation recorded. However, administrators did respond in their surveys to several questions relevant to this issue. Survey responses were generally reported at the administrator level rather than by observation. Thus, we can see from the data reported in Table 6 that administrators did three times as many observations as informal walk-throughs. However, the surveys do not reveal the proportion of all of these observations that were walk-throughs.

[7] Districts were given code numbers; these numbers are not linked to previous tables where participant districts were laid out alphabetically

[8] In this district, some teachers were graded on up to 11 criteria on the Danielson (2007) framework, but only 8 were supported by their online data management tool. Most of the teachers were therefore graded on 8 criteria.

[9] 2 if tenured, 3 if non-tenured.

Table 6 – *Selected Survey Results About Observation Implementation From Second Administrator Survey*

Self-reported Data as per RU GSE Administrator Survey	August 2012
Number of observations completed (Median)	30
District range:	11–45
Number of walk-through observations completed (Median)	10
District range:	0–75
Hours to complete observation (Median)	3
District range:	2–5
Percentage of observations that included collection of artifacts (Mean)District range:	50% 21%–72%
Percentage of observations discussed with the teacher (Mean)	87%
District range:	75%–93%
Percentage of administrators who did not collect classroom artifacts during any observation	7%
District range:	0%–13%
Percent of administrators who collected artifacts every time	19%
District range:	6%–45%
Percentage of administrators who reported that they discussed all of their observations with the teacher for feedback	71%
District range:	43%–88%

Most observers did multiple observations. As shown in Table 7, the median number of ratings per observer in the observation data ranged from 5 in one district to a high of 56 in another. For the other 6 districts, the median number of ratings per observer ranged between 13 and 33. The number of self-reported observations was not always consistent with the number of observations that were included in the respective district's data. Often, self-reported frequencies of observations were greater than was apparent in the district observation data. Whether this was due to false perceptions by administrators or other potential sources of confusion is not clear.

Table 7 – *Median Number of Observations per Rater by District*

District	Survey Self-Report	Observation Data
1	40	13
2	11	N/A
3	35	N/A
4	22	30
5	36	19
6	28	12
7	45	56
8	11	5
9	28	15
10	27	33

To what extent is there evidence that the rubric differentiates among teachers within the system? Even though different teacher evaluation rubrics were used, most used a 4-point scale that was anchored at the high end by terms such as *highly proficient* and *distinguished*. Scale points of 1 were anchored with terms such as *unsatisfactory*. In order to communicate the distribution of scores, the proportions of scores are rated from 1 to 4 in Table 8. Please note that although 9 districts provided some observation data, only 7 districts included actual observation scores.
Overall, there was a significant range in how districts utilized the rubric. For example, in one district, 60% of teachers received the highest score, while in other districts, only 6% received the top score. However, the modal score was 3 in all but one district, indicating that the vast majority of teachers were judged using a term such as *proficient*. Few teachers in most districts received a score of 2 or below. In most districts, scores are clustered at 1 or 2 points that are relatively high on the scale. However, there are a substantial number of scores given in the lower part of the scale, especially for one district.

Table 8 – *Distribution of Observation Scores by District*

District	Observation Scores Scale			
	1	2	3	4
1	0%	10%	82%	8%
3	1%	14%	71%	13%
4	2%	8%	84%	6%
5	0%	4%	74%	22%
7	0%	3%	70%	27%
8	0%	1%	39%	60%
9	0%	35%	50%	15%

To what extent is there evidence that the protocols are being used as they were designed and to identify strengths and weaknesses of teachers? We examined some specifics of the scoring to gauge how faithfully evaluators were following the prescribed protocols. One piece of evidence concerned the extent to which the full protocol was used during observations. As shown in Table 5 above, in several of the pilot districts, teachers were not evaluated on all criteria or evaluators did not rate teachers on the same set of criteria.

However, we do note that the scores appear generally higher than those observed in recent research studies. For example, scores on the instructional domain for Framework for Teaching in the MET project generally were more likely to be in the *Basic* (2) range (Bill and Melinda Gates Foundation, 2012). In the pilot districts, teachers composite, end-of-year scores are more likely found in the *Proficient* (3) range.

Summary

The pilot districts made substantial progress in implementing the classroom observation component of the evaluation rubric this year. The work really began mid-year, and there were a great deal of start-up efforts that needed to be undertaken. Thus, these findings should not be viewed as being

critical of any aspect of the new system. Rather, the findings identify areas that are important to resolve as the state and districts move toward an operational system. Key findings are:

- A large proportion of eligible teachers were observed at least once. Participation rates varied substantially across districts, however.

- Most teachers were observed only one time. This is likely a function of both the timing of the pilot and the capacity within the district. There is a very substantial demand on potential observers to carry out all requisite evaluations.

- To varying extents, districts made use of formal and informal observations. The informal observations are less time-consuming and involve less work for both teacher and observer.

- There were relatively few cases of multiple independent observers participating in an observation. This makes it problematic to determine the level of agreement in the system (i.e., inter-rater reliability). Without such data, it is also impossible to determine whether particular observers hold different standards than other observers and what this would mean for the accuracy of the evaluation rubric.

- The protocols were applied somewhat unevenly. In some districts, not all scale points were used and in some districts, observers had some latitude in using different scale points within the observation protocol. Other districts applied the protocols as designed.

- There is some differentiation of teachers across protocols and districts. The majority of teachers are assigned scores that are relatively strong but not at the top of the scale. However, a large proportion of teachers receive scores at the top of the scale. A smaller, but substantial, number of teachers receive at least some ratings that are below the midpoint of the respective observation scales, and very few receive scores at the bottom. Thus, there are indications that the rubric is providing more differentiation than has existed under traditional evaluation systems.

- The scores tend to be higher than scores assigned in research studies for which observers have no relationship with teachers, schools, or districts (Bill and Melinda Gates Foundation, 2012). However, higher scores are also being observed in many rubrics that are now being implemented across states (Consortium on Chicago School Research, 2011; Tennessee Department of Education, 2012).

Perceptions of Rubric Quality

This section begins by exploring the extent to which there is an understanding of and support for the pilot Teacher Evaluation program. In addition, we explore how different stakeholders (e.g., administrators and teachers) perceive the evaluation rubric and how consistent these perceptions are across districts. To the extent that there are differences, it is useful to understand the circumstances that lead to more favorable perceptions. To that end, our assessment team asks:

- What criteria do teachers and administrators use to judge the quality of a teacher teaching practice evaluation instrument, and how similar are the criteria used by the two groups?

- To what extent do teachers and administrators view aspects of the evaluation rubric favorably (and unfavorably)?

To the extent that stakeholders view the pilot Teacher Evaluation program (and more specifically the district-selected teacher evaluation framework) favorably, they are more likely to use it in ways envisioned by state policy—that is, to guide professional development and inform personnel decisions. In addition to judging the intended design of the rubric, stakeholders also judge the quality of implementation and the likelihood that it will be implemented effectively. Therefore, we also ask:

- What have districts done to support the use of data in making personnel and professional development decisions?

Criteria for Judging Teacher Evaluation Rubrics

To learn how educators in the pilot districts understood the selected teacher evaluation frameworks, we asked teachers and administrators a similar set of survey questions. As described earlier, our assessment team designed a subset of questions to address specific issues. We report data by focusing on questions from each of the main survey categories and include comparisons of teachers' and administrators' responses. Given that we summarized data across all respondents, more weight was given to larger districts as they have more participants and therefore, we also report the range of district averages to show across-district variation. We then use the data from interviews and focus groups to provide greater insight into the evaluation rubric. These help identify why stakeholders might perceive the evaluation system in one way or another.

Teachers and Administrators' Perceptions of the Teacher Evaluation Rubrics

When we examined how teachers and administrators perceived the quality of their evaluation rubrics, it became clear that administrators were more comfortable with the rubric than teachers. As seen in Table 9, more administrators than teachers agreed with statements that the observation rubrics were fair, accurate, and useful for guiding professional development, separating accomplished from unaccomplished teachers, and making tenure and promotion decisions, among other things. For one item – the statement that the pilot Teacher Evaluation program consumed resources that could be better used elsewhere – more teachers than administrators showed agreement with the statement. While these averages hide substantial district-to-district variation in perceptions of the overall adequacy of the teaching practice evaluation instrument. Administrators' perceptions were consistently more positive than teachers' perceptions. In our surveys, questions assessing perceptions were presented in a Likert scale format with 3 to 7 response choices,

depending on the question. In data analysis, similar response categories such as strongly agree and agree were collapsed for ease of presentation and analysis.

Table 9 – *Major Criteria: Administrators' and Teachers' Perceptions of Utility, Fairness, and* Accuracy *of the Teacher Evaluation Rubric*

	Administrators (August 2012)	**Teachers**
	Percent agreement	
The district's rubric for assessing teachers generates accurate assessments.		
Percent Agree[10]	74%	32%
District Range	38%–100%	20%–52%
In my experience, the district's rubric for assessing teachers is fair.[11]		
Percent Agree	80%	39%
District Range	56%–100%	21%–54%
The district's rubric for assessing teachers generates assessments that help provide individual feedback and design professional development.		
Percent Agree	75%	53%
District Range	33%–100%	37%–70%
The district's rubric for assessing teachers is well aligned with the district curriculum.		
Percent Agree	69%	45%
District Range	44%–100%	26%–55%
The district's rubric for assessing teachers clearly separates accomplished from unaccomplished teachers.		
Percent Agree	64%	30%

[10]Response categories are collapsed across several response options. *Agree* includes *Strongly agree* and *Agree*. *Disagree* includes *Strongly disagree* and *Disagree*.
[11] Teacher's version does not have *In my experience*.

District Range	17%–100%	11%–49%
The district's rubric for assessing teachers fits well with other school/district initiatives.		
Percent Agree	76%	39%
District Range	50%–100%	22%–60%
The district's rubric for assessing teachers provides a firm basis for making teacher tenure and promotion decisions and weeding out weak teachers.		
Percent Agree	54%	28%
District Range	17%–100%	12%–45%
The district's rubric for assessing teachers helps this district meet its accountability requirements under NCLB and other external mandates.		
Percent Agree	70%	43%
District Range	33%–100%	30%–56%
The district's rubric for assessing teachers helps improve student achievement.		
Percent Agree	53%	31%
District Range	22%–100%	16%–40%
The district's rubric for assessing teachers consumes resources that could be better spent on promoting key district improvement initiatives.		
Percent Agree	29%	52%
District Range	0%–71%	38%–72%

To find out how much these perceptions of the program represented changes from previous practice we asked administrators and teachers how the current rubric being implemented as part of the pilot program compared to the district's previous system on five dimensions: formalization, ease of use, grounding in research, intuitiveness, and usefulness for providing guidance to teachers. The responses to questions about change were similar to their overall assessment of the teacher

evaluation rubrics, as can be seen in Table 10. On all dimensions, administrators were much more inclined than teachers to agree that the new rubrics were an improvement over the old. Even on the dimensions that the most teachers agreed showed improvement--grounding in research and providing guidance to teachers--30 to 45 percent more administrators agreed that there was improvement. The other similarity was in the substantial difference among districts in their views about improvement. Four times as many administrators agreed that the new rubric was easy to use in one district as in another, and almost three times as many teachers thought the newer rubrics provided more useful guidance in one district as another.

Table 10 – *Perceptions of New Teacher Evaluation Rubric in Comparison to Previous System*

	Administrators (August 2012)	**Teachers**
	Percent who agree current rubric is better[12]	
Formalization (clear rules, steps, procedures, reporting forms) District range:	78% 56%-100%	38% 22%-57%
Ease of use District range:	61% 17%-82%	24% 9%-31%
Grounding in research District range:	87% 57%-100%	42% 33%-58%
Intuitiveness District range:	73% 25%-100%	29% 15%-41%
Usefulness for providing guidance to teachers District range:	75% 50%-100%	41% 23%-61%

The second administrator survey included questions similar to the teacher survey items about how effective administrators were as evaluators so that comparisons between teacher and administrator responses could be made. According to Table 11, more administrators agree that they have the characteristics of good evaluators than teachers do. For example, 54% of teachers think that their evaluators have the required knowledge to evaluate them while almost 95% of administrators do. Almost nine-tenths of the administrators versus about three-fifths of teachers report that the feedback provided in post-conferences is useful. In addition, 57% of teachers report that their feedback focuses on suggestions for improvement, while 94% of administrators report that their feedback does so. Quality of evaluators is another area where there are notable differences across districts, but even considering them, administrators are consistently more positive than teachers about the quality of those conducting the evaluations. Nevertheless, when one compares Tables 9 and 11, teachers are more positive about their evaluators than about the teacher evaluation rubric.

Table 11 – *Perceived Quality of Evaluators*

[12] Response categories "much better" and "better" are combined.

	Administrators (August 2012)[13]	Teachers
	Percent agreement	
The evaluation process at my school allows teachers to explain decisions and actions.		
Percent Agree	86%	58%
District Range	72%–100%	47%–84%
I am given useful feedback by the evaluator.		
Percent Agree	89%	58%
District Range	67%–100%	43%–80%
I feel that the evaluators in my school have the required knowledge and competencies to appraise teachers.		
Percent Agree	94%	54%
District Range	75%–100%	42%–63%
I feel that the evaluators in my school have received adequate training to perform their job.		
Percent Agree	78%	49%
District Range	46%–100%	40%–64%
In general, I think that the feedback that I am given focuses upon suggestions for improvement.		
Percent Agree	94%	57%
District Range	67%-100%	42%-72%

When asked about the overall impact of the evaluation rubric (see Table 12), teachers are generally much less positive about the effects of the new rubric than administrators. More than three-quarters of administrators report positive impacts of the observations, while teachers are much more reserved in their judgments; only one third to two-fifths of teachers report positive effects. Some of this may be because administrators have received considerably more training on the district-selected teacher evaluation framework as is described below.

Table 12 – *Overall Impact of Observation Rubric*

[13] Administrator version is worded differently; please consult Appendix 2.

Rubric has positive effect on:	Percent Reporting Positive Effect Administrators (August 2012)	Teachers
Your professional development [14]		
Percent Positive	78%	42%
District Range	60%–100%	20%–51%
Collaboration with others		
Percent Positive	79%	35%
District Range	50%–100%	25%–42%
Your school		
Percent Positive	77%	32%
District Range	38%–100%	15%–42%

Teachers' Perceptions of the Observation Rubric

While teachers were not as positive about the Teacher Evaluation pilot as administrators, they generally recognized the need for evaluation, as Table 13 indicates. For this set of questions, teachers were asked to agree or disagree (or neither) with a set of statements about teacher evaluation in general. Overall, teachers valued quality observations and recognized the advantages of teacher evaluation. They agreed that teacher evaluation was necessary to raise standards of teaching and that it could provide useful information to improve their practice. On the other hand, they disapproved of evaluations that primarily supported managerial decisions and aimed at meeting minimum standards.

Table 13 – *Teachers' Attitudes Toward Evaluation*

	Agree[15]	**Neither Agree nor Disagree**	**Disagree**
Teacher evaluation is essential to raise the standards of teaching and learning.	84%	10%	6%
Teacher evaluation should primarily focus on the identification of my professional development needs.	66%	22%	12 %
Teacher evaluation aims at meeting the minimum standards.	27%	29%	45%
Teacher evaluation aims at providing useful	79%	10%	11%

[14] In the administrator version, "Your" is omitted.

[15] For the following two tables, response categories are collapsed across several response options. *Agree* includes *Strongly agree* and *Agree*. *Disagree* includes *Strongly disagree* and *Disagree*.

information for teachers to improve their performance.			
Teacher evaluation should be based on a list of professional competencies or behaviors.	76%	17%	7%
As a professional, I am entitled to have my performance appraised.	86%	12%	2%
Teacher evaluation should aim primarily at making managerial decisions.	16%	34%	49%
Teacher evaluation aims to enhance teachers' reflection on their practice.	81%	10%	8%
Teacher evaluation should be used both for professional development and accountability purposes.	68%	17%	15%

The majority of teachers thought that evaluation had overall benefits although it raised some issues (Table 14). While most teachers said it encouraged them to reflect on their teaching practice and many said it made them more aware of their strengths and weaknesses, the majority reported that it also created tension among staff.

Table 14 – *Teachers' Perceptions of Impacts*

	Agree	**Neither Agree nor Disagree**	**Disagree**
The evaluation rubric encourages me to reflect on my teaching.	65%	23%	12%
The evaluation rubric has made me more aware of my strengths and weaknesses as a teacher.	48%	32%	21%
The evaluation rubric has led to tensions among staff.	58%	30%	12%

Use of Observation Data

This section reports on how stakeholders described their use of observation data. There was very little evidence of use of the observation data in the first year—either for planning of professional development or for personnel decisions—but there was a great deal of consideration of one aspect of accuracy and how to improve it: inter-rater reliability. The evidence stems largely from interviews.

Professional development. Districts were quite vague about how they might use observation data to plan or target collective professional development. Several districts implied that they could identify weaknesses in the teaching staff by quantifying teacher observation ratings. One district reported that it had used observation data to address a common issue among teachers. Another believed that the observation data, especially in the post-observation conference, helped teachers

develop their professional growth plans (PGPs).[16] Whatever process and decision rules might have been used, problems in simply getting the data management systems to aggregate data and generate reports would have made it very difficult to use the observation data for program or individual improvement during most of the first year (see section on Barriers and Facilitators).

District-level administrators in all districts were optimistic that the observation protocols would help identify weak areas in individual teachers or groups of teachers, and all the districts unanimously saw potential for linking teacher observations to professional development. They said they were just beginning to harness the potential of this data and that they were not conducting observations punitively so much as to identify ways to help teachers "become more responsive to children". Still, observation data had not been used to guide collective professional development—e.g., for school and/or district workshop planning. Even in the districts that led in collecting data, usage of this data to guide professional development was limited and diffuse.

While the use of observation data to plan professional development remained largely theoretical, it was used to support teachers through the one-on-one post-observation conferences. Administrators in several districts reported that the new observation protocols (based on district-selected teacher evaluation frameworks) were generating an increased amount of discussions between evaluators and teachers. This positive usage fit with the Teacher Evaluation pilot program's expectation that the observation process could quickly be used to generate changes in teaching practice. In some cases, teachers were asking questions about how they might improve their evaluation scores. One principal described how some teachers "will ask, 'well, if I'm *proficient* now, how do I, how can I be a *distinguished*?'"

In describing how teachers responded to being observed, a superintendent explained that they are better prepared to actively participate in discussion about their work:

> *"They come ready and prepared to have a discussion about their practice versus me sitting there. I go, 'and then you did this and then you did that and then you did this that.' They [now] come and they tell me, 'Okay I know this about my kids. I have this; I have that; I have that.' They come better prepared to have that in-depth conversation about practice."*

A few similar comments described the increased discussion between teachers and administrators. While administrators mentioned that discussions increased, teachers also reported that their post-observation discussions were changing to reflect the direct connection to the evidence collected in the observation process.

Personnel decisions. The pilot districts operated under New Jersey's regulations governing the annual review of teacher performance during the first year. Some used an earlier system for teacher observation at the beginning of the year before shifting to the new teacher evaluation rubrics adopted through the pilot program later on. As a result, most districts emphasized that they will use the first year of the pilot program for learning to collect observation data, more than for making personnel decisions. Several respondents referenced the idea that this was "only a pilot year." In

[16] state-mandated, individual professional development plans

fact, during this year, many aspects of the system were adjusted to support implementation and decrease unease with the new rubric, as described by one superintendent:

> *"I would say just involve teachers. Make sure they are well-trained and that your administrators are all on the same page and that they go through that same process of inter-rater reliability and that they understand that this is a pilot year and that yes, everyone is learning as we go along but we are looking for growth we are not looking for a gotcha model and that its just a learning process and the goal is for everyone to grow professionally."*

With at least two local teacher associations among the pilot districts challenging the legitimacy of the new system for teacher evaluation, the use of the new program for dismissal or other personnel decisions would likely be extremely tenuous for administrators this year. There was also evidence that administrators, especially principals, truly believe that all of their teachers are effective and that personnel decisions need not to be made. This mindset would also limit the use of the teacher evaluation tool for personnel decisions.

Still, administrators are aware that the evaluation protocols can be used for personnel decision making, and some favor this use, as indicated by one principal who stated (prior to the passing of the New Jersey revised tenure law):

> *"I'm in favor of ramping up the teacher evaluation. If the legislature does not have the, how can I say it, the fortitude to change the tenure laws. Alright so now we have to do this which is great because it is going to improve teaching and learning. I get it, there's a lot of politics involved here but on the ground level, it will improve teaching and learning and I subscribe to that."*

This statement shows that administrators are subtly using and articulating the evaluation process and purpose in two ways. One is that teachers will improve with better quality evaluation, a key argument for advocates of this program. The other, though stated indirectly, emphasizes improvement by teacher dismissal rather than through instruction improvement feedback. Some administrators use this coded speech around personnel to avoid ruffling the feathers of sensitive teachers. Another administrator said:

> *"One of the biggest benefits I would say, as we've implemented this particular framework, is that I slowly began to see a more objective way of collecting evidence leading to recommendations and decisions that obviously were very clearly defined, or identified, I should say, and leading to new steps or next steps."*

Overall, superintendents were careful to emphasize the formative purpose of evaluation first. Student growth and teacher development were often referenced.

Accuracy. When discussing accuracy issues, attention focused on inter-rater reliability and "subjectivity." Some administrators are aware of and plan to address inter-rater reliability, while others believe that it will not be an issue. Teachers especially feel that different evaluators score them differently. Additionally, while many administrators see the new protocols as objective or increasing the objectivity of the evaluation, teachers disagree.

Inter-rater reliability was an issue of great significance across pilot districts. These comments were of two distinct types. The first was about districts' plans or procedures to address inter-rater reliability. Three districts had procedural comments. Second, in every district, someone—either a teacher or an administrator—described inter-rater reliability as a problem. In addition, 14 different interviewees addressed the issue of whether the new teacher evaluation rubric was more or less objective. In general, teachers found the rubric subjective, with eight focus groups, including at least one from each district, claiming that the rubric was still subjective. Only one teacher focus group talked about increased objectivity. Administrators in four districts who addressed the issue of objectivity of the new rubric for teacher evaluation unanimously believed that it was more objective.

Finally, the assessments of the overall accuracy of the observations were split. In three teacher focus groups, some teachers believed that the new observations were accurate, while in other four, teachers were skeptical about the accuracy of the observations.

Since all evaluators had been trained on inter-rater reliability, the term did not surprise anyone. Some districts focused on it more than others. The timing of the new teacher evaluation implementation affected inter-rater issues. As one Teacher Evaluation project director said,

> "*We are planning on significant inter-rater reliability measures, meaning a morning where you spend some time training around coming back to the rubrics and the use of the rubrics and then looking and rating and doing our own evaluation of where they stand. Our challenge is that we wanted people to start using it before we did inter-rater reliability because we wouldn't have been able to have a strong data field.*"

The same project director acknowledged "a big struggle about the quality of evaluations" because "there are some that are very strong and others that are very weak. At our round tables, we've done informal inter-rater reliability checks." Initial training in this district did not achieve adequate inter-rater reliability. Allowing observers to try the teaching practice evaluation instrument out before agreement was high was an intentional training strategy (learning by doing) but may have undermined seriousness about achieving high reliability. One principal interpreted the district's progress more positively than the Teacher Evaluation project director, saying that the data "looked pretty good" and that improvement was "just a matter of focusing."

Another training strategy that might have put off achieving high inter-rater reliability, but might also have effectively helped observers master the cognitive complexity of these rubrics, was used by a superintendent who took

> "*one [codable dimension] out of [one area] and one [codable dimension] out of [another area]. I didn't put their names on. I just put the evidence, and I brought it to our administrative meeting and shared that and said, 'Look, are we looking at this one [dimension], are we all looking at it and citing evidence in the same way, and what do you notice and what should we do the same or differently?'*"

Selecting just a few categories to focus on may have allowed administrators to direct their concentration until the accuracy of their observations increases, while they address only those parts of the observation protocol that they feel comfortable with and assume that familiarity among faculty would breed similar scoring.

Leaders in another district were confident that inter-rater reliability would not be an issue, due to regular meeting times to talk about teaching and learning. A leader in that district said,

> *"I'm not terribly worried about the inter-rater reliability here because we are a group who come together, like I said, weekly… And then in pockets of schools that need extra support from central office we meet with them again weekly so there's a lot of communication, there's a lot of group time together to talk about teaching and learning."*

While central administrators were more or less certain that inter-rater reliability problems were solvable, some problems were apparent. For example, although most respondents talked about the importance of inter-rater reliability, some key leaders did not fully understand its meaning and significance. One superintendent said:

> *". . . we're all different people with different talents and different views and I think it's important that there is some diversity in what it looks like. I don't want to get so much that everybody should be exactly the same."*

This superintendent thought that when two raters were within two points of each other on a 4-point scale, they were close enough—a view that is not psychometrically defensible. Teachers did notice inconsistency on raters' part:

> *"When my colleague, for example, had been observed by one of our administrators, she got in two of the categories outstanding and she got some feedback from the administrator. Three days later, she had another observation, incorporated that feedback to the same exact lesson, and the second administrator observed her. It was kind of a drastic difference. . . . Two exact same lessons with feedback incorporated I would imagine only could get you a little bit of a better observation, but that one was lower so [it]really makes me question . . . "*

Administrators also recognized differences among observers. One said, "I would be concerned that elementary people would have different reliability than secondary. . . . And our principals are expected to be instructional leaders so we don't have that huge drop off but there is . . . definitely a different focus."

Finally, one teacher association representative questioned how sustainable high reliability might be:

> *"We have been incredibly diligent in the manner in which we have implemented the whole process, but I don't see how we can maintain that diligence long-term and increase the number of observations that we are going to do next year. Unless we start to cut corners and maybe only observe certain domains, or certain elements, in which case, how do you make a high-stakes decision about a teacher's performance? . . . And the big question I have . . . .is how they are*

ensuring inter-rater reliability and preventing drift and making sure that everybody is on the same page."

Summary

In summary, taken together, the survey and interview data suggest the following conclusions about perceptions of implementation in the first year:

- Teachers and administrators use similar criteria for considering the quality of the teaching practice evaluation instrument and want a rubric that is accurate, fair, and useful for both professional development and for personnel decisions. In addition, teachers have clear ideas about what constitutes a high-quality evaluator.

- Administrators generally view the new teacher evaluation rubric more positively than teachers do.

- Districts have not done much yet to prepare to use data for personnel decisions.

- Districts have done somewhat more to use data for professional development purposes—especially through one-on-one coaching—but there is still much to accomplish in this direction.

- Overall, the thinking about accuracy focuses on inter-rater reliability, and administrators are more optimistic about achieving reasonable rates of reliability than teachers.

- We point to some real issues that the interviews suggest might inhibit achieving such reliability, which include not viewing it as an issue, training quality or lack thereof, focusing on one dimension to the exclusion of others, and the overall complexity of the teaching practice evaluation instrument.

The survey data, in particular, point to important differences in perceptions across the districts. The evidence so far makes it difficult to know how much these differences reflect the different teacher observation rubrics (Danielson vs. McREL vs. Marzano vs. James Stronge) and how much they reflect differences in district demographics or leadership.

Barriers and Facilitators

This section uses survey and interview data to describe the barriers and facilitators to the implementation of the Teacher Evaluation rubric that were relevant in the pilot districts. It addresses four specific issues that might affect the quality of implementation:

- What time demands did implementation of the teacher evaluation rubric generate in the pilot districts, and how were those demands similar or different for administrators and teachers?

- What training was offered to teachers and administrators in the pilot districts, and how was it seen as helping prepare staff for implementing the pilot Teacher Evaluation program?

- What advantages did the data management tools that were part of each teacher evaluation system afford the districts, and what learning challenges did districts face in taking advantage of these tools?

- What sources of resistance to the pilot Teacher Evaluation program did teachers describe?

For the survey data, we report means for all respondents and then provide medians and ranges across districts.

Time

Time was one of the most frequently raised concerns in the interviews. Administrators, in particular, reported that they spent a considerable amount of time on the new teacher evaluation program. Table 15 provides administrators' reports from the second survey (administered in August) about how their allocation of time across tasks has changed since the introduction of the pilot Teacher Evaluation program. Not surprisingly, about 90% of administrators report spending more time conducting observations and entering data related to those observations since the program started. In addition, just over 45% of administrators report spending more time on other tasks. These data suggest that the requirements of the Teacher Evaluation program substantially increased the demands on administrators' time. In some cases—most notably when discussing "other administrative tasks"— administrators' responses varied considerably from district to district.

Table 15 – *Administrator Reports on How Workload Has Changed (August 2012)*

	More time[17]	**About the same time**	**Less time**
Conducting observations (% agreement)	89%	9%	2%
District range:	83%–100%	0%–16%	0%–7%
Entering data (% agreement)	91%	5%	4%
District range:	73%–100%	0%–20%	0%–14%
Other administrative tasks or job responsibilities (% agreement)	46%	34%	20%
District range:	25%–86%	0%–60%	0%–50%

[17] Response categories are collapsed across several response options. *More time* includes *Much more time* and *More time*. *Less time* includes *Much less time* and *Less time*.

Two teacher survey questions addressed how the new Teacher Evaluation program influenced their work. As Table 16 indicates, most teachers did not think that the pilot intensified their work or added a great deal to the work they had to do overall, but 60% agreed that it increased their bureaucratic work. This response varied quite a bit across districts, showing up almost twice as often in some districts as others (44% vs. 82% agree). The phrasing of these items suggests that the new teacher evaluation program did not greatly increase the amount of work teachers did, but it did require them to engage in activities they did not find meaningful.

Table 16 – *Teacher Perceptions of How Workload Has Changed*

	Agree	Neither Agree nor Disagree	Disagree
The evaluation rubric has led to an intensification of my work.	39%	39%	22%
District range:	23%–48%	25%–54%	9%–39%
The evaluation rubric has increased the bureaucratic work at school.	60%	33%	7%
District range:	44%–82%	12%–52%	1%–9%

Survey data did not provide the most comprehensive information to assess time allocation. The interview data provides a more complex (if incomplete) picture of how administrators use their time and how that use is valued. Superintendents acknowledged that the observations increased time demands but were ambivalent about it, seeing the value of the observations. One explained that:

> *". . . tenured teachers now are required to have two observations whereas before they were required to have one and additionally the pre-conference requirement for both tenured and non-tenured teachers is an additional time constraint. I understand and strongly believe in their effectiveness and their importance in the process but it wasn't anything that we were requiring people to do before so for our administrators we essentially tripled the amount of time that they are spending on teacher evaluation."*

Teachers and administrators alike described the additional work required, although not clearly enough to quantify it. This work included:

- Added observations of tenured teachers;
- The new requirement for a certain number of pre-conferences; and
- The time required to complete the necessary documentation.

In response to a question in the second administrator survey, administrators reported that the average time to complete one observation (from pre-conference to finish) was three hours, but there was a substantial range, with some taking as long as six hours.

Consequently, various other administrative responsibilities were given less attention and time. One district administrator described how his staff was asking, "When do we get you back [name], when do we get you back?" A teacher explained that:

> *"The discipline problems have increased and they're just not getting [handled]. And the staff does not blame the assistant principal that's in the building because they know she's working but it just seems like so much energy being put into this, it takes away from the other things."*

Several things compounded the time problem in the first year of the pilot Teacher Evaluation program. One was the pilot program's late start-up. Part of the problem occurred when the districts learned that they had been awarded the pilot grants. One superintendent said:

> *"Our biggest issue with training was the tight timelines in which we had to get it all accomplished. . . . Being notified that we received a grant at the very end of August, having to select our leadership team . . . , have them trained on the framework, and then have the rest of the administrators trained."*

Among the four districts where late start-up or late changes in state requirements were explicitly mentioned, actual observations began between November and February, rather than in early fall as would normally have been the case.

Another factor consuming time in the first year was learning how to conduct the observations, including actually conducting the observations, filling out the forms, and handling the software that accompanied every district-selected framework for teacher evaluation. In two districts, observers noted that the process went more quickly over time, saying:

> *"It got down to closer to two hours, which will also include my class periods are . . . 60–65 minutes. And so my second and third evaluation timeline went a little better."*

Supporting this view is the evidence from the first administrator survey (March 2012) where 75% of the administrators agreed that, "I feel comfortable using my district's rubric for assessing teachers,"; agreement on the same item increased to 86% in the second administrator survey (August 2012) (see Table 17).

Table 17 – *Comparison of Administrator Comfort in Using Evaluation Rubric in March 2012 and August 2012*

	Agree	Neither Agree nor Disagree	Disagree
I feel comfortable using my district's rubric for assessing teachers. (March 2012)	75%	14%	11%
District range:	33%–100%	0%–33%	0%–33%

I feel comfortable using my district's rubric for assessing teachers. (August 2012)	86%	11%	4%
District range:	67%–100%	0%–33%	0%–14%

The pilot districts approached differently the issue of increased time demands for administrators in particular. For example, they used some regularly scheduled meetings to train observers on how to conduct observations (see section on training). In addition, one superintendent hired temporary administrators to fill in for the regulars who were going through online training.

Moreover, in at least one district, the superintendent clarified priorities. Because the recording process took so long and because of other competing commitments, a few teachers complained about the lag between the observation and the post-observation conference:

> *"I received [the written evaluation] but then we didn't actually sit down to conference and sign it off and everything else. By the time we actually did, of course, it was an emergency, she was pulled out and I didn't get the chance to actually speak . . . Weeks later I said to her, 'do you remember anything?' She said, ' no.' I said, 'Okay, I don't remember anything either.' I signed it."*

In addition, scheduled observations might not be held as planned:

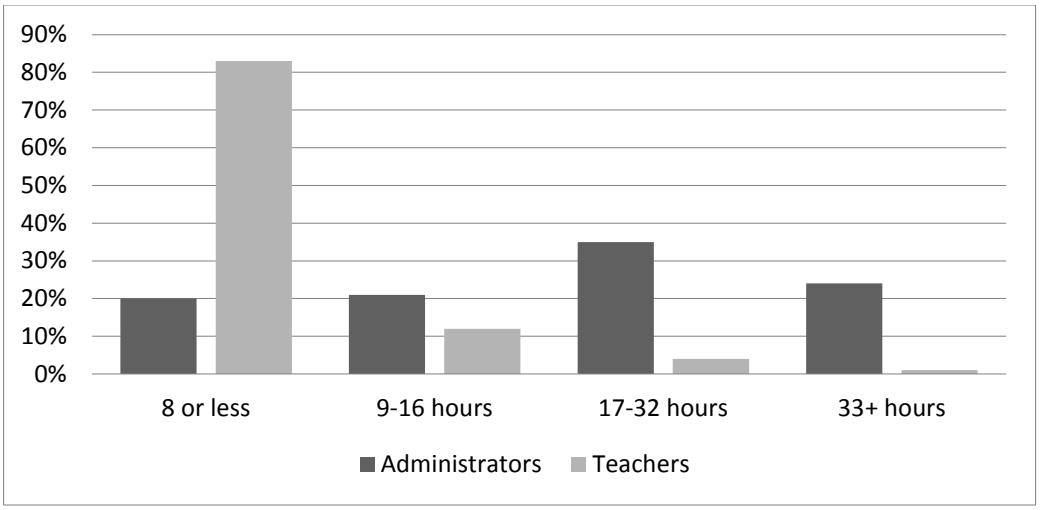
> *"We had several teachers who had the same thing happen to them where they spent a lot of time answering those questions . . . And we were answering the questions correctly and then no one came to observe them or from a 90-minute block they only came for 45 minutes. And it was just kind of disheartening to have to do that."*

Scheduling problems were more common when the priority given to completing observations was not made completely clear. They would respond to unhappy parents or discipline emergencies even if they had to break an appointment for an observation. However, in one district, the superintendent made clear to his staff that, "the most important thing we do on our day is the observations and if we had that scheduled, nothing should get in the way. . . ." Other administrators accepted this view, and the district appeared to complete most of the required number of observations.

Training

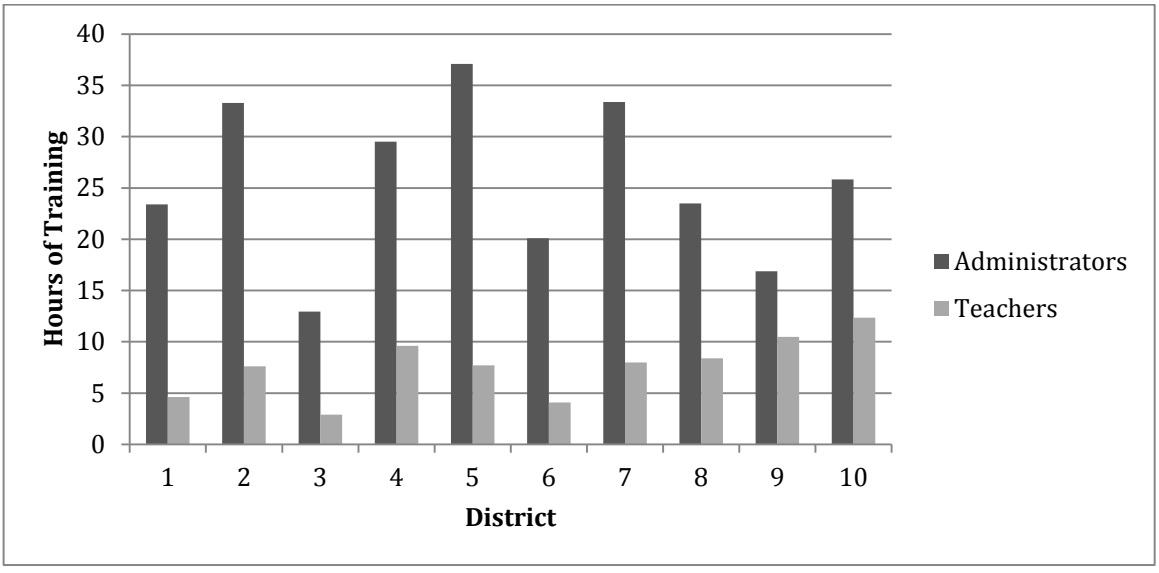
Teachers and administrators had a very different perception of the provided training on the newly selected teacher evaluation frameworks, perhaps because teachers received significantly less training than administrators (see Figure 1). Four-fifths of the teachers received fewer than eight hours of training, including one third that received only one to two hours of training. By contrast, 20% of administrators received fewer than eight hours of training, and just over half received from 9–24 hours or more than one and up to three days of training. Administrators received more training because they actually conducted the observations. Nevertheless, the added training may have increased their understanding of the program and therefore given them a more positive perspective.

Figure 1. Hours of Training



As illustrated in Figure 2, there was a significant degree of variation in hours of training across districts for both administrators and teachers. In one district, teachers received only three hours of training on average and administrators received only 13 hours of training. In contrast, the district with the most training for teachers offered ten hours, while the district with the most training for administrators offered 37 hours of training on average. Overall, hours of training for teachers ranged from 3–12 hours on average. For administrators, training ranged from 13–37 hours.

Figure 2. Variation in Hours of Training by District for Administrators (March 2012) and Teachers



This difference in access to training is apparent in the survey data on perceptions of the quality of training (see Table 18). We align related questions in the table. Administrators were almost twice as likely as teachers to report that their training helped them understand the new rubric for assessing teachers. This was the largest difference between teachers and administrators. This is another topic where there was a substantial district-to-district variation, however. In some districts all administrators thought the training had met a goal, while in others, none thought so. Other large

differences concerned assessments of how well the training helped make judgments about teaching quality and assess teachers' instructional practices, but even the smaller differences—assessing planning practices or giving and understanding feedback after an observation—were relatively large.

Table 18 – *Perceptions of Training Quality*

	Administrators Accomplished (March 2012)		Teachers Accomplished
Help you understand your district's rubric of assessing teachers	80%	**Help you understand your district's rubric of assessing teachers**	41%
District range:	0%–100%	District range:	30%–68%
Help you reach thorough, well-grounded judgments of teacher quality	70%	**Help you to understand what underlies judgments of teacher quality**	37%
District range:	0%–100%	District range:	26%–53%
Help you to assess teachers' instructional practices	78%	**Help you to understand the criteria for assessment of teachers' instructional practices**	46%
District range:	0%–100%	District range:	35%–67%
Help you be aware of potential biases in the way you evaluate teachers	63%	**Help you to understand potential biases in the way teachers are evaluated**	32%
District range:	17%–100%	District range:	17%–52%
Help you provide effective feedback to teacher after observation	72%	**Help you to understand the feedback after an observation**	46%
District range:	17%–100%	District range:	36%–68%
Help you to assess teachers' planning practices	57%	**Help you understand the criteria for assessment of teachers' planning process**	40%
District range:	17–100%	District range:	23%–73%

The interviews identified problems with the initial training that introduced most staff to the teacher evaluation rubric selected by the district. This was often the most substantial or the only training that teachers received. This training was often seen as rushed. As one administrator noted, they were "building the plane while flying it." A principal observed that, "training was rushed and had to be in order to get the training in so we could get moving." A teacher observed that, "We're very confused in a lot of things 'cause there was so much shoved down our throats at once. . . . Some of these could have been very good but by the second day your brain is fried."

The speed with which training was initiated led to other problems. For example, teachers reported that the training formats were excessively didactic. One said, "You had a presenter each day, just giving all sorts of information in the classroom. . . . It was just basically everything just thrown at you." Content was not always appropriate. Another teacher said the trainer spent too much time showing how to change the colors and fonts in a program and not enough dealing with substantive information. When training was organized quickly and delivered by different sources, the content was not always consistent. As one district administrator noted, "The challenge there was [the training] had two different trainers and so the two groups received two different pieces of information." Nor were all trainers fully prepared to lead their sessions. At least one admitted to being uncomfortable "turn-keying" information when he was not the expert.

One strategy that offered more long-term learning opportunities to teachers was to identify some expert teachers as trainers of others and to give them special preparation so they could help their peers learn the new evaluation rubric. In a world of Professional Learning Communities and extensive collaboration, this seemed like a useful strategy, with at least one district adopting it However, in another district, it raised concerns that some teachers would get special advantages through extra training that were not available to others:

> *"I felt that the quality of training I received is excellent, but that's because I was one of the teacher leaders who got to sit on the evaluation training. . . . There were teachers who went up to the teachers [with this special training] and they said, 'You have an unfair advantage over me.'"*

In a third district, the administration refused to use teacher leaders because the district was "in a sensitive contract year" and they did not want to be accused of giving anyone an advantage over someone else.

Follow-up training helped administrators develop a more positive attitude toward the new Teacher Evaluation program. Sometimes this training came through well-orchestrated joint walk-throughs.

> *"We've been doing . . .. instructional rounds. We were following the Harvard model from Richard Elmore and that's been very good for us. . . . We huddle after we leave a classroom and say, 'Okay, what did you see? . . .' So that's been good for conversation but now we're into something much deeper and we're still doing the instructional rounds using the [district model] framework."*

Sometimes it came through administrator meetings or in other settings:

> *"we established what we call a [job embedded] coaching model where [name] consultant in our case [name], has been coming back into the district every other month or so. And it is… the opportunity for us to be able to, have a conversation with her on the work that we're doing…. to have her with us, and conduct observations, provide feedback, have a conversation and the collection of evidence and the decisions that are being made, and more importantly to be a support to the system as the training and as the implementation obviously continues throughout the year"*

In one district, the principals extended this extra training to their teachers, but that practice seemed exceptional and challenging to accomplish.

Data Management Tool

One important element of the adopted teacher evaluation systems is the data management tools it requires. These are used to store initial observation data, collate information about one observation from multiple sources (and about multiple observations for one teacher), store data about large numbers of teachers, and analyze data to identify patterns that are crucial for assessing accuracy of observations, teacher quality, and to develop further steps that should follow from teacher observations. Initial observation data may be recorded in the classroom on a laptop, iPad or another portable device.

For the district then, key decisions revolve around the selection of a teacher evaluation framework and appropriate data management system. Key implementation tasks require more training and experience with the technology. Almost all of the information on technology comes from interviews. Respondent comments on technology almost equally balanced between positive (17 comments) and negative (18 comments). However, comments from district-level administrators were the most positive (11 of 18 were favorable comments) and comments from teachers were least positive (1 of 7 were favorable comments). Principals had evenly mixed comments during these interviews. Teachers generally had less interaction with the technology, except for those protocols that required input of lesson plans, reflections and other observation-related documentation.

Technology improves the data collection process and ultimately helps to identify patterns in the data. Since districts were just learning how to collect observation data, individuals were most interested in that. For that purpose, some viewed the technology to be essential. As one district administrator said,

> *"You can't do this with pencil and paper. I tried to do it when I was a principal and supervisor 'cause I used [name of system] before. You can't do it with pencil and paper; its mind boggling after a while . . . . The collection tools are important and being able to push a button and get reports out and look at district data, school data, grade level data, teacher data."*

In four of the six districts, administrators talked about how collecting data with laptops, iPads, or at different stages of the process with a district web site, facilitated the process:

> *"All the administrators have laptops . . . and that's what we were using. Plugging those in because it wasn't wireless and you know setting it up. And now with iPads here it's more [portable] and more convenient to blend."*

Although some administrators view bringing technology into the classroom as potentially intrusive (and some teachers think it distracts the observer), it increases convenience of recording substantially. It also has other advantages. For example, data analysis can be used to increase observer accuracy as one district administrator notes:

> *"Could we actually identify the observer? Not just how each person was scored in each domain . . . That would be really interesting because then we could look in the patterns to say person A always is scoring harder or easier or pretty much right on the mark with everyone else."*

Using technology for inputting data was something that administrators felt well prepared for by the end of the year. Table 19 shows that in August 2012, 71% said their training had prepared them to "understand how to input data" very well, and 64% said their training had prepared them to "understand how to provide feedback for a single teacher" very well.

Table 19 – *How Well Technology Training Accomplished Goals (Administrators, August 2012)*

	(Very well) accomplished	Somewhat accomplished	Not(at all) accomplished
Help you to understand how to input data	71%	27%	2%
Help you to understand how to provide feedback for a single teacher	64%	33%	4%
Help you to understand how to retrieve data to understand patterns of strengths and weaknesses in groups of teachers	43%	38%	19%

In addition, technology can help integrate the observation data with other kinds of data teachers need to collect. Administrators mention two kinds of information that are stored with observation data: state-mandated professional growth plans and lesson plans. Presumably, over time, integrating these elements will help educators ensure that the growth plans build on strengths and weaknesses noted in observations and that observations reflect the lessons teachers teach.

As these examples illustrate, the currently available data management systems are facilitating data availability and its use. As one district administrator notes, referring to conversations facilitated by her district's teacher observation rubric:

> *"It's a level of transparency now around evaluation data. Now people are having conversations or looking at what other evaluators have rated a teacher and it's caused some challenges, which is great. I mean . . . I'm really happy to say that 12 months ago nobody said, 'I don't believe in the inter-rater reliability between me and somebody else,' that's fabulous*."

Yet, others believe the use of data systems to analyze data, identify patterns, and consider different courses of action will not develop seriously until they and their colleagues are more proficient in using the rubrics they have:

> *"I think [identifying patterns] will come later, after we . . . we're still kind of muddling through this but I would think after the end of the year when we kind of regroup and look at, 'hey here*

was the areas that we saw a lot of strengths and these are the . . . ' and then we can incorporate that into our district goals; that will probably come at a later time."

In the second administrator survey (August 2012), 43% of administrators said the training they had received helped them understand how to retrieve data to understand patterns of strengths and weaknesses in groups of teachers very well (see Table 19).

The technology challenges the districts faced were the start-up problems that often occur when using a new system or a system designed for a different purpose. For example, in some rubrics either a teacher or an administrator could not update information once data had been entered, which inhibited the conferencing process that is supposed to accompany the observations and provide a learning opportunity for teachers:

"Hitting finish! Once you hit finish, you can't change anything, so that's nobody ever wants to hit that button, there was that kind of . . . .When it comes to the reflection, [expert leading the system] says, 'The post-conference shouldn't occur for after the teacher has reflected on what has happened in the classroom, because if you wait, then they're going to reflect on what you said, not necessarily on what they did.' So with that you couldn't, I couldn't end my observation."

More generally, some programs in some districts generated problems in saving data. A teacher explained that, "you sit down for an hour composing something at a computer and you go to press save and it's all gone, it happened to me it's very frustrating." This problem recurred frequently in that district and was related to the fact that information was stored on the provider's server. Other districts had less dramatic connectivity issues as described by a principal who said:

"that really hasn't been an issue except for the fact that I have to be connected while I'm doing it. . . . I can't work on it some place else, where in the past I could do that. I could do an observation while I'm sitting waiting for a doctor's appointment."

Another challenge was the time required to learn to use the actual data management systems. Some administrators put this challenge in perspective, saying, "It's such a learning curve but ultimately this whole experience has really enabled us to start that progression of improvement." Others found more specific elements to be confusing. For example, some found that the layout of data made the tool difficult to understand:

"You get a rating for each component and whereas before you'd have the rating and then the evidence, right now we're seeing all of the ratings for the different components on one page and then the evidence in the back. Format-wise it's a little awkward and I think it just lined up nicer when you could see the evidence underneath the rating."

Teachers were more concerned that the system for recording data distracted observers in the classroom. One said:

"I feel like they lose a majority of what's going on in the classroom because they're so focused on typing everything, and it's a huge distraction to them and the students. . . . If they were able to just sit back and perhaps take notes . . . the observations would be a lot better."

Finally, although some district administrators could see the data analysis possibilities of the new data systems, many of these were far into the future. In the short run, some tools had serious gaps in their capacity to help districts conduct basic analyses. One superintendent complained that:

> *"we should be able to run these reports . . . and [name] assured me that whatever [name] needs we are going to have as far as reports I said, 'well quite frankly I don't care what [name] needs I need it so.'"*

Still, by August 2012, considerable progress had been made. Ninety percent of administrators felt at least somewhat prepared to use the technology available (54% well prepared, 36% somewhat prepared). Table 20 shows how difficult it was to use different features of the online data management tools, and Table 21 shows the problems most frequently encountered. Most administrators found it easy to accomplish most functions including two-thirds who found it easy or very easy to access data to prepare reports and to manage already uploaded data. One fifth or more of administrators were still dealing with issues like getting kicked out of the system and losing unsaved data, as well as crashing and time-outing the tool, but almost half (44%) reported that they rarely encountered significant technical problems. It appeared, however, that administrators in some districts were learning the technology much more effectively than others.

Table 20 – *Difficulty Using Various Features of the Online Data Management Tool (Administrators, August 2012)*

	Very Easy	Easy	Hard
Log into the system and input information	49%	43%	3%
District range:	0%–75%	25%–75%	0%–19%
Use of the tool during actual teacher observations	24%	49%	19%
District range:	0%–44%	22%–100%	0%–38%
Use of the tool for communication and exchange of information	22%	55%	18%
District range:	0%–56%	0%–80%	0%–50%
Access already stored data to identify patterns of practice in the school or district	22%	53%	14%
District range:	0%–44%	22%–100%	0%–38%
Access already stored data to prepare reports on teacher evaluation data for various audiences other than the individual teacher being observed	17%	49%	16%
District range:	0%–44%	20%–63%	0%–38%
Management of already uploaded information	20%	45%	19%
District range:	0%–56%	20%–100%	0%–40%

Table 21 – *Frequent Issues with the Data Management System (Administrators, August 2012)*

	Percentage of Respondents	District Range	District Median
Short log-in period (getting kicked out and losing data before saving)	26%	0%–63%	25%
Problems saving information while working online	24%	11%–44%	24%
Problems with crashing/time-outing of the tool	19%	0%–69%	16%
Personal account information issues—lengthy password changes, updates of identification information, etc.	9%	0%–24%	3%
Poor accessibility of previously uploaded information	12%	0%–50%	7%
Ineffective online help feature/technical support	10%	0%–43%	7%
I have rarely encountered any significant technical problems	44%	0%–67%	40%

Resistance

One frequent finding from past implementation efforts is that the implementation of new practices often generates resistance (Evans, 1996; Fullan, 2001). Changes that affect personnel decisions especially have been known to generate resistance (Brandt, 1990; Firestone & Bader, 1991). Such resistance may be especially likely in a program like the Teacher Evaluation pilot, where one goal is "to improve the effectiveness of our educators (as defined by professional practice and student outcomes) through a system that a) clarifies the expectations for teacher practices and the metrics that will be used in their evaluation; and b) provides meaningful feedback to teachers to clearly identify strengths and weaknesses that will result in a relevant growth plan for teachers." (New Jersey Educator Effectiveness Task Force, 2011a). Using data for formative feedback requires that teachers be open to the use of the evidence provided.

In at least three districts, administrators described evidence of some level of resistance among teachers:

> *". . . teachers are much more guarded now and they're misinterpreting some statements made in the paper and like I said through the department or through trending a bad . . . you know if you get this rating, just a consecutive time, you know . . . you lose your . . . there are walls already built and it's unfortunate, it's totally unfortunate, because some really rich discussions about improving teaching and student achievement resulted from the pilot."*

This concern of teachers is not unusual, but it comes in conjunction with the consistent finding noted above that teachers viewed the fairness, accuracy, and usefulness of the teacher evaluation program less positively than administrators and that they were less convinced of the adequacy of available evaluators. These findings prompted us to follow Fullan's (2001) advice to explore what resisters

think because they often have insights that advocates may have missed. Teachers' critiques of the program centered on its subjectivity, other factors leading to inaccurate observations, changing definitions of distinguished practice, failure to receive formative feedback, and other communication issues.

The most frequently mentioned issue was that "it's very subjective." Teachers made many variations of this observation across the districts:

> *"I think any evaluation process is going to have subjectivity. I mean I really do . . . I don't think there's anyone, unless you have the same person doing every single staff member. That's the only way you're going to get the same validity I think, because of course everybody brings their own . . ."*

Some teachers acknowledged that the pilot teacher evaluation rubrics were less subjective than the preceding ones, but many strongly believed that some level of subjectivity was inherent in the observation process. This was not a view articulated by administrators. In addition to general concerns, teachers' doubts related directly to evidence of inconsistency among observers:

> *"Its kind of subjective so one administrator everyone knows this particular administrator gives a lot of high evaluation points and another one hardly gives out any. So your evaluation is kind of hoping to get a certain evaluator."*

Teachers think several factors limit administrators' agreement. One is the limited administrator training in the first year. One teacher explained that her administrator said, ""we're still working it out." I know that but you know what, this is my career. This is my evaluation; this is going in my file, you know."

Teachers are also aware of how administrators are under time pressures:

> *"The amount of time that the process is taking, and to truly do it with integrity, and to do it right, it is a burdensome process. It's just not something that can be maintained long-term, and so the natural tendency to be able to find ways to complete it is to cut corners. . . . If there's not going to be consistency, the data is not reliable and then how are you going to use the data to make high stakes decisions?"*

Teachers also questioned observers' recording practices. One said:

> *I remember my principal, after my observation, she said to me, "You were talking so fast." And I said, "I really wasn't" 'cause she was just trying to type everything so quickly and I felt like most of the things that I was trying to do and teach got lost, because she was so trying to type everything into this thing. I didn't know if all the principals were using the laptop, but I find it to be a huge distraction.*

Another teacher noted observers who did not fully pay attention: "I've had people [say] I got to take a phone call in the back. Really? Are you kidding me?"

Another issue of great concern in at least two districts is the expected change in distribution of evaluation scores. One teacher reported:

> *"And the one thing that stands out for me . . . was when the presenter said that a teacher would never receive distinguished. Do you remember that? . . . To receive distinguished was almost an impossible thing. And I thought, with the administrators there present, you were already setting in their mind like you are not to give a teacher distinguished because this is not something that, you know, we're allowed to receive."*

In addition, teachers worried about what they saw as inappropriate evaluation criteria. In this case, the observation protocol of the teacher evaluation rubric placed a very high premium on using technology.

> *"The objection seems to me more did I implement technology. . . . That's not going to happen six periods a day, 186 days a year. So yes am I doing those things right in the course of the year . . . , but should that be what separates a good lesson from a bad lesson from an observational stand point?"*

Sometimes the criterion was believed to be inappropriate because it was not necessary for every lesson. Sometimes the inappropriateness was because the criterion was about the source of materials expected to be used, which the teacher thought was arbitrary. Other times it was because the observation scheme did not give credit for something required by another district program. Finally, some programs raised questions about what the relative weighting should be for criteria applied in a lesson versus those applied for the teacher's out-of-class work.

By contrast, districts mention communication as a factor that could reduce resistance when carried out correctly. For example, a union representative explained:

> *"I was totally against this until I sat there and was convinced . . . . I had an awful lot of concerns because there is no trust here. And that really made me feel a lot better and my members that went with me but there were only five of us there . . . . So, and we came back and we expressed that at our rep council that . . . we felt a lot more comfortable with the state's intention . . . . So I think if there had been more explanation in each building to the staff about how it was going to work it would be better received."*

Summary

To summarize, the survey data from the ten pilot districts and interview data from the sample of six pilot districts at which we conducted site visits, suggests conclusions about the following barriers and facilitators to implementation of the teacher evaluation pilot program:

- **Time.** The greatest time demands of the pilot Teacher Evaluation program were placed on administrators. These were mostly requirements to conduct additional observations and complete documentation. Some demands stemmed from start-up problems, including the late start of the pilot and the extra time required to learn how to conduct observations, which will not be repeated or as extreme in subsequent years. These demands inhibited the delivery of the Teacher Evaluation pilot in the first year.

- **Training.** There was substantially more training provided to administrators than to teachers since administrators have to conduct the actual observations. All participants were involved in initial training that was often in large-scale collective formats. Some participants identified problems with the design and delivery of the initial training, although our evidence does not indicate how these problems were distributed:

    - It was rushed.
    - It was offered in didactic formats that did not facilitate learning.
    - It sometimes spent too much time on less-relevant, mechanical content.

  Administrators benefited from follow-up training through joint walk-throughs and repurposed administrator meetings that helped them develop greater comfort with procedures and training criteria. As a result, more administrators than teachers reported in the surveys that the training met its goals.

- **Data Management Tools**. The teacher evaluation rubrics require the support of data management tools for the collection, storing, management, sharing, and analysis of observation data that will support later personnel and professional development decisions. The learning problems associated with the data management tools were typical of those learning to use many new data management systems. Some people found these problems very irritating, but they worked through them. Training helped most administrators improve their familiarity with the tools for data collection. Fewer administrators reported that the training helped them use data to analyze patterns in the data. By the end of the year, most administrators had developed considerable comfort with the data management tools. In general, administrators reported greater strength in using tools for data collection and storage than for data analysis.

- **Resistance.** The data collected do not allow us to assess the strength and distribution of resistance to the new teacher evaluation program, but they do highlight some issues or practices that generated or alleviated teacher concern. Some teachers find the teacher evaluation rubrics subjective. They worry about differences among raters. Some teachers object to what they see as a quota on *distinguished* or highest-level ratings. Some teachers are concerned about what they see as inappropriate criteria applied to their observations. Where communication about the program is more extensive, teachers are more comfortable with the Teacher Evaluation pilot program.

These data suggest that there may be important differences in barriers and facilitators across districts. The team has not yet explored the nature and explanation of these differences.

Conclusions and Challenges

This section brings together findings across our three objectives. It first summarizes conclusions from the first year of the study. Then it then uses these findings to identify challenges for the state to address moving forward. Finally, our team identifies research issues that need further attention in coming years.

Conclusions

The first year of the pilot teacher evaluation project was a learning year for districts, vendors, and the NJ DOE. To facilitate that learning, RU GSE conducted an assessment of the implementation process in the ten participating districts. This assessment relied on surveys, site visits, and an analysis of the observation data generated during the first year. The districts were a diverse, but not necessarily representative, sample of New Jersey's more than 600 school districts. Many issues that arose among the initial pilot districts may not occur when the whole state of New Jersey begins implementing the new teacher evaluation requirements, in part because of experience gained from the pilot districts. Here the assessment team summarizes what they learned with regard to actual implementation activities, participants' perceptions, and issues affecting implementation.

Implementation activities.

1. *Districts successfully selected teacher evaluation rubrics and the data management tools to accompany them and then provided initial training to all teachers and administrators.*

Most (six) districts selected Charlotte Danielson's Framework for Teaching, with four of these districts using the Teachscape data management tool to collect, store, analyze, and report data. The two other districts used iObservation, a system similar to Teachscape used with the Danielson system in the past. Two other pilot districts selected the McREL Teacher Evaluation System, which has its own data management system. One pilot district chose Marzano's Causal Teacher Evaluation Model, supported by iObservation, and another pilot district selected the James Stronge Model, which also has its own data management system. All districts provided administrators and teachers the required training to get started by January 2012.

2. *Districts observed most teachers at least once, although the number of observations per teacher varied across districts and observations were sometimes unevenly distributed across schools and grades within districts.*

The observation data, as well as the interview and survey data reviewed in this report, suggest that this achievement represents a good faith effort by the districts to comply with NGO requirements, in spite of a late start and the need for everyone to learn how to conduct observations and how to navigate and take advantage of the data systems supporting these observation models.

3. *Because of the press to learn how to conduct observations, inadequate steps were taken to help districts and the whole state assess the quality of observations conducted.*

Two gaps are apparent in the first-year data. Too few observations were conducted with multiple observers to reliably assess agreement among raters, though this was not an explicit requirement of the NGO. In addition, too few teachers were observed more than once to assess the stability of ratings of individuals. The available information suggests that, in some cases, these gaps stem from the difficulty in completing enough observations. However, it is also the case that not all of the teacher evaluation rubrics provided adequate guidance on how to calculate agreement among raters or over time.

Perceptions of the evaluation rubrics.

1. Administrators are more positive about the teacher evaluation rubrics in use than are teachers.

While teachers and administrators agreed on what criteria should be used to evaluate the teaching practice evaluation instruments, they applied those criteria quite differently. More administrators than teachers agreed that the teacher evaluation rubrics in use generated accurate assessments, did so fairly, provided teachers with useful feedback for improving their practice, and separated more and less accomplished teachers. More administrators than teachers also agreed that observers had the knowledge required to appraise teachers and actually gave accurate feedback. This difference may simply reflect the tendency of evaluators to view evaluation rubrics more positively than do those who are evaluated. It may also reflect the fact that the leaders of these districts volunteered to be in the pilot. If they are more positively disposed toward teacher evaluation rubrics than the average administrator, these differences may be larger than would occur in the future.

2. Not surprisingly, the districts did very little to prepare to use teacher observation data to make personnel decisions or to plan collective professional development. The nature of the observation process ensures that districts are practicing giving teachers individual feedback after formal evaluations.

The NGO did not require districts to use the teacher observation data for actual personnel decisions during the first year of the pilot (or to save the data for use in future years). Given the start-up challenges of the first year, which were to be expected, this was a wise decision. As a result, districts focused on getting the teacher evaluation systems up and running and generating sufficient numbers of observations and de-emphasized ultimate uses of the data. Moreover, as might be expected, more attention was given to providing professional development on the teacher evaluation rubric than using teacher observation data to plan future professional development activities. On the other hand, the increased emphasis on formal evaluations with post-observation conferences gave observers greater experience with providing feedback to teachers. For the most part, the comments we heard about this feedback were positive.

Barriers and facilitators.

1. The new evaluation rubrics make great time demands on administrators. So far, actual time demands have stayed the same for teachers.

Administrators reported increased demands from the added observations required of them and the work involved in providing more detailed documentation, as well as the increased demands that come from pre-conferences as well as post-conferences. Teachers and administrators both observed that the time pressure constrained the accuracy of administrators' observations and the adequacy of their documentation and feedback. It also meant that other administrative tasks went undone, were delegated to others (where others existed), or were done after hours. As important as administrative supervision is, school leaders must attend to other issues as well in order for schools to function. Generalizing from the pilot year to the future is difficult. Learning to do the observations and just operate the data management tool took a great deal of time. Yet, most districts did not actually conduct all of the observations that will be required in the future. Every indication is that time management will remain a major problem for administrators. Administrative time constraints have occurred in other districts and states that have implemented teacher evaluation rubrics and seem to have been addressed (to a degree) through redistributing observation work, redistributing other administrative tasks, and working nights and weekends (Boser, 2012; Curtis, 2012; Milanowski and Kimball, 2003).

So far, teachers have been less concerned about the total amount of time that observations take than that they do not see preparing for observations helping them to teach better. However, when each participates in more observations, their time management problems may increase as well.

2. *More administrators than teachers agreed that the training they received helped them to understand the evaluation rubric. The extensive time administrators spent in training may have contributed to their greater appreciation.*

Generally, more administrators than teachers agreed that the training they received helped them understand the overall purpose and approach of their teacher evaluation rubric and the specific techniques for observing, recording, and providing feedback required by the rubric. Several factors may have contributed to their greater appreciation of the training. First, they just received more, as much as four times more, training than teachers, as befits the people who actually conducted the observations. Second, more of their training came in formats that were more conducive to learning. Everyone participated in rushed, didactic training sessions at the beginning of the year when it was important to get the basic facts. However, administrators participated in follow-up sessions in the form of walk-throughs, administrative meetings, and other opportunities that provided more focused coaching and opportunities for collaborative, interactive learning.

3. *Data management tools are essential for implementing teacher evaluation rubrics. Learning to use them was a major challenge that was largely met during the first year.*

Teacher evaluation rubrics generate a large amount of data. When they work well, electronic data management tools help with the actual observations. They also support storing data, communicating results (both to individual teachers and district decision makers), and identifying patterns in those data. However, they also take time to learn, and that process can be extremely frustrating, especially as it happened this year, if the data management system vendors are still learning what their clients need. It appears that most of the learning takes place in the first year.

4. *Teachers' "resistance" in the first year reflected several concerns about the potential accuracy and appropriateness of observation scores that cannot yet be addressed.*

While those who are observed frequently question evaluation rubrics, several concerns that teachers raised could be addressed much more specifically. Teachers frequently argued that the observation scores were *subjective*. While this observation was often taken as an axiom rather than an observable fact, teachers also noted situations where they believed that observers gave the same teachers quite different scores. The absence of evidence of inter-rater agreement within districts in the first year before joint observations were required means that state and district administrators lack a defense for this perception. A second concern is that a new and—to teachers in some districts, arbitrary—*cap on very high ratings* has been put in place. This concern stems from announcements reported in several districts that the number of distinguished ratings would be severely reduced. A third concern is that some criteria used to judge lessons are *inappropriate* for several different reasons. For example, some criteria appear to be required in all lessons when their suitability varies from across student ages, subjects, and instructional activities. Generally, these concerns are less extreme when communication with teachers about the program is more extensive and open. They may also be reduced where teacher-administrator conversations about instruction that occur as part of the evaluation process help teachers to improve their practice. Some administrators have already noted an improvement in the quality of these conversations.

Challenges Moving Forward

Ultimately, for New Jersey's teacher observation requirements to be deemed successful, they must contribute to increased student learning in a manner that makes scaling up to all districts in the state feasible. To do that, they should meet standards of rigor and accuracy for external credibility and also convince their users—that is, teachers and administrators—of their utility so that any recommendations that stem from observation data will be enacted. While it is difficult to know exactly what steps must be taken to reach these goals, three challenges must be addressed:

1. **Time management.** It is hard to see how the new expectations for teacher observation can be sustained and still meet the intent of the law if each observation is as time consuming as has been the case, if extensive numbers of observations are required, if the pool of observers is limited to current school administrators, and if those administrators must continue to do everything they are now doing. Some economies and efficiencies will undoubtedly be found as districts and individual administrators become more adept. However, new approaches are needed. Some are currently being developed among the pilot districts. In the spirit of broadening the discussion of the issue, we offer these suggestions for exploration without endorsing any of them. Some of these are already being tried within the state:

- Videotaping lessons so they can be scored later by a trained observer and on occasion even by two or more to assess inter-rater agreement.

- Hiring out-of-district observers to conduct some observations—this would require that the observers actually be expert at this craft, understand issues specific to the district in which they work, and have face validity with those who are observed.

- Using specially-trained peer teachers to conduct some observations—some districts around the country are experimenting with this option, which, under the right conditions, could help further professionalize teaching.

2. Accuracy. The assessment team lacks the data to assess accuracy of observations on several key dimensions for reasons having more to do with the challenges of implementing the new teacher evaluation programs than because of ill will on anyone's part. Moving forward, several steps could be taken to improve or simply document the accuracy of the teacher evaluation rubrics in use. Doing so, however, will increase the time demands on district observers. Such steps would include:

- Conducting enough joint observations of lessons (whether by having two people in a room or by using videotape) to document the level of agreement among observers will go a long way to demonstrate to teachers and outsiders that the rubric is not subjective. Where data do not support that conclusion, corrective action can be taken. As New Jersey moved into the second year of the teacher evaluation pilot and developed regulations for state-wide teacher evaluation, NJDOE promulgated regulations requiring such joint observations.

- Similarly, conducting enough observations of individual teachers over time to assess the stability of judgments would also help to make (or improve) the case for accuracy of the rubric.

- A number of steps would help to ensure that the distribution of high, medium, and low ratings of teachers documented in the state is, in fact, appropriate. One would be just to make information on what the distribution by expert observers in different kinds of schools looks like available. Another would be to ground judgments about *distinguished*, *proficient*, and *less than proficient* teachers in clearly understood criteria so teachers would know what to aspire to and so the public would know what to expect.

- Finally, it might be important for the authors of different teacher evaluation models to clarify which observation criteria are universal and which should only be observed under certain conditions and make clear what those conditions are.

One cross-cutting issue has to do with the number of accepted frameworks for teacher evaluation in the state. Allowing at least eight different teacher evaluation frameworks to be used places a premium on school district autonomy. In the short run, it allows for considerable local experimentation. However, it also constrains the state from providing enough support to any one framework to truly refine it and use resources from across New Jersey to address the problems raised above.

3. Communication. Indications are that teachers simply do not understand the teacher evaluation rubric as well as administrators do. Constant communication is needed to help teachers see potential benefits and understand how the rubric is supposed to work. This can happen through:

- ongoing administrator communication about the program;
- additional follow-up training for teachers; and
- experience with the program that provides useful formative feedback or opportunities to improve teaching.

Future Program Evaluation Needs

While the assessment of the pilot Teacher Evaluation program conducted this first year has helped to clarify some issues, it has raised others. These include:

1. Exploring differences among districts. On many variables, the differences across districts were substantial. Although not described, the site visit data also noted substantial differences across districts. Past research identifying the importance of local factors to the effectiveness of policy implementation suggests that it is important to understand these differences. We see three possible sources of differences, each of which might lead to a different policy response. For example, some of the differences might stem from the teacher evaluation rubrics that districts selected. If that were the case and some proved to be more accurate and easier to implement, it would suggest that they should be given preference at least by districts, and possibly by the state. Some differences may result from the demographics, finances, or other objective features of school districts. This is very likely given the substantial disparities in size and wealth of the districts in the pilot Teacher Evaluation program. Identifying those differences would help policy makers differentiate the teacher evaluation requirements so they could be designed in ways to be most effective for the local context. Finally, some could reflect local leadership differences. Then, understanding those differences could suggest recommendations for training so that the more effective leaders could share their practices with other districts.

2. Exploring the distribution of observation ratings. Another important task would be to create a document for use within the state that compares more systematically the distributions of observation ratings obtained from pilot districts with those obtained under more exemplary conditions—-for example, in settings where well-trained raters are available. Such a document might clarify for both teachers and administrators such issues as why very thorough training on how to apply observation categories is needed and why the distribution of ratings currently being achieved in New Jersey is too high, if that proves to be the case. With supplementary work, it could also help clarify the meaning of different observation categories and ratings, especially *distinguished*. Finally, it will be important to know if different teacher evaluation rubrics generate different distributions of ratings—that is, some rubrics lend themselves to more or fewer *distinguished* ratings. This will be important information for the state as it considers whether to continue supporting

multiple frameworks for teacher evaluation and, if so, which ones to support, and, if not, which one or which hybrid to adopt.

References

Bill and Melinda Gates Foundation (2012). *Learning about teaching: Initial findings from the Measures of Effective Teaching project*. Washington, DC.

Boser, U. (2012). Race to the Top: What Have We Learned from the States So Far? A State-by-State Evaluation of Race to the Top Performance. Center for Americal Progress. Retrieved from: http://www.americanprogress.org/wp-content/uploads/issues/2012/03/pdf/rtt_states.pdf

Curtis, R. (2012). Building it together: The design and implementation of Hillsborough County Public Schools' teacher evaluation system. Washington, DC: The Aspen Institute.

Consortium on Chicago School Research (2011). Rethinking Teacher Evaluation in Chicago Lessons Learned from Classroom Observations, Principal-Teacher Conferences, and District Implementation , University of Chicago: Urban Education Institute.

Danielson, C. (2011). *The framework for teaching evaluation instrument.* Princeton, NJ: The Danielson Group.

Evans, R. (1996). *The human side of school change*. San Francisco, CA: Jossey-Bass.

Firestone, W. A., & Bader, B. D. (1991). Professionalism or bureaucracy? Redesigning teaching. *Educational Evaluation and Policy Analysis, 13*(1), 67–86.

Fullan, M. (2001). *Leading in a culture of change*. San Francisco, CA: Jossey-Bass.

Marzano R. J. (2011). The Marzano Teacher Evaluation Model, Retrieved from: http://pages.solution-tree.com/rs/solutiontree/images/MarzanoTeacherEvaluationModel.pdf

Milanowski, A., & Kimball, S. M. (2004). The framework-based performance assessment systems in Cincinnati and Washoe. Madison, WI: CPRE.

New Jersey Department of Education (2012). *Notice of Grant Opportunity: Excellent Educators for New Jersey (EE4NJ) Pilot Program Teacher Effectiveness Evaluation System Cohort 2A.* Trenton, NJ: New Jersey Department of Education. Retrieved from: http://education.state.nj.us/events/details.php?recid=17045New Jersey Educator Effectiveness Task Force (2011a). *Notice of grant opportunity. Excellent Educators for New Jersey (EE4NJ) pilot program*. Trenton, NJ. Retrieved from http://www.state.nj.us/education/grants/docs/11-CO01-S01.doc

New Jersey Educator Effectiveness Task Force (2011b). Interim report. Trenton, NJ. Retrieved from http://www.state.nj.us/education/educators/effectiveness.pdf

Podgursky, M., & Springer, M. G. (2007). Credentials versus performance: Review of the teacher performance pay research. *Peabody Journal of Education, 82*(4), 551–573.

Stronge, J. H. (2006). Teacher evaluation and school improvement: Improving the educational landscape. *Evaluating teaching: A guide to current thinking and best practice*, 1–23.

Tennessee Department of Education (2012). Teacher evaluation in Tennessee: A report on first year implementation. Nashville, TN: Author.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project. Retrieved from http://widgeteffect.org/downloads/TheWidgetEffect.pdf

Wright, S. P., Horn, S. P., & 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.

New Jersey Teacher Evaluation,

RU-GSE External Assessment,

Year 1 Report, January 2013

Appendices

Appendix A: RU-GSE Site Visit Guide

Appendix B: RU-GSE Administrator Survey

Appendix C: RU-GSE Teacher Survey

Appendix A

RU-GSE Site Visit Guide

Spring 2012

SITE VISIT GUIDE

Following Patton (2002), this is an unstructured, open-ended interview guide. Questions are intended to cover the major topics to be addressed. The interviewer is expected to adjust question wording and order to maximize rapport with the respondent and ensure that full information is provided. Following the principles of the site visit guide (Yin, 1989), this guide covers all the questions to be asked of each site since the goal of the research is to describe the situation in each school district, and individual perceptions are primarily important for clarifying district conditions. Interviewers will have discretion to ask questions that are appropriate to the respondent's position--i.e., different questions will be selected for principals and district student data coordinators--and to make sure that adequate information is collected across the district if all questions cannot be answered of every respondent.

Individual Interview Guide

TEACHER OBSERVATION SYSTEM

1. What is the district doing to ensure that it gets the requisite number of teacher observations completed for every teacher?
2. What is the district doing to ensure that it gets *accurate* teacher observation scores?
3. What makes it hard to collect teacher observation data?
4. What makes it easy to collect teacher observation data?
5. What helps or hinders your district's efforts to store and retrieve teacher observation data?
6. What have been your most important sources of knowledge about how to collect, analyze, and use teacher observation data**?**
7. What contributes to the accuracy of your teacher observation system?
8. What undermines the accuracy of your teacher observation system?
9. In what ways is the teacher observation system useful for planning supervision, professional development, changes in the curriculum or other things?
10. What could be done to improve the usefulness of the teacher observation system?
11. How is the teacher observation system facilitating or impeding collaboration among educators in this district?

STUDENT GROWTH SCORES (TESTED SUBJECTS)

1. What is the district doing to ensure that it has student growth scores for teachers in tested subjects?
2. What is the district doing to ensure that it has *accurate* growth scores for teachers in tested subjects?
3. What makes it hard or easy to get teacher growth scores?
4. What helps or hinders your district's ability to store and retrieve teacher growth scores?
5. What contributes to the accuracy of your teacher growth scores?
6. What undermines the accuracy of your teacher growth scores?
7. In what ways are your teacher growth scores useful for planning supervision, professional development, changes in the curriculum or other things?
8. What could be done to improve the usefulness of the teacher growth scores?
9. How are teacher growth scores facilitating or impeding collaboration among educators in this district?

TEST SCORES IN UNTESTED SUBJECTS

1. What is the district doing to ensure that it has measures of student growth that it can link to

teachers from untested subjects?
2. What is the district doing to ensure that it has *accurate* measures of student growth that it can link to teachers from untested subjects?
3. What helps or hinders your district's ability to store and retrieve measures of student growth in untested subjects? Measures of student growth linked to teachers?
4. What contributes to the accuracy of your growth measures in untested areas?
5. What gets undermines the accuracy of your growth measures in untested areas?
6. (Only if experience over a year) In what ways are growth measures in untested areas useful for planning supervision, professional development, changes in the curriculum or other things?
7. (Only if experience over a year) What could be done to improve the usefulness of the teacher growth scores?
8. How are teacher growth scores facilitating or impeding collaboration among educators in this district?

Teacher Focus Groups Interview Guide

QUESTIONS ABOUT OBSERVATIONS
1. Thinking about the teacher observations that you have had this year, what made them better or worse than the observations you had last year?
2. What have been your most important sources of knowledge about how to collect, analyze, and use teacher observation data?
3. In what ways is the teacher observation system useful (or not) for planning supervision, professional development, changes in the curriculum or other things?
4. How do you judge the expertise of the person who observes you?
5. How well do you know the people who have observed you this year?
6. How is the teacher observation system facilitating or impeding collaboration among educators in this district?
7. What could be done to improve the usefulness of the teacher observation system?

QUESTIONS ABOUT GROWTH SCORES
8. What progress has the district made in creating a system of growth scores for your students and classes?

STUDENT GROWTH SCORES (TESTED SUBJECTS)
9. How are teacher growth scores facilitating or impeding collaboration among educators in this district?

TEST SCORES IN UNTESTED SUBJECTS

10. How are measures of progress in untested areas facilitating or impeding collaboration among educators in this district?

Appendix B

RU-GSE Administrator Survey

Summer 2012