The Educator Evaluation for Excellence in Teaching and Learning (E3TL) Consortium Evaluation Report

American Institutes for Research
1000 Thomas Jefferson St., NW
Washington, DC 20007

September, 2015

Table of Contents

INTRODUCTION

Teachers are the most important in-school factor with respect to increasing student achievement (Hanushek, 1992; Kane et al. 2006; Clotfelter, Ladd, & Vigdor, 2006), particularly for disadvantaged students (Nye, Konstantopoulous & Hedges, 2004). Surprisingly, minimal research exists on how to identify, assess, and further improve such teachers' skills. Traditionally, state-mandated evaluation systems have not helped identify effective teachers, nor do they foster growth in instructional practice (Brandt et al., 2007; Ellet & Garland, 1987; Loup et al., 1996; Weisberg et al., 2009). However, with the introduction of state and federal programs (e.g., Race to the Top, School Improvement Grants, and the Investing in Innovation Fund, ESEA Flexibility Waivers) that promote evidence-based practices such as use of performance-based teacher evaluation systems, this has begun to change.

The current report is the independent evaluation of one such program, examining the implementation of an Investing in Innovation (i3) grant to the American Federation of Teachers (AFT) to develop and implement a performance-based teacher evaluation system in 10 districts in New York and Rhode Island. In addition to improving teachers' instructional practices and, in turn, student achievement at the school and district levels, the ultimate goal of this project was to identify components of quality implementation with respect to performance-based teacher evaluation systems to assist in scale up of such systems across the country.

AIR Evaluation

In addition to the development and implementation of the performance-based teacher evaluation system, AFT's i3 grant also called for an independent evaluation. The American Institutes for Research (AIR) was tasked with conducting formative and summative evaluations of the teacher evaluation system. Throughout the four-year project, AIR has systematically gathered data from various stakeholders on different aspects of implementation. This report is the final installment of AIR's work: the summative evaluation of AFT's Educator Evaluation for Excellence in Teaching and Learning (E3TL) Consortium development grant.

In collaboration with the Consortium, AIR developed the following research questions to guide the evaluation:

1. To what extent are the teacher evaluator and stakeholder trainings implemented with fidelity to the Danielson model?[1]

2a. To what extent does training reflect best practices in training adults for a professional role?

2b. To what extent do the participants perceive that the training has met their needs?

3. To what extent is the new teacher evaluation system being implemented with fidelity (to the framework) across all study districts?

4. To what extent do districts implement all features of the evaluation system?

5. Do teacher evaluators demonstrate increased accuracy in identifying effective practice and effective teachers?

6. Do teachers in participating districts improve their practices?

7. Does student achievement in participating districts improve?

Throughout the life of the project, these research questions and their underlying constructs have guided the development of data collection instruments. Research questions 1, 2a, and 2b also served as a starting point for AIR's earlier work: a formative evaluation of implementation during the pilot and first year of full scale implementation.

Formative Evaluations

Prior to the current report, AIR conducted a formative evaluation of implementation, developing two reports that provided AFT with information that could be used in refining the system and implementation itself throughout the remainder of the grant.

As part of this process, AIR collaborated with AFT to develop a conceptual framework or logic model of the E3TL Consortium development grant. The model includes all elements of the teacher evaluation system, showing how these contribute to desired project outcomes. The model also includes the data sources and evaluation activities to inform AIR's formative and summative evaluations. Over the course of the four-year grant, we revised the logic model to accurately reflect changes in project scope. Appendix A includes the final iteration of the model as well as a detailed explanation of its components.

The first AIR report focused on the initial, small-scale rollout of the teacher evaluation system during the 2010-2011 school year (the "pilot" year). During the pilot year, New York State United Teachers (NYSUT) and Rhode Island Federation of Teachers (RIFT) began system implementation in a sample of schools from each of the 10 participating districts. The pilot year report presented data on stakeholder trainings on the new evaluation system as well as successes,

[1] During this first year of widespread implementation, most evaluators were trained by the same experts, who provided one centralized initial training for each state. Due to changes made at the grant-administration level, training was no longer based on the FfT model but still followed a structure based off the Danielson framework, which was tailored to each state's standards. This research question was dropped in subsequent years.

challenges, and recommendations for future rollout and continued implementation of the system. Subsequently, AIR developed the second formative report following Year 1 (2011-2012 school year), the first year of full scale implementation. To inform the Year 1 report, AIR analyzed a subset of the data collected to examine teachers' and principals' perceptions of the evaluation system, expanding upon benefits of the system as well as recurring issues that hindered on model implementation. (Information on all data collection activities are presented in detail below.)

In particular, AIR's Year 1 and 2 formative reports presented findings related to trainings provided to teachers, evaluators, and other stakeholders (i.e., specific to research questions 1, 2a, and 2b). We also reported on perceptions of the new teacher evaluation system; successes and challenges encountered; and factors that supported or impeded implementation. The current report draws upon data gathered during all four project years, presenting a comprehensive examination of system implementation from the initial pilot year rollout through Year 3.

Study Sample

Ten urban and suburban districts in New York and Rhode Island implemented the teacher evaluation system. During the pilot year, the AIR team gathered data from a subset of participants from nine of the 10 participating districts, in parallel with the phased-in rollout of the evaluation system.[2] In subsequent Years 1 through 3, we expanded our scope to mirror that of project implementation, incorporating additional schools into the sample. To the extent possible, data collection activities remained the same across years. In instances where changes in implementation and project scope affected data collection, we indicate this below. These changes are also noted in Appendix A.

The AIR team collected data from a range of participants in various roles, including teachers, principal evaluators, and i3 district-level coordinators. Because the performance-based teacher evaluation system was implemented in 147 schools across the two states, it was not feasible to gather data from all stakeholders. For the principal interviews and online teacher survey, the AIR team used a stratified sampling strategy to identify a sample of participants from elementary, middle, and high schools across the 10 districts. Exhibit 1 provides counts and participation rates for several of the data collection activities by year. After this, we list and describe all of the data collection activities.

Exhibit 1. Summary of Data Collection Activities and Participation Rates, by Year

Activity	Implementation Year		
	Year 1 (2011-2012)	Year 2 (2012-2013)	Year 3 (2013-2014)
Evaluator Focus Groups	15 teachers 10 administrators 37 principals	N/A	N/A
Principal Interviews	79 (73%)	67 (63%)	61 (59%)

[2] For focus groups conducted during the pilot year, Providence was not represented because the IRB process had not yet been completed. Providence was the only district that had its own IRB process. The protocols went through IRB at AIR prior to use.

Activity	Implementation Year		
	Year 1 (2011-2012)	Year 2 (2012-2013)	Year 3 (2013-2014)
i3 Coordinator Interviews	10 (100%)	10 (100%)	10 (100%)
Teacher Survey	496 (33.8%)	753 (51.3%)	737 (50.2%)

Data Collection Activities

The AIR evaluation team used both qualitative and quantitative methods to collect data and developed an array of data collection protocols. The team designed (1) an observation and fidelity protocol to document features of the professional development provided to teacher evaluators and teachers during the pilot; (2) a focus group protocol for evaluators; (3) a protocol for semi-structured telephone interviews with school principals; and (4) an online survey instrument for teachers. All interview protocols and surveys were adapted from instruments previously used in large-scale research studies on program implementation, professional development, and school reform. AIR's Institutional Review Board approved all data collection instruments.

Principal Interviews

During the pilot year and Years 1 through 3, AIR team members conducted annual telephone interviews with principals from a sample of schools implementing the teacher evaluation system. Using a stratified sampling strategy, AIR identified and invited elementary, middle, and high school principals from across the 10 districts to participate. The purpose of the interviews was to gather principals' impressions of the implementation of the performance-based system at their respective schools.

The AIR team developed and tailored the semi-structured interview protocol during each year of the project. During the pilot year and Year 1, we gathered information on the schools' then current evaluation systems, principal background information, principal involvement in the i3 process and new system development, teacher training, pilot implementation, and perspectives about the future of the new system. As the project progressed, the protocol also included questions regarding principals' impressions of ongoing implementation, whether or not there was ongoing training, experienced or expected challenges in implementation, and suggested changes to trainings and the system overall. See Appendix B for the final version of the principal interview protocol used during Year 3.

Observation of Teacher Evaluator and Stakeholder Training

AIR team members observed training provided to evaluators, teachers, and other stakeholders on the new teacher evaluation system during the pilot year and Year 1. In February 2011 and again in the summer of 2011 (in preparation for full-scale rollout in Year 1), we attended and observed the weeklong training of teacher evaluators and stakeholders. For this purpose, the AIR team developed a field notes template to document the training features known to characterize high-quality professional development (e.g., form, duration, collective and fidelity of implementation to the planned training activities) as well as enter other important qualitative notes. Observers also recorded basic information each day, including the number of trainees in attendance, the

number of speakers/presenters, and room set up. At the end of each day, observers also reflected on whether or not trainees had an opportunity to provide feedback, the quality of presentations, level of trainee engagement with the content, and the use of technology and materials. See Appendix B for the observation form.

Due to project changes that affected training delivery, the AIR team did not observe trainings on the teacher evaluation system following Year 1. However, principal evaluator interviews as well as responses to the teacher survey gathered data on stakeholders' perceptions of training, provided in varied forms and formats at the district and school levels in Years 2 and 3. For additional detail on how training activities changed throughout the project, see Chapter 2. Training.

i3 Coordinator Interviews

AIR team members also conducted annual phone interviews with 10 i3 District Coordinators. The semi-structured interview included questions about the coordinator role, training provided on the evaluation system in his or her particular district, implementation of the system, and concerns moving forward. See Appendix B for the i3 Coordinator interview protocol.

Teacher Survey

During each year of the grant, a sample of teachers from participating elementary, middle, and high schools from the 10 districts was also invited to participate in an online teacher survey. The brief survey was tailored over time to reflect teachers' initial and then ongoing involvement in the new evaluation system. The survey asked questions regarding teachers' previous experience with evaluation, participation in and reaction to training on the new system, and perceptions of the new evaluation system. See Appendix B for the survey instrument.

Teacher Evaluator Focus Groups

During the pilot year and Year 1, AIR team members conducted focus groups with a sample of teacher evaluators following the fourth day of a week-long evaluator training. Each participating district was represented, and district representatives chose evaluators to participate in the focus groups. As these evaluators were new to the system, the purpose of the focus groups was to document participants' prior experiences with evaluation, gather impressions of the training, and establish a baseline regarding their understanding of the new evaluation process. Data collectors also elicited evaluators' attitudes and beliefs around the new system as well as anticipated challenges in adoption and implementation. The focus group protocol is presented in Appendix B.

District Reports and Student Achievement Data

As part of the summative evaluation, the AIR team requested information from the 10 participating districts. We gathered data on school and student characteristics from each district's central office, including student demographics as well as percentages of students receiving free and reduced price lunch (FRL), students with disabilities (SWD), and limited English proficient (LEP) students.

In addition, for the summative evaluation, we collected state assessment data, aggregated at the district level, for each year of the grant (2010-2011 to 2013-2014 school years). The purpose of this was to examine the potential impact of the teacher evaluation system on student performance

by analyzing assessment scores over time. Perceptions of impact on student achievement—as a result of improvement in teachers' instructional practices due to the performance-based evaluation system—were captured in the principal interview and teacher survey.

Teacher Performance Data

As mentioned, the initial grant proposal also called for collection of teacher performance data gathered using the new evaluation system. The purpose of this was to examine teacher performance over time. The AIR team also planned to compare trends in teacher performance over the four years to student achievement, examining standardized assessment scores of students in sampled teachers' classes across the four years of the grant. Unfortunately, access to teacher performance data was not possible due to privacy concerns voiced by the state-level teaching unions. However, perceptions of impact of the evaluation system on teachers' instructional practice were captured in the principal interview and teacher survey.

Organization of the Report

This report is AIR's summative evaluation of AFT's i3 E3TL grant that supported development and implementation of a performance-based teacher evaluation system in 10 districts in New York and Rhode Island. In addition to results presented at the school and district levels, we also identify higher-level quality implementation components that can provide guidance for others engaged in scale up of performance-based teacher evaluation systems across the country.

The current chapter provided background of the grant and an overview of the AIR formative and summative evaluation. In Chapter 2, we discuss the training component of the grant, examining the extent to which the evaluator and teacher trainings were implemented with fidelity; reflect best practices in training; and met the needs of the trainees. Here, we answer research questions 1, 2a, and 2b. In Chapter 3, we examine implementation of the various components of the evaluation system over the four years of the project, presenting successes of the system as well as challenges, including factors that hindered on-model implementation. Here, we answer research questions 3 and 4. Next, in Chapter 4, we report on impact of the performance-based teacher evaluation system, examining stakeholders' perceived impact with respect to teachers' instructional practice and, in turn, on student achievement. Here, we answer research questions 6 and 7.

Last, in Chapter 5, we identify aspects of quality implementation and lessons learned that can be applied more generally in scale up and implementation of performance-based teacher evaluation systems. These components stem from findings presented in earlier chapters, which have been analyzed further to distinguish particular supports for and best practices in adoption and implementation of similar evaluation systems, which aim to both improve teacher practice and, in turn, student performance.

CHAPTER 2. TRAINING

Evaluators, teachers, and stakeholders participated in trainings on the teacher evaluation system over the course of the AFT E3TL Consortium Project. To gather information about these trainings as well as gauge the extent to which they have prepared those involved for implementation, AIR conducted training observations, principal interviews, teacher surveys, and stakeholder focus groups. These data collection activities were specifically designed to answer the following research questions:

1. To what extent are the teacher evaluator and stakeholder trainings implemented with fidelity to the Danielson model?

2a. To what extent does training reflect best practices in training adults for a professional role?

2b. To what extent do the participants perceive that the training has met their needs?

Throughout the life of the grant, changes in the structure and format of trainings occurred. As a result, AIR modified its training-related data collection activities over time. These changes also affected the extent to which we can answer the research questions set forth in the initial proposal.

In the following chapter, we describe trainings offered related to the evaluation system followed by an overview of AIR's training-related data collection activities. Then, we present findings specific to training delivered during the pilot year. Next, we share results related to evaluator and teacher trainings, including outlining changes in both the nature of these trainings and stakeholders' perceptions throughout the project.

Training Overview

The format of trainings changed from year to year in response to district needs. In the pilot year (2010-2011) and Year 1 (2011-2012), initial training was essential to the roll-out and implementation of the evaluation system. During the pilot year, evaluators and teachers from NY and RI came together as a consortium to be trained. All 10 districts were represented in the pilot, but only a small number of teachers from a few schools in each district participated.

In Year 1 (2011-2012), the number of participating districts remained the same, but additional schools within each district began to adopt the evaluation system. This roll-out was gradual and varied by district. Many districts only implemented the system in specific grade-levels, expanding the scope over the duration of the grant. Subsequently, in Year 2 (2012-2013) and Year 3 (2013-2014), the evaluation system was fully implemented in every classroom and school in each of the 10 participating districts. During the final two years of the grant, districts provided shorter follow-up trainings and recalibrations for experienced evaluators and initial training for new evaluators as needed. In addition, in Year 2 both NY and RI implemented a master coder training during which the highest performing and most consistent evaluators came together to dissect observation videos and pull out exemplars for each rating of each standard on the rubric.

Teacher trainings have typically been provided at the district-level by district union representatives (often i3 coordinators) who were trained during the initial consortium trainings. Principals also provided teachers with ongoing training and support related to the system.

Data Collection Overview

In the pilot year, the AIR team observed the centralized, week-long training for all evaluators and teachers held in Albany, NY as well as a follow-up training for RI participants. In preparation for Year 1, we observed state-level evaluator training in RI and NY. For additional detail and the observation protocol used to capture information during these trainings, see Appendix B. As mentioned, large-scale initial evaluator trainings ended after Year 1.

In addition to observations, AIR staff conducted focus groups with evaluators during the pilot and Year 1 trainings. The purpose of these focus groups was to gather evaluators' perceptions of the evaluation system and trainings themselves. These protocols are included in Appendix B as well.

In Year 2, AIR staff attended the master coder training for select evaluators in NY. As with other trainings, observers used a field notes template tailored to the agenda provided. This observation form is also available in Appendix B.

In the remainder of this chapter, we present a breakdown of training-related findings. First, we report findings from the pilot year training. We purposely discuss these separately from those that pertain to subsequent years of the grant. The pilot was different in that a significantly smaller number of schools participated, and evaluators and teachers were trained together on Charlotte Danielson's Framework for Teaching (FfT). Next, we present subsequent findings within *Evaluator Training* and *Teacher Training* sections. During Years 1 through 3 of the project, members of these stakeholder groups participated in separate trainings, which were no longer based on the FfT. This organization allows for discussion of how training offered to each group changed over the three years of widespread implementation.

Pilot Training

Description

During the pilot year (2010-2011), staff from Teaching and Learning Solutions (TLS) delivered an initial training on the new evaluation system. The trainers were experts in Charlotte Danielson's FtT[3], on which the new NY and RI evaluation systems were based. School administrators and teachers participating in the pilot implementation attended the week-long evaluator training in Albany. Some also attended follow-up training sessions in NY or RI tailored to respective state standards.

[3] Retrieved from http://danielsongroup.org/framework/

The centralized pilot training demonstrated a majority of what the training literature defines as characteristics of a good training[4] and were appropriate for adult learners. On average, the training kept to the planned timeframe, but trainers were also flexible if participants needed more time. Trainers gave participants time to work individually, in pairs or table groups, and as a collective whole. They encouraged discussion of experiences, questions, and concerns, which participants could voice during the day or in a daily survey via email.

Stakeholder Views

During the pilot training week, AIR conducted a focus group with evaluators, which were primarily principals administrators and teachers. During those focus groups, participants reacted positively to the pilot training, considering it challenging but worthwhile. The most common concern was that there was too much information presented in a limited time. Evaluators felt adequately trained to implement the system. One principal in particular explained that the training helped him grow as an educator, an evaluator, and a practitioner.

Discussions with pilot training participants helped identify training successes and aspects that could potentially be improved moving forward. In particular, they noted:

- Training videos were helpful, but participants would have liked more attention focused on how to deal with classroom realities such as disruptive students and interruptions.
- The training provided materials that supported the learning outcomes, although participants wished the binder was better organized.
- Participants appreciated the availability of trainers for private consultation, but they wanted even more interaction (e.g., in the form of personal feedback, observation and evaluation of a mock lesson). They also felt that ongoing trainings would prove beneficial but that actual experience using the rubric would be most helpful. After learning about the different aspects of the process, evaluators were nervous about how they would be able to complete all evaluations given the amount of work necessary for each one.

While evaluators were satisfied with the initial pilot training, some felt that the follow-up training did not provide sufficient coverage of important topics nor provide them with necessary feedback regarding their accuracy in making evidence-based scoring decisions. More training around the scoring process, summative evaluation, evidence categorization, use of multiple student achievement measures, goal writing, professional conversations, and aspects other than the observation protocol were mentioned as desired topics.

Teachers also reacted positively to the pilot training, which helped them look more carefully at their students' cognitive engagement. Pilot teachers felt that they were able to voice concerns and were a part of the development process. Despite these overall positive feelings, teachers did

[4] Garet M. S., Porter A. C., Desimone L., Birman B. F., Yoon K. S. *(2001). What makes professional development effective? Results from a national sample of teachers. American Educational Research Journal, 38(4), 915-945.*

.

express reservations regarding the time consuming nature of the evaluation process; they were unaware of this before attending the training. In one school, two of the five participating teachers quit the pilot because they had not realized how much time it would require. Teachers also wanted to have additional conversations to address their questions, specifically around evidence alignment, how the scoring process works, and timeline of the roll out. Principals expressed related concerns, including that that teachers would not be adequately trained because processes were not yet finalized.

Evaluator Training

Description

In Year 1 (2011-2012), the initial training was rolled out to a broader set of evaluators from all schools within the participating 10 districts. Modeled off of the pilot year training, the purpose of the initial training was to introduce and familiarize evaluators with the system, the rubric and its standards, and how to collect objective, bias-free evidence. All trainings were held in the summer and fall. In the summer, Teaching and Learning Solutions (TLS) staff delivered a centralized, five-day initial training in each state. At these centralized trainings, "apprentices" (evaluators who participated in the pilot) assisted trainers, shared their experiences, and built additional knowledge and skills to eventually take over training in their respective districts. After these trainings, decentralized initial trainings were held for those who did not attend the state-level sessions. In RI, the apprentices began delivering the initial evaluator training, while TLS staff continued to do so in NY. Last, ongoing training during Year 1 varied by state and district. Many NY principals did not participate in ongoing training, while principals in RI received ongoing training and support from district-level union staff, including the district-level i3 coordinator, to whom they could turn with related questions or concerns.

In Years 2 and 3, the training focus shifted from initial training and certification to ongoing training and recalibration. In both states, district-level staff provided ongoing evaluator training, although in RI the state consortium offered some evaluator training as well. District staff provided initial training for new principals as needed. Ongoing training content focused on indicators, scoring, changes to the rubric, and evaluation system logistics. Principals from all districts participated in recalibration exercises held at the district level, though what these involved varied: evaluators in Albany conducted co-observations; in North Syracuse, they participated in informal recalibration work during biweekly meetings; in Providence, they recalibrated formally through Teachscape [5](data collection platform) three times per year and also collaborated with other principals to conduct mock evaluations; and in West Warwick and Woonsocket, they recalibrated as part of their work with so-called "master scorers." Despite such variation in ongoing training, most principals did not receive training around how to evaluate classrooms with SWDs and/or LEP students; some watched a video of effective instruction delivered in classrooms with LEP students and/or SWDs.

Additionally, during Year 2, the consortium conducted one-time, state-level "master coder" trainings. The "master coders" were evaluators were specifically chosen district leaders because they had proven themselves to be the most accurate and consistent observers in their district over

[5] https://www.teachscape.com/

the past two years. The purpose of the master coder training was to (1) create a cohort of evaluators in each state and district who applied the rubric in the same and intended manner, and who could serve as expert evaluators and (2) identify exemplar video clips for each rubric indicator.

Stakeholder Views

Principals described the initial evaluator training as being intense, overwhelming, and including a large amount of information. However, they also acknowledged that such training was essential for quality implementation and ultimately found it helpful. One NY principal recalled, "My initial reaction was, 'This thing is hard, huge and cumbersome.' It was overwhelming. As sessions went on I felt more comfortable in terms of managing the document, but it's lengthy and different, and anything different creates a certain amount of angst." According to one principal, involving the evaluators who participated in the pilot was valuable, as they provided first-hand experiences and insights into implementation.

During the focus groups, evaluators voiced several concerns throughout implementation that contributed to their not feeling fully prepared. Lack of feedback regarding their performance and accuracy was a major issue during both initial and ongoing training. One principal explained, "I got conditionally certified but I didn't know why. We need more information about how we did and the areas we need to work on." Evaluators in NY and RI also stressed the need for more recalibration exercises, even though all took part in some form of these each year. Participating in ongoing training underscored the importance of continued calibration to enhance consistency and accuracy, given that evaluators often interpreted the same evidence in different ways.

Evaluators also noted that they struggled with subjectivity in scoring. They noted that trainers were not always able to tell them specifically why certain lessons were scored certain ways; instead, it was always a discussion. Further, changes made to the platform and evaluation rubric, both during and between implementation years, also presented a challenge. The timing of training was also not always ideal. For example, in RI, new principals were trained immediately before schools started testing – a difficult time for any principal to be out of school, let alone new principals. Lastly, principals expressed the desire for additional guidance on the use of evidence not found in the classroom, summative conferencing, and the evaluation of non-classroom teachers (e.g., librarians, physical education teachers).

Those evaluators who attended master coder training provided positive feedback, lamenting the fact that all evaluators could not participate. One evaluator stated, "This has really helped [boost] my confidence, and I really think we should get together as a group once a month and watch a video together. I was very overwhelmed before at the thought of conducting observations next week." Participants praised the detailed discussions specific to each indicator, which provided them with much needed clarity. They also appreciated the collaborative nature of the training, allowing them to discuss the intricacies of the rubric with others in and outside of their districts. Principals also mentioned wanting more informal collaboration with others regarding determining ratings based on evidence.

Despite the aforementioned concerns, in general most evaluators noted that they felt well-supported and adequately trained, particularly after gaining first-hand experience using the system. Evaluators in NY who participated in the pilot attributed their preparedness to their

experience on the ground more so than to formal training received. Throughout implementation, principals across the states similarly noted that they learned the most by "muddling through" and completing evaluations in the field. Principals, particularly in RI, also mentioned that their respective i3 coordinators were helpful. For this reason, a higher proportion of principals felt that they were adequately prepared to conduct evaluations each subsequent year. Teachers' perceptions corroborated this: approximately 80 percent of teachers agreed that their evaluators were appropriately and adequately trained each year.

Teacher Training

Description

Teacher training on the evaluation system looked different in Year 1 than in Years 2 and 3. Year 1 was a gradual implementation year, thus principals did not evaluate all teachers and those they did evaluate varied by district. For example, principals used the i3 model to evaluate only teachers in grades 2, 3, and 4 while, in another, they focused on teachers without tenure. As a result, not all teachers were trained in the system until Year 2. The purpose of these was to introduce teachers to the new system and rubrics as well as provide general logistical information. Across years and states, most teacher training was provided at the district and/or building level. District-level trainings, provided by district or union staff, were optional, and many teachers did not attend. This resulted in most (if not all) training being delivered by principals. In Year 3, for example, teacher trainings in NY tended to focus on standards and the evaluation rubric, while those in RI tended to be on the system as a whole and student learning objectives (SLOs).[6] As with evaluators themselves, many indicated that teachers learned most about the system through first-hand experience; like with evaluators, those who participated in the pilot felt the most prepared.

Stakeholder Views

According to data from principal interviews and teacher surveys, the degree to which teachers were adequately trained on the evaluation system varied. One reason for this was that not all teachers attended training since it was often not mandatory. Those who did attend had mixed reactions, though generally in-house training was better received than district training. Some teachers reacted positively because the training alleviated anxiety many felt related to the evaluation, providing them with information about the system, process, and use of ratings. Teachers who responded positively indicated that the training helped them improve their practice and engage in open conversations with evaluators. Others were overwhelmed by the amount of information provided and anxious about how ratings might be used. This was not surprising, given that teacher trainings typically did not cover the rubrics in detail. Those who responded negatively described the training as unhelpful or simply were not interested because they saw the initiative as yet another initiative that would not last. As was the case with some evaluators, timing of teacher training was also an issue: some were trained long before implementation began. Each subsequent year, principals indicated their teachers were better prepared for the system, this likely the result of more exposure and experience with it more than any other factor.

[6] While SLOs were not part of the teacher evaluation system, whether or not they were met was considered an indicator of effective instruction and, in turn, used to determine scores.

Last, training teachers received as a result of their evaluation scores also varied by state and district. In Year 2, the majority of evaluators indicated that there was no specific formal PD given to teachers after receiving "developing" or "ineffective" scores in a particular area, although principals across states noted that development of a formal plan was underway. However, the majority of principals across states indicated that teachers did receive some sort of support based on evaluation scores (one-on-one mentorship, coaching, classroom visits, etc.) For example, Albany provides Peer Assistance Review teachers who provide support to those found to be ineffective or developing, while North Syracuse provides a consultant teacher. Evaluators in RI used scores to provide teachers with recommendations for additional professional development. Across states, several principals noted that they considered teachers' evaluation scores to make strategic decisions regarding what PD to offer the following year. Relatedly, principals in Plattsburgh expressed concern regarding the amount of money needed to provide PD to teachers who were found to be ineffective.

Overall, stakeholders appeared fairly satisfied with trainings offered on the evaluation system. Evaluators stressed the need for additional calibration opportunities to ensure that all were implementing the system in the same and intended way. Teacher trainings were often not mandatory, resulting in differences in their knowledge of the system which, in turn, may have contributed to anxiety and lack of buy-in around the evaluation system. Also, given that most teacher training was delivered by principals at the school level, these sessions varied extensively from site to site. Throughout the pilot and three years of the project, it became increasingly clear that experience and use of the evaluation system was the factor that contributed most to stakeholders' knowledge of and comfort around use of the system.

CHAPTER 3. IMPLEMENTATION

In this chapter, we report on implementation of the teacher evaluation system in the 10 participating districts. Specifically, we address the following two research questions:

3. *To what extent is the new teacher evaluation system being implemented with fidelity across all study districts, reflecting standards set for selecting evidence of practice (i.e., teaching) and procedures set for collecting evidence?*

4. *To what extent do districts implement all features of the evaluation system (e.g., system is rigorous, transparent, and fair and uses multiple rating categories and multiple measures of effectiveness)?*

Below we present findings related to fidelity and the extent to which district and school staff implemented the system according to the intended model. Then, we discuss stakeholders' perceptions of whether implementing the system as intended provided evaluators with an accurate representation of teachers' everyday instruction.

Fidelity of Implementation

To determine the extent to which the teacher evaluation system was being implemented with fidelity across all study districts, the AIR team gathered various implementation data. We define fidelity in a number of ways, including whether all classroom teachers in participating schools within the 10 districts were evaluated; whether all teachers implemented instructional best practices; and the extent to which evaluators used the rubric to evaluate teachers with SWDs and ELL students. After this, we discuss two aspects of the model that proved to be problematic in regards to on-model implementation: scripting challenges and the multi-part, time-consuming nature of the process itself.

Evaluation of All Classroom Teachers

The proportion of classroom teachers evaluated per school increased over the course of the grant. In Year 2, 69 percent of New York principals and 93 percent of Rhode Island principals indicated that the evaluation system was fully implemented at their respective school. In Year 3, all teachers across the 10 districts were evaluated. However, some principals reported conducting fewer observations per teacher because they could not find the time to complete all required steps. In these cases, principals conducted the formal observation, evidence from which is used to determine the teacher's effectiveness rating, but would skip one or both informal observations. Despite this deviation, nearly 90 percent of teachers across states and years reported that they received timely feedback from their evaluators. (For additional details regarding time as a factor that hindered on-model implementation, see below.)

Effective Use of Best Practices

Principals also emphasized another aspect of fidelity: implementation of instructional best practices such as effective engagement of all students to meet or exceed learning standards and teachers using appropriate formal and informal assessment strategies with individuals and groups

of students to determine the impact of instruction on learning. Many identified teachers' failure to deliver effective lessons during the observations as a major reason the system would not realize instructional changes nor boost student achievement to the extent possible. As intended, implementation of the system helped identify potential areas for improvement and overall informed teachers' PD needs to a certain extent, although the percentage of teachers surveyed who agreed with this statement dropped from 63 percent to 48 percent from Year 1 to Year 3. Similarly, when asked about the specific feedback they received from their evaluator, approximately 65 percent of teachers across states agreed that this personally helped them identify PD needs. As some principals explained, identification of needs is the first step towards providing teachers with the PD that they need to effectively implement best practices.

Some principals noted that while changes made since the simultaneous adoption of the evaluation system and CCSS were certainly on the right track in terms of effecting change with respect to improved student outcomes, much still needed to be done. Several principals explained that once teachers truly engage and understand the elements of the evaluation system they will be able to implement methods as intended and, as a result, this will boost student achievement. While evaluator-teacher conferences provided some support, many noted that more extensive and ongoing PD was needed to ensure teachers had the requisite knowledge and skills to integrate best practices into their instruction. Until teachers received more in-depth PD, fidelity in this regard could not be achieved and no widespread change among teachers nor students would be evident.

Evaluating Teachers of Special Populations

The extent to which principals followed the rubric to evaluate teachers of SWDs and LEP students varied considerably. A few principals felt more prepared using this system than any previous tool to evaluate teachers serving SWDs and/or LEPs because the rubric removes subjectivity, they explained. However, many others did not feel prepared because they did not receive guidelines specific to evaluation for teachers of these populations, and the rubric does not accommodate instruction geared towards such students. For example, regarding use of the rubric with teachers of SWDs, the level of students' disability dictated the type of instructional activities teachers could use, which in turn affected their ratings. About 11 of the 61 principals noted that use of the rubric with no adjustments for instruction of such populations was unfair, principals explained, and put these teachers at a disadvantage in terms of their ratings. Moreover, the rubric was not designed to capture the effective aspects of instruction these particular teachers were implementing, which may have been appropriate for special populations but would not effectively serve those in general education classrooms. For example, one principal explained, "The nature of those constraints doesn't lend itself to most of the rubric, especially for LEP[s]. They struggle with language not intelligence. The entire system is based on the assumption that everyone understands the language. You could have effective questioning but they don't understand the language so it doesn't mean anything."

Because evidence from classrooms with SWDs and ELLs does not always correspond to the rubric, certain principals noted that they have not been able to hone their skills in evaluating such instruction depending on their school population. Further, some principals explained that they do not score these teachers on the rubric itself, make unofficial modifications, and/or use professional judgment because it is unfair and inappropriate to use the rubric with these teachers.

Most evaluators have not received any training regarding how to use the system with these populations but desired this support. Because of this, evaluators have not become more comfortable or prepared using the tool in classrooms that include SWDs and/or LEP students.

Thus, the extent to which the system was implemented with fidelity varied depending on how this was defined. In addition, there were aspects of the system that proved particularly challenging for execution of on-model implementation. These included scripting and the multi-step evaluation process itself, particularly the administrative aspect, which proved to be very time-consuming.

Challenges to Full Fidelity

Scripting Difficulty

Participants reported that the extensive note taking process required for accurate and informative evaluations was time-consuming and inefficient. During the formal observations, evaluators must take extensive notes about the teacher's instruction to ensure that their evaluation reflects what they saw. Evaluators described this note-taking process, known as "scripting," as problematic throughout the initiative. Over half of the principals had difficulty scripting a comprehensive account of instruction during the observation because they were unable to write or type fast enough. One principal, who was told to practice typing faster, said "I'm an administrator for a reason, not a secretary." These principals reported that scripting challenges impeded their ability to record all relevant observation data, but without gathering all potentially-useful evidence, evaluators' ratings may not have been fully reflective of the instruction they witnessed.

Time-Consuming, Multi-Step Process

The most cited challenge with respect to implementation is the time required to complete the evaluation because it includes many time-consuming steps (including setting up a preconference, conducting the observation and conducting a post conference). The process presents a large administrative burden that detracts from other principal duties. Principals reported that completing the full evaluation process for each teacher was extremely difficult without either abandoning other principal responsibilities or not completing all aspects of the evaluation. One principal explained, "There is no feasible way to run a building and do more than one evaluation a day. It is insane and it makes it very challenging." Principals reported that it took up to 20 hours to go through the full process for one teacher, and some principals had 20 teachers to evaluate. As a result, principals reported that completing such a time-consuming process for each teacher was not feasible, and without additional staff to conduct a portion of the evaluations, principals knew they would not be able to implement the system as intended. For example, time constraints precluded some principals from conducting all of their summative conferences, despite the fact that nearly all evaluators viewed this as a very important part of the process.

The heavy administrative demands this evaluation placed on evaluators was frustrating and impeded principals' other efforts to realize instructional change. Principals have spent large amounts of time working to meet administrative demands of the evaluation, which has detracted from frequent instructional support they previously provided to teachers. One such principal stated, "I am not doing my job as a principal. I am now a full time evaluator." Instead of engaging in rich conversations on strategies and ways in which teachers can tailor their lessons,

areas in which many principals have expertise, they have had to greatly reduce their involvement to ensure the time-consuming administrative pieces of the evaluation are completed. Many principals commented that the excessive time commitments were frustrating and stressful. One opined, "I think it's a waste of time and effort. I've lost contact with my teachers and kids, and I have no personal life."

Despite the fact that principals and teachers alike described pre- and post-conferences as beneficial in that they provided a forum for thoughtful discussion (see Chapter 4. Impact for additional information), many principals felt that their new evaluation duties precluded them from serving as instructional leaders. Perhaps this is why teachers were divided regarding whether or not they thought the system held principals accountable for fulfilling their role (i.e., meeting with teachers, providing feedback, responding to teacher needs). Throughout the duration of the initiative, approximately 60 percent of teachers agreed that the system held principals accountable. However, in Year 3, teachers in NY and RI differed regarding the extent to which they agreed with this: while approximately two-thirds of NY teachers surveyed agreed that the system held principals accountable, only about a half of RI teachers did.

Accurate Reflection of Teachers' Work

Another theme that emerged was the extent to which the system was fair and accurately reflected teachers' instruction. In Year 1, most evaluators thought the evaluation process allowed them to apply the system to all teachers equally. Using the same process, rubric, and standards provided a standardized way to rate teachers of all subjects. By contrast, others (albeit fewer) expressed concerns that the evaluation system did not treat teachers equally across grades, subject areas, or classrooms with special student populations. This was a growing concern over the course of the grant, which was reflected both in principal interview data and teachers' survey responses. In Year 1, approximately 73 percent of teachers across NY and RI agreed or strongly agreed that the system was consistent and objective. However, by Year 3, only about 51 percent of teachers thought of the system in this way—a notable decrease of 22 percentage points, roughly one-fifth of all teachers who participated in the survey.

Similarly, principals grew more vocal with their concerns about the rating system over the course of implementation. Principals across both states initially thought the rating system accurately reflected teachers' work. For example, principals said the system highlighted strong and weak practices, such as if a teacher lacks classroom control or has a classroom environment that fosters learning. However, in the later years of the initiative many principals explained that teachers were receiving low scores because it was almost impossible for them to touch on all parts of the rubric during a single lesson (i.e., one-hour formal observation) and that the rubric's 39 indicators were far too many. Despite their concerns regarding the number of indicators, principals did not think the rubric quite captured all important teacher qualities, such as the social emotional work teachers do. Therefore, teachers would receive low ratings on the rubric, which some principals explained did not take into consideration important aspects of instruction. The inability of the system to capture all important aspects of teachers' work may have contributed to their response to a particular survey question: when asked whether the evaluation was an accurate reflection of their work, only about half of teachers surveyed thought this was the case (ranging from 59 percent to 44% across the three years of full implementation). Still, most teachers thought the system did measure important instructional skills. Agreement ranged from

87 percent to 73 percent in Years 1 and 3, perhaps due to teachers' growing familiarity about specifics of the rubric over time and realization that it indeed did not include all of what they do in the classroom.

Principals also expressed their dissatisfaction with other aspects of the rubric. One concern was related to ambiguity in wording and potential differences in evaluators' interpretation of the terms "frequently" versus "sometimes." Because the system did not provide definitions for these frequency ratings, evaluators may not have applied these consistently. During Year 2, some principals commented that whether a teacher received an ineffective or developing rating depended on how the evaluator interpreted the meaning of these terms, which had little to do with the instruction they observed.

Another common concern was related to use of the same standards and criteria across teachers of all subjects. Principals did not think that the same standards could not be used for both art and history teachers whose lessons vary tremendously. For example, a physical education teacher's lesson and content may not lend themselves to use of 21st century skills like those of an ELA teacher. However, these issues were all secondary to principal's concern that the system did not allow evaluators to capture and evaluate teachers' typical instruction.

Formal Observations Not Representative of Typical Instruction

Principals reported seeing major differences in instruction when conducting announced formal observations compared to unannounced informal observations. Principals soon realized that for formal observations teachers were preparing specific lessons that would yield effective scores on the rubric. The majority of lessons principals formally observed were well planned and closely aligned with the rubric. Some noted that students behaved better than normal. By contrast, during informal observations, they often did not witness these excellent lessons, effective teaching practices, and good student behavior.

Although principals voiced serious concerns about the time consuming nature of the evaluation process as a whole, many explained that merely one formal observation was not an accurate representation of a teacher's everyday instruction. Many principals claimed that prior to using this tool they knew who their strong and weak performing teachers were. However, in some instances they would observe a strong teacher who was having an off day or was sick, which affected the delivery of their lesson. As a result, even though this one observation was not representative of that teacher's practice, the teacher would likely receive a lower than expected rating. For these reasons, many principals recommended having more informal and fewer formal observations so that observations, and in turn their ratings, would be more indicative of teachers' daily practice.

Last, these reasons for less-than-accurate evaluations, articulated by principals, may also have contributed to teachers' opinions regarding the accuracy of the evaluation system. As mentioned, only approximately half of teachers surveyed agreed that the system accurately reflected their work in the classroom. Interestingly though, a substantially higher percentage of teachers across years (e.g., approximately 76% across states in Year 3) reported that their own self-assessments were either usually or always consistent with their evaluation ratings.

Concerns about SLOs Not an Adequate Measure of Effectiveness

Student learning objectives (SLOs) are measurable, year-long instructional goals set by teachers for groups of particular students, which principals then must approve. SLOs are one way in which principals assess the effectiveness of teachers' instruction via an examination of student growth. Establishing these goals is a state mandate and not connected to the evaluation system. However, the evaluation system incorporates teachers' respective SLOs as an indicator. Because the SLOs provide a concrete measure of student performance, principals are able to use these data to assess teachers' ability to promote student growth. As such, some principals have mid-year conferences with teachers that are used as check-ins to determine if they are on track to meet their SLOs. By year's end, the principals are able to determine if teachers have met or exceeded their SLOs and can incorporate this information in the teacher's overall effectiveness rating.

While some principals expressed satisfaction with use of the SLOs to inform their assessments because the SLOs allow teachers to easily track students' progress, this was often problematic because many teachers were not setting rigorous objectives. Principals noted that teachers were reluctant to set challenging SLOs because failure of students to meet these goals impact their effectiveness rating. With non-rigorous SLOs, meeting or even exceeding such goals does not provide meaningful information regarding whether or not the teacher delivered effective instruction. In response, some principals suggested that until the subjectivity can be removed from the SLOs, they should not be part of the evaluation system because the way in which they are used is not only inconsistent but unfair.

Overall, the extent to which the evaluation system was implemented with fidelity varied. By Year 3, all teachers across the 10 districts were being evaluated, and the proportion of teachers evaluated increased each year of the grant. However, some principals spoke about fidelity with respect to implementation of instructional best practices, featured on the rubric itself. In many cases, this was not happening to the degree they would have liked. Over the course of implementation, principals voiced growing concern that the system did not allow for a fair evaluation of teachers serving SWDs and/or LEPs. These teachers often had to adjust instruction to best meet these students' needs, but in these cases they received lower evaluation scores based on contents of the rubric. Overall though, evaluators felt that the system allowed them to accurately reflect teachers' instruction given its systematic nature and multi-faceted rubric. However, the one-time, announced observations as well as subjective, non-rigorous SLOs (used as a criterion) were potential reasons why the system did not provide as accurate of a reflection of instruction as possible.

CHAPTER 4. PERCEPTIONS OF IMPACT

In this chapter, we report on stakeholders' perceptions of changes in teaching and learning in the ten districts in NY and RI that implemented the E3TL evaluation system. Prior to the start of the project, AFT identified the following project outcomes: positive changes in stakeholder attitudes regarding the purposes and potential uses of teacher evaluation and buy-in from stakeholders; increased accuracy in identifying effective practices and teachers; an increase in the percentage of teachers meeting the standards over time; and an increase in student achievement and a closing of achievement gaps between student groups.

It is important to note that many of the other drivers of change were similar across schools, districts, and the two states. With respect to teachers' skill building, principals mentioned school- and district-level PD unrelated to the evaluation itself, school-based professional learning communities (PLCs) and data teams, and administrator-provided instructional support. Of course, concurrent rollout of the CCSS played a large role in changing teacher practice, and it was through use of the evaluation rubric that this change could be measured to a certain extent. Principals named other structural changes that contributed to improvement as well, including adoption of the response to intervention (RtI) model and introduction of team teaching, both of which have led to increased differentiation in particular schools. Further, many of these schools and districts have been pushing widespread shifts regarding specific best practices in instruction (e.g., differentiation, data-based decision making) for several years, prior to the introduction of the evaluation system or CCSS. While many (if not all) of the changes correspond to elements of the rubric, more generally they are components of good teacher practice and, because of this, were school- and district-level foci before adoption of the system.

To guide the evaluation and measure whether the evaluation system produced the desired outcomes, AIR developed three research questions with respect to impact. These included:

5. *Do teacher evaluators demonstrate increased accuracy in identifying effective practice and effective teachers (e.g., more precision in using the evaluation framework and in conducting observations and conferences)?*

6. *Do teachers in participating districts improve their practices during the implementation of the evaluation system (e.g., increased percentage of teachers receiving higher ratings by teacher evaluators)?*

7. *Does student achievement in participating districts improve (e.g., change in the level of student achievement as measured by state assessments in ELA and mathematics)?*

This chapter is organized into three major sections: impact on evaluators, teachers, and students. Results are presented based on an analysis of teacher survey, principal interview, and student

achievement data.[7] In many instances, recurring and common themes were found across both states. In the sections below, we discuss these findings at a high, cross-state level when applicable and make state-specific distinctions as needed.

Impact on Evaluators

One goal of the project was to train teacher evaluators to accurately assess teaching performance and assist teachers in improving their practice. The findings reported in this section are based on principal interview and teacher survey data regarding perceptions of evaluator accuracy.

General Perceptions of Accuracy

By Year 3, principals in both states felt as though their skill in identifying effective practices and teachers had improved since adoption of the evaluation system. One NY principal explained, "[The system] had a dramatic impact on my evaluative skill set." As of Year 3, approximately two-thirds of principals interviewed indicated that their participation in the evaluation system helped them, in some ways if not overall, to better recognize effective teachers. This was consistent across both states.

Evaluators noted that actually gaining first-hand experience implementing the system was critical to increased accuracy in ratings. Continued use has also helped evaluators become more adept at navigating the rubric and familiar with language of the tool itself. Over the past three years, evaluators have become more comfortable with the system, streamlining their processes and conducting what they describe as objective evaluations that provide actionable feedback. Collaboration with other administrators and evaluators as well as periodic calibration exercises contributed to increased accuracy as well, providing evaluators and teachers within and across sites with a common lens through which they view and discuss instruction.

Evaluators also spoke about specific aspects of the system and model that have contributed to their effectiveness in conducting evaluations. Over 20% of principals explained that the process of recording evidence and aligning it with the rubric allows them to better capture the lesson, which in turn facilitates accurate scoring. Further, use of the rubric has made evaluators more attuned to elements of effective instruction, and they are now more focused on what to look for during the lesson (e.g., questioning techniques, strategies that foster student collaboration, formative assessments). While evaluators may have had a sense of which teachers were very effective, the system has helped quantify what many administrators already knew and provided a foundation for productive discussion. Conversely, some principals also mentioned that the rubric enables them to better identify less effective teachers, helping tease out areas of weakness on which to focus.

In addition to increased accuracy, evaluators noted that the use of the system allowed them to learn more than before about each teachers' strengths and weaknesses. Principals explained that they got a more nuanced sense of each teacher's capabilities, for example, identifying areas of need in teachers they otherwise thought were very effective. Overall, implementation required principals, as evaluators, to spend time in classrooms, which a number of principals hailed as a

[7] There were some limitations with respect to planned data collections, which limited the extent to which we could respond to the original research questions. These limitations are noted where they apply.

benefit of the system. (Interestingly, other principals felt that the time-consuming administrative aspect of the evaluation precluded them in general from spending as much time in classrooms as they previously had.) Overall, principals explained that they had a better sense of what was happening in classrooms because of the evaluation system. Despite increased skill level in identifying effective teacher practice, many still mentioned room for growth, explaining that they are consistently identifying ways to improve as evaluators as they implement the system with increasing competence.

Conversely, a handful of evaluators did not think use of the system helped them more accurately identify good instruction nor effective teachers. Long-time principals in particular noted that they already knew which teachers were effective and did not feel they needed the system to determine this. Evidence from walkthroughs, student data, test scores, attendance, and referrals informed their perceptions. Nevertheless, some of these principals acknowledged that the system provided an additional set of criteria to help paint a more nuanced picture of instruction than they previously had.

Evaluating Teachers of SWDs and ELL Students

Another goal of the teacher evaluation system was to develop and incorporate standards of effective practice for teaching SWDs and LEPs in general education classrooms. Many principals spoke about how best practices featured on the rubric are nonetheless beneficial for SWDs and ELLs. A majority of principals noted that increasing familiarity with the rubric has better equipped them to evaluate teachers in general, including those in general education classrooms who serve SWDs and ELLs students. Asked about evaluating instruction in those classrooms, one principal explained, “Good teaching is good teaching, regardless of the student population.” Of those principals interviewed in Year 3, 18 percent felt more prepared using the current evaluation system than previous tools when conducting evaluations in classrooms with LEPs and SWDs; an additional 28 percent felt more prepared in some ways but not in others. A few principals also mentioned that their respective districts organized professional development on use of the rubric in classrooms with these populations and how teachers can modify practices in mixed classrooms to perform well on the evaluation. Sessions focused on how certain elements of the rubric (e.g., questioning, students’ self-assessment) should look with these populations. While this is an ongoing process, administrators in the few districts where such discussions have begun report experiencing marked progress on this front.

However, many principals did not feel they had acquired the knowledge and skills required to evaluate teachers of SWDs and/or ELL students because they have only used the rubric in its current form, which does not accommodate instruction geared towards those types of learners. Of those interviewed in Year 3, 41 percent did not feel more prepared using the current system than they had using previous tools. (For a discussion of how this affected fidelity, see *Chapter 3. Implementation*). Principals’ opinions varied based on their school populations and experience.

Impact on Teachers

The evaluation system was designed to improve teachers’ instruction and, in turn, increase student achievement. We investigated the extent to which teacher knowledge of and attitudes regarding the evaluation system have changed and perceptions about whether instruction itself

has improved. In this section, we present findings from the Teacher Survey and principal interviews to address these outcomes, identified at the start of the project. First, we report on teacher attitudes and changes in how they thought about the system across years. Then, we discuss the extent to which evaluators perceive an improvement in teachers' instruction, factors that may have contributed to this change in practice, and possible impediments to this change.

Changes in Teacher Attitudes

Teachers' self-reported knowledge and understanding of the evaluation system was fairly high and remained relatively stable throughout the three years of the initiative.[8] Across states and implementation years, approximately three-quarters of teachers surveyed felt as though they had received sufficient information about the evaluation system. With respect to NY teachers, 16 percent more teachers agreed or strongly agreed that they had received sufficient information in Year 3 than in Year 1, possibly indicating an increase in communication around the system in NY, or that teachers simply were more familiar with the system after multiple years of implementation. Overall, most teachers also reported that they understood the various components of the system: data suggest a slight increase in understanding from Year 1 to Year 3, as 74 percent and later 82 percent of teachers agreed or strongly agreed that they understood the components. At the state level, these numbers were fairly stable for RI teachers but point to a possible increase in understanding among NY teachers (agreement with this particular survey item increased 15 percent from Year 1 to Year 3).

Overall, approximately 70 percent of teachers surveyed reported understanding the uses of the evaluation system, and this number remained fairly stable across years. In RI specifically, however, teachers' agreement that they understood the uses decreased by approximately 11 percentage points from Year 1 to Year 3, which may be indicative of a slight decrease in teacher awareness in that state. Further, when asked about specific uses of the system, teachers were also generally consistent in their endorsement of particular uses across years, as seen in Exhibit 2.

Exhibit 2: Understanding Regarding Use of the Evaluation System

The evaluation system in my district is used for…	2011-12 (N=490)	2012-13 (N=749)	2013-14 (N=735)
Informing/improving instruction	83.3%	78.1%	73.7%
Improving student learning/achievement	70.8%	72.9%	63.1%
Creating reflective teachers	64.3%	66.4%	59.3%
Informing PD for teachers	33.9%	37.1%	38.8%
Informing salary decisions	1.0%	1.7%	0.3%
Informing bonus/monetary, non-salary decisions	0.6%	0.7%	0.5%
Deciding on non-renewal of teachers	35.5%	35.9%	29.7%

[8] As noted throughout the course of the evaluation, response rates on the Teacher Survey were low across years and particularly in Year 1 (33.8%). As such, we cannot assume these or other survey data are representative of the districts or the states as a whole but suggest teachers' attitudes regarding aspects of the evaluation system throughout implementation.

The evaluation system in my district is used for…	2011-12 (N=490)	2012-13 (N=749)	2013-14 (N=735)
Deciding on teacher promotion/tenure	38.2%	24.0%	21.2%
Identifying teachers for leadership	7.1%	5.3%	6.8%
Other	4.3%	7.5%	7.5%
Don't know	6.3%	7.2%	11.8%

Teachers in both states endorsed "informing/improving instruction" and "improving student learning/achievement" as the two most prominent uses of the evaluation system, and these numbers were relatively stable across years. This is noteworthy, as these were and are the two main purposes of the initiative. In terms of the survey data, there was a slight decrease from Year 1 to Year 3 in teachers' endorsement of these uses. However, neither decrease is particularly large. They suggest that teachers may have been less aware of these key uses as time went on, or that the system was not effectively being used to improve instruction or student learning. In RI specifically, approximately 12 percent fewer teachers agreed or strongly agreed that the system was used to improve student learning and achievement in Year 3 as compared to Year 1.

Lower levels of agreement regarding uses of the evaluation system for making decisions about teacher promotion and monetary rewards were also in line with the intended uses (and non-uses) of the system. Teachers were generally knowledgeable about things for which the system was not used, as evidenced by these data. Specifically, notable decreases from Year 1 to Year 3, such as that regarding use of the system to make promotional and tenure-related decisions, suggest that teachers learned about and grew more familiar with the system and its uses over time. This may also be an area in which more or clearer communication on the specific uses (and non-uses) of the system could have been shared from the start of implementation. In addition, while most responses were very similar across states, agreement with this statement regarding use of ratings to make promotional and/or tenure decisions varied considerably by state in Year 3: 32 percent of NY teachers surveyed thought that the evaluation was used to decide on promotions or tenure status, whereas only 12 percent of RI respondents thought ratings were used in making such decisions.

Overall, teachers' survey responses reflected the teacher's view of the intended uses and non-uses of the evaluation system. According to the survey, identifying changes in instructional practice was reported as the number one use. Nearly 80 percent of teachers agreed that evaluator feedback helped improve their instruction. Interviews conducted with these evaluators provided additional detail regarding how teachers' instruction changed over the course of the grant, findings from which we present next.

Evidence of Instructional Change

Increased Awareness of Best Practices

Since adoption, teachers in NY and RI have become increasingly familiar with the detailed evaluation rubric, which presents effective instructional practices and provides tangible examples. In Years 2 and 3, principals note that the rubric itself is a useful tool because it breaks down and clearly presents elements of effective instruction, and teachers have benefited from such a comprehensive, actionable inventory of best practices.

Moreover, having a shared set of standards and evaluation criteria from which teachers and evaluators work has established shared expectations. Teachers are more mindful of best practices and aligning their activities to established standards. Teachers' consistent and high level of agreement that the system clearly addressed their performance based on established goals (82-85% across years) and clearly communicated standards (82-77% across years) speaks to this positive aspect of the system as well. Further, pre- and post-observation meetings have allowed for collaborative goal setting between evaluators and teachers, reinforcing and fostering practice in the use of effective instructional strategies. In Years 2 and 3, continued use of the system has helped clarify the differences between characteristics of an "effective" versus "highly effective" teacher as well. While principals reported variation in the extent to which these practices, new for many teachers, were being implemented, it is clear that adoption of the system has engendered a major shift in mindset and acknowledgement of the importance of these best practices in increasing student achievement.

Professional Conversations

Teachers and evaluators engage in one-on-one, pre- and post-observation conferences. During the pre-observation conference, they discuss instructional plans. During the post-observation conference, they discuss whether the teacher's instruction incorporated various elements of the rubric, met standards, and how s/he might improve. In both NY and RI, principals overwhelmingly mentioned the reflective nature of the discussions as one of the main ways in which the system has begun to improve teacher practice. One principal stated that the new process is "more reflective than any other evaluation that we've done." Many principals believe the greater professional conversations between the evaluator and teacher are a result of the heightened accountability due to the systems' emphasis on behavioral evidence.

According to principals, adoption of the system has jump started productive dialogue around teacher practice and standards that may not have been occurring in many places. The system provides teachers and evaluators with a common language, a shared set of standards, and clear expectations. Pre- and post-observation conversations are particularly beneficial because they are teacher specific, revolving around the specific lesson (the plan, evidence from the observed instruction, etc.) and how this reflects components of the rubric. Together, the teacher and evaluator set goals that are specific to the teacher, adopting a personalized rather than uniform approach many used in the past. Teachers are therefore able to see evidence – elements of their recent instruction – and how this informs their evaluation. This evidence has driven conversations around strengths and challenges, fostering productive conversations not just around aspects of instruction that could be improved but also specific, actionable ideas regarding how to do so. Because they are grounded in evidence, discussions are more rigorous and objective than they had been in the past. Overall, the process is more transparent because everyone is aware of, focused on, and working from the same set of standards and rubric. This has contributed greatly to improvements in instruction.

Many principals described these conversations as professional, non-confrontational, and collaborative. This comfortable, supportive environment has fostered productive discussion to help teachers move their instruction forward. Although pre- and post-observation conferences are not very frequent, the few formal, professional conversations teachers and evaluators have had have required that teachers be more reflective about their instruction. This process also sparked

subsequent informal discussions (e.g., among fellow teachers) and an exchange of ideas around how to implement particular strategies in the classroom, translating to improved practice.

In addition to the pre- and post-observation conferences, the summative conference allows principals the opportunity to present teachers with a finalized effectiveness rating for the school year, an analysis of their teaching ability according to the rubric. Many principals reportedly enjoy the rich conversations they have with teachers about their teaching practice during the summative conference. In some occurrences, based on teachers' areas of opportunity, principals recommend teachers attend specific PD, such as how to engage students in the active learning process.

Instructional Planning

Principals also thought that the evaluation system contributes to an improvement in teachers' instructional planning. Teachers have had to be more thoughtful in ensuring that elements of the rubric are reflected in their plans, particularly for formally observed lessons. The system makes teachers think critically about how particular activities will serve students' needs as opposed to simply going through the motions for the sake of documentation. In both states, most teachers reported that the pre-conference, during which they presented and discussed the lesson plan to their evaluator, helped them prepare for the observation (e.g., 77% of teachers in Year 3) and gave them the opportunity to explain their lesson plan and teaching artifacts (e.g., 89% of teachers in Year 3). These percentages were stable across implementation years. Further, examining elements of the rubric, which are grounded in Common Core State Standards (CCSS), has made teachers more aware of best practices and has changed teacher's opinions regarding what constitutes best practices. Because of this change in instructional planning, principals explained, they have seen improvement in teacher practice.

While this shift towards more thoughtful instructional planning is undoubtedly positive, administrators from both states questioned the extent to which this time-consuming process would carry over from preparation for formal observations to planning for lessons delivered every day. Administrators did report seeing more instructional elements that required in-depth planning outside of formal observations (e.g., higher-level questioning, informing students of learning goals, and grouping students by level), citing this as evidence that the quality of teachers' planning and, in turn, instruction had generally improved. However, many spoke about the noticeable difference in instruction delivered during formal versus informal observations. One reason best practices have not been carried over to daily instruction is that they require time consuming, rigorous planning. One RI principal described the formal observations that require extensive, one-time planning as doing teachers (and students) a "disservice," in that because they are so infrequent that they are not driving any lasting change in planning or delivery that would translate to increased student learning.

Higher-Order Questioning

Throughout the i3 grant, principals in NY and RI noted major improvements in teachers' use of strategic questioning to promote higher-level thinking and cognitive engagement in students. Some thought this was one of the biggest changes with respect to teacher instruction since the adoption of the evaluation system. Teachers have shifted from asking more close-ended, recall-based questions to posing fewer but more engaging questions. One administrator in NY referred to the use of purposeful questioning as "low hanging fruit," in that it was a relatively easy

teacher practice to modify, but such changes have the potential to significantly boost students' cognitive engagement and ostensibly contribute to student learning. In this regard, the evaluation rubric has served as a guide for teachers in revamping lesson plans to incorporate progressively more challenging lines of questioning.

Student Collaboration

Particularly in the past few years, teachers in NY and RI have begun moving from teacher-centered to student-centered instruction with activities that promote student collaboration. Much importance is placed on student collaboration given this is recognized as a best practice and featured in the CCSS standards and evaluation rubric. Further, collaborative group work helps ensure that all students—not just a few—participate in the lesson, and has resulted in increased student engagement. Some principals directly attributed this increase in student collaboration to changes in teachers' instructional methods as a result of the evaluation system.

As of Year 3, implementation of practices that support student collaboration varied considerably by school and by teacher across states. At some school sites, principals noted that their schools are taking part in discussions of the theoretical purpose of such a shift. Elsewhere, principals noted that teachers have implemented fairly simple student collaboration activities (e.g., turn and talk), while others (seemingly fewer in number) are further along in their use of this strategy, arranging student collaboration (e.g., small-group work designed to foster purposeful discussions) with the goal of fostering critical, higher-level thinking. Despite this variation, principals noted that there is a clear trend that teachers are more aware of the importance of student collaboration and have begun building it into their approach.

Student Engagement

There has also been a dramatic increase in student engagement, principals across both states noted, attributing this to various instructional changes. Many principals attributed increased student engagement to the new system, while others acknowledged the system was one of multiple factors that contributed to this change. Some principals ventured to say this increase in engagement would most likely boost student achievement but admitted to not yet having the data to back up that claim.

Most principals are noting that the major contributor to the marked increase in student engagement is a shift toward student-centered instruction and learning. In general, there is an ongoing shift towards a "workshop" model, involving transfer of responsibility and ownership to students for their own learning. Teachers are relinquishing some of their traditional role as lecturer and instead serving in a facilitator role, at least for a portion of the lesson. This release of responsibility to the students themselves has increased participation and promoted active learning. Of course, the extent to which this has occurred varies and is dependent on teacher comfort with and knowledge of more independent, student-centered activities. While a complete shift to the workshop model still remained a goal for most, awareness of the importance of student-centered learning and incorporation of various elements denote a clear trend in gradual adoption of this instructional approach.

Principals also noted that another way in which teachers have promoted student engagement is through discussion of specific learning standards, instilling students with a sense of ownership of their own learning. Many teachers now tell students what they will be doing at the beginning of

the lesson; then, at the end, they work together to assess whether or not they accomplished their goals. In NY, this was happening in middle and secondary schools. One principal in RI mentioned that teachers in her school were talking about standards in both regular and special education classrooms. In speaking about emphasizing standards and shared responsibility, another RI administer explained: "[The system] provides students with an awareness of their expectations, and teachers are aware of what they need to teach. They both have the same goal in mind, success in learning and teaching, and they are working together to reach it."

Some principals in NY also mentioned teachers' efforts to better relate content to students' lives to boost engagement. More than ever before, teachers are integrating students' race, culture, and socioeconomic status in their delivery of content and planned activities. This was particularly relevant for LEP students, including recent immigrants, for which engagement was a critical first step in making learning gains.

Differentiated Instruction

As with other best practices, inclusion of differentiated instruction in the evaluation system has led to heightened awareness, increased buy-in, and understanding of the importance of this practice. In some cases, principals have seen change among teachers who were not previously differentiating. Approximately 82 percent of teachers in Years 2 and 3 also agreed that the pre-conference in particular afforded them the opportunity to discuss with their evaluator how to differentiate within the planned lesson. Despite these conversations, most administrators still identified differentiation as an area of desired PD, noting most teachers did not know how to differentiate and/or were not doing so as effectively as they could.

Some principals attributed this shift towards differentiation to adoption of both CCSS and the new evaluation system. Other administrators did not consider the system as the impetus for differentiated instruction, but inclusion of differentiation in the rubric has certainly brought increased attention and sparked collaborative discussions on tailoring lessons to meet students' needs. In RI, SLO goals in particular were mentioned as a driving force behind use of differentiation.

Across NY and RI, principals noted that there is a clear yet gradual shift towards incorporating differentiated instruction, but the extent to which teachers were implementing this approach, or elements of it, varied considerably. In most schools, principals noted that modest changes in differentiated practices were being seen, but they were quick to point out that this is an ongoing process. For example, at some sites discussions around differentiated instruction were solely theoretical in nature; in others, principals shared concrete evidence of such changes (e.g., switch in students' seating arrangement from rows to small groups; grouping of students by level; use of varied assessments based on student data; allocation of specific differentiation time). As one principal hypothesized that ultimately "the one-size-fits-all instruction will dissipate."

Because differentiation requires extensive planning and professional development, getting teachers to effectively incorporate differentiated activities is difficult. Many principals noted that teachers did not have sufficient training to implement such practices proficiently. Moreover, while some teachers differentiate to a certain extent, they do so only by student level but do not modify their support for LEP students or SWDs. Elsewhere, teachers may modify instruction for SWDs but the differentiated activities do not necessarily serve students at the higher end of the

spectrum. Such examples highlight a strong need for extensive training on differentiation practices for students at all levels and varied needs. Inclusion of differentiation as an evaluation criterion has brought extra and necessary attention to this widespread area of need.

Fewer principals remarked that differentiation was happening effectively before the evaluation system, and thus adoption has not changed this. One elementary school principal explained that differentiation in early grades is critical, and that good instruction drives the use of differentiation, not the evaluation. Elsewhere, use of differentiation (which started before adoption of the system) was attributed to principals continually stressing this approach and providing teachers with development opportunities to hone their skills—factors unrelated to the evaluation system.

Use of Assessments

Across both states, principals spoke of a dramatic increase in awareness and use of assessments as teaching tools, which many attributed directly to the evaluation system. Principals noted a change in the use of summative and, in particular, formative assessments. In most schools, teachers were already doing the former but only recently have started conducting quick, informal assessments. Now, they use these throughout the lesson to quickly gauge students' understanding. They also think more about the purpose and placement of each assessment during instructional planning. As a result, principals report that teachers are more sensitive to students' varied needs, making use of assessment data to support all, not just the lowest performing, students.

Given the general recognition across the county of data-based decision making, most teachers are now aware of the link between instruction, assessment, and student progress, which has led teachers to adopt varied evaluation strategies. As one principal explained, assessments are no longer one-size-fits all. Examples include using response cards and exit tickets; having students turn and talk; writing answers on individual whiteboards; indicating understanding with thumbs up or down; and writing in daily journals. Teachers also use results in a number of ways. For example, they use data from the end of a lesson (e.g., exit slips) to tailor the next day's plans accordingly; inform individualized support; make need-based grouping decisions; and identify topics for re-teaching. Teachers who are now implementing such assessment strategies regularly speak highly of the great advantages they afford. Integrating regular formative assessments provides teachers with a standardized way to track student progress between summative assessments, and it force teachers to think critically about their instructional practices with respect to whether they are moving individual students forward. In RI in particular, principals have noticed an increase in teacher collaboration and more informal conversations around different types and use of assessments. Renewed focus on assessment brought about by the evaluation has provided teachers with a "common language" to share and hone strategies together.

The change in awareness around and use of assessments signifies a major shift, but many principals noted that teachers were still in the early stages of implementation, cautioning that resultant student improvement would be a ways off. Still, many expressed optimism about the future and how such practices were developing. Many teachers were still adding more assessment options to their repertoires, familiarizing themselves with how to conduct these, and learning to use data in the moment to adjust instruction to meet students' needs. The change is

also evident in teachers' focus: before, it was on a planned lesson schedule, whereas now it is on student learning, which may necessitate a deviation from the planned lesson. While the change is clear it is also ongoing, and integration of informal assessments into instruction is not yet dynamic nor seamless. Consequently, the ultimate goal—the ability to effectively target instruction to maximize learning—was still, for most, something to strive for.

Despite this widespread change, principals in both states emphasized the need for continued PD on use of assessments and use of these data to tailor instruction. As one principal in NY explained, the rubric has given school-staff a vision for assessment, but it has not helped clarify the differences between and how to use summative and formative assessments—this is something teachers still need. Further, a few principals explained that although there had been some change in use of assessments, there was not as large of a shift as they had hoped. One principal explained that in general there are a lot of data from various sources, which need to be better aligned to facilitate teacher use.

In both states, fewer principals did not think that there had been a real change in use of assessments. On one hand, a few principals explained that teachers were already conducting formative assessments prior to adoption of the system. In these cases, this was the result of principals having placed great importance on assessments, for example, providing teachers with PD opportunities to help make this part of the school culture. Still, although the evaluation system may not have changed the assessments used, it has been advantageous in enforcing standardized documentation of their use and of student data. Elsewhere, a handful of principals noted informal assessments were implemented by pockets of teachers but not school-wide. In many places, these were not mandated, and use during formal observations, occurring only a couple of times per year, was often not reflective of ongoing implementation.

Other Contributors to Change

While the vast majority of principals noted some change in teachers' instructional practices since the adoption of the evaluation system, the extent to which they attributed this to the evaluation system varied. At most, principles acknowledged that the system contributed to change but would not attribute change solely to the system itself, mentioning other factors that have also contributed. Most schools and districts were in a state of flux, with changes in assessments, district and state requirements, adoption of CCSS, and other factors all having a potential impact on both teacher and student performance.

Aside from positive changes it has engendered, the evaluation system is first and foremost a tool that reflects the ongoing changes in instruction nationwide. It serves as a lens through which they can take a close look at teachers' instruction. Most principles agreed that the system itself has not served as the impetus for broad instructional changes. At best, it has served to heighten awareness, push practices forward, and require practitioners to fine tune their approaches. Principals mentioned that these changes were part of a longer-term shift in adoption of best practices and 21st century skills supported at the school, district, and state level(s) prior to the adoption of the evaluation system. When they mentioned a primary contributor to change, it was often ongoing PD on specific instructional approaches and strategies that have truly guided improvement in teachers' practices. While the effects of this support in terms of improved teacher practice would, in theory, be captured using the evaluation rubric, the PD itself was independent from the evaluation system.

Impediments to Instructional Change

Evaluation Anxiety among Teachers

Particularly during Years 1 and 2 of the grant, some principals in NY and RI spoke about anxiety and stress teachers felt as a result of the evaluation system, possibly due to a lack of knowledge about the system, which may have hindered its effectiveness. Experience with the system over time may have helped reduce this anxiety in Year 3, which may also be why there was a 15 percent increase from Year 1 to Year 3 in teachers' agreement that they had received sufficient information regarding the system.

Although the system has made teachers more reflective, some teachers found the system overwhelming. Teachers have had to spend more time on evaluation-related work because the rubric was long, detailed, and difficult to understand, time which they could otherwise have devoted to planning or direct work with students. Fear connected to the ratings has placed a burden on teachers that, yet another principal described, has particularly gotten veteran teachers overwhelmed and stressed out about receiving ineffective scores and, consequentially, worried about their job status and security. This anxiety contributed to negative moral and lack of buy-in for the system among some teachers. Further, it may point to a lack of communication on the part of districts and administrators regarding the purpose and use of the system, as the system was never designed to effect teachers' job status.

Lack of Formal Training

As mentioned, principals acknowledged that the evaluation system heightened awareness of specific elements of quality instruction—the necessary first step in effecting change. However, some principals did not think the system had brought about the large-scale changes many had hoped. Speaking about the gap between awareness and actual implementation of best practices, one principal explained that the evaluation tool assumes teachers have knowledge that, in many cases, they simply do not possess. Further, many principals noted that targeted PD on specific practices was the critical factor in moving instruction forward, commenting that this support would not come from the evaluation system itself. While use of the system helped identify key areas for PD, the PD needed to affect teacher practice across the board did not seem to be happening in most places due to reasons unrelated to the evaluation system (e.g., available funding). This may explain why teachers' endorsement of the system ratings as informing district PD were fairly low (between 34% and 39% across years) whereas agreement that the system helped *identify* PD needs was notably higher (between 65% and 70% across years).

Impact on Students

In this section, we discuss the potential impact of the teacher evaluation system on student learning. First, we share findings based on stakeholder perceptions of impact and evidence they cite as demonstrations of increased student achievement. Second, we analyze extant student achievement data to determine if districts that implemented the E3TL evaluation system made gains in ELA and math compared to predicted scores in the absence of the system.

Stakeholder Perceptions

Principals in NY and RI stated that they could not determine whether student performance had improved due to use of the evaluation system based on students' scores on state assessments.

Even at the end of Year 3, they thought that it was too early to tell if there had been student growth as a result of the evaluation system. In NY, a change in the standardized test to more closely align with CCSS precluded comparison across years as well, with principals noting that they expected a dip in scores due to higher difficulty level of the test. Moreover, any increase in student achievement may or may not have been due to the evaluation system itself, as many principals cited other student-level factors (SES, stress, etc.) and ongoing initiatives (e.g., CCSS and SLO adoption; see *Other Contributors to Change* above) that may have supported student growth. Measuring the specific impact of the evaluation system on student achievement (or improvement in teachers' instruction), they explained, was virtually impossible given the various other factors that might also play a role. However, it is worthwhile to note that many commented on the widespread positive change in teachers' awareness and understanding of the importance of best practices, prompted by the adoption of the evaluation system and CCSS. Principals in both states spoke of significant changes they had observed in teachers' instructional practice which could in theory contribute to gains in student achievement.

Principals across both states cited evidence of student progress other than standardized test scores. Anecdotally, many noted observing progress during classroom visits (e.g., in the level of students' questioning and engagement) as well as receiving positive reports from teachers themselves regarding their students' growth. Over two-thirds of teachers surveyed each year also identified improving student learning/achievement as one of the primary uses of the evaluation system. In RI, student growth was apparent upon examination of SLO measures, but once again this progress could not be attributed solely to the evaluation system itself. Similarly in NY, many principals cited "soft" data and indication of student growth on local school- or classroom-level assessments, but these data showed changes for some but not all students. These types of measures were the extent of the available evidence. Nonetheless, the upward trend on some progress monitoring and other informal assessments was promising.

Analysis of Extant Data

The research team also conducted additional analyses of participating districts' state assessment scores to determine if districts that employed the E3TL teacher evaluation system made gains in ELA and math achievement compared to their predicted scores in the absence of the evaluation system. To answer this question, we used an interrupted time-series (ITS) design to examine the impact of E3TL on students' standardized test scores. The study estimated impacts on elementary and middle school students' reading and math achievement. The ITS design does not use a comparison group of schools. Instead, it uses the participating schools as their own comparison group by comparing the trajectory of student scores before and after an intervention is introduced to see if there is a significant difference in the trajectories. If the E3TL evaluation system had no impact on student achievement, we would expect the trajectories after the start of the program to match those before the program started. A change in the trajectory of student scores for the years after the start of the program, however, could indicate that the change in student achievement was associated with the E3TL evaluation system. Although a strong research design, this analysis does not indicate causality.

Data Sources

Schools from the five RI districts and the five NY districts were included in the study. Within these districts, schools with outcome data available for all of the study years (2006-07 through

2013-14) were included in the analysis. In total, there were 104 unique schools included in the Grade 4 analysis (39 in NY, 65 in RI) and 31 unique schools included in the Grade 8 analysis (13 in NY, 18 in RI).[9]

We obtained publically available data from each state's department of education website.[10] Available data included school-level average math and reading outcomes as well as school-level enrollment; percentage LEP; percentage FRL; and percentage minority. We collected data for 2007 through 2014 (five years before the start of E3TL implementation and three years post implementation).

Methods

To implement this design we ran hierarchical linear models (HLMs) to adjust for the multiple observations across time within each school. We ran eight separate HLMs: one for each of four outcomes (school average test scores for Grades 4 and 8 in both math and ELA) in each state (NY and RI). The estimates reported from the model represent whether or not there was a change in the trajectory of student scores in the post-implementation years.[11]

Results

The standardized estimates indicating a change in trajectory of Grade 4 and Grade 8 student scores in math and ELA in post-implementation years are reported in Exhibit 3.

Exhibit 3: Standardized Estimates by District, Grade, and Subject

District	Grade	Subject	Standardized Estimate	P-value
New York	Grade 4	ELA	-0.14	0.02*
		Math	-0.09	0.17
	Grade 8	ELA	0.08	0.20
		Math	0.05	0.54
Rhode Island	Grade 4	ELA	0.00	0.99
		Math	-0.03	0.44
	Grade 8	ELA	0.00	0.97
		Math	0.04	0.45

Note: * indicates significant effect (p-value less than .05).

We found no significant estimates for either outcomes or grades in RI, indicating that the trend in student scores from the 2006-07 school year through the 2011-12 school year was not significantly different from the trend from the 2011-12 school year through the 2013-14 school

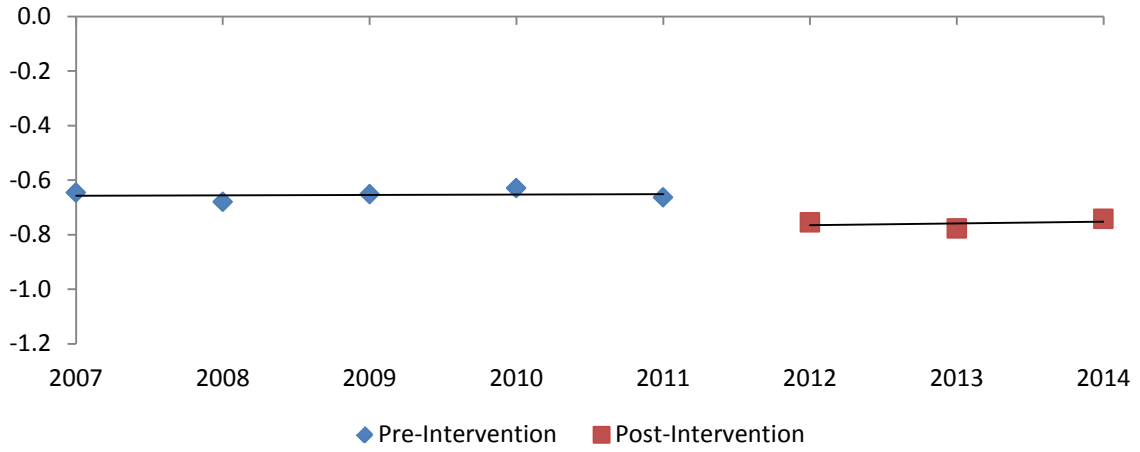
[9] All schools had both math and English-Language Arts data.

[10] Rhode Island data retrieved from: http://www.eride.ri.gov/FileExchange/FredPublic.aspx; New York data retrieved from: http://data.nysed.gov/downloads.php.

[11] In each of these models we included an indicator for year, an indicator for the start of implementation of the evaluation system (0 before 2011-12 and 1 after), and an interaction between the year and implementation indicator. We also included district fixed effects as well as controls for enrollment, percentage LEP, percentage FRL, and percentage minority.

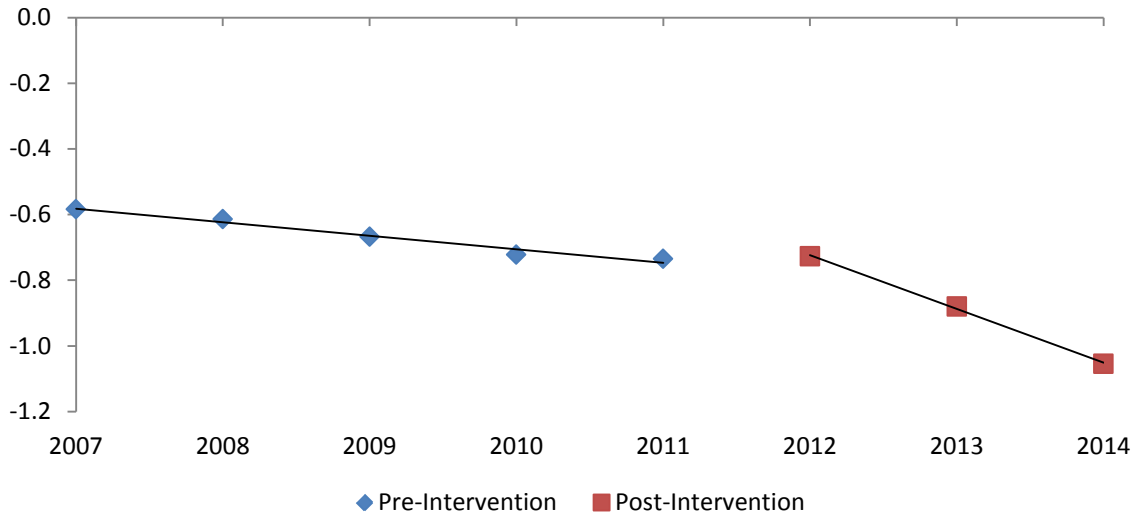
year (i.e., when the teacher evaluation system was implemented in the districts). This pattern is demonstrated graphically in Exhibit 4. The points on this graph represent the expected average Grade 4 ELA scores in participating RI schools. The trend during the pre-intervention years is fairly flat over time, and this trend remains flat during the intervention years.

Exhibit 4: Expected Yearly Achievement: Rhode Island, Grade 4, English-Language Arts



In New York, we found one significant negative estimate for the Grade 4 ELA outcome. All other estimates were not significant. This significant estimate is demonstrated graphically in Exhibit 5. As before, each point on this graph represents the expected average Grade 4 ELA score in participating New York schools. During the years before implementation of the evaluation system, there is a slight downward trajectory; however, during implementation years, the trajectory of expected scores drops substantially.

Exhibit 5: Expected Yearly Achievement: New York, Grade 4, English-Language Arts



It is important to examine this change in trajectory with caution. While the change in scores could be associated with the start of the E3TL evaluation system, it could also be due to a

number of other factors that occurred during the same set of years. For example, in NY, the standards and state tests underwent a substantial change in the 2011-12 school year, which is also the year when implementation of the teacher evaluation system began. This could explain why we observed a significant negative drop in Grade 4 ELA scores after the 2011-12 school year. Further studies using comparison schools would be needed to determine the true cause of the identified trend in Grade 4.

Overall, implementation of the teacher evaluation system has led to widespread awareness and, to a varied extent, improvements in instruction with respect to best practices. Stakeholders across NY and RI spoke about marked improvements in instruction and the integration of strategies featured on the evaluation rubric, such as higher-order questioning, activities that foster student collaboration and engagement, differentiated instruction, targeted use of formative assessments, and use of data to tailor instruction. There has been a major shift in how teachers think about and plan their instruction as a result of implementation of the new system. The extent to which teachers have changed their practice varied across and within schools. As one principal noted, teachers are not yet "fluent" in many of these practices, but general consensus is that all are moving in the right direction.

Nonetheless, the evaluation system has undoubtedly brought about many positive changes in the participating districts. As noted, awareness of best practices has skyrocketed, and teachers now acknowledge the importance of these to meet the varied needs of today's students. The system has established a sense of accountability specifically for student learning, whereas before many teachers were simply focused on getting through each lesson. Moreover, the system helps identify teachers' specific PD needs. It also provides teachers and evaluators with a common language to discuss instructional practice, sparking conversations that in many places were not happening, as well as provides a systematic, more objective way to measure teachers' improvement over time.

While our analysis of districts' state assessment scores before and during implementation of the evaluation system did not indicate an upward trend during the implementation period, this could be due to a variety of factors unrelated to the system. Nevertheless, principals' perceptions of student growth—based on classroom visits, meeting of SLOs, etc.—during implementation were generally positive.

CHAPTER 5. LESSONS LEARNED

Implementation of the evaluation system in 10 districts—differing in size, student demographics, and urbanicity—across two states offered considerable variation in participating school sites, providing us with diverse perspectives on the system and its implementation.

This variation in sample districts and schools was an asset in that it allowed us to identify common aspects of quality implementation of the teacher evaluation system. Data collection efforts, conducted across years and with different stakeholders, were purposely designed to yield findings related to roll-out and implementation. In this chapter, we translate these into lessons learned that can be applied to other districts and schools implementing performance-based teacher evaluation systems across the country. Lessons learned are organized here under training and implementation.

Training

Findings yielded several lessons learned specific to training of evaluators and teachers. These are included here.

Evaluator Training

Lesson Learned: Hands-on practice and feedback on evaluation accuracy are both critical elements of initial and ongoing evaluator training.

Principals from both states expressed concern that they did not receive enough feedback on their performance as evaluators during the initial training. Principals were unsure about their ratings because they did not receive any input on whether or not they were applying the rubric as intended. Therefore, principals were not made aware of areas for improvement or ways in which they could increase their rating accuracy. Further, principals stressed the need for more practice and "real life" training through scenarios and paired observations, during both the initial and ongoing training. Learning the various elements of the rubric was an important first step, but applying these to everyday instruction, having the opportunity to discuss the intricacies of the rubric, and getting tailored feedback from experts is necessary to really engage with, understand, and use the system effectively.

Lesson Learned: Recalibration is essential and should be conducted periodically and in a uniform fashion across districts to ensure evaluators are applying the rubric in the same way.

The need for additional recalibration was a major concern expressed by principals in both states throughout the duration of the grant. All participated in some recalibration activities, but these varied significantly by district. Such exercises helped ensure evaluators within districts were using the rubric in the same way. For example, recalibration was useful for newly trained and experienced evaluators to align their practice, particularly since the training format and providers changed over time.

Teacher Training

Lesson Learned: Districts and/or schools need to provide teachers with ongoing professional development to build and refine skills identified by the evaluation system as areas for improvement.

Even at the end of Year 2, most principals noted there was no formalized support for teachers who received low ratings but said that a formal plan was being developed. Some principals were concerned about the resources required to provide teachers with the level of professional development needed, this need reflected in some teachers' low ratings. By the end of the grant, teachers generally received some type of training or support as a result of their evaluation, but what that entailed varied considerably by district and school.

Lesson Learned: Teacher trainings on the evaluation system should be mandatory and provide consistent information.

Teacher trainings on the evaluation system were not mandatory. Because some teachers attended and others did not, this resulted in differences in their knowledge of the system, its purpose and intended uses, and specifics of the rubric itself. Moreover, regarding teacher trainings that were offered, evaluators noted they did not convey consistent information and that this depended on who was delivering the training (the state, district, union, etc.). Together, these two factors greatly affected the extent to which teachers possessed a shared understanding of the system.

Lesson Learned: Teacher trainings should include an in-depth examination of the evaluation rubric so that teachers are aware of and understand all criteria.

Teacher trainings did not cover aspects of the rubric in-depth, although some principals suggested this would be helpful because the tool includes examples of instructional best practices that pertain to each indicator. Teachers have had to spend more time on evaluation-related work because the rubric was overwhelming and difficult to understand, time which they could otherwise have devoted to planning or direct work with students. Further, teachers did not understand their evaluation scores in large part because they were not familiar with the rubric itself.

Lesson Learned: The training on the evaluation system should be offered within a reasonable timeframe with respect to roll-out or other related aspects of implementation.

Teachers identified timing of their training on the evaluations system as an issue. Some were trained far before the system was implemented, creating a gap between when they learned about the system and when implementation began. Time between training and implementation should be minimized to ensure knowledge and skills gained are carried over and applied to the fullest extent possible.

Lesson Learned: Training on the evaluation system must include guidance on external measures of student learning that are ultimately factored into evaluation ratings.

Principals explained that other metrics of student learning (e.g., state-mandated SLOs) are used as an indicator in the evaluation system, which in turn affects teachers' evaluation rating. Principals in both states desired better training around these assessments themselves as well as how they should be used with respect to the evaluation system and conferring a rating.

Lesson Learned: There must be a standardized method to set equitable and rigorous goals for student progress if these data are used to inform teachers' evaluation rating.

Principals noted that some teachers were hesitant to set challenging student learning objectives (SLOs) because whether or not these were met affected their effectiveness evaluation rating. Teachers typically set SLOs for their class at the beginning of the school year, which were then approved by the principal. Whether or not they met their specific SLOs was factored into their evaluation. Setting non-rigorous goals enabled students to meet these fairly easily. However, this did not necessarily yield meaningful information as to whether or not the teacher provided effective instruction, which was the purpose of incorporating SLOs into the evaluation system

Implementation

Lesson Learned: Clear and consistent communication regarding the intended purpose and uses of the evaluation system is vital prior to and throughout implementation.

Some teachers experienced anxiety as a result of the evaluation system, which may have hindered its effectiveness. Fear connected to use of the ratings made some teachers feel overwhelmed and stressed about receiving ineffective scores and, consequentially, worried about job status and security. This anxiety contributed to negative morale and lack of buy-in for the system. Further, it may point to a lack of communication with teachers regarding the purpose of the system, as it was never designed to affect teachers' job status.

Lesson Learned: There must be established guidelines regarding how to evaluate teachers of students with disabilities and limited English proficiency.

Many evaluators did not feel prepared to use the system to evaluate teachers of special populations because the rubric does not accommodate instruction geared towards those learners. Teachers of SWDs and LEP students make adjustments to instruction to meet their students' needs, but the appropriate level of instruction for these students often does not warrant high ratings. Because of this, evaluators have not been able to hone their skills with respect to evaluation of teachers serving such populations. Some called use of the current rubric unfair with such populations. Other evaluators acknowledged making unofficial modifications and using their own professional judgment to rate teachers because they were not trained nor had guidelines to follow.

Lesson Learned: There must be established guidelines regarding how to evaluate non-classroom teachers and those outside of the primary content areas.

Evaluators desired additional support regarding use of the system with non-classroom teachers such as librarians, art, and physical education teachers. For example, a physical education teacher's lesson and content may not lend themselves to use of 21st century skills like those of an ELA teacher.

Lesson Learned: Appointing other personnel to conduct a portion of the evaluations may greatly reduce the time-consuming burden the system presents for principals and in turn contribute to on-model implementation.

The time-consuming nature of the evaluation process was the most common challenge voiced by principals charged with conducting evaluations. The process could take up to 20 hours for one teacher, mostly due to the large administrative burden, and some principals had 20 teachers to evaluate. Because of this, some principals could not complete all aspects of the process as intended, including not being able to conduct all end-of-year summative conferences. One common suggestion was to have additional personnel conduct some of the evaluations to reduce the burden on principals, who have various other responsibilities to fulfill. Some principals received such support from district staff during Year 1 and found it very helpful.

Lesson Learned: Increased awareness of instructional best practices due to the evaluation system does not necessarily translate to implementation. Additional, ongoing support is critical in changing teacher instructional practice.

Adoption of the system was perceived to have engendered a major shift in mindset and awareness of instructional best practices as vital in increasing student achievement. While this is a necessary first step towards lasting instructional change, awareness does not equate to implementation when it comes to best practices. However, changing teacher practice is a gradual process that requires in-depth and ongoing PD. Many principals acknowledged this was a critical factor in improving instruction, but the evaluation system itself does not provide this necessary support. While use of the system helped identify key areas for PD, the support needed to effect and sustain instructional change must also be taken into account and provided to maximize system benefits

APPENDIX A: OVERVIEW OF EVALUATION FRAMEWORK

To map the activities and relationships among elements of the teacher evaluation system, and to illustrate how these were designed to deliver project outcomes, AIR developed a conceptual framework in the form of a logic model.

A logic model provides a pictorial map of the project elements. These maps and/or diagrams provide a way for evaluators to organize information pertinent to a specific project evaluation, which they can then use to facilitate stakeholders' discussions on changes, challenges, and achievements. The logic model for the AFT E3TL consortium project has been developed as a roadmap and discussion tool; it has many working pieces. In this chapter, we present this logic model and its components. We also describe various modifications made to the model to reflect changes in scope over the duration of the four-year project.

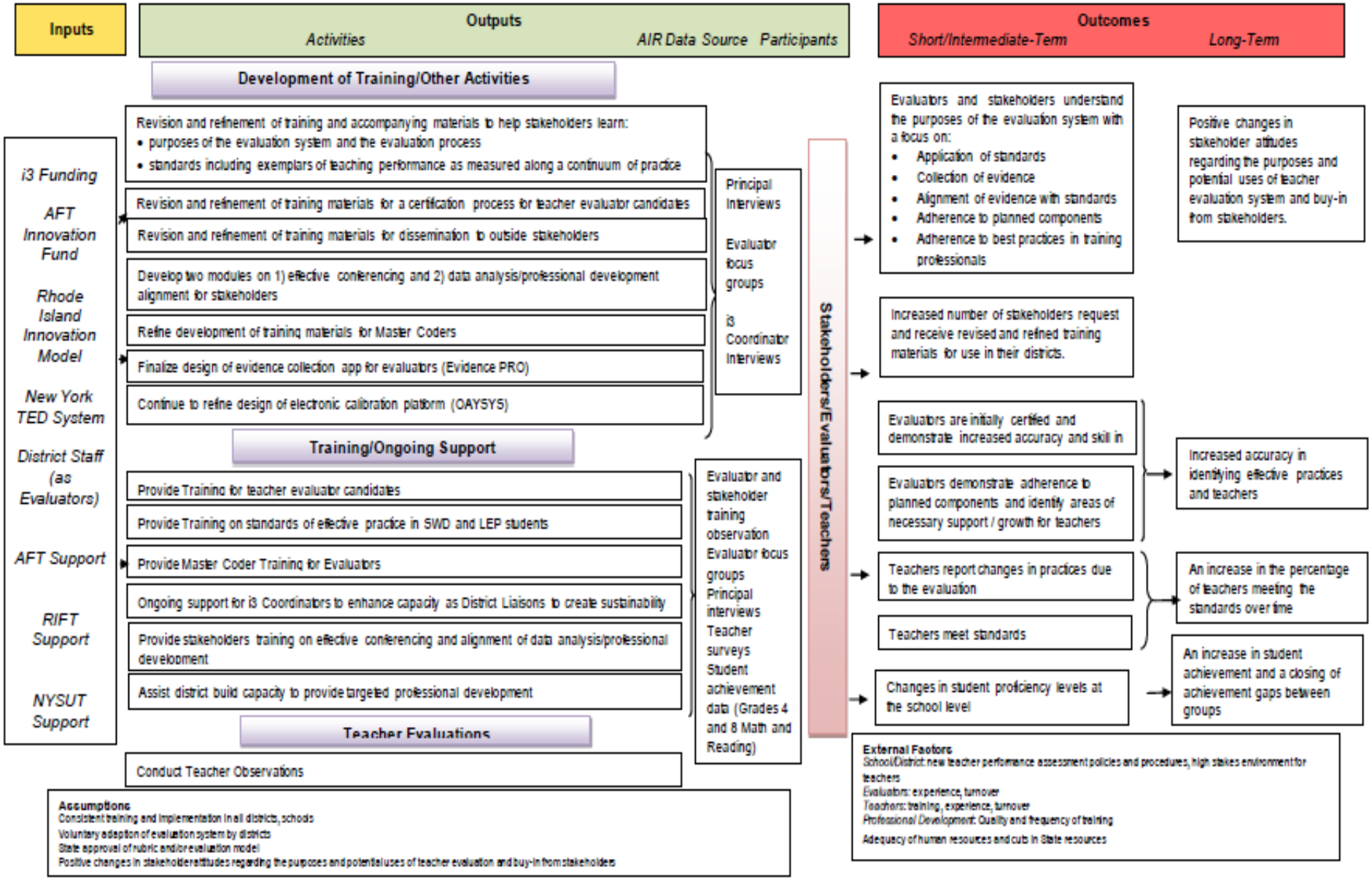
Overview of E3TL Logic Model

Following established conventions, the E3TL evaluation logic model includes many working pieces. These components include:

- *Assumptions:* Expected and agreed upon conditions for the initiative.
- *AIR Data Source:* A data collection activity used to provide evidence for the evaluation.
- *Participants:* People who are involved with or directly implementing the teacher evaluation system.
- *Inputs:* The funding, human resources, and support needed for program implementation.
- *Outputs:* Training and ongoing support, evaluation activities, and materials developed with the resources and as a result of the initiative.
- *Outcomes:* Products and results of the initiative.
- *External Factors:* Elements outside the influence of the project that may influence project results.

In Exhibit 1, we present the final version of the logic model, representing Year 4 of the E3TL consortium project.

Exhibit A-1. E3TL Evaluation Logic Model (Year 3; 2013-2014)


Inputs
Outputs
Activities
AIR Data Source Participants
Outcomes
Short/Intermediate-Term
Long-Term
Development of Training/Other Activities
Revision and refinement of training and accompanying materials to help stakeholders learn:
• purposes of the evaluation system and the evaluation process
• standards including exemplars of teaching performance as measured along a continuum of practice
Revision and refinement of training materials for a certification process for teacher evaluator candidates
Revision and refinement of training materials for dissemination to outside stakeholders
Develop two modules on 1) effective conferencing and 2) data analysis/professional development alignment for stakeholders
Refine development of training materials for Master Coders
Finalize design of evidence collection app for evaluators (Evidence PRO)
Continue to refine design of electronic calibration platform (OAYSYS)
Training/Ongoing Support
Provide Training for teacher evaluator candidates
Provide Training on standards of effective practice in SWD and LEP students
Provide Master Coder Training for Evaluators
Ongoing support for i3 Coordinators to enhance capacity as District Liaisons to create sustainability
Provide stakeholders training on effective conferencing and alignment of data analysis/professional development
Assist district build capacity to provide targeted professional development
Teacher Evaluations
Conduct Teacher Observations
i3 Funding
AFT Innovation Fund
Rhode Island Innovation Model
New York TED System
District Staff (as Evaluators)
AFT Support
RIFT Support
NYSUT Support
Principal Interviews
Evaluator focus groups
i3 Coordinator Interviews
Evaluator and stakeholder training observation
Evaluator focus groups
Principal interviews
Teacher surveys
Student achievement data (Grades 4 and 8 Math and Reading)
Stakeholders/Evaluators/Teachers
Evaluators and stakeholders understand the purposes of the evaluation system with a focus on:
• Application of standards
• Collection of evidence
• Alignment of evidence with standards
• Adherence to planned components
• Adherence to best practices in training professionals
Increased number of stakeholders request and receive revised and refined training materials for use in their districts.
Evaluators are initially certified and demonstrate increased accuracy and skill in
Evaluators demonstrate adherence to planned components and identify areas of necessary support / growth for teachers
Increased accuracy in identifying effective practices and teachers
Teachers report changes in practices due to the evaluation
Teachers meet standards
An increase in the percentage of teachers meeting the standards over time
Changes in student proficiency levels at the school level
An increase in student achievement and a closing of achievement gaps between groups
Positive changes in stakeholder attitudes regarding the purposes and potential uses of teacher evaluation system and buy-in from stakeholders.
Assumptions
Consistent training and implementation in all districts, schools
Voluntary adoption of evaluation system by districts
State approval of rubric and/or evaluation model
Positive changes in stakeholder attitudes regarding the purposes and potential uses of teacher evaluation and buy-in from stakeholders
External Factors
School/District: new teacher performance assessment policies and procedures, high stakes environment for teachers
Evaluators: experience, turnover
Teachers: training, experience, turnover
Professional Development: Quality and frequency of training
Adequacy of human resources and cuts in State resources

As noted, the components of the E3TL logic model have changed as the project has progressed. These next sections are designed to highlight each component and the changes that have taken place by year.

Assumptions

An assumption, for the intents and purposes of this project, are expected and agreed upon conditions for the initiative. Assumptions were set at the beginning of this project and have remained consistent throughout all four years. The assumptions that undergird this project include:

- Consistent training and implementation in all districts and schools
- Voluntary adoption of evaluation system by districts
- State approval of rubric and/or evaluation model
- Positive changes in stakeholder attitude regarding the purposes and potential uses of teacher evaluation and buy-in from stakeholders

Participants

Participants in the E3TL evaluation have been consistent throughout the project. Participants are defined as people who are involved with or contributing to the E3TL project and evaluation. Their roles in the education system may be also be defined as:

- *Stakeholders:* People impacted by or affected through implementation of the initiative. Teachers and Principals are included in this group as the primary users of the system. Other interested parties who would be impacted by the system and identified as stakeholders include district, regional and state administration, parents, business leaders and other interested community members.
- *Evaluators:* Personnel assigned to evaluate teacher performance using the E3TL observation system. This category includes active and retired principals as well as district and regional employees who have been certified as evaluators. Evaluators observe teachers instruction and follow the identified protocol throughout the year.
- *Teachers:* Educators whose performance is evaluated using the E3TL observation system. The category of teachers includes classroom teachers, language teachers, or anyone who sees students. Teachers are highly involved in the entire observation process, from pre-conferencing and submitting lesson plans to discussions of evidence and goals, the process is collaborative.

AIR Data Sources

AIR collects data from a variety of sources to gather information on each of the activities. These sources changed throughout the project. Exhibit 2 lists the data sources.

Exhibit A-2. Data Sources for AIR's Evaluation

Evaluator and Stakeholder Training Observation	Evaluator Focus Groups	Principal interviews
Teacher Surveys	Student Achievement Data (Grades 4 and 8 Math and Reading)	Document Reviews[12]
i3 Coordinator interviews[13]		

External Factors

External factors are events or circumstances outside of the E3TL project which may influence project outcomes. These include:

- *School/District:* new teacher performance assessment policies and procedure, high-stakes environment for teachers
- *Evaluators:* experience, turnover
- *Teachers:* training, experience, turnover
- *Professional Development:* quality and frequency of training, adequacy of human and cuts in state resources

Inputs to the E3TL Project

At the start of the E3TL project, seven distinct entities were identified as having "input" or providing essential support for the project. They included: i3 funding, The American Federation for Teachers (AFT) support, the AFT Innovation Fund backing,, , the New York State United Teachers (NYSUT) support, the Rhode Island Federation of Teachers (RIFT) support, Charlotte Danielson trainers and materials, as well as district staff to serve as evaluators.

During the Pilot Year, one input was the Charlotte Danielson model. This model, with its accompanying trainers and materials, was used to build knowledge in participants in both states. While preparing for school year 2011-2012 (i.e., Year 1), the use of the Charlotte Danielson model was discontinued. At the time of discontinuation, each state began the rollout and implementation of their state-specific model. In New York, this was the NY TED System, and in Rhode Island, this was the RI Innovation Model. During the successive school years, states have continued to follow these models, as captured in the final Year 3 logic model above. There were no other changes in inputs during the course of the project.

[12] Used in the Pilot Year (2010-2011).
[13] Used Year 1 (2011-2012), Year 2 (2012-2013), and Year 3 (2013-2014).

Outputs

Over the four years of the project, the changing logic model captured the three distinct activities listed within the Outputs section: 1) Development of Training/Other Activities, 2) Training/Ongoing Support, and 3) Teacher Evaluations. Here, we discuss these three outputs in detail, followed by changes that warranted modifications to the model.

Development of Training/Other Activities

Development of Training/Other Activities is the first component of the Outputs section. This component focuses on the development and refinement of materials needed for the growth of the initiative. As such, it is crucial to the development and success of the E3TL program.

It is important to note that this is the component of the logic model that saw a great deal of change throughout the four years of the program. The Pilot Year established a baseline for the component:

Pilot Year (2010-2011)

- The development of training and materials to help stakeholders learn 1) the purposes of the evaluation system in the evaluation process and 2) standards including exemplars of teaching performance as measured along a continuum of practice.
- A state evaluation model and/or rubric was to be developed and adopted in year one.
- Materials and a certification process for teacher evaluators were included.
- The priority of the identification of instructional considerations for SWD and LEP students in inclusive environments were specifically highlighted in the Pilot Year.

After the Pilot Year, changes in training and other activities occurred, necessitating modifications to the model in subsequent years:

Year 1 (2011-2012)

- Began training and materials revisions for participation of i3 Coordinators.
- Introduction of master coders, establishing the need for a new training and accompanying materials.
- Focus placed on development and design of electronic tools to assist in the evaluation process; specifically, and evidence collection tool called Evidence PRO, and an electronic calibration platform called OAYSYS.

Year 2 (2012-2013)

- Development of training on required Instructional Considerations for SWD and LEP students in inclusive environments in Year 2.

Year 3 (2013-2014)

- Development of two modules: 1) effective conferencing and 2) data analysis/professional development for stakeholders was added to the outputs section of the logic model in Year 3.

Training/Ongoing Support

A second set of activities within the Outputs section is the Training/Ongoing Support component. This component builds upon and extends on the Development of Training/Other Activities section. Initial activities are from the Pilot Year ((2010-2011) and the subsequent changes are noted below.

Pilot Year (2010-2011)

- Establishment of training for teacher evaluator candidates.
- Creation of professional development for evaluators with a focus on teacher evaluation and improved practice.
- On-site training to assess evaluator fidelity to observation model protocols.
- i3 Coordinators trained on the evaluation system.

Year 1 (2011-2012)

- Addition of master coding training for evaluators.
- Begin RIDE professional practice ratings.

Year 2 (2012-2013)

- Training was instituted on successful practice with SWD and LEP students.
- Establishment of an initiative to test, finalize, and train people on the Evidence PRO and OAYSIS systems.

Year 3 (2013-2014)

- Additional focus of assisting the district build capacity through provision of targeted professional development.
- Begin Stakeholder training on effective conferencing and alignment of data analysis/professional development

Teacher Evaluations

Teacher evaluations are the third and final activity of the Outputs section. In the pilot year (2010-2011), teacher observations began, conducted by an evaluator paired with an expert consultant. During subsequent school years, evaluators themselves conducted teacher observations as a part of the project.

Outcomes

The Outcomes section of the logic model provides the expected end results of the project. These results are organized as *Short/Immediate* term goals that lead to linked *Long* term goals. This entire section has been used to facilitate and organize discussions on the E3TL evaluation in addition to tracking the expected progress of the initiative. A variety of changes can be observed in these components during the four year cycle.

Short/Intermediate and Long Term Goals

Short/intermediate term and long term goals are the main components of the Outcomes section. In this description, information is provided on the original goals and the changes throughout the project. In order to facilitate the reading of this section, goals that are constant throughout the project are presented first in Exhibit 3.

Exhibit A-3. Short/Intermediate and Corresponding Long-Term Goals Unchanged

Short/Intermediate Term Goals	Long Term Goals
Evaluators and stakeholders understand the purposes of the evaluation system with a focus on: • Application of standards • Collection of evidence • Alignment of evidence with standards • Adherence to planned components Adherence to best practices in training professionals	Positive changes in stakeholder attitudes regarding the purposes and potential uses of teacher evaluation system and buy-in from stakeholders.
Evaluators are initially certified and demonstrate increased accuracy and skill in ratings	Increased accuracy in identifying effective practices and teachers
Evaluators demonstrate adherence to planned components and identify areas of necessary support/ growth for teachers	
Teachers report changes in practices due to the evaluation	An increase in the percentage of teachers meeting the standards over time
Teachers meet standards	
Changes in student proficiency levels at the school level	An increase in student achievement and a closing of achievement gaps between groups

In addition to the goals presented above, an additional Short/Intermediate term goal was added for Years 2 (2012-2013) and 3(2013-2014). Exhibit A-4 presents this change.

Exhibit A-4. Short/Intermediate and Corresponding Long-Term Goals added for Years 2 and 3

Short/Intermediate Term Goals	Long Term Goals
Increased number of stakeholders received request and receive revised and refined training materials for use in their districts.	

Scope Changes in Project

The E3TL evaluation has undergone scope changes throughout the life of the project and, accordingly, some of these scope changes have altered data collection. One of the larger scope changes has to do with the observation of the training of evaluators. During the Pilot Year (2010-2011),of the project, the Charlotte Danielson model, with accompanying trainers and materials, was the trainer for both states and centralized training was held with approved trainers and materials. AIR sent observers to this training for observation and data collection purposes. Beginning in Year 1 (2011-2012), the use of the Charlotte Danielson model, with accompanying trainers and materials was discontinued. Separate state models, with localized trainings often held at the region or district level were instituted. This change resulted in copious trainings being held in both states. Subsequently, these trainings were unable to be observed by AIR staff and, therefore, data was collected through interviews.

Sustainability of the evaluation model and highly trained evaluators became a focused output of the project through the introduction of master coders in Year 2 of the project. These essential personnel contribute much to the stability and sustainability of the model. Information on master coders was gathered through interviews and focus groups. Master coders were not an original part of the evaluation design; however, it would be remiss not to include these positions.

APPENDIX B: DATA COLLECTION INSTRUMENTS

Year 4 Teacher Survey

Please provide us with information about your teaching experiences.

1. How many years of teaching experience do you have in each of the following settings? (Include any full-time teaching assignments, part-time teaching assignments, and long-term substitute assignments.)

   For each row:
   Enter the number of years in whole numbers only.
   Count the current school year as one year.

   a. Years of teaching in total: ___
   b. Years of teaching at this school: ___

2. Are you a tenured teacher? ○ Yes ○ No

3. Your district has an evaluation process currently in place. The following questions ask what you know about the features of that system.

   Indicate the characteristics of the teacher evaluation system in your district for each of the categories below.

   Observations of teaching (check all that apply)
   a. My evaluation consisted of:
   - ☐ **Single announced observations**
   - ☐ **Single unannounced observation**
   - ☐ **Multiple announced observations**
   - ☐ **Multiple unannounced observations**
   - ☐ **Observations by the principal or other school administrator**
   - ☐ **Observations by a mentor or master teacher**
   - ☐ **Observations by a peer**
   - ☐ **Pre-observation conference**
   - ☐ **Post-observation conference**
   - ☐ **Don't know**
   - ☐ **I was not observed last year**

   Self-reflection (check all that apply)
   b. **As part of my evaluation, I am asked to reflect on my teaching practice:**
   - ☐ **Orally (e.g. in a conversation with my principal)**
   - ☐ **Written reflection**
   - ☐ **Not required to self-reflect**
   - ☐ **Don't know**

   Student work and professional activities (check all that apply)

c. **Evaluations include attention to the following:**
- ☐ **Samples of student work**
- ☐ **Informal evaluation of student performance**
- ☐ **Rating for general responsibility (e.g. attendance at teacher meetings)**
- ☐ **Professional development sessions attended**
- ☐ **Graduate courses attended**
- ☐ **Other, please specify: _____**
- ☐ **Don't know**

Feedback (check all that apply)

d. **Feedback for evaluations:**
- ☐ **Is not provided**
- ☐ **Is provided orally**
- ☐ **Is provided in a written format**
- ☐ **Is provided once per year**
- ☐ **Is provided multiple times per year**
- ☐ **Don't know**

Student achievement (check all that apply)

e. **Student achievement data is measured through:**
- ☐ **SLOs**
- ☐ **Growth scores/value-added measure**
- ☐ **Other, please specify: _____**
- ☐ **Don't know**

4. Please select all that apply. The evaluation system in my district is used for…
- ☐ Informing and improving instruction
- ☐ Improving student learning and/or achievement
- ☐ Creating a more reflective teacher workforce
- ☐ Informing professional development for teachers
- ☐ Informing salary decisions
- ☐ Informing bonus and other monetary, non-salary decisions (e.g., additional pay based on performance)
- ☐ Deciding on non-renewal of teachers
- ☐ Deciding on teacher promotion and/or tenure status
- ☐ Identifying teachers for leadership roles
- ☐ Other (Please specify) _____
- ☐ Don't know

5. Next, please think about your own evaluation experience.

How many times have you been **formally** evaluated at **this school, this school year**?

- ○ Not formally evaluated
- ○ Once
- ○ Twice

○ Three times
○ More than three times
(please specify) _____

5.1 Do you have a formal evaluation planned for this school year?
o Yes
o No

6. How many times have you been **informally** evaluated this school year?
o Not informally evaluated
o Once
o Twice
o Three times
o More than 3 times
(please specify) _____

For questions 7 through 10, please think about the **first formal** evaluation you had **this school year**.

7. Did you have a pre-conference with your evaluator as part of your **first formal** evaluation this year?
○ Yes
○ No

7.1 Please rate your agreement with the following statements about the pre-conference you had with your evaluator.

The pre-conference…	Strongly disagree	Disagree	Agree	Strongly agree
a. Helped me prepare for my observation.	○	○	○	○
b. Gave me an opportunity to explain my lesson plan and teaching artifacts.	○	○	○	○
c. Gave me an opportunity to tell my evaluator things s/he may not otherwise know by just observing (e.g., students that are on behavior plans, ways I will assess students, etc.).	○	○	○	○
d. Gave me an opportunity to discuss with my evaluator how to differentiate the lesson for multiple kinds of learners in my classroom.	○	○	○	○

8. Were you observed as part of your **first formal** evaluation this year?
○ Yes ○ No

8.1 How many minutes did your planned lesson take? (leave blank if you don't know)
_____

8.2 How many minutes did the observer sit in on the lesson? (leave blank if you don't know)
_____

9.	If you were observed **AND** received a pre-conference…		
	a. Was the evaluator the same for both the pre-conference and your observation?	○ Yes	○ No
	b. Did the evaluator observe the lesson you discussed during your pre-conference?	○ Yes	○ No
10.	Did you have a post-conference with your evaluator as part of your **first formal** evaluation this year?	○ Yes	○ No
11.	Did you receive feedback from your evaluator about your teaching?	○ Yes	○ No

11.1 Please rate your agreement with the following statements about the feedback you received from your evaluator.

The feedback I received…	Strongly disagree	Disagree	Agree	Strongly agree
a. was provided in a timely manner	○	○	○	○
b. helped me to improve my instruction	○	○	○	○
c. was provided in a respectful manner	○	○	○	○
d. provided information that I can apply to all of my classes	○	○	○	○
e. clearly addressed my performance based on established goals	○	○	○	○
f. helped me to identify professional development needs	○	○	○	○

		Not at all	To a small extent	To a moderate extent	To a great extent
11.2	To what extent did the evaluation feedback change your use of student assessments in your classroom?	○	○	○	○
11.3	To what extent did the evaluation feedback sessions lead to changes in your instructional practice?	○	○	○	○

11.4	Did you participate in any professional development as a result of your evaluation feedback sessions?	○ Yes	○ No
11.5	Did you receive other support or training as a result of your evaluation feedback sessions?	○ Yes	○ No
12.	Were you given the opportunity to reflect on	○ Yes	○ No

your own teaching?

13. Are you a peer evaluator? ○ Yes ○ No

13.1 Did you receive specific or specialized training to prepare you for being a peer evaluator? ○ Yes ○ No

13.2 Have you had the opportunity to evaluate others this school year as part of your role? ○ Yes ○ No

13.3 Please rate your agreement with the following statements about your experience as a peer evaluator.

	Strongly Disagree	Disagree	Agree	Strongly Agree
a. I am comfortable evaluating my peers.	○	○	○	○
b. My position as peer evaluator is respected by the people I am evaluating.	○	○	○	○
c. The feedback I give my peers is appreciated and welcomed.	○	○	○	○

14. Were you evaluated by a peer evaluator? ○ Yes ○ No

14.1 Please rate your agreement with the following statements about your experience being evaluated by a peer evaluator.

	Strongly Disagree	Disagree	Agree	Strongly Agree
a. I was comfortable with being evaluated by a peer.	○	○	○	○
b. I respected my peer's position as my evaluator.	○	○	○	○
c. The feedback I received from my peer evaluator was helpful and welcomed.	○	○	○	○

The next series of questions asks about your perceptions about evaluations of teaching in general and in your district. Some questions ask about your feelings about evaluations of your own teaching. Your identity and your responses will be kept strictly confidential and we thank you in advance for your candor.

15. Please rate your agreement with the following statements about this evaluation system in your district.

The evaluation...	Strongly disagree	Disagree	Agree	Strongly agree
a. Measures important instructional skills.	○	○	○	○
b. Is based on clearly communicated standards.	○	○	○	○
c. Is consistent and objective.	○	○	○	○
d. Criteria are developed after sufficient input	○	○	○	○

from teachers.

e. Criteria provide an accurate reflection of my performance as a teacher.	○	○	○	○
f. Is conducted by appropriately trained evaluators.	○	○	○	○
g. Is beneficial in improving the professional climate at school.	○	○	○	○
h. Is beneficial in informing the professional development needs of teachers.	○	○	○	○
i. Holds principals accountable for fulfilling their role, e.g., meeting with teachers, providing feedback, responding to teacher needs.	○	○	○	○

16. Is your assessment of your own performance consistent with the ratings provided in your evaluations?
    - o Never
    - o Sometimes
    - o Usually
    - o Always

17. To what extent do you agree with the following statements about the role of student achievement data in teacher evaluations?

Student achievement data should play a role in...	Not at all	To a small extent	To a moderate extent	To a great extent
a. Teacher evaluations, *in general*	○	○	○	○
b. Staffing decisions (e.g., recruiting, hiring, placement, nonrenewal or dismissal) *in general*	○	○	○	○
c. Determining professional development for teachers, *in general*	○	○	○	○
d. Determining *my* contract renewal	○	○	○	○

	Not at all	To a small extent	To a moderate extent	To a great extent
18. To what extent did your Student Learning Outcomes (SLOs) impact your teaching?	○	○	○	○

The next section concerns the teacher evaluation system being implemented in your district.

19. Please rate your agreement with the following statements regarding the communication you received about this evaluation system.

	Strongly disagree	Disagree	Agree	Strongly agree
a. I have received a sufficient amount of information about this evaluation system.	○	○	○	○

b.	I know who to ask if I have questions.	○	○	○	○
c.	I understand the components of this evaluation system.	○	○	○	○
d.	I understand what this evaluation system is being used for.	○	○	○	○

20. Are there any other thoughts or comments you have about your experience with the teacher evaluation system in your district?

If you would like to receive your $20 gift card reward for completing this survey, please add your **home address** below. We cannot send gift cards to your school because we want to keep your identity completely confidential. Your address will not be shared with any outside party and will be used only to send the gift card.

Thank you!

_____
_____
_____

Principal Phone Interview

Principal ID:
Date:
Interviewer:

Introduction

Thank you for taking time to speak about your experience with the teacher evaluation system as part of the i3 grant awarded to the American Federation of Teachers. The American Institutes for Research, or AIR, is a research firm conducting an independent evaluation of the implementation of the teacher evaluation system. Your insights will be critical in helping AIR gather accurate information for its reports. This interview will take about 45 minutes.

You are not required to answer any questions that you don't wish to answer, and you are permitted to discontinue your participation in this interview at any time.

This conversation is confidential, meaning your name will not be attached to any information you provide and the only people able to access to this interview will be the AIR staff in charge of analyzing and reporting data for this project.

Do you have any questions? Do you consent to participate in this interview?

Background

1. How long have you been a principal?
    - In this district?
    - At this school?
    - What grades are you a principal of? [***NOTE TO INTERVIEWER: want to know if elementary, middle school, or high school principal***]

Involvement

I'd like to get a sense of your involvement with the evaluation system.

2. Are you an evaluator at your school?
    - Are you the only evaluator at your school? If not, who are the other evaluators?

3. For the purposes of this interview, full implementation means that every classroom teacher in the building is being evaluated. Is the evaluation system fully implemented at your school?

4. Please describe for me what your experience as an evaluator has been like.
    - What are the challenges or successes your school has faced using this evaluation system this year?

Training

Next, I'd like you to tell me about the training around this evaluation system for evaluators and teachers.

Evaluator Training

5. ***[NOTE TO INTERVIEWER: Only ask <u>new</u> principals this question]*** Have you received any evaluator training?
    - Who provided the training?
    - What was the content of this/these training(s)?
    - Can you describe your satisfaction with the quality of this training?

6. Please tell me about ongoing training have you received this school year around the evaluation system?
    - What was the content of this/these training(s)?
    - How adequate was/were the training(s)?
    - ***[NOTE TO INTERVIEWER: Be sure to ask this probe]*** Did you receive any training for calibration and inter-rater reliability?

7. To what extent do you feel adequately trained for this evaluation system?
    - ***[NOTE TO INTERVIEWER: Be sure to ask this probe]*** What additional training would better prepare you for this system?

Teacher Training

8. What ongoing training have your teachers received throughout this school year to prepare them for this evaluation system?
    - What was the content of this/these training(s)?
    - Who provided the training (the evaluator, district, state, union, etc.)?
    - How prepared do you think your teachers are for this evaluation system?
    - What additional supports would they benefit from?

9. How have teachers reacted to the training(s) they received?
    - What additional training have teachers recommended to better prepare them for participating in this evaluation system?

10. What type of additional support or training do teachers receive as a result of the evaluations you conduct? *[Example: PD they receive in areas they receive low ratings in.]*
    - **[*NOTE TO INTERVIEWER: If additional support or training is received ask]*** What are teachers' reactions to this additional support or training they receive?

Instructional Practice

Now I'm going to ask you a series of questions about how you think this evaluation system has impacted teacher instructional practices.

11. What kind of change have you seen in the extent to which teachers are proposing activities that differentiate instruction based on learning style, culture, disability, etc. due to this evaluation system?

12. Please describe any change in your teachers' instructional practices. *[Examples include: lesson planning, asking questions and challenging their students in the classroom, engaging students, having students work collaboratively and solve problems in groups]*
    - To what extent do you think this evaluation system improves teacher instructional practice? ***[NOTE TO INTERVIEWER: Only ask if not sure]***

13. How has the evaluation system changed the extent to which teachers are using informal or formal assessments to gauge how students are progressing?
    - To what extent do you think this evaluation system improves student achievement outcomes?
    - How have SLOs or Student Growth Measures informed your assessment of teachers' professional practice?

14. What's your assessment of how good a job the rubric does of providing an accurate reflection of the work teachers do in their classrooms?

15. Please describe any summative conferences you have had with teachers.

Lessons Learned

16. Has participating in this evaluation system enabled you to better recognize effective teachers? If so, how?
    - To what extent do you feel more prepared using this evaluations system, than previous tools, to evaluate classrooms with English language learners and students with disabilities?

17. Overall, what have you learned while implementing this evaluation system into your school and district? [*Probe: These may be lessons learned about implementation, training, communication, etc.*]

18. In what ways has your experience as an evaluator with this system changed over time?
    - Has there been a change in the amount of time it takes you to complete an evaluation
    - Has it become easier? If so, why?

19. To what extent do you think labor management partnerships have impacted implementation? *[Example: union and district participation]*

20. What recommendations do you have to improve the use of this evaluation system?

21. Is there anything else you would like to share?

Thank you for your time. Feel free to contact me with any further question.

Evaluator Focus Group

As you know, the New York and Rhode Island chapters of the American Federation of Teachers were awarded a grant to develop a performance-based teacher evaluation system in a group of districts across the two states. These districts formed the Educator Evaluation for Excellence in Teaching and Learning (E3TL) Consortium to assist with implementation of the new teacher evaluation systems. This focus group is part of AIR's evaluation of the implementation and outcomes of the E3TL project. We're going to ask the group some questions about your experiences and feelings about evaluation and the training you have been receiving this week.

As you already know, because this is a focus group context, what you say to us won't be entirely confidential (since there are other people here with you). Please be assured, though, that we will not repeat anything you say in this meeting today to anyone other than our research team who is responsible for analyzing the data and writing a report about our findings. We would also like to ask that you please do not share what you hear with anyone outside of this session as well. Anything you tell us will be reported anonymously, which means we will never use your name or say you said anything.

So that we can make sure that we understand what everyone is saying, we ask that only one person speak at a time, but think of this as a conversation--when you have something to add, just chime in.

Does anyone have any questions?

1. Let's start with a quick round of introductions. Please tell us your name and what district you'll be working with, and why you decided to become an evaluator for the E3TL consortium.

2. Does anyone have any prior experience conducting teacher evaluations?
    a. When you were conducting evaluations, were you acting as a school administrator, district administrator, or in another position?
    b. Was there a particular format that you had to use for the evaluation? Please describe it briefly (were there ratings on a scale or a dichotomous (satisfactory/unsatisfactory) rating; how many observations; other data collected; feedback given).
        i. Were observations a part of the evaluation process?
        ii. Did you provide teacher feedback or mentoring as part of the evaluation process?
    c. Has anyone ever used the Danielson framework before?

3. Outside of the evaluation process, does anyone have any prior experience conducting teacher observations or teacher feedback or mentoring conferences?
    a. In what context (such as an instructional coach)?

4. What do you think about your district's current evaluation?
    a. Do you think it gives an accurate picture of who is an effective teacher?
    b. Do you think it is fair? (Why or why not?)

5. What is your understanding of the new evaluation system?
    a. What is the purpose of this system?
    b. What are the major components of this system?
    c. How is it different from your current system?
    d. Do you think this system will be an improvement and why?

6. What challenges do you anticipate in your role as evaluator?
    a. Time
    b. Teacher resistance
    c. School culture/school climate
    d. Difficulty of implementing the evaluation system
    e. Other

7. Going into the training this week, what were you expecting?
    a. What did you hope to learn?
    b. What sort of content were you expecting would be covered?

8. Now that we're about halfway through the week, how do you think the training is going?
    a. Is the training meeting your expectations?
    b. Do you think the training will meet your needs in terms of what you need to know to conduct the evaluations?
    c. What would you like to learn that hasn't been covered yet?
    d. What do you think about the formats that have been used (lecture presentations, discussion groups)?
    e. Were there any presentations or activities that you thought were particularly helpful/meaningful?
    f. Were there any that you thought were lacking?
    g. What do you think about the materials provided?
    h. What do you think about the form you use for the ratings?

9. Is there anything else anyone would like to tell me about the teacher evaluation system or the training this week?

ABOUT AMERICAN INSTITUTES FOR RESEARCH

Established in 1946, with headquarters in Washington, D.C., American Institutes for Research (AIR) is an independent, nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance both domestically and internationally. As one of the largest behavioral and social science research organizations in the world, AIR is committed to empowering communities and institutions with innovative solutions to the most critical challenges in education, health, workforce, and international development.

LOCATIONS

Domestic

Washington, D.C.

Atlanta, GA

Baltimore, MD

Chapel Hill, NC

Chicago, IL

Columbus, OH

Frederick, MD

Honolulu, HI

Indianapolis, IN

Naperville, IL

New York, NY

Portland, OR

Sacramento, CA

San Mateo, CA

Silver Spring, MD

Waltham, MA

International

Egypt

Honduras

Ivory Coast

Kenya

Liberia

Malawi

Pakistan

South Africa

Zambia

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000 | TTY 877.334.3499

http://www.air.org

Making Research Relevant