Center on
GREAT TEACHERS & LEADERS
at American Institutes for Research

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

ASK THE TEAM

BY ROSHNI MENON, ALEX BERG-JACOBSON,
TIM FIELD, and BRENDAN YORKE

February 2015

Get the Information You Need: How to Design Educator Evaluation Studies for Continuous Improvement

Question From the Field

How are states using pilot and implementation studies to continuously improve educator evaluation systems?

In the past four years, 49 states and the District of Columbia have passed new policies or regulations with respect to their evaluation systems for teachers and principals (Center on Great Teachers and Leaders, 2014). As with most large-scale policy changes, it is critical to the success of the new systems that they are implemented using a continuous improvement process. The cycle of collecting ongoing feedback and evaluating system impact to continuously improve and refine the educator evaluation system is crucial to keeping the system appropriate, meaningful, and informative to all educators. Pilot and implementation studies are one resource supporting continuous improvement and can provide important information to state education agencies (SEAs), including:

- Documentation of the evaluation system design and resources
- Evidence of how the system works in practice
- Strengths and weaknesses of the design, supporting resources, and implementation process
- Stakeholder feedback and perspectives
- Quality control checks on fidelity of implementation, quality of evaluation data, etc., across districts and schools

In this *Ask the Team* brief, we draw on reports from 13 states[1] that have already undertaken educator evaluation studies. Using these reports, we highlight strategies and examples for designing your own evaluation study as part of a continuous improvement process. Evaluation reports from these states offer a wealth of information that other states can use when planning for their own implementation processes.

[1] The states, their reports, and pertinent information about their evaluation systems are listed in Tables 4 and 5, under Bonus Resources.

How We Selected Educator Evaluation Implementation Studies

We used the following steps to conduct our review of state educator evaluation reports:

1. We selected states that implemented new educator evaluation systems within the past five years as part of pilots or full-scale implementations.

2. To identify evaluation implementation studies, we scanned the Center on Great Teachers and Leaders (GTL Center) Databases on State Teacher and Principal Evaluation Policies and conducted a Web search for pilot implementation studies available for states that released evaluation pilot studies since 2011.

3. We conducted direct follow-up with states when needed. For example, because information on Tennessee's 2013–14 pilot implementation was not readily available online, we interviewed a Tennessee SEA representative to gain insight into the state's evaluation program.[2]

4. Our final list included reports from 13 states. Of these,

    a. Ten states conducted teacher evaluation pilots, and eight states conducted principal evaluation pilots, often with a subset of districts or schools, to test and refine their initial system designs and implementation supports before rolling out the new evaluation systems statewide.

    b. Three states conducted multiyear studies or annual studies that evaluated new systems over a longer period of time, usually as the state scaled up from pilots to full, statewide implementation.

Definitions

Pilot Implementation—In a pilot implementation, the new evaluation system is implemented on a trial basis to see how well it is working and to determine what changes need to be incorporated to make it more effective before full implementation. Staff evaluations from the pilot implementation are typically not used for personnel decisions.

Interrater Agreement—The degree to which two raters, using the same scale, give the same rating in identical situations (e.g., while observing teaching practice or analyzing an artifact).

Fidelity of Implementation—The degree to which implementation is executed in accordance with the rules, procedures, and intended spirit of an implementation plan (i.e., faithfully executed).

Large-Scale Evaluation Study—Involves a greater number of participants as well as an expanded purpose and scope in terms of the research questions asked and the information gathered, which can increase the generalizability of results and the type and complexity of the data analysis and, therefore, the cost of the study.

Continuous Improvement Process—An implementation process based on iterative cycles of policy development, practical implementation, and assessment designed to improve upon itself with each iteration. Crucially, such a process has no identified end point but, rather, short-term and long-term benchmarks of success.

[2] Interview with Paul Fleming, Ed.D., deputy assistant commissioner and executive director of leader effectiveness at the Tennessee Department of Education. Dr. Fleming indicated that a report on the Tennessee principal evaluation pilot would be published and publicly available later in the summer of 2014.

1. START WITH THE QUESTIONS: WHAT DO YOU WANT TO LEARN?

Before selecting data collection methods, work with educators, researchers, and policymakers to fully identify the research questions that will guide the study. What is it that your SEA most needs to know and understand about pilot or full implementation? What aspects of the implementation process do you most need immediate information about? It may also be useful to prioritize research questions based on salience, funding availability, timeline, and capacity to complete the study. Some states have worked with external partners on development of the study. This could include some low- or no-cost options, such as local institutions of higher education, regional educational laboratories (RELs), or technical assistance centers, such as the GTL Center.

To get a better sense of the types of questions SEAs have been trying to answer through a pilot or implementation study process, we identified a list of common research questions that states have included in their pilot or implementation studies. It is worth noting that only eight of the 13 states explicitly identified their research questions in their study reports.

Table 1. Common Research Questions States Included in Their Pilot or Implementation Studies

Common Research Questions	Number of States That Included This Question
District and school fidelity of state model implementation	6 out of 8
Communication clarity between states, districts, and schools	5 out of 8
Correlation among evaluation system components (e.g., observation, school performance, artifact review, etc.)	3 out of 8
Stakeholder (e.g., evaluators, participants, unions, etc.) engagement and satisfaction with evaluation process and results	3 out of 8
Perceived effectiveness of training and support	6 out of 8
Impact on teacher or principal practice and effectiveness	8 out of 8
Intended and unintended consequences of evaluation system changes	4 out of 8
Sustainability of the evaluation system	4 out of 8

State Spotlight

Washington. The Washington state teacher evaluation pilot report asked the following research questions:

1. Which aspects of the legislative requirements are districts aware of, and where are the areas of misunderstanding or confusion?
2. What stage are districts at in terms of implementing and communicating about these changes?
3. How much variation exists across Washington districts in terms of implementation plans and timelines?

Source: Fetters, Behrstock-Sherratt, & Zhu, 2013

2. DATA AND DESIGN: CONSTRUCTING A STUDY DESIGN THAT ANSWERS YOUR QUESTIONS

After you have focused on the research questions you would like to be answered, you need to identify the best way to gather data to answer your questions. During this phase, you decide on the best design and data collection method(s) for your study.

Selected Study Designs

Some types of studies often implemented by states are focused studies, pilot studies, scaled studies, and annual studies. These four types of studies are defined in more detail below.

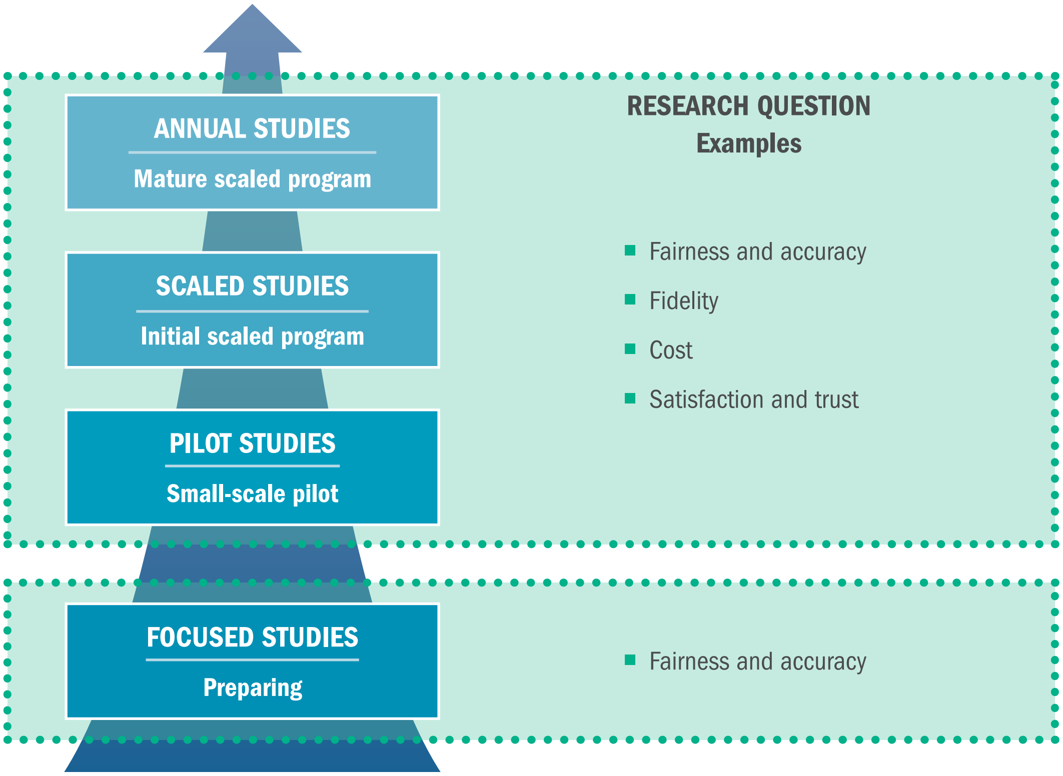
Focused Studies—Focused studies are similar to case studies in that they provide researchers an opportunity to evaluate system performance in a limited number of cases, or a diverse set of cases, by examining implementation effects of some aspect(s) of an evaluation system in various settings. These studies provide information about (1) educator interests and needs for a new evaluation system, (2) appropriateness of standards and content, and (3) accuracy of measures. Focused studies are essential to building trust among educators and the public that the new evaluation system will be fair and will inform design.

Pilot Studies—Pilot studies involve implementation of a few aspects, preferably all aspects, of an evaluation system and provide state or district task forces information about initial system implementation under the most optimal conditions. In the pilot phase, individual evaluation results generally do not count. Pilot studies often address questions about (1) training quality, (2) implementation fidelity, (3) time and other costs of implementation, (4) quality of measures, (5) use of results, (6) educator satisfaction, and (7) challenges to implementation.

Scaled Studies—Scaled studies generally involve implementation of the entire system at scale and are often multiyear studies (usually lasting one to three years) that are conducted early in the process of scaling up the new educator evaluation system. Scaled studies can expand on questions addressed by pilot studies but may also address interrater agreement questions at scale and the effects of implementation on educators' work, schools, and students. In general, these studies monitor implementation of principal and teacher evaluation systems simultaneously to give districts a sense of combined effects and implementation challenges.

Annual Studies—Annual studies occur after new educator evaluation systems have matured or have become routine. Annual studies typically involve analysis of data that local education agencies (LEAs) are required to report to the state by SEA staff or subcontractors. When hiring external evaluators, you should make sure that the evaluators are turning over data collection instruments, methods, etc., to you for ease of replication.

Selected Types of Studies



Source: Clifford, 2013

Data Collection Methods

Certain data collection methods lend themselves to answering certain categories of questions. Your design could include qualitative or quantitative data collection methods or both. The following is an overview of the data collection methods mentioned in the evaluation study reports from the 13 states that were reviewed for this brief.

Practitioner Surveys—Across the 13 states, the average response rate for online educator surveys was approximately 50 percent. Survey questions included both multiple-choice and open-ended responses. All states that conducted surveys administered them to educators after they received their summative evaluations at the end of the year. The majority of states also conducted midyear surveys before summative evaluations were available, and several states conducted surveys before the start of their pilots.

Focus Groups/Interviews—Eight states used focus groups and/or individual interviews with principals, teachers, and district administrators to gather input. Depending on the scope of implementation and available SEA capacity, some states conducted focus groups with all participating school districts, while others chose to focus on small subsets of districts and schools that could provide representative samples. States generally used annual focus groups to obtain finer detail on how to implement specific changes, such as human resources (HR) reforms.

Observations—In our review of 13 state pilot study evaluations, we identified five states that used observations as part of the evaluation study. Observations were more common when the system evaluation was conducted by an external organization.

Analysis of Evaluation Data—States often include quantitative and qualitative analyses of principal and teacher evaluation results. Out of 13 state evaluation reports reviewed, seven states had included analysis of evaluation data; the overall objective of this analysis was to analyze the validity, reliability, and equity of the evaluation system and to identify trends when conducted over multiple years. However, if you include this data collection method in your study, be sure to assess the accuracy of the evaluation data. The most common foci of these evaluations included:

1. Distribution of summative results (e.g., percentage of teachers receiving ratings of effective or highly effective)
2. Qualitative reviews of evaluation components (e.g., review of teacher and principal student learning objectives [SLOs] for rigor and alignment)
3. Correlation studies to compare observation scores with student achievement/growth scores
4. Trends across all elements of the evaluation system to assess differences across demographics, subject areas, geography, and other key components

Tables 2 and 3 highlight the data collection methods that states used for pilot implementation.

Table 2. Summary of Data Collection Methodology: Teacher Evaluation Systems

State	Practitioner Surveys	Focus Groups/ Interviews	Observations	Analysis of Evaluation Data
Colorado	✓			✓
Connecticut	✓	✓		✓
Delaware				✓
Maryland		✓		
New Jersey	✓	✓	✓	✓
Ohio	✓		✓	
Pennsylvania	✓	✓		✓
Rhode Island		✓	✓	✓
Tennessee	✓	✓		✓
Washington	✓	✓		

Table 3. Summary of Data Collection Methodology: Principal Evaluation Systems

State	Practitioner Surveys	Focus Groups/ Interviews	Observations	Analysis of Evaluation Data
Colorado	✓			✓
Georgia	✓	✓	✓	✓
Maryland		✓		
Minnesota	✓	✓		
New Jersey	✓	✓		✓
Rhode Island	✓			✓
Tennessee	✓	✓		
Wisconsin	✓	✓	✓	

State Spotlight

Ohio. The Ohio teacher evaluation pilot study used three data collection methods—practitioner surveys, observations, and case studies. Survey results at the beginning of the year allowed districts to select a representative sample (based on geography, their chosen assessment methodology, and type of district) of LEAs to observe more closely (Zoller, 2012). Ohio's initial and summative survey questions for administrators, union leaders, evaluators, and educators can be found here.

State Spotlight

Maryland. The Maryland teacher and principal evaluation pilot studies used focus groups as one data collection method. LEA focus group leaders led discussions with central office personnel, principals, and teachers separately and then combined their feedback to inform their pilot implementation (Dolan, 2013). These sessions were held once in the spring for each focus group.

3. CONTEXT, CONTEXT, CONTEXT: BALANCING PRACTICALITY AND RIGOR IN STUDY DESIGN

Regardless of the type of study under consideration, SEAs face a balancing act in designing an evaluation study. In an ideal scenario, you create thorough, rigorous, and intensive studies; in reality, however, you must balance this desire for rigor with the practicalities of pilot implementation in your own state context. Several important contextual considerations include:

- **Implementation timelines: For pilots, to what extent does the pilot implementation timeline allow for the new evaluation system to be truly piloted so that early or nonvalidated results will not influence human capital decisions?** Although many states are implementing evaluation systems with timelines dictated by state legislation or federal grant requirements, several pilot studies highlight the benefits of delaying links between educator evaluation systems and high-stakes personnel decisions until evaluations have been validated for accuracy, reliability, and equity.

- **Scale-up process: Is the state scaling up all components of the evaluation system at once, or are certain components (e.g., student growth) being phased in over time?** States should anticipate that teachers and principals may be wary of the validity and fairness of student-related measures. Some states have chosen to phase in the use of student achievement measures over time or to use abbreviated versions of their student achievement measures with select members of a school's staff so that educators can experience the process in a nonevaluative manner.
- **Consistency of HR databases: To what degree are the state's data systems capable of capturing and maintaining educator performance data necessary for the study to be carried out (e.g., what educator performance evaluation data are LEAs required to electronically capture and report to the state, to what extent are evaluation results tied to HR decisions)?** There are legal and logistical concerns to consider when using unrefined evaluation systems for dismissal or transfer decisions. Evaluations during the pilot should not be used for personnel decisions, but they can be used to gauge how evaluation results will influence summative ratings.
- **Funding: Does the state have sufficient funds, and the staff, to carry out the data collection and analysis required for completing the study (or to hire an external contractor to do so)?** States often employ grants and private foundation funding to support implementation of their educator evaluation systems. Recognizing the short-term nature of these funding streams, some states used consultants and local education organizations to provide training and implementation support rather than hire staff that could not be sustained beyond the duration of grant funding. Short-term funds can also be used to develop tools, training materials, and other resources that do not require ongoing funding to sustain. For more ideas on funding this work, see the GTL Center's related Policy Snapshot, *Evaluating Evaluation Systems: Policy Levers and Strategies for Studying Implementation of Educator Evaluation*.

State Spotlight

Pennsylvania. The Pennsylvania Department of Education conducted a teacher evaluation implementation pilot study over three years from 2010–2013. The initial study included 10 districts, and this number rose to 300 over the three years of the study (McGuinn, 2012). In addition, it included quantitative and qualitative analyses of the evaluation results. These analyses were used to analyze the validity, reliability, and equity of the evaluation system (Lipscomb, Chiang, & Gill, 2012).

State Spotlight

Colorado. The Colorado Department of Education conducted both teacher and principal evaluation implementation pilot studies between 2011–2012 and 2012–2013. Each study involved more than 20 pilot districts. The themes that emerged from initial closed- and open-ended survey questions were used to track the progress of implementation in subsequent surveys.

Source: Colorado Department of Education, 2014

4. TAKING THE LONG VIEW: EVALUATION STUDY DESIGN FOR THE LONG HAUL

One component of a successful evaluation study is a plan for the long term. Such a study allows the new evaluation system to be assessed continuously and may identify ways that it can be improved over the implementation timeline and beyond. How can such a study be designed? What are some important elements of a long-term study design?

- **Implement multiyear pilots with sustained support:** Although most states experienced significant improvements in educator awareness and comfort of evaluation systems after the first year of implementation, many states determined that one year of pilot implementation was insufficient to establish practices and systems that provide acceptable levels of accuracy, reliability, and equity. In Year 1, educators will likely be consumed with the process of the evaluation systems and cannot reasonably be expected to fully understand observation rubrics and SLO systems. For most educators, operationalizing the indicators will require several years of experience. If implementing multiyear pilots is not possible, the results of the single year pilot study should be viewed in the context of the limitations described above.
- **Facilitate sustained cross-district communication of evaluation study results:** Pilot studies have led to the development of training and support materials for initial implementation of evaluation systems. In addition, high-capacity districts have used these studies to develop resources and tools that can be shared across their states. Over time, as states develop new resources and tools, continued cross-district communication can contribute to long-term continuous improvement across the state.

State Spotlight

Tennessee. Tennessee included quantitative and qualitative analyses of principal and teacher evaluation results. The overall objective of this state's data analyses was to analyze the validity, reliability, and equity of the evaluation system and to identify trends when conducted over multiple years. In addition, Tennessee used implementation pilots to make modifications to their observation rubrics. For example, it consolidated the number of indicators on its principal observation rubric from 22 to 17 during the pilot year and modified the weighting of the two required, annual observations—prioritizing end-of-year observations to focus on growth and mastery of standards throughout the school year. Additional information on Tennessee's educator evaluation can be found here.

Bonus Resource

Tables 4 and 5 indicate which states were analyzed for this brief and provide basic information about pilot implementation.

Table 4. Teacher Evaluation Implementation Studies

State	Pilot Years	Pilot Scope	Key Resources
Colorado	2011–2012, 2012–2013	26 local education agencies (LEAs)	Pilot report Survey data Implementation guidance
Connecticut	2012–2013	14 districts	Pilot report Evaluation resources (e.g., SLO handbook, sample surveys)
Delaware	2011–2012	19 districts	Pilot report
Maryland	2011–2012, 2012–2013	7 LEAs (2011–2012) 22 LEAs (statewide field test in 2012–2013)	Pilot report
New Jersey	2011–2012, 2012–2013	30 LEAs (both pilots)	Pilot report Teacher survey and interview
Ohio	2011–2012	139 LEAs	Pilot report
Pennsylvania	2010–2013	10 districts (2010–2011) 100 districts (2011–2012) 300 districts (2012–2013)	Pilot report Evaluation rubric Report on value-added models
Rhode Island	2011–2012	32 districts	Pilot report Step-by-step instructions on creating student growth measures
Tennessee	2010–2011	30 districts	Pilot report Evaluation system Evaluation tools
Washington	2011–2012	16 LEAs in 2011–2012	Pilot report (1) Pilot report (2) Glossary of terms used in evaluation

Table 5. Principal Evaluation Implementation Studies

State	Pilot Years	Pilot Scope	Key Resources
Colorado	2011–2012, 2012–2013	23 LEAs, 241 principals (2012) 21 LEAs, 410 principals (2013)	Pilot report (2012–2013) Feedback survey results State website with resources
Georgia	2011–2012	26 LEAs	2012 pilot evaluation report
Maryland	2011–2012, 2012–2013	7 LEAs (2011–2012) 22 LEAs (statewide field test in 2012–2013)	Teacher and principal evaluation field test
Minnesota	2012–2013	17 LEAs, 102 principals	Pilot study findings and resource page
New Jersey	2012–2013	13 LEAs	Pilot final report Lessons from educators Principal surveys
Rhode Island	2012–2013	Statewide pilot	Pilot year report State website with resources
Tennessee	2013–2014	10 LEAs, 250 principals	Administrator evaluation website
Wisconsin	2012–2013, 2013–2014	115 LEAs (2013) 225 LEAs (2014, statewide)	Pilot Year 1 summary Overview of pilot evaluation design

I WANT TO KNOW MORE!

Center on Great Teachers and Leaders. (2014). *Databases on state teacher and principal evaluation policies.* Washington, DC: Author. Retrieved from http://resource.tqsource.org/stateevaldb/

Cheney, G. R., Davis, J., Garrett, K., & Holleran, J. (2010). *A new approach to principal preparation: Innovative programs share their practices and lessons learned*. Retrieved from http://www.anewapproach.org/download.html

Clifford, M. (2013). *Designing educator evaluation systems: Empirical approaches to system improvement.* Presentation given at the U.S. Department of Education Waiver Flexibility Workshop.

Dolan, M. (2013). *Maryland teacher and principal evaluation field test.* Retrieved from http://msde.state.md.us/tpe/Exhibit_W_Field_Test_Rpt_Dolan.pdf

Fetters, J., Behrstock-Sherratt, E., & Zhu, B. (2013). *Washington Teacher/Principal Evaluation Project (Year 3) report: Final project report to the Office of State Superintendent of Public Instruction.* Retrieved from http://tpep-wa.org/wp-content/uploads/AIR%20TPEP%202012-2013%20Report.pdf

New Leaders. (2012). *Improving principal preparation: A review of current practices & recommendations for state action.* New York, NY: Author. Retrieved from http://www.newleaders.org/newsreports/publications/improving-principal-preparation/

Turnbull, B., Riley, D., & MacFarlane, J. (2013). *Cultivating talent through a principal pipeline—Building a stronger school principalship: Volume 2.* New York, NY: The Wallace Foundation. Retrieved from http://www.wallacefoundation.org/knowledge-center/school-leadership/principal-training/Documents/Building-a-Stronger-Principalship-Vol-2-Cultivating-Talent-in-a-Principal-Pipeline.pdf

Zoller, S. (2012). *Ohio Department of Education Teacher Evaluation System (OTES) pilot.* Olympia, WA: MGT of America. Retrieved from http://education.ohio.gov/getattachment/Topics/Teaching/Educator-Evaluation-System/Ohio-s-Teacher-Evaluation-System/Additional-Information/5150-OTES-Final-Report-Appendices.pdf.aspx

For more examples of or information about educator evaluation studies, please e-mail gtlcenter@air.org.

Roshni Menon, Ph.D., is a researcher in the Education program at American Institutes for Research, where she works on multiple research and evaluation projects revolving around educator quality. She also provides technical assistance support for the GTL Center.

Alex Berg-Jacobson, M.P.P., is a technical assistance associate in the Education program at American Institutes for Research. He works on research projects revolving around educator quality and provides technical assistance support for the GTL Center.

Tim Field, M.B.A., is a senior consultant with Public Impact. He consults and leads project teams to provide clients with research-based guidance on a wide range of policy and school management issues, such as educator effectiveness, evaluation, retention, and compensation.

Brendan Yorke is a research analyst with Public Impact, collecting, analyzing, and communicating school outcome and student achievement data.