What's Happening

November 2014

Principal and teacher perceptions of implementation of multiple-measure teacher evaluation systems in Arizona

Stephen J. Ruffini
Reino Makkonen
Jaclyn Tejwani
Marycruz Diaz
WestEd

Key findings

School districts in Arizona that piloted new teacher evaluation models in 2012/13 used multiple measures. Information from surveys and focus groups indicates that teachers and principals in these districts view standards-based teacher observations as the most credible form of evaluation but are open to incorporating student performance and stakeholder survey measures if they are comprehensible, fair, and consistent. Some 39 percent of surveyed teachers from pilot districts agreed that their final classification under the new system accurately reflected their performance, 32 percent indicated that it did not, and 30 percent were undecided.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S. Department of Education

REL WEST
Regional Educational Laboratory At WestEd

REL 2015–062

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

November 2014

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-12-C-0002 by Regional Educational Laboratory West (REL West) at WestEd. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Ruffini, S., Makkonen, R., Tejwani, J., & Diaz, M. (2014). *Principal and teacher perceptions of implementation of multiple-measure teacher evaluation systems in Arizona* (REL 2015–062). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from http://ies.ed.gov/ncee/edlabs

This report is available on the Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

Summary

Recent years have seen a massive effort to overhaul teacher evaluation methods. Nearly two-thirds of states have made changes to their teacher evaluation policies since 2009. Many states require annual evaluations, often based on the results of multiple measures of performance. As states continue to wrestle with developing and implementing evaluation systems, only a small number of studies have sought to contextualize these reforms and explore how their implementation might provide insights to improve policy.

This descriptive study explores how new teacher evaluation systems were put in practice in 10 volunteer districts in Arizona after a shift in state policy. A collaboration between the Regional Educational Laboratory West and the Arizona Department of Education, this study examines the initial implementation of new teacher evaluation models aligned with the Arizona Framework for Measuring Educator Effectiveness. In fall 2012 five Arizona local education agencies (four school districts and one charter school, referred to collectively here as five pilot districts) volunteered to test the Arizona Department of Education's new teacher evaluation model, and five other districts (referred to here as partner districts) volunteered to share feedback about the initial implementation of their locally developed but Arizona Framework–aligned models of teacher evaluation. Data were provided to the study team by the Arizona Department of Education and were collected through an end-of-year teacher survey and multiple focus groups with teachers and principals.

The study's goals were to describe challenges or unintended consequences from the first year of implementation, teacher perceptions of the accuracy and usefulness of the piloted evaluation measures, and perceived changes in teachers' instructional practices or in collaboration among teachers and administrators. Participants' perceptions can provide valuable information to state and district leaders about the extent to which new measures are being implemented as intended and are providing useful information to teachers and principals, thus complementing more-empirical analyses of new measures' reliability and validity.

Some similar themes emerged across pilot and partner districts—particularly about the time demands of the systems and perceptions of traditional measures of teacher effectiveness compared with the new measures. However, the pilot and partner districts were implementing different systems in different contexts, and their results often differed.

Time constraints limited implementation of the teacher evaluation models, but principals indicated that online resources helped reduce the time burden. Participating teachers and principals reported mixed feelings about their training in preparation for implementation.

Perceptions of the accuracy and usefulness of evaluation measures

- Some 39 percent of teacher survey respondents from pilot districts agreed that the final summative performance classification they received from the new teacher evaluation process accurately reflected their overall performance in 2012/13, 32 percent indicated that it did not, and 30 percent were undecided.
- Responding teachers and principals from pilot and partner districts viewed classroom observation as the most credible form of evidence about teacher effectiveness.
- Post-observation conferences provided meaningful feedback on how to improve instruction, according to teacher survey respondents from pilot districts.

- Teachers in two pilot and three partner districts expressed concerns about consistency in classroom observation ratings by principals (inter-rater reliability), and teachers in three pilot districts expressed concerns about the number and type of observations needed to accurately rate teacher performance.
- While teacher survey respondents from pilot districts were generally supportive of using student assessment data in teacher evaluations, focus group participants expressed concerns about the accuracy and fairness of the pilot-year methodology used to incorporate the student assessment data into teacher ratings.
- Respondents had mixed views about stakeholder surveys as a measure of teacher effectiveness. In particular, problems with the administration of these surveys in 2012/13 contributed to negative opinions about their credibility.

Perceptions of changes in work behaviors following initial implementation

- Teachers from all five pilot districts reported being more reflective about their instructional practice and professional development during the initial pilot implementation year (2012/13) than in previous years; principals from pilot and partner districts reported corresponding perceptions.
- Principals from 6 of the 10 study districts reported that their instructional leadership abilities had improved.
- Principals from all five pilot districts reported that their interactions with teachers were more collaborative in 2012/13, but responding teachers from pilot districts did not share this perception.
- Teachers and principals from four pilot districts suggested that participating teachers appeared to be working together more collaboratively in 2012/13, but principals from two of these pilot districts attributed this to work on the upcoming implementation of Common Core State Standards.

School and district capacity issues emerged in implementing the evaluation models, and time constraints, inter-rater reliability concerns, ongoing training and support, and the use of technology were key issues that may need to be addressed. Teacher respondents viewed traditional performance assessments more favorably and were more skeptical about incorporating results from student assessments and stakeholder surveys into teacher evaluations. Correspondingly, study participants had mixed perceptions of the new teacher evaluation systems.

Contents

Why this study?

Responding to new state laws and federal grant requirements, most states have changed the way they evaluate teachers in the past several years. Nearly two-thirds of U.S. states have changed their teacher evaluation policies since 2009 (Jerald, 2012), and 43 states now require annual evaluations of all new teachers (National Council on Teacher Quality, 2012). Federal grant applications for Race to the Top in 2009 and 2010 required states to design comprehensive evaluation systems with multiple measures of teacher performance, including student achievement growth measures, student reflections and feedback, and teacher observations (Overview Information: Race to the Top Fund, 2010). Applications for flexibility in meeting Elementary and Secondary Education Act provisions require states to describe their plans to reform teacher and principal evaluation and support systems to focus on quality of instruction and student results (U.S. Department of Education, 2012).

Since 2012/13 state law has required Arizona school districts to evaluate teachers annually using an evaluation instrument that meets the requirements of the state board–approved Arizona Framework for Measuring Educator Effectiveness

Yet states that have been awarded Race to the Top grants are still struggling with implementing teacher and principal evaluation systems (McNeil, 2013). As states wrestle with this issue, it is important to understand the reforms in the light of "what is implementable and what works for whom, [and] where, when, and why" (Honig, 2006, p. 2). As states rush to implement new standards-based, multiple-measure teacher evaluation systems, researchers have raised concerns that policymakers have left little time for the important trial-and-error process necessary for successful implementation (see, for example, Mead, Rotherham, & Brown, 2012).

The best way to ensure the efficacy and sustainability of any new teacher evaluation system is to systematically monitor its initial performance and examine whether stakeholders value and understand the system and how it affects teacher practice (Goe, Holdheide, & Miller, 2011). Earlier studies explored how implementation can reshape policy (McLaughlin, 1990; Fowler, 2004). However, the few studies that have sought to contextualize these reforms have raised concerns. In Chicago, for example, although teacher observation ratings were correlated with student performance on achievement tests, teachers complained that principals often talked for most of the post-observation conference rather than interacting with them (Sartain, Stoelinga, & Brown, 2011). In Seattle principals supported the district's new teacher observation framework but reported a lack of solid calibration training, leading to speculation about inconsistencies in teacher evaluation ratings (Flaherty, Tejwani, & Rodriguez, 2012). Tennessee's evaluation of its new statewide system found that local capacity for high-quality feedback and targeted professional development (ideally based on teacher evaluation results) varied considerably across school districts, and many administrators bemoaned the burdensome and time-consuming data entry requirements (Tennessee Department of Education, 2012).

In line with national policy trends, Arizona adopted a law (Senate Bill 1040) in May 2010 requiring school districts and charter schools to evaluate teachers annually.[1] The bill empowered the Arizona State Board of Education to adopt a teacher evaluation framework that includes quantitative data on student performance. Since the 2012/13 school year, state law has required Arizona school districts to evaluate teachers annually using an evaluation instrument that meets the requirements of the state board–approved Arizona Framework for Measuring Educator Effectiveness. By 2013/14 districts were to specify how teacher and principal performance classifications will be used in employment-related decisions.

The Arizona Framework has three components: classroom academic progress data, teaching performance, and school data. Each may contain subcomponents, including measures of student academic progress (classroom academic progress data), observations of instructional practice in classrooms (teaching performance), and stakeholder surveys (school data).

Arizona officials requested this study, conducted by Regional Educational Laboratory West in collaboration with the West Comprehensive Center, to explore teacher and principal experiences with the teacher evaluation models in several districts to help refine state guidance and policy in advance of broader implementation. Other Arizona districts may want to use these findings to modify their evaluation systems or implementation strategies. State officials outside Arizona working to design or implement teacher evaluation systems can use the study to better understand what type of policy guidance might support school districts in a pilot year.

What the study examined

This study examines the initial implementation of new teacher evaluation models aligned with the Arizona Framework for Measuring Educator Effectiveness

This descriptive study examines the initial implementation of new teacher evaluation models aligned with the Arizona Framework for Measuring Educator Effectiveness.[2] In response to the Arizona Department of Education's outreach to school districts across the state in summer 2012, five Arizona local education agencies (four school districts and one charter school, referred to collectively here as five pilot districts) volunteered to test the department's new teacher evaluation model (box 1),[3] and five other districts (referred to here as partner districts) volunteered to share feedback about initial implementation of their own teacher evaluation models that were locally developed but aligned with the Arizona Framework.[4] Understanding how these teacher evaluation models were implemented during the initial year and how participating teachers and principals perceived the implementation process offers the state the opportunity to modify training, guidance, and policy.[5]

This study addresses three research questions:

- What challenges or unintended consequences did teachers and principals perceive in the initial implementation of the evaluation systems?
- How did teachers and principals perceive the effectiveness of the piloted evaluation measures?
- How did teachers and principals perceive changes in participating teachers' instructional practice and in knowledge sharing and collaboration among teachers and administrators following initial implementation of the new evaluation systems?

The Arizona Department of Education provided data from focus group transcripts and end-of-year teacher surveys to the researchers for analysis. Because of a low response rate from partner districts, survey data are reported for pilot districts only. Only comments that were raised in more than one focus group are included. Prevalence of an opinion within a group could not be determined since the transcripts did not distinguish among participants.

The study methodology is described in box 2, with more detail in appendix A.

Box 1. The 2012/13 Arizona Department of Education teacher evaluation model

The Arizona Department of Education's teacher evaluation model is based on the Arizona Framework for Measuring Educator Effectiveness and includes three components: teaching observations; surveys of students, parents, and peer teachers; and measures of student academic progress. A teacher's composite evaluation score is a weighted average of the three component scores, which were assigned weights of 50 percent (observation), 17 percent (surveys), and 33 percent (academic progress). The Arizona Department of Education contracted with Teachscape (an organization affiliated with Charlotte Danielson, the creator of the Framework for Teaching) for its observation instrument, associated training, and online platform and tools. Participating pilot teachers were observed by their supervising administrator, who conducted two classroom observations over the 2012/13 school year.

The department also administered online end-of-year surveys to students in grades 3–12 and to participating parents and peer teachers. The student survey was based on public domain items from Cambridge Education's Tripod Student Perception Survey and asked students to rate their teacher on the extent to which the teacher engages and challenges students. The school-level parent survey asked parents to rate the quality of their child's school, its teachers, and the administration on an A–F rating scale. The 15-question peer survey asked teachers to rate their peers' performance on a four-point ordinal scale.

Due to variations in the nature and extent of available test data, the Arizona Department of Education model relies on a number of formulas to calculate student academic progress depending on the grade level and subject area taught, and the state has published more than 40 data tables for various teacher role groups. Updated versions of the state's teacher data tables are available at http://www.azed.gov/teacherprincipal-evaluation/teacher-rating-tables/, and the state's teacher evaluation process for pilot districts in 2013/14 is online at http://www.azed.gov/teacherprincipal-evaluation/files/2012/10/teacher-evaluation-v4.0-website-update-11_22_13-sl.pdf.

Source: Arizona Department of Education.

Overall, 39 percent of surveyed teachers from pilot districts agreed that their final summative performance classification accurately reflected their overall performance in 2012/13, 32 percent indicated that it did not, and 30 percent were undecided

What the study found

Teachers from pilot districts considered observations to be a useful form of performance evaluation when based on the Danielson Framework for Teaching rubric, along with a post-conference interview with their administrator. Teachers were open to measures involving student performance and stakeholder surveys, but only if they perceived the metrics to be fairly and consistently applied. Teachers and principals reported that teachers had become more reflective about their teaching practice and had instituted new practices during the initial implementation year. Overall, 39 percent of teacher survey respondents from pilot districts agreed that their final summative performance classification accurately reflected their overall performance in 2012/13, 32 percent indicated that it did not, and 30 percent were undecided. The findings are discussed in detail below.

What challenges or unintended consequences did teachers and principals perceive in the initial implementation of the evaluation systems?

Time constraints limited implementation of the teacher evaluation models, but principals indicated that online resources eased the time burden. In focus groups, principals from all pilot and partner districts expressed concern about the time required to implement

Box 2. Data and methods

Selecting study participants

To study the implementation of new teacher evaluation systems, the Arizona Department of Education collected information from a group of pilot and partner districts. In the pilot group, which adopted the state department's model, the department sought to include a large school district in a large county, a small school district in a large county, a school district in a small county, and at least one charter school, given the large number of charter schools in the state. In the partner group, which implemented locally designed but Arizona Framework–aligned models, the department wanted a set of large school districts. Five pilot and five partner districts participated in the study. The pilot districts (including one charter) had 297 participating teachers across 12 participating schools, and the partner districts had 3,139 potential participating teachers across 70 schools.

Data collection and analysis

The Arizona Department of Education engaged West Comprehensive Center researchers to conduct focus groups with participating principals and teachers. It also developed an online end-of-year (May 2013) survey for participating teachers in pilot and partner districts to gather data on their perceptions and experiences with the new evaluation models. The study team analyzed the focus group transcripts and the 2013 teacher survey data. The table summarizes the study's data sources and analysis methods. A more detailed description is in appendix A.

Data sources and analysis methods

Data source	Description	Analysis
Interview and focus group transcripts	Arizona Department of Education consultants conducted a series of semi-structured interviews and focus groups with small samples of participating teachers and principals from pilot and partner schools and districts. These discussions were recorded and transcribed. Thirteen principals and 46 teachers from pilot districts and 41 principals and 46 teachers from partner districts participated.	A set of codes was developed and refined to review all transcripts. The initial codes were based on the main concepts of the focus group protocols; additional codes addressed new concepts that emerged from initial transcript reviews. Two reviewers coded each transcript. Inter-rater reliability was established on an initial principal and teacher transcript and again after the reviewers coded approximately two-thirds of the transcripts. At each point and for each transcript, inter-rater reliability met or exceeded the required minimum level of 0.8. Each reviewer then summarized the coded passages for different research questions, from which themes were identified.
Teacher survey	The department emailed participating teachers a link to an online survey. The raw survey data were given to the study team. A total of 165 teachers from pilot districts (61 percent) participated in the survey; responses by 157 were included in the analysis. Links to the online survey were also provided to approximately 2,650 teachers in partner districts, and 622 (23 percent) responded; due to the low response rate, no information from the surveys of teachers in partner districts is reported here.	The study team determined the response rates then summarized the survey data. Based on early anecdotal evidence from the state, it also explored differences between teachers with varying years of experience as well as between elementary and secondary teachers. Only statistically significant group differences ($p \leq .05$) were reported. See appendix C.

the teacher evaluation models, citing, for example, less time for in-depth conversations with teachers and less time to visit classrooms informally than they had previously had. However, principals from three pilot and five partner districts also cited the time benefits of online platforms that allowed them to enter comments while observing a teacher and then gave them the option to share the feedback with the teacher electronically. Teachers from two partner districts indicated in focus groups that online feedback sometimes replaced face-to-face conversations, and several teachers noted that they had difficulty interpreting the online feedback. Such difficulty is an important factor to weigh against the technology's potential efficiencies.

Participating teachers and principals had mixed feelings about their training in preparation for implementation. Principals from all five pilot districts commented positively on the training they received from the state (box 3), calling it comprehensive and beneficial and said they were able to apply what they had learned. They stressed the value of the training for evaluating the Classroom Environment and Instruction domains of the Danielson Framework for Teaching (Danielson Group, 2011). These principals reported feeling less prepared to evaluate their teachers on the Danielson framework's nonobservational domains (Planning and Preparation and Professional Responsibilities).

Principals from all pilot and partner districts expressed concern about the time required to implement the teacher evaluation models, citing, for example, less time for in-depth conversations with teachers and less time to visit classrooms informally than they had previously had

Teachers from three pilot districts reported in focus groups that the training they received in fall 2012 lacked information about the evidence they needed to submit in support of their ratings in the Danielson Framework's nonobservational domains. They also felt underinformed about how the new evaluation model's various measures would contribute to their summative rating. Teacher survey respondents from pilot districts echoed these concerns. Some 39 percent of teacher survey respondents from pilot districts agreed that their training was adequate to effectively participate in the first year of the pilot, and 32 percent disagreed. (Appendix B summarizes the survey responses for teachers from pilot districts, while appendix C presents the four significant differences by teacher subgroup in the survey results.)

Box 3. Support for initial implementation of the Arizona teacher evaluation model, 2012/13

Training for the implementation of the Arizona Department of Education's teacher evaluation model began in October 2012, with an orientation for administrative leaders of each pilot district. Participants received an overview of the new Arizona Framework for Measuring Educator Effectiveness, the Danielson Framework for Teaching, sample parent and student surveys, a description of the necessary data entry and collection work, and a calendar for the 2012/13 school year. Teacher orientations on the model were held in each pilot district in November and December 2012. Also in November, classroom observers received three days of in-person professional development on the Danielson Framework for Teaching, 30–40 hours of video practice in observation techniques (primarily online), and a (practice) online assessment exam. Staff from the department's research and evaluation division trained data specialists (who had experience with data collection, management, and analysis) from each pilot district in November 2012, walking them through the processes for collecting and tabulating teacher evaluation data. Follow-up data training was provided in May 2013. Finally, the department hosted weekly calls with key contacts in pilot districts to gather formative feedback throughout the 2012/13 school year.

Source: Arizona Department of Education personnel.

In the partner districts, training varied in delivery and content, and focus group participants' end-of-year perceptions of their training were mixed. Some partner districts relied on external experts or online webinars to provide training directly to participants, while others relied on a train-the-trainer model in which information from professional development sessions was brought back to schools by attendees, who then trained participants. Focus groups indicated that some partner districts embedded training in teachers' weekly meetings, with master teachers going through the observation rubric in detail over the course of a semester. In focus groups principals from four partner districts said that their trainings were comprehensive but not always sufficient, and principals from three partner districts complained that they needed more detailed information about implementation (for example, a timeline to keep their evaluations on track during the school year).

How did teachers and principals perceive the effectiveness of the piloted evaluation measures?

Teachers and principals from all five pilot districts indicated that classroom observations, coupled with feedback, were the most beneficial components of the new teacher evaluation model for improving teacher practice

Responding teachers and principals viewed classroom observation results as the most credible evidence of teacher effectiveness. Teachers and principals from all five pilot districts indicated in focus groups that classroom observations, coupled with feedback, were the most beneficial components of the Arizona Department of Education's new teacher evaluation model for improving teacher practice. Moreover, principals from all five pilot districts and teachers from three pilot districts cited the benefits of the Danielson Framework for Teaching rubric. Principals noted that the framework provided a common language that enabled them to converse with teachers about practice; that it yielded more objective, evidence-based feedback for teachers than their previous system; and that it allowed them to identify specific areas of teacher difficulty.

In addition, teachers from three pilot districts pointed out that the level of evidence and required documentation for the Danielson Framework for Teaching rubric yielded less subjective ratings than those used in the past, and principals from all five pilot districts reported that the framework's reliance on evidence increased teacher support for the new evaluation system.[6] Some 60 percent of teacher survey respondents from pilot districts agreed that their post-observation conference provided meaningful feedback on how to improve their instruction (23 percent disagreed).

Most teacher survey respondents from pilot districts expressed confidence in administrator observations as a performance measure. Some 64 percent indicated that they had confidence in their evaluator's ability to accurately rate their instructional practice. In addition, 79 percent reported that formal classroom observations by their principal could provide either a moderately or a highly accurate measure of their teaching performance.

Focus group participants raised concerns about inter-rater reliability and about the number and type of observations needed to accurately rate teacher performance. In focus groups, teachers from two pilot districts expressed concern over the lack of calibration among their evaluators, with some noting that the quality of evaluations and feedback varied by rater (this point was also mentioned by principals from two pilot districts). Principals and teachers from three partner districts also raised general concerns in focus groups about inter-rater reliability. Focus group participants from these two pilot districts and three partner districts suggested that principals need further training to evaluate teachers consistently.

Teachers and principals from three pilot districts also questioned the credibility of relying on only two formal, scheduled observations to rate teachers' performance, suggesting that incorporating more unscheduled observations would yield a more authentic (unscripted) view of teacher practice and could provide feedback to teachers throughout the school year. Some 51 percent of teacher survey respondents from pilot districts agreed that the number of formal observations they received was adequate to assess their performance (26 percent disagreed).

Although teacher survey respondents from pilot districts were generally supportive of using student assessment data in teacher evaluations, they expressed concerns about the accuracy and fairness of the pilot-year methodology. Principals and teachers from all five pilot districts expressed concerns about using student performance data to rate teachers in the first pilot year of the Arizona Department of Education's evaluation model. For example, in focus groups teachers from two pilot districts questioned the fairness of rating teachers based on student test results from the prior year. Principals from four pilot districts noted that they did not fully understand how to interpret the data tables and cutscores and questioned the fairness of using different assessments and formulas to calculate scores for teachers in different grades and content areas.

Despite some concerns more than half of responding teachers from pilot districts reported that current-year standardized test scores from their classrooms could provide a moderately or highly accurate assessment of their teaching performance

Despite these concerns with the initial-year methods, which were set to change in 2013/14,[7] survey results indicated that more than half of responding teachers from pilot districts supported using standardized test scores (if not the Arizona Department of Education's 2012/13 methodology). Specifically, 57 percent of teacher survey respondents from pilot districts reported that current-year standardized test scores from their classrooms could provide a moderately or highly accurate assessment of their teaching performance, while 60 percent felt that way about standardized schoolwide test scores.[8]

Focus group participants from partner districts expressed skepticism about the varied use of student academic progress data in teacher evaluations. Teachers from four partner districts questioned the fairness of using different calculations to assess the performance of teachers with classroom-level standardized test data and those without such data. Teachers from three partner districts noted that their schools enrolled many high-need students and suggested that growth on standardized tests may not be their highest instructional priority. Furthermore, principals from a partner district that used a state-generated school letter grade for all teacher evaluations worried that their teachers would move to schools that received better letter grades to boost their overall evaluation rating. Teachers from this district also expressed concern that using a schoolwide grade might not reflect their classroom performance and could be detrimental to their overall rating. For these reasons they did not view schoolwide results as a true measure of their performance.

Respondents had mixed views about the credibility of stakeholder surveys as a measure of teacher effectiveness, especially after problems with survey administration in 2012/13. The opinions of teacher survey respondents from pilot districts were split about the accuracy of student surveys in assessing teacher performance. Half felt that such surveys could provide either a moderately or highly accurate assessment of their performance, and half felt that such an assessment was not accurate or had only low accuracy. In focus groups teachers expressed negative perceptions of student surveys used for teacher evaluation. Teachers from four pilot districts viewed 2012/13 student surveys as less credible than surveys of parents or peer teachers because they did not think their

students fully understood the implications of the results and might have marked responses indiscriminately.

Survey and focus group assessments of the use of peer review surveys also differed. Some 70 percent of teacher survey respondents from pilot districts reported that peer teacher surveys could provide either a moderately or highly accurate assessment of their performance, compared with 50 percent for student surveys and 46 percent for parent surveys. However, in focus groups teachers from all five pilot districts cited problems with the logistics of peer review in the initial year, with some reporting discomfort in responding honestly because of a perceived lack of anonymity. Others noted that they could not accurately answer certain questions because they had not observed their peers' practices or were unaware of their professional development activities or memberships in professional organizations.

Survey and focus group assessments of the use of peer review surveys differed

Teachers and principals also expressed concerns about using parent survey results in teacher evaluations. Some 46 percent of teacher survey respondents from pilot districts indicated that parent surveys could provide a moderately or highly accurate assessment of teacher performance. In focus groups principals from two pilot districts noted that they did not have enough computers for parents to complete the surveys online and that parents who did not have computers or Internet access at home were unable to complete the survey (thus response rates were low). Teachers from another pilot district noted in focus groups that instead of emailing survey links to parents, some schools administered the survey during a scheduled parent event at the school (using the computer lab), which they felt worked well.[9]

Focus group participants from partner districts had mixed perceptions of the credibility and usefulness of stakeholder surveys. Teachers and principals from one district reported that they did not think the student survey was a credible measure of teacher effectiveness because they did not believe students took the survey seriously (a concern also cited by focus group participants from the pilot districts). However, a focus group participant from another partner district saw value in using peer and student surveys, though only after receiving substantial training on the survey and its results. Focus group participants from another partner district viewed their district's online parent survey as unreliable because of the low response rate and the lack of computer access for many parents (also noted by focus group participants from pilot districts).

Teachers from pilot districts raised concerns about the accuracy of the final performance classifications they received on the new teacher evaluation model. In focus groups teachers from all five pilot districts expressed concerns about the accuracy of the summative results of their evaluations. They raised questions about the fairness of the model's weightings of students' scores on the Arizona Instrument to Measure Standards (a state standards-based assessment for measuring student proficiency) and on supplemental assessments (such as the Stanford 10 and Galileo tests), as well as the formulas principals used to calculate scores for teachers working in different grades and subjects. Overall, 39 percent of teacher survey respondents from pilot districts agreed that their final summative performance classification accurately reflected their overall performance in 2012/13, 32 percent indicated that it did not, and 30 percent were undecided (see table B10 in appendix B).

How did teachers and principals perceive changes in participating teachers' instructional practice and in knowledge sharing and collaboration among teachers and administrators following initial implementation of the new evaluation systems?

In focus groups teachers from all five pilot districts felt that they were more reflective about their instructional practice and professional development; principals from pilot and partner districts reported corresponding perceptions of their teachers. Teachers from all five pilot districts reported in focus groups that they had reflected more about their instructional practice during the initial pilot year and were seeking further professional development opportunities to address areas of weakness. Teachers from one pilot district noted that reflecting on their practice—assessing their pedagogy against the Danielson Framework for Teaching expectations—is becoming the norm in their schools. This finding was supported by principals in focus groups and by teacher survey respondents from pilot districts. Principals from three pilot districts said their teachers seemed more aware of their own practice since the pilot model was implemented. For example, teachers were asking about ways to improve their practice and trying different instructional strategies without worrying about being unsuccessful, as well as finding resources to support their work. Some 54 percent of teacher survey respondents from pilot districts agreed that their post-observation conferences with their principal helped identify needs for professional growth.

Teachers from pilot districts reported in focus groups that they had reflected more about their instructional practice during the initial pilot year

In focus groups principals from partner districts reported perceptions that paralleled those of their pilot district counterparts. Principals from four partner districts perceived an increase in teacher self-awareness and reflection since implementation of the new evaluation models had begun. For example, in 2012/13 their teachers seemed more able to deeply explore ways to motivate and engage their students and seemed more adept at identifying their own strengths and areas for improvement and working with their colleagues to improve their teaching practice.

Principals from 6 of the 10 study districts reported in focus groups that their instructional leadership abilities had improved. Principals from four pilot districts noted that their instructional leadership abilities had improved since implementing the pilot system, as exemplified by having a better definition of an effective teacher and being able to identify specific areas of practice (Danielson Framework for Teaching components) that teachers should address in professional development. Principals from two partner districts stated that their instructional leadership abilities improved during 2012/13, in part because the district's new evaluation rubric allowed them to more specifically identify and target the professional development needs of their teachers for the coming school year.

Some perceived changes in participating teachers' instructional practices in 2012/13 might not have been shaped by the new evaluation practices. Principals from all five pilot districts and from three partner districts reported observing changes in teachers' instructional practices during their classroom visits and lesson plan reviews in 2012/13. For example, principals noted in focus groups that teachers were using data more often in planning and instruction, using group work more effectively, improving their questioning techniques, and organizing the classroom to better support student learning. In addition, in focus groups teachers from all five partner districts pointed out new instructional practices they employed in 2012/13, including an increased use of technology, more deliberate lesson planning, and more focus on student participation and engagement. However,

teacher survey respondents from pilot districts indicated that these perceived changes might not have been due to new evaluation practices: 42 percent agreed that the new teacher evaluation process led them to improve their instructional practice.

Although in focus groups principals from all five pilot districts felt that their interactions with teachers were more collaborative in 2012/13, teachers tended not to share this perception. Principals from all five pilot districts reported in focus groups that their interactions with teachers were more collaborative and evidence-based than before 2012/13, noting, for example, that they spent more time listening to teachers explain their lessons and less time talking during their interactions with teachers. However, in focus groups teachers from the pilot districts reported little change in their interactions with their administrators, and only 27 percent of teacher survey respondents from pilot districts agreed that the quality of their instructional interactions with their administrator had improved as a result of the new teacher evaluation process.

Principals from pilot districts reported in focus groups that their interactions with teachers were more collaborative and evidence-based than before 2012/13, but teachers from pilot districts reported in focus groups little change in their interactions with their administrators

The sentiments of principals from partner districts were similar to those of principals from pilot districts. Principals from four partner districts reported that their conversations with teachers were more collaborative in 2012/13, as exemplified by teachers' increasing use of the new observation rubrics and resulting feedback to make reflective comments about their teaching and identify areas for improvement. Unlike teachers from pilot districts, teachers from three partner districts reported in focus groups that their conversations with administrators had become more focused on reflection and growth in 2012/13.

Teachers and principals thought teachers were working more collaboratively, but principals from two pilot districts attributed this to work on a project outside the new evaluation system. In focus groups, teachers and principals from four pilot districts and principals from three partner districts noted seeing more teacher collaboration in 2012/13—characterized, for example, by more teachers visiting each other's classrooms to observe and model instruction, engaging in data-centered peer conversations, and sharing instructional resources. However, principals from two pilot districts attributed these perceived changes to work related to the adoption of new Common Core State Standards and not necessarily to the new teacher evaluation models. Consistent with this sentiment, 24 percent of teacher survey respondents from pilot districts agreed that the quality of their instructional interactions with their colleagues had improved as a result of the new teacher evaluation process, and 11 percent agreed that the new teacher evaluation process had improved the climate and culture at their school in 2012/13.

Principals from all five pilot districts reported varying levels of acceptance of the new system among teachers, indicating that newer teachers were generally more receptive than were veteran teachers. This discrepancy was also evident on end-of-year teacher surveys: 44 percent of the teacher survey respondents from pilot districts with more than 10 years of experience disagreed that the new teacher evaluation system was an improvement over the prior system, whereas 29 percent of teachers with 5–10 years of experience and 18 percent with fewer than 5 years of experience disagreed.

Implications of the study findings

Teachers and principals from pilot districts indicated that more training would help promote greater understanding of the new evaluation process and maintain or improve

rater consistency in classroom observations. The state responded to this feedback in summer 2013 by training principals from pilot districts in time management and leading instructional conversations. All classroom observers must now pass an online proficiency exam prior to observing teachers in 2013/14.

Study respondents viewed the more traditional evaluation measures—classroom observations and conferences—more favorably than they did the newer, less familiar measures, such as the use of student academic progress data or stakeholder survey results. This preference suggests that more training or support is needed to help principals and teachers appreciate the value of including the newer measures in the evaluation process. The Arizona Department of Education reported revising many stakeholder survey questions that respondents found to be problematic and in the summer and fall of 2013 provided additional guidance to pilot districts on survey administration and use of results.[10] In response to concerns expressed in focus groups by teachers from all five pilot districts about differences in how student academic progress data were incorporated into teacher evaluations in 2012/13, the department plans to use a student learning objective process in its 2013/14 teacher evaluation model to make evaluation more consistent across teachers. The state intends to gather evidence about the use of student learning objectives, participants' perceptions of the process, and the comparability of student learning objectives across teachers.

The Arizona Department of Education plans to track some of the findings of this study in future years

The Arizona Department of Education plans to track some of the findings of this study in future years. For example, respondents from pilot and partner districts noted a higher incidence of teachers assessing their own instructional practice. Thus, the department plans to further monitor this practice in 2013/14 and 2014/15. Additionally, in response to reports that some principals used the evaluation results to target professional development opportunities for their teachers, the department hopes to use data from future implementation to further explore how teachers' evaluation outcomes are being used to inform the professional learning opportunities they are offered. Moreover, in response to teachers' concerns about the potential detrimental impact of schoolwide student assessment data on their evaluation ratings, the department plans to examine the relationship between schoolwide student achievement and teachers' summative performance scores.

Limitations of the study

Although this study identified capacity challenges—related to time constraints, training, and technology—it did not attempt to characterize the fidelity or quality of the implementation of the new teacher evaluation systems under study. Future research, perhaps involving direct observation of practice, might assess the fidelity of system implementation in relation to the requirements of the Arizona Framework for Measuring Educator Effectiveness or the Arizona Department of Education's model, as well as the factors that influence implementation fidelity. Such research could identify areas for additional training that might lead to more efficient implementation of teacher evaluation systems.

This study was limited by the data collection processes, the selection of districts and focus group participants, and survey response rates. Its findings are based on secondary analyses of focus group and survey data collected by the Arizona Department of Education. The focus group transcripts did not distinguish among speakers, so the study team could not ascertain how many participants held the same opinions. The prevalence of focus group

sentiments is thus conveyed solely by the number of pilot or partner districts in which a particular sentiment was expressed.

Since districts and teachers within those districts volunteered to participate, rather than being randomly selected, the results cannot be generalized beyond the participating districts. Also, teacher response to the end-of-year survey was uneven across districts. The response rate for the sample of five pilot districts was 61 percent, but none of the teachers from one (smaller) pilot district responded to the survey, and the largest pilot district accounted for 53 percent of responses. (Bias related to this differential response rate was found for one survey item.) Results might have differed had teachers from other pilot districts responded. (The study data were insufficient to carry out a nonresponse analysis comparing survey respondents and nonrespondents.) Also, the inability to analyze the results for teacher survey respondents from partner districts (because of the low response rate) precluded the opportunity to corroborate information from the focus groups with participants from partner districts.

Appendix A. Research questions, data collection, and analysis

This descriptive study examined the initial implementation of new teacher evaluation models aligned with the Arizona Framework for Measuring Educator Effectiveness during the 2012/13 school year in 10 Arizona volunteer pilot or partner local education agencies (districts and charter schools). It addressed three research questions:

- What challenges or unintended consequences did teachers and principals perceive in the initial implementation of the evaluation systems?
  - What factors were reported to influence implementation?
  - How did the identified challenges and unintended consequences related to implementation vary across pilot and partner districts?
- How did teachers and principals perceive the effectiveness of the piloted evaluation measures?
  - Were particular measures viewed as clearer or more credible or useful than others?
  - Was feedback seen as helpful for targeting subsequent professional development?
- How did teachers and principals perceive changes in participating teachers' instructional practice and in knowledge sharing and collaboration among teachers and administrators following initial implementation of the new evaluation system in the following areas:
  - Did teachers increasingly reflect about their practice?
  - Were there more instruction-focused interactions between teachers and administrators?
  - Were there changes in instructional practice?
  - Was there more teacher collaboration and knowledge sharing?

To address each of these questions, the study team analyzed focus group transcripts and survey data provided by the Arizona Department of Education.

Focus group transcripts

Focus groups were conducted by West Comprehensive Center researchers enlisted by the Arizona Department of Education according to state guidelines. While acknowledging that the focus groups would need to be populated by volunteers, the department suggested ways to achieve a balanced set of teacher perspectives by recommending that each focus group include:

- Teachers with classroom student achievement data and teachers without such data.
- Two to three teachers with less than 5 years of experience in the district, two to three with 5–10 years, and two to three with more than 10 years.
- Two to three elementary school teachers, two to three middle school teachers, and two to three high school teachers.
- At least one teacher of English language arts, math, science/social studies, and arts/physical education/languages.
- Teachers from both small and large schools.
- Teachers from both low- and high-performing schools.

In the pilot districts 46 teachers participated in 7 focus groups, and 13 principals participated in 10 interviews or focus groups. In the partner districts 46 teachers participated in 7 focus groups, and 41 principals participated in 11 focus groups.

Two Regional Educational Laboratory West reviewers each coded two focus group transcripts, one for discussions with teachers and one with principals. The principal investigator developed a spreadsheet to compare the results and calculate the reliability level for each transcript. The inter-rater reliability was 0.4 for the initial coding of the transcripts. After discussing the results, reviewers coded the same transcripts a second time. Inter-rater reliability for the teacher transcript improved to 0.8, and the inter-rater reliability for the principal transcript improved to 0.7. Following additional discussion of discrepancies in the principal transcript coding and a third coding, inter-rater reliability improved to 0.9. After the reviewers completed coding approximately two-thirds of the transcripts, they followed a similar process to re-establish inter-rater reliability. When both reviewers coded a teacher and a principal transcript, inter-rater reliability was found to have slipped to 0.3 for the teacher transcript and 0.5 for the principal transcript. Following discussions about the discrepancies, the reviewers coded each transcript a second time. Inter-rater reliability improved to 0.9 for the teacher transcript and 0.7 for the principal transcript. (Discussion between reviewers and the principal investigator revealed that this was a technical error and not a conceptual error.) The one reviewer coded the teacher transcript for the third time, which improved the inter-rater reliability to 0.8.

After coding the transcripts, each reviewer focused on a different research question to organize and analyze the transcript data. The reviewers organized the data by codes, which enabled them to identify patterns and themes as well as to track common themes across districts. More specifically, reviewers considered responses relevant to the research questions and then assessed the prevalence of responses across districts—a theme raised by teachers or principals was included as a finding if the theme was raised in more than one pilot or partner district. (Prevalence within districts could not be determined since there is no way to know from the transcript data how many individual participants held the same opinions.)

Teacher survey

The Arizona Department of Education developed an online end-of-year survey to gather data on teachers' perceptions and experiences with the new evaluation systems. Links to the survey were provided to pilot and partner district leaders in early May 2013 for distribution to participating teachers. A total of 787 teachers from 7 of the 10 school districts in this study responded to the survey—165 from pilot districts (61 percent response rate) and 622 from partner districts (23 percent response rate). Because of the low response rate for partner districts (no responses were received from teachers from three of the five partner districts), no results for the teacher survey were reported for partner districts.

Additional exploratory analysis revealed that teachers from the largest pilot district accounted for 53 percent of respondents from pilot districts. Comparing this district's responses with those of other pilot districts yielded only one significant difference: Teachers from the largest pilot district were less likely than teachers from other pilot districts to report that standardized schoolwide test scores could provide an accurate assessment of their performance (54 percent from the largest pilot district compared with 66 percent

elsewhere). Exploratory analyses were conducted to identify the level of responses within each group of survey items, determining that the low level of missing responses presented no challenges for the analysis.

Appendix B. Detailed responses to the teacher survey from pilot districts

The Arizona Department of Education developed an online end-of-year survey for participating teachers from pilot and partner districts to gather data on their perceptions and experiences with the new evaluation systems. The response rate for teachers from pilot districts was 61 percent, and these results are shown here (tables B1–B11). Because the response rate for teachers from partner districts was only 23 percent, their responses were not included in the results.

Table B1. What is the primary subject area that you teach?

Respondents	Elementary, all	Reading/ English language arts	Math	Science	Social studies	Special education	Other
Number	46	29	16	7	14	18	27
Percent	29.3	18.5	10.2	4.5	8.9	11.5	17.2

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Table B2. What is the primary grade level that you teach?

Respondents	Elementary	Middle	High school
Number	61	56	40
Percent	38.9	35.7	25.5

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Table B3. Counting this school year, how many years have you been teaching at your current school?

Respondents	< 1 year	1–3 years	3 5 years	5–7 years	7–9 years	9 10 years	10+ years
Number	21	57	23	19	14	[a]	21
Percent	13.4	36.3	14.7	12.1	8.9	[a]	13.4

a. Not reported because fewer than five responses.

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Table B4. Counting this school year, how many years have you been teaching overall?

Respondents	< 1 year	1–3 years	3 5 years	5–7 years	7–9 years	9 10 years	10+ years
Number	10	24	17	12	15	11	68
Percent	6.4	15.3	10.8	7.6	9.6	7.0	43.3

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Table B5. How many times were you observed in the classroom by your evaluator during this school year?

Observation type	1 time		2–3 times		4–5 times		6 or more times		Total	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Formal	24	15.3	130	82.8	[a]	[a]	[a]	[a]	157	100
Informal	47	29.9	56	35.7	18	11.5	36	22.9	157	100

a. Not reported because fewer than five responses.

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Table B6. To what extent do you agree that this number of formal classroom observations was adequate to assess your performance?

Observation type	Disagree		Neither disagree nor agree		Agree		Total	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Formal	41	26.1	36	22.9	80	51.0	157	100

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Table B7. How many times did you participate in a pre-/post-observation conference with your evaluator?

Conference type	0 times		1 time		2–3 times		4–5 times		6 or more times		Total	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Pre-observation	15	9.6	28	18.0	111	71.2	[a]	[a]	[a]	[a]	156	100
Post-observation	11	7.1	41	26.3	98	62.8	6	3.9	0	0	156	100

a. Not reported because fewer than five responses.

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Table B8. What was the duration/average length of your pre-/post-observation conference(s)?

Conference type	Less than 15 minutes		15–30 minutes		31–45 minutes		46–60 minutes		More than 1 hour		Total	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Pre-observation	32	22.7	63	44.7	35	24.8	10	7.1	[a]	[a]	141	100
Post-observation	29	19.9	66	45.2	37	25.3	14	9.6	0	0	146	100

a. Not reported because fewer than five responses.

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Table B9. To what extent do you agree that the pre-observation conference(s) fully prepared you for what to expect and the post-observation conference(s) provided you with meaningful feedback on how to improve your instruction?

	Disagree		Neither disagree nor agree		Agree		Total	
Value	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Pre-observation: fully prepared	27	19.2	38	27.0	76	53.9	141	100
Post-observation: meaningful feedback	34	23.3	25	17.1	87	59.6	146	100

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Table B10. To what extent do you agree with the following statements based on your participation in the new process this year?

	Disagree		Neither disagree nor agree		Agree	
Process/outcome	Number	Percent	Number	Percent	Number	Percent
The new teacher evaluation process represents an improvement over prior teacher evaluations at my school	46	30.3	68	44.7	38	25.0
The new teacher evaluation process is fair	47	31.3	52	34.7	51	34.0
The training I received on the new teacher evaluation process was adequate for me to effectively participate in the process this year	48	31.8	44	29.1	59	39.1
The criteria on which I was evaluated were made clear to me	33	21.7	64	42.1	55	36.2
The new teacher evaluation process has provided a common language for professional practice in my school	34	22.7	52	34.7	64	42.7
My post-observation conference(s) helped identify needs for my professional growth	31	22.0	34	24.1	76	53.9
The new teacher evaluation process helped me engage in professional growth opportunities targeted to my needs	50	32.9	58	38.2	44	29.0
The new teacher evaluation process led me to improve my instructional practice	35	23.0	54	35.5	63	41.5
The quality of my instructional interactions with my administrator improved as a result of the new teacher evaluation process	40	27.0	68	46.0	40	27.0
The quality of my instructional interactions with my colleagues improved as a result of the new teacher evaluation process	46	30.3	70	46.1	36	23.7
I have confidence in my evaluator's ability to accurately rate my instructional practice	23	15.1	32	21.1	97	63.8
The new teacher evaluation process has improved the climate and culture in my school	61	40.1	74	48.7	17	11.2
My participation in the new teacher evaluation process has benefited my students	41	27.0	68	44.7	43	28.3
The new teacher evaluation process will lead to continuous school improvement	31	20.7	58	38.7	61	40.7
The final summative performance classification I received from the new teacher evaluation process accurately reflected my overall performance this year	48	31.6	45	29.6	59	38.8

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Table B11. To what extent do you feel that the following activities/measures can provide an accurate assessment of your performance as a teacher?

Measure	No accuracy		Low accuracy		Moderate accuracy		High accuracy	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Formal classroom observation(s) by my principal	10	6.6	22	14.6	68	45.0	51	33.8
Informal classroom observation(s) by my principal	15	9.9	20	13.3	64	42.4	52	34.4
Formal classroom observation(s) by someone other than my principal—for example, an instructional coach, mentor, or trained outside observer (based off-site)	27	18.0	22	14.7	71	47.3	30	20.0
Informal classroom observation(s) by someone other than my principal—for example, an instructional coach, mentor, or trained outside observer (based off-site)	28	18.5	24	15.9	65	43.1	34	22.5
Student learning objectives for the year, established through consultation with my principal	29	19.2	26	17.2	71	47.0	25	16.6
Standardized test scores from my classroom(s) of students this year	27	17.9	38	25.2	75	49.7	11	7.3
Standardized schoolwide test scores	23	15.2	37	24.5	83	55.0	8	5.3
Student surveys	30	19.9	45	29.8	65	43.1	11	7.3
Parent surveys	30	19.9	52	34.4	65	43.1	[a]	[a]
Peer teacher surveys	19	12.6	27	17.9	85	56.3	20	13.3

a. Not reported because fewer than five responses.

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Appendix C. Significant differences by subgroups in teacher survey results from pilot districts

To further understand the teacher survey results from pilot districts, the demographic information provided on the survey's background items was used to separately analyze the 15 survey items exploring the evaluation process and perceived outcomes, as well as the 10 survey items on perceptions of different teacher effectiveness measures.

For one analysis, respondents were grouped by experience: relatively new teachers (up to 5 years reported experience), teachers with moderate experience (5–10 years), and highly experienced teachers (10 or more years). This analysis produced significant differences for three items (tables C1–C3). Generally, veteran teachers put less stock than new teachers in the new evaluation measures. Another analysis explored differences based on teachers' reported primary grade span taught (elementary, middle, or high school). This analysis showed no significant differences. Finally, to test for the differential impact of responses from the largest pilot district, which accounted for 53 percent of all responses from teachers from pilot districts, an analysis compared responses for teachers from the largest pilot district with responses for teachers from the other pilot districts. This analysis indicated one significant difference: Teachers from the largest pilot district felt that schoolwide student assessment data were less reliable as an indicator of teacher performance than did teachers from other districts (table C4).

Table C1. To what extent do you agree that the new teacher evaluation process represents an improvement over prior teacher evaluations at your school, by years of teaching experience?

	Disagree		Neither disagree nor agree		Agree	
Years of teaching experience	Number	Percent	Number	Percent	Number	Percent
Up to 5 years	13	25.5	32	62.8	6	11.8
5 up to 10 years	7	18.4	18	47.4	13	34.2
More than 10 years	26	41.3	18	28.6	19	30.2

$n = 152$, $chi^2 = 17.6362$, degrees of freedom = 4, p value = .001.

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Table C2. To what extent do you agree that the new teacher evaluation process led you to improve your instruction, by years of teaching experience?

	Disagree		Neither disagree nor agree		Agree	
Years of teaching experience	Number	Percent	Number	Percent	Number	Percent
Up to 5 years	12	23.5	20	39.2	19	37.3
5 up to 10 years	8	21.1	6	15.8	24	63.2
More than 10 years	15	23.8	28	44.4	20	31.8

$n = 152$, $chi^2 = 11.8274$, degrees of freedom = 4, p value < .02.

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Table C3. To what extent do you feel that student learning objectives, established in consultation with your principal, can provide an accurate assessment of your teaching performance, by years of teaching experience?

Years of teaching experience	No accuracy		Low accuracy		Moderate accuracy		High accuracy	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Up to 5 years	12	23.5	8	15.7	25	49.0	6	11.8
5 up to 10 years	6	15.8	[a]	[a]	15	39.5	13	34.2
More than 10 years	11	17.7	14	22.6	31	50.0	6	9.7

n = 151, chi^2 = 13.1432, degrees of freedom = 6, p value < .05.

a. Not reported because fewer than five responses.

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Table C4. To what extent do you feel that standardized schoolwide tests can provide an accurate assessment your teaching performance, largest pilot district versus other pilot districts?

Districts	No accuracy		Low accuracy		Moderate accuracy		High accuracy	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Largest pilot district	11	13.8	25	31.2	43	53.8	[a]	[a]
All others	12	16.9	12	16.7	40	56.3	7	9.9

n = 151, chi^2 = 8.714, degrees of freedom = 3, p value< .04.

a. Not reported because fewer than five responses.

Source: Authors' analysis of teacher survey data provided by Arizona Department of Education (2013).

Notes

1. State officials indicated that Arizona did not have statewide teacher evaluation requirements prior to the 2010 passage of Senate Bill 1040; evaluation was left to district discretion, with teacher workforce laws focused on employment retention priorities (tenure) and removal procedures.
2. The state's teacher evaluation process for pilot districts in 2013/14 is online at http://www.azed.gov/teacherprincipal-evaluation/files/2012/10/teacher-evaluation-v4.0-website-update-11_22_13-sl.pdf. (Only the evaluation of teachers is studied in this report.)
3. The four pilot school districts implemented the model in 11 public schools. To be included in the study, the volunteering pilot districts were required to meet a set of criteria. In addition to implementing the evaluation components specified in the Arizona Department of Education's model, pilot districts must provide release time for educators to be trained, designate a person familiar with working with data to collect and submit data related to the pilot, and participate in ongoing communication with the Arizona Department of Education (providing feedback on the quality of the department's model, identifying challenges and unintended consequences during implementation, and suggesting refinements).
4. These five partner evaluation systems, while all aligned with the state board–approved Arizona Framework for Measuring Teacher Effectiveness, present a diverse array of features that cannot be explored deeply in a single descriptive study. Seventy schools participated in the partner districts.
5. Presentations by participating district leaders indicated that prior to 2012/13, the pilot and partner districts generally resembled one another in that they all evaluated teachers using classroom observations and did not tend to incorporate information from stakeholder surveys or student achievement measures into teachers' professional performance ratings.
6. Teachers and principals from the pilot district that adopted the Danielson Framework for Teaching prior to 2012/13 noted in focus groups that although their familiarity with the rubric eased the implementation of the Arizona Department of Education's model, their instructional conversations became more in-depth and focused than they had been before the pilot.
7. The Arizona Department of Education is incorporating a student learning objective process—essentially classroom target setting, assessment, and monitoring by teachers, overseen by their evaluator—into its model for all participating teachers in 2013/14. Teacher survey respondents from pilot districts were supportive of student learning objectives as a performance measure: 64 percent reported that they felt that student learning objectives could provide either a moderately or highly accurate assessment of their teaching performance.
8. Perceptions of the accuracy of schoolwide test scores as a teacher performance measure varied slightly within the sample of teachers from pilot districts—54 percent of teacher survey respondents from the largest pilot district viewed schoolwide test scores as either a moderately or highly accurate assessment, while 66 percent of respondents from the other pilot districts reported feeling this way (see table C4 in appendix C).
9. To increase the accessibility of parent surveys and promote a higher response rate, the Arizona Department of Education allowed parent surveys to be administered at school events in 2013/14.
10. The most recent guidance documents from the Arizona Department of Education are available at http://www.azed.gov/teacherprincipal-evaluation.

References

Arizona Department of Education. (2013). *Results of online end-of-year survey for participating pilot and partner district teachers.* Phoenix, AZ.

Danielson Group. (2011). *The Framework for Teaching.* Retrieved April 22, 2013, from http://www.danielsongroup.org/article.aspx?page=frameworkforteaching

Flaherty, J., Tejwani, J., & Rodriguez, F. (2012). *Evaluation of the Seattle Public Schools (SPS) Teacher Incentive Fund (TIF) program: Year one report.* Los Alamitos, CA: WestEd.

Fowler, F. (2004). Looking at policies: Policy instruments and cost effectiveness. In *Policy studies for educational leaders: An introduction* (2nd ed.). Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

Goe, L., Holdheide, L., & Miller, T. (2011). *A practical guide to designing comprehensive teacher evaluation systems.* Washington, DC: National Comprehensive Center for Teacher Quality. http://eric.ed.gov/?id=ED520828

Honig, M. I. (2006). *New directions in education policy implementation: Confronting complexity.* Albany, NY: The State University of New York Press.

Jerald, C. D. (2012). *Movin' it and improvin' it! Using both education strategies to increase teaching effectiveness.* Washington, DC: Center for American Progress. Retrieved August 7, 2012, from http://www.americanprogress.org/issues/education/report/2012/01/23/10982/movin-it-and-improvin-it/

McLaughlin, M. W. (1990). Embracing contraries: Implementing and sustaining teacher evaluation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation.* Newbury Park, CA: Sage.

McNeil, M. (2013). Race to the Top winners make progress, face challenges, Ed. Dept. reports. *Education Week*, *32*(20). Retrieved August 7, 2012, from http://www.edweek.org/ew/articles/2013/02/01/20rtt.h32.html

Mead, S., Rotherham, A. J. & Brown, R. (2012). *The hangover: Thinking about the unintended consequences of the nation's teacher evaluation binge.* Teacher Quality 2.0 Special Report 2. Washington, DC: American Enterprise Institute.

National Council on Teacher Quality. (2012). *State of the states 2012: Teacher effectiveness policies.* Washington, DC: National Council on Teacher Quality. http://eric.ed.gov/?id=ED536371

Overview Information: Race to the Top Fund; Notice inviting applications for new awards for fiscal year 2010. *Federal Register, 75*(68), 18171.

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

implementation. Chicago, IL: Consortium on Chicago School Research. http://eric.ed.gov/?id=ED527619

Tennessee Department of Education. (2012). *Teacher evaluation in Tennessee: A report on year 1 implementation*. Nashville, TN: Author. http://eric.ed.gov/?id=ED533726

U.S. Department of Education. (2012). Laws & Guidance, Elementary & Secondary Education, ESEA Flexibility. Retrieved February 2012 from http://www.ed.gov/esea/flexibility

The Regional Educational Laboratory Program produces 7 types of reports

Making Connections
Studies of correlational relationships

Making an Impact
Studies of cause and effect

What's Happening
Descriptions of policies, programs, implementation status, or data trends

What's Known
Summaries of previous research

Stated Briefly
Summaries of research findings for specific audiences

Applied Research Methods
Research methods for educational settings

Tools
Help for planning, gathering, analyzing, or reporting data or research