Making Connections

October 2014

# Properties of the multiple measures in Arizona's teacher evaluation model

**Valeriy Lazarev**
**Denis Newman**
**Alyssa Sharp**
Empirical Education Inc.

## Key findings

This study of Arizona's pilot teacher evaluation model explored the relationships between component measures (teacher observations, student academic progress, and stakeholder surveys) and investigated how well the model differentiated between high- and low-performing teachers.

- Most teachers were rated "proficient" on most observation items.
- Observation scores correlated with student academic progress primarily in domains outside the classroom (Planning and Preparation and Professional Responsibilities).
- The strength of correlation between observation items and student academic progress differed for higher and lower scoring teachers.
- Student academic progress correlated with observation or survey results only among teachers whose results derived from statewide math and reading tests their students took for other classes.

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 2015–050

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.

October 2014

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-12-C-0002 by Regional Educational Laboratory 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.



Lazarev, V., Newman, D., & Sharp, A. (2014). *Properties of the multiple measures in Arizona's teacher evaluation model* (REL 2015–050). 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

The Arizona Department of Education piloted a multiple-measure teacher evaluation model in five school districts in 2012/13. Using results from teacher observations, measures of student academic progress, and surveys of students, parents, and peers (including a self-assessment), the model calculated a composite score for each teacher. This study examines the model's effectiveness in differentiating between higher and lower performing teachers. The study also explores the relationships among the model's components, focusing on the properties of the teacher observation results.

The observation instrument and rubric used in the model divide teaching into four domains: two evaluated through classroom observation (Classroom Environment and Instruction) and two assessed through information gathered outside the classroom (Planning and Preparation and Professional Responsibilities). Each domain includes five or six observation items. The observation scores were typically assigned by the teacher's principal.

Key findings from the 2012/13 pilot year were:

- Teachers were scored "proficient," the second highest score on a four-point scale (unsatisfactory, basic, proficient, and distinguished), on 62 percent of observation items.
- Positive correlations were evident between all observation items, but only a few significant correlations were found between observation items and student academic progress.
- Observation domain scores correlated with student academic progress only in domains observed outside the classroom (Planning and Preparation and Professional Responsibilities).
- Observation results captured several aspects of teaching performance (rather than a single "teacher effectiveness" trait).
- The strength of correlation between observation items and student academic progress differed for higher and lower scoring teachers.
- Student academic progress correlated with observation or survey results only among teachers whose results derived from statewide math and reading tests their students took for other classes.

The study findings suggest several considerations:

- If observers were to receive more thorough training than the optional pilot training without assessment, the teacher observation results might more accurately differentiate between higher and lower performing teachers, particularly in the Instruction domain.
- A single aggregated observation score (the sum of the observation item scores across all four domains) might not adequately measure independent aspects of teacher performance.
- The model's complex structure could benefit from some simplification. Specifically, it may be useful to eliminate or substantially revise the peer survey, which showed no correlation with observation results or student academic progress and had low weight in the aggregated survey component score.
- It might be useful to develop different scoring schemas that appropriately account for teachers' specific teaching environment (for example, whether classroom-level standardized test scores are available for their content area).

# Contents

## Tables

## Why this study?

States and school districts across the country are overhauling teacher evaluation models. Recent Race to the Top federal grant applications required states to design comprehensive evaluation systems with multiple measures of teacher performance (U.S. Department of Education, 2010). And in recent applications for Elementary and Secondary Education Act waivers, states had to describe their plans to reform teacher evaluation and support systems to focus on instruction quality and student results (U.S. Department of Education, 2012). In turn, nearly two-thirds of U.S. states have made changes to their teacher evaluation policies since 2009 (Jerald, 2012).

A growing number of studies have analyzed the new teacher evaluation systems. For example, the Measures of Effective Teaching project, funded by the Bill & Melinda Gates Foundation, has yielded empirical evidence of correlations between various teacher effectiveness metrics, including scores from several widely used classroom observation instruments, student surveys, and estimates of teachers' value-added contributions to student test achievement (Kane & Staiger, 2012). One of the project's culminating reports also examined different approaches to combining these metrics into a composite score of teacher effectiveness (Mihaly, McCaffrey, Staiger, & Lockwood, 2013).

*This study provides information about the variability of the three component scores and the teacher effectiveness composite score used in the teacher evaluation model piloted by the Arizona Department of Education in 2012/13*

But states and districts still lack concrete findings about how best to interpret, combine, and use these metrics in practical decisionmaking (Rothstein & Mathis, 2013). As states begin to implement new teacher evaluation systems, they need empirical support and practical feedback on procedures and policies.

This study represents an effort to begin this type of applied research by examining the teacher evaluation model piloted by the Arizona Department of Education in 2012/13. The study yields information about the variability of the model's three component scores and the teacher effectiveness composite score. It also yields information about consistency across the component scores and the relationship between observation items and stakeholder survey scores and student academic progress. These findings may help the department refine its model by modifying scoring guidelines, rescaling items, changing component weightings, or excluding redundant items.

This study was carried out in partnership between the Regional Educational Laboratory (REL) West and the Arizona Department of Education, within the context of the REL West's regional Educator Effectiveness Alliance. The intended core audience for this study includes the Arizona Department of Education and the Arizona State Board of Education and state legislature. The study findings and methodology may also interest Arizona local education agencies and other state education agencies that are developing or implementing new multiple-measure teacher evaluation systems.

## What the study examined

This exploratory study analyzed the statistical properties of the components of the teacher evaluation model piloted in five Arizona school districts in 2012/13 (Arizona Department of Education, 2012; see appendix A for details of the model). The study explored the extent to which these components distinguished between higher and lower performing teachers and yielded internally consistent results (that is, ratings that correlated positively with one

another) in the pilot year. The study findings may help the Arizona Department of Education consider adjustments to the model and its scoring schemas.

While this study analyzed all the components of Arizona's model (teacher observation, student academic progress, and surveys), analysis focused on the observation instrument adopted by the Arizona Department of Education—the Danielson Group's Framework for Teaching[1]—for three reasons (Danielson Group, 2011). First, Arizona's educator evaluation regulations emphasize measuring teaching performance,[2] and teacher observation measures involve interactions between the observer (usually the school principal) and teacher. Second, observation instruments can be used to collect performance data repeatedly throughout a school year, unlike metrics that rely on data collected once (typically at the end of the school year). Third, the Framework for Teaching observation instrument has been researched in various contexts (including the Measures for Effective Teaching project studies), so the results of its use in the Arizona pilot can be compared with those from existing analyses.

*The larger goal of the study is to inform the construction of an internally consistent teacher evaluation model that distinguishes effectively between more and less successful teachers*

### Research questions

The study addresses several questions about the Arizona teacher evaluation model, divided into three research areas:

- Statistical properties of the teacher observation instrument:
  - How are observation scores (domain scores and observation item scores within each domain) distributed? Do some scores tend to be more or less dispersed than others?
  - What are the correlations among scores within the teacher observation instrument?
  - Can observation results be represented effectively by a single aggregated observation score?
- Relationships between observation scores (domain scores and observation item scores within each domain) and student academic progress:
  - What are the correlations between observation scores and the student academic progress component?
  - For each observation item, are there significant differences in average student academic progress between teachers scoring at each level?
- Statistical relationships among the teacher observation, student academic progress, and survey components:
  - What are the correlations between these components?
  - Are there significant differences in the average observation and survey component scores of teachers with high student academic progress and teachers with low student academic progress?
  - What are the differences in student academic progress scores between teachers in the higher and lower quartiles of the distributions of observation and survey scores?

The larger goal of the study is to inform the construction of an internally consistent teacher evaluation model that distinguishes effectively between more and less successful teachers. The analysis of the distributions of observation items and their correlations will help identify observation items that have limited or no potential to differentiate between higher and lower performing teachers. And this analysis will identify observation items

that do not add useful information because of their very strong correlation with other observation items.[3] The analysis will also help establish the extent to which the model was internally consistent during its pilot-year implementation.

### Data and analysis

Data were collected from five local education agencies—four public school districts and one charter agency—in 2012/13 (see table B1 in appendix B for local education agency demographics). The data included information from all three components of the evaluation model: item-level results of teacher observations from two observation cycles; student academic progress calculations; and summative results from student, parent, and peer surveys (box 1). All teacher evaluation data were aggregated to the teacher level. The final dataset included information on 297 teachers in 12 schools.

*The analysis will help establish the extent to which the model was internally consistent during its pilot-year implementation*

The study relied primarily on descriptive statistics of component teacher evaluation metrics and analysis of correlations among these components. A variety of tests of statistical significance were used. One research question was addressed through principal component analysis, a more advanced technique described in detail in appendix C.[4] In addition to component-level analysis, the study team undertook an in-depth item-level analysis of the observation scores, benchmarking the results against the Measures of Effective Teaching database of Framework for Teaching observation scores, the largest existing set of teacher evaluation data collected in multiple districts using the same instruments.

Analyses involving the student academic progress component were performed separately for two groups of teachers, defined by whether classroom-level statewide test data were available for the teacher's content area (see box 1).

## Box 1. Components of the 2012/13 Arizona Department of Education teacher evaluation model

The three components of the 2012/13 Arizona teacher evaluation model are described below.

#### Teacher observation

The teacher observation component uses the Danielson Group's Framework for Teaching, which divides teaching into four domains: Planning and Preparation, Classroom Environment, Instruction, and Professional Responsibilities (Danielson Group, 2011). Each domain includes five or six observation items scored by observers on a four-point scale (0 = unsatisfactory, 1 = basic, 2 = proficient, 3 = distinguished), for a total of 22 observation item ratings (see table). Two domains were evaluated through in-person classroom observations (Classroom Environment and Instruction), and two using information gathered outside the classroom (Planning and Preparation and Professional Responsibilities). Observers conducted two classroom observations for each teacher and then assigned a single final score for each observation item. Observers could assign half-points (such as 1.5) for the final observation item scores. Domain scores were calculated as the sum of the observation item scores, and the total teacher observation score was calculated as the sum of the four domain scores.

*(continued)*

## Box 1. Components of the 2012/13 Arizona Department of Education teacher evaluation model *(continued)*

*Danielson's Framework for Teaching*

| Domain 1: Planning and Preparation | Domain 2: Classroom Environment |
|---|---|
| 1a Demonstrating Knowledge of Content and Pedagogy | 2a Creating an Environment of Respect and Rapport |
| 1b Demonstrating Knowledge of Students | 2b Establishing a Culture for Learning |
| 1c Setting Instructional Outcomes | 2c Managing Classroom Procedures |
| 1d Demonstrating Knowledge of Resources | 2d Managing Student Behavior |
| 1e Designing Coherent Instruction | 2e Organizing Physical Space |
| 1f Designing Student Assessments | |
| **Domain 3: Instruction** | **Domain 4: Professional Responsibilities** |
| 3a Communicating with Students | 4a Reflecting on Teaching |
| 3b Using Questioning and Discussion Techniques | 4b Maintaining Accurate Records |
| 3c Engaging Students in Learning | 4c Communicating with Families |
| 3d Using Assessment in Instruction | 4d Participating in a Professional Community |
| 3e Demonstrating Flexibility and Responsiveness | 4e Growing and Developing Professionally |
| | 4f Showing Professionalism |

**Source:** Danielson Group, 2011.

### Student academic progress

During the pilot year's first evaluation conference, each teacher was assigned a particular rating table to be used for the student academic progress component (see sample rating table in appendix A). Due to variability in the types of student test data available, study teachers were divided into two groups. Group A teachers had classroom-level student standardized test data that were appropriate for their content areas. According to the Arizona Framework for Measuring Educator Effectiveness, classroom-level results on Arizona's Instrument to Measure Standards (AIMS), the standardized state assessment, had to be used for group A teachers if available. As a result, group A comprised mainly math and reading (English language arts) teachers.

Group B teachers had no standardized test data for their content area, so student achievement was assessed using other criteria, which could include math and reading AIMS data for their students or aggregated results from Stanford 10, Advanced Placement, International Baccalaureate, Cambridge International, or ACT assessments, as well as student graduation rates. Most pilot teachers from high schools (65 percent) and middle schools (51 percent) were in group B. (See table B2 in appendix B for more detailed information on the composition of groups A and B.) In both groups, student achievement indicators were converted to the same 40 point summary student academic progress metric to make them comparable across all teaching environments.

### Surveys

The survey component included student, peer, and parent surveys, each with a different weight in the aggregated survey component score.

Separate student surveys were administered to students in grades 3–5 (34 items) and grades 6–12 (37 items), assessing (through closed-response Likert-style items) the extent to which the student agrees that the teacher Captivates Students, Cares about Students, Challenges Students, Clarifies Lessons, Confers with Students, Consolidates Knowledge, Controls Behavior, and Engages Students. The surveys were based on publicly available items from Cambridge Education's Tripod Student Perception Survey field tested by the Colorado

*(continued)*

**Box 1. Components of the 2012/13 Arizona Department of Education teacher evaluation model** *(continued)*

Department of Education in 2011/12. Surveys were administered online by the Arizona Department of Education. Students submitted their surveys anonymously. For this study the department provided only averaged teacher-level survey results. The student survey was weighted at 75 percent of the survey component score and constituted 15–20 percent of the teacher's composite evaluation score.

Peer surveys were administered for each participating pilot teacher. The 15-question survey asked teachers to rate their peer's performance on a four-point ordinal scale from strongly agree to strongly disagree. Three peer surveys were collected for each participating teacher; two peers were chosen by the principal and one by the teacher. A total peer survey score was calculated for each teacher and constituted a maximum of 3 percent of the evaluation.

A 16-question, school-level survey was also administered to parents of students in participating schools. These online surveys were voluntary and anonymous. Parents were asked to rate the quality of their child's school, teachers, and administration using an A–F rating scale. The parent survey results constituted a maximum of 4 percent of the teacher's evaluation.

## What the study found

***The distributions of the observation item and domain scores for all teachers in the pilot study were heavily concentrated around the median ratings and positively skewed toward the higher ratings***

This study of the 2012/13 Arizona pilot teacher evaluation model focused on the statistical properties of the model's components and the relationships between them, relying largely on descriptive statistics and correlation analysis.

### Most teachers were rated proficient on most observation items

The observation item and domain scores for all teachers participating in the pilot study exhibited similar characteristics: their distributions were heavily concentrated around the median ratings and positively skewed toward the higher ratings. All observation items had median and modal (most frequent) values of 2 (proficient) on the rubric's 0–3 scale, and more than half the scores on every item were 2 (table 1). Most teachers were rated proficient on most observation items. Few teachers received an unsatisfactory score (the lowest score) on observation items in the domains of Planning and Preparation and Classroom Environment. Among all the observation item scores awarded during the pilot year, only about 2 percent were unsatisfactory.

The degree of concentration of scores differed across observation items, with standard deviations ranging from .42 to .65 (see table 1). Several observation items involving formal aspects of teachers' classroom interactions with students—2c, Managing Classroom Procedures; 2d, Managing Student Behavior; 2e, Organizing Physical Space; and 3d, Using Assessment in Instruction—were the most concentrated, with standard deviations of 0.5 or less and with proportions of proficient scores that were as high as 78 percent (for 3d, Using Assessment in Instruction; see table 1). The observation domain scores (table 2) reflected the same positive skew and low variability.

These results suggest that the observation instrument was not used in a manner that effectively differentiated among levels of teacher practice. On average, 62 percent of the scores for observation items were proficient and only 2 percent were unsatisfactory. Assuming

**Table 1. Descriptive statistics of teachers' observation item scores in the Arizona pilot teacher evaluation model, 2012/13**

| Domain and observation item | Minimum | Mean | Standard deviation | Frequency of mode (percent of teachers receiving proficient score) |
|---|---|---|---|---|
| 1. Planning and Preparation | | | | |
| 1a: Demonstrating Knowledge of Content and Pedagogy | 1 | 2.4 | 0.53 | 57 |
| 1b: Demonstrating Knowledge of Students | 0 | 2.3 | 0.53 | 65 |
| 1c: Setting Instructional Outcomes | 1 | 2.2 | 0.55 | 63 |
| 1d: Demonstrating Knowledge of Resources | 1 | 2.3 | 0.51 | 65 |
| 1e: Designing Coherent Instruction | 1 | 2.3 | 0.53 | 62 |
| 1f: Designing Student Assessments | 0 | 2.1 | 0.47 | 74 |
| 2. Classroom Environment | | | | |
| 2a: Creating an Environment of Respect and Rapport | 1 | 2.3 | 0.54 | 60 |
| 2b: Establishing a Culture for Learning | 0 | 2.2 | 0.53 | 66 |
| 2c: Managing Classroom Procedures | 1 | 2.2 | 0.48 | 73 |
| 2d: Managing Student Behavior | 1 | 2.2 | 0.50 | 71 |
| 2e: Organizing Physical Space | 0 | 2.3 | 0.50 | 72 |
| 3. Instruction | | | | |
| 3a: Communicating with Students | 0 | 2.3 | 0.55 | 61 |
| 3b: Using Questioning and Discussion Techniques | 0 | 2.0 | 0.52 | 69 |
| 3c: Engaging Students in Learning | 0 | 2.1 | 0.56 | 68 |
| 3d: Using Assessment in Instruction | 0 | 2.0 | 0.42 | 83 |
| 3e: Demonstrating Flexibility and Responsiveness | 0 | 2.1 | 0.55 | 71 |
| 4. Professional Responsibilities | | | | |
| 4a: Reflecting on Teaching | 0 | 2.2 | 0.65 | 61 |
| 4b: Maintaining Accurate Records | 0 | 2.2 | 0.55 | 72 |
| 4c: Communicating with Families | 0 | 2.1 | 0.52 | 74 |
| 4d: Participating in a Professional Community | 0 | 2.2 | 0.59 | 63 |
| 4e: Growing and Developing Professionally | 0 | 2.1 | 0.51 | 72 |
| 4f: Showing Professionalism | 0 | 2.2 | 0.58 | 65 |

**Note:** Data are for 297 teachers. The model uses a 0–3 scale for observation scores: 0 = unsatisfactory, 1 = basic, 2 = proficient, and 3 = distinguished. For all observation items the maximum score was 3, and the mode was 2. Mean scores above the midpoint of the 0–3 scale demonstrate the observers' positive bias. A minimum score of 1 means that none of the teachers were rated unsatisfactory on the corresponding item.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

that teaching performance truly varied in the study sample, such heavy concentration of scores at one level does not supply decisionmakers with the information needed to distinguish between the highest and lowest performers for purposes of professional development or administrative decisions.

### *Comparison with scoring data from the Measures of Effective Teaching project*

Framework for Teaching scoring data from the Measures of Effective Teaching project offer some informative comparisons with the Arizona pilot teacher evaluation model.[5] First, the Arizona score distributions are more concentrated: the frequencies of the modal score ranged from 60 percent to 84 percent in the Arizona study data but from 50 percent to 72 percent in the Measures of Effective Teaching data (figure 1).

**Table 2. Descriptive statistics of teachers' observation domain scores in the Arizona pilot teacher evaluation model, 2012/13**

| Domain | Minimum | Maximum | Median | Mean | Standard deviation |
|---|---|---|---|---|---|
| 1. Planning and Preparation | 6 | 18 | 13 | 13.6 | 2.4 |
| 2. Classroom Environment | 4 | 15 | 11 | 11.1 | 2.0 |
| 3. Instruction | 2 | 15 | 10 | 10.4 | 2.2 |
| 4. Professional Responsibilities | 2 | 18 | 12 | 13.1 | 2.6 |

**Note:** Data are for 297 teachers. The model uses a 0–3 scale for observation scores: 0 = unsatisfactory, 1 = basic, 2 = proficient, and 3 = distinguished. Domain scores are the sum of all observation item scores within each domain. Domains 1 and 4 have six items each and therefore have score ranges of 0–18. Domains 2 and 3 have five items each and therefore have score ranges of 0–15. Median and mean scores above the midpoints of the corresponding scales demonstrate the observers' positive bias.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

***The heavy concentration of the observation scores at one level does not supply decisionmakers with the information needed to distinguish between the highest and lowest performers***

Second, the Arizona score distributions were more skewed. The mean scores were higher than in the Measures of Effective Teaching data, in which proficient was not always the modal category and where items in the Instruction domain tended to have a lower mode (basic) and score distributions differed most. Thus, in the Measures of Effective Teaching project the Framework for Teaching scores seemed to better differentiate between high and low classroom teaching performance.

There were some differences in how the observations were conducted in the two studies. The Measures of Effective Teaching observations were conducted by specially trained external observers who viewed videotaped lessons, while the Arizona administrators, according to Arizona Department of Education officials, received some training but were not required to pass the inter-rater reliability assessment in using the Framework for Teaching prior to observing teachers in person.

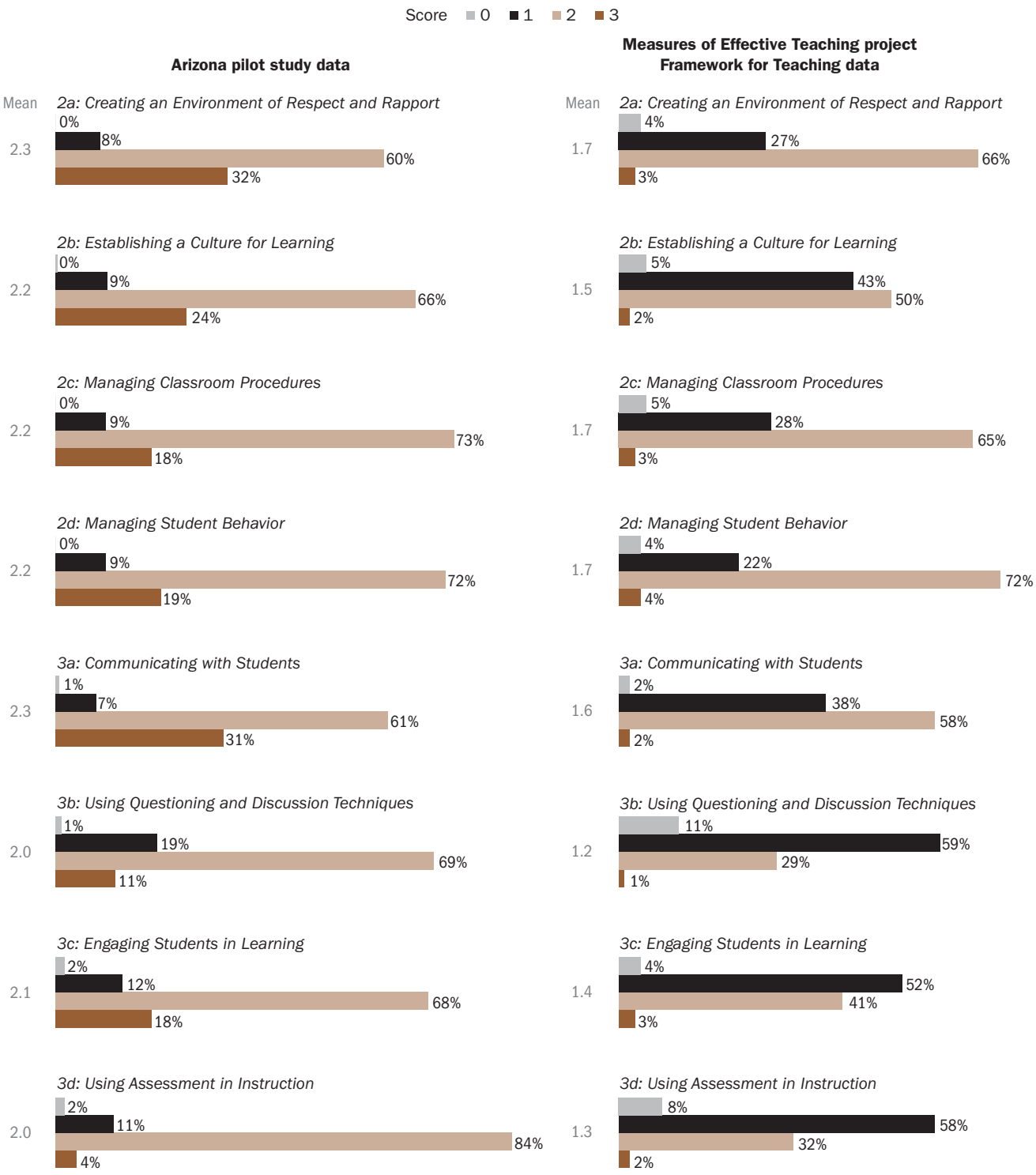
### *Observations by school principals tend to produce inflated scores*

There is evidence that observations by school principals tend to produce inflated scores—in particular, many teachers with students with low academic progress receive high observation scores. This concentration of scores at the top of the scale has occurred in other recent state implementations (see, for example, the Race to the Top Year 2 Reports for Maryland and Tennessee; U.S. Department of Education, 2013a,b). States are dealing with this problem by providing more training (U.S. Government Accountability Office, 2013). Thus, it is possible that the concentration of Arizona score distributions resulted from limited observer preparation and that the measurements could better differentiate teaching practices if observers received more training. It is also possible that the difference in score distributions is due to differences in incentive structures for principals and external observers. However, no published studies directly compare scores by different types of observers.

#### Positive correlations were evident between all observation items

Positive and statistically significant correlations were evident across all domains (table 3) and all observation items (table 4). Domain scores tended to correlate strongly with one another, with one exception. The highest correlation (.80) was observed between the two

**Figure 1. Classroom observation item scores in the Arizona pilot teacher evaluation model in 2012/13 were more concentrated than those in the Measures of Effective Teaching Project model**




**Note:** The Arizona pilot teacher evaluation model uses a 0–3 scale for observation scores. Measures of Effective Teaching data uses a conventional 1–4 scale. In this figure Measures of Effective Teaching scores are shifted down by one for comparability. The Arizona distributions include data from all pilot study teachers. Ratings were based on two observations for the Arizona pilot study and two or more observations for the Measures of Effective Teaching data. Since each study used different aggregation procedures, the scores were rounded to the nearest integer for comparability. Percentages may not sum to 100 because of rounding.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education and Measures of Effective Teaching Project (2010).

**Table 3. Correlations between observation domain scores in the Arizona pilot teacher evaluation model, 2012/13**

| Domain | 1. Planning and Preparation | 2. Classroom Environment | 3. Instruction | 4. Professional Responsibilities |
|---|---|---|---|---|
| 1. Planning and Preparation | × | | | |
| 2. Classroom Environment | .74 | × | | |
| 3. Instruction | .71 | .80 | × | |
| 4. Professional Responsibilities | .77 | .56 | .51 | × |

**Note:** Positive correlations imply that teachers who have higher scores on one domain will have higher scores on other domains as well. All correlation coefficients are significant at the .05 level.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

domains scored through classroom observation, Classroom Environment and Instruction. The Planning and Preparation domain, which is based on information gathered outside the classroom, also had a strong correlation (.71 to .77) with all other domains. By contrast, the Professional Responsibilities domain, which is also based on observation outside the classroom, had lower correlations with the Classroom Environment (.56) and Instruction domains (.51).

*The correlations between the 22 observation items ranged from .10 to .67, and all but one were statistically significant at the .05 level*

The correlations between the 22 observation items ranged from .10 to .67, and all were statistically significant at the .05 level (except the correlation between item 3d, Using Assessment in Instruction, and item 4b, Maintaining Accurate Records, which was significant at the .10 level; see table 4). About two-thirds of the pairwise coefficients ranged from .3 to .5. These findings suggest that the observation instrument yielded internally consistent results (all correlations are positive) and that no elements provide redundant information (no correlations are close to one).

Item-to-item correlations within domains were all around .5, on average. The average correlations between items in different domains exhibited the pattern demonstrated in table 3—stronger correlations for domains 1–3 and lower correlations for domain 4. Item 3d, Using Assessment in Instruction, tended to have lower correlations with other items and domains, due in part to its low variability (see table 1).

The average correlation in the Arizona data (.58) was slightly higher than the average correlation in the Measures of Effective Teaching data (.53). The correlations were about the same within domain 2 but tended to differ within domain 3 and between the observation items in domains 2 and 3 (see table D1 in appendix D). These differences are consistent with the finding described earlier: a lower variability of scores in the Arizona data than in the Measures of Effective Teaching data and a greater similarity between the score distributions of the two studies within domain 2. Rating teacher performance in the instruction domain may thus present greater difficulties for observers who have not received extensive training in the Framework for Teaching.

### Observation results captured multiple independent aspects of teaching performance

If the observation items measure a single underlying aspect of teaching effectiveness, a single aggregated observation score obtained by summing or averaging item scores would be a valid measure of teaching. If observation items measure multiple independent aspects

**Table 4. Correlations between observation items in the Arizona pilot teacher evaluation model, 2012/13**

| | Domain 1 | | | | | | Domain 2 | | | | | Domain 3 | | | | | Domain 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1a | 1b | 1c | 1d | 1e | 1f | 2a | 2b | 2c | 2d | 2e | 3a | 3b | 3c | 3d | 3e | 4a | 4b | 4c | 4d | 4e |
| 1a | | | | | | | | | | | | | | | | | | | | | |
| 1b | .41 | | | | | | | | | | | | | | | | | | | | |
| 1c | .58 | .53 | | | | | | | | | | | | | | | | | | | |
| 1d | .55 | .46 | .53 | | | | | | | | | | | | | | | | | | |
| 1e | .59 | .51 | .66 | .40 | | | | | | | | | | | | | | | | | |
| 1f | .53 | .50 | .63 | .58 | .52 | | | | | | | | | | | | | | | | |
| 2a | .46 | .52 | .54 | .45 | .44 | .47 | | | | | | | | | | | | | | | |
| 2b | .61 | .42 | .57 | .56 | .51 | .56 | .63 | | | | | | | | | | | | | | |
| 2c | .48 | .39 | .53 | .48 | .44 | .52 | .56 | .59 | | | | | | | | | | | | | |
| 2d | .48 | .43 | .49 | .39 | .45 | .50 | .61 | .59 | .57 | | | | | | | | | | | | |
| 2e | .35 | .21 | .31 | .32 | .32 | .32 | .24 | .36 | .42 | .28 | | | | | | | | | | | |
| 3a | .62 | .46 | .60 | .46 | .55 | .47 | .52 | .64 | .52 | .51 | .36 | | | | | | | | | | |
| 3b | .51 | .41 | .51 | .49 | .53 | .53 | .49 | .64 | .56 | .50 | .40 | .62 | | | | | | | | | |
| 3c | .46 | .38 | .44 | .45 | .48 | .45 | .50 | .66 | .52 | .53 | .41 | .61 | .67 | | | | | | | | |
| 3d | .34 | .35 | .32 | .23 | .31 | .34 | .38 | .39 | .33 | .35 | .28 | .41 | .39 | .38 | | | | | | | |
| 3e | .49 | .40 | .45 | .41 | .41 | .50 | .54 | .60 | .53 | .51 | .41 | .57 | .57 | .62 | .59 | | | | | | |
| 4a | .52 | .49 | .44 | .39 | .52 | .48 | .31 | .37 | .31 | .39 | .24 | .41 | .41 | .28 | .18 | .33 | | | | | |
| 4b | .38 | .45 | .40 | .52 | .36 | .54 | .35 | .37 | .34 | .33 | .14 | .31 | .31 | .28 | .10 | .32 | .48 | | | | |
| 4c | .38 | .52 | .39 | .45 | .44 | .49 | .32 | .37 | .36 | .32 | .21 | .31 | .31 | .25 | .11 | .25 | .46 | .64 | | | |
| 4d | .49 | .55 | .48 | .47 | .50 | .57 | .45 | .45 | .39 | .43 | .25 | .41 | .42 | .38 | .34 | .39 | .48 | .52 | .43 | | |
| 4e | .46 | .49 | .51 | .50 | .45 | .50 | .39 | .41 | .32 | .35 | .25 | .46 | .43 | .29 | .21 | .33 | .51 | .49 | .45 | .61 | |
| 4f | .38 | .54 | .39 | .40 | .44 | .45 | .39 | .34 | .29 | .35 | .19 | .36 | .35 | .28 | .25 | .32 | .58 | .50 | .42 | .63 | .51 |

**Note:** See table 1 for a list of the observation items by number and name. Positive correlations imply that teachers who have higher scores on one observation item will have higher scores on other observation items as well. All coefficients of .11 or above are significant at the .05 level. Correlation of .10 between items 3d and 4b is significant at the .10 level.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

of teaching effectiveness, more complex compositing methods should be developed—for example, awarding separate scores for each aspect of teaching.

Correlation analysis indicated that all observation items were interrelated, but these results cannot indicate whether observation items measure a single aspect or multiple aspects of teaching effectiveness. While a thorough examination of this issue would require analysis beyond the scope of this exploratory study, a principal component analysis was performed (see appendix C for details) to better understand the underlying teacher effectiveness construct. Of particular interest was the data variation explained by the first principal component (a combination of items that explains the largest proportion of total variance). This statistic summarizes the underlying structure of observation item correlations. In a dataset where all items are closely correlated (that is, where all items could be measurements of a single teacher effectiveness construct), the proportion of total variance in the data explained by the first principal component would be large (well in excess of 50 percent[6]), or the difference between the proportion of variance explained by the first and all other components would be large, while the differences among remaining components would be small.[7]

In this study, less than half (46 percent) of the total variance in the Arizona observation scores was explained by the first principal component (table 5). The second principal component accounted for a substantial proportion of the total variance, almost 10 percent. Only at the fourth principal component was the proportion of explained variance small (and relatively unchanged for subsequent components). This suggests that the Arizona observation instrument may have captured several independent aspects of teacher performance. Further exploration of alternative methods for developing composite observation scores is recommended.

### Observation domain scores correlated with student academic progress only in domains observed outside the classroom

Analysis of correlations between domain-level scores in the observation component and the student academic progress component indicated that only the Professional Responsibilities domain had a statistically significant correlation with student academic progress for all teachers (.14; table 6). The Planning and Preparation and Professional Responsibilities

*Only the Professional Responsibilities domain had a statistically significant correlation with student academic progress for all teachers*

**Table 5. Principal component analysis of observation items in the Arizona pilot teacher evaluation model, 2012/13**

| Principal component | Proportion of variance | Cumulative proportion |
|---|---|---|
| 1 | 46.1 | 46.1 |
| 2 | 9.9 | 55.9 |
| 3 | 5.6 | 61.5 |
| 4 | 4.0 | 65.5 |
| 5 | 3.6 | 69.1 |
| 6 | 3.4 | 72.5 |
| 7 | 2.8 | 75.4 |
| 8 | 2.6 | 77.9 |

**Note:** Only the first eight principal components are shown (each remaining component explains approximately 2 percent of total variance or less).

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

**Table 6. Correlations between observation domain scores and student academic progress scores in the Arizona pilot teacher evaluation model, 2012/13**

| Domain | All teachers (*n* = 297) | Group A[a] (*n* = 104) | Group B[b] (*n* = 193 ) |
|---|---|---|---|
| 1. Planning and Preparation | .10 | .02 | .15** |
| 2. Classroom Environment | .09 | .10 | .07 |
| 3. Instruction | .04 | .01 | .04 |
| 4. Professional Responsibilities | .14** | .08 | .18** |

** Significant at the .05 level.

**Note:** Because of the differences in defining and calculating student academic progress for the different groups of teachers in the Arizona pilot study, analysis results are given separately for all pilot teachers, group A teachers, and group B teachers.

**a.** Teachers for whom classroom-level student standardized test data were available.

**b.** Teachers for whom no classroom-level student standardized test data were available for their content area, so their results derived from math and reading tests their students took for other classes.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

domains correlated significantly for group B teachers, the group whose results derived from statewide math and reading tests their students took for other classes. There were no significant correlations between the Classroom Environment and Instruction domains and the student academic progress component. This finding suggests that observations made outside the classroom could be a more reliable predictor of student academic progress than formal classroom observations.

### Few correlations were found between observation items and student academic progress

There were apparent differences in the patterns of correlation between observation item scores and student academic progress between the two groups of teachers (table 7). For group A teachers (those for whom standardized classroom-level achievement data were available for their content area), the only observation item that correlated significantly with student academic progress was Managing Classroom Procedures, from the Classroom Environment domain.[8]

***For teachers for whom standardized classroom-level achievement data were available for their content area, the only observation item that correlated significantly with student academic progress was Managing Classroom Procedures, from the Classroom Environment domain***

For group B teachers, significantly correlated items were found only in the other three domains, primarily in the two domains based on information collected outside the classroom (Planning and Preparation and Professional Responsibilities). Considering that group B teachers lacked data on standardized classroom-level student achievement, it may not be surprising that student academic progress had a low correlation with domains based on classroom observation data (Classroom Environment and Instruction). Items associated with productive interactions with students (such as Using Questioning and Discussion Techniques and Demonstrating Knowledge of Students) were among those that had the highest correlations with student academic progress for group B teachers, while the more "directive" aspects of teaching assessed by such items as Managing Classroom Procedures, Managing Student Behavior, and Setting Instructional Outcomes had higher correlations for group A teachers.[9]

The differences between the two groups of teachers may be due in part to the fact that a larger proportion of group B teachers worked in middle and high schools, while a larger proportion of group A teachers taught in elementary schools. The Measures of Effective Teaching data showed that there were significant differences across grade levels in the patterns of correlation between observation items and student academic progress (Lazarev & Newman, 2013). In particular, for teachers in elementary grades correlations were higher with items associated with classroom and student behavior management, while for teachers in middle school grades correlations were higher with items such as Establishing a Culture for Learning. These differences in correlation patterns of observation items and student academic progress metrics could be associated with developmental changes affecting the patterns of effective teaching at different grade levels. Arizona results appear consistent with this interpretation, although the small size and heterogeneity of groups A and B do not allow for definitive conclusions.

The findings suggest that observation and student academic progress scores were consistent for the Arizona pilot teacher evaluation model (see table 7). They also suggest that, due to differences in patterns of associations between student academic progress and observation scores for groups A and B, separate compositing formulas may be needed for the two groups of teachers.

**Table 7. Correlations between observation item scores and student academic progress scores in the Arizona pilot teacher evaluation model, 2012/13**

| Domain and observation item | All teachers (*n* = 297) | Group A[a] (*n* = 104) | Group B[b] (*n* = 193 ) |
|---|---|---|---|
| 1. Planning and preparation | | | |
| 1a: Demonstrating Knowledge of Content and Pedagogy | –.02 | –.05 | .02 |
| 1b: Demonstrating Knowledge of Students | .09 | .04 | .15** |
| 1c: Setting Instructional Outcomes | .10 | .14 | .07 |
| 1d: Demonstrating Knowledge of Resources | .04 | –.08 | .13 |
| 1e: Designing Coherent Instruction | .09 | –.04 | .19** |
| 1f: Designing Student Assessments | .17** | .11 | .17** |
| 2. Classroom Environment | | | |
| 2a: Creating an Environment of Respect and Rapport | .05 | .06 | .06 |
| 2b: Establishing a Culture for Learning | .01 | –.03 | .01 |
| 2c: Managing Classroom Procedures | .12** | .22** | .06 |
| 2d: Managing Student Behavior | .12** | .13 | .10 |
| 2e: Organizing Physical Space | .08 | –.01 | .10 |
| 3. Instruction | | | |
| 3a: Communicating with Students | .06 | .03 | .06 |
| 3b: Using Questioning and Discussion Techniques | .10 | –.06 | .19** |
| 3c: Engaging Students in Learning | .09 | –.01 | .13 |
| 3d: Using Assessment in Instruction | .09 | .05 | .10 |
| 3e: Demonstrating Flexibility and Responsiveness | .10 | .07 | .08 |
| 4. Professional Responsibilities | | | |
| 4a: Reflecting on Teaching | .05 | .04 | .06 |
| 4b: Maintaining Accurate Records | .11 | .03 | .13 |
| 4c: Communicating with Families | .09 | .13 | .06 |
| 4d: Participating in a Professional Community | .12** | .00 | .21** |
| 4e: Growing and Developing Professionally | .13** | .03 | .17** |
| 4f: Showing Professionalism | .13** | .11 | .16** |

** Significant at the .05 level.

**a.** Teachers for whom classroom-level student standardized test data were available.

**b.** Teachers for whom no classroom-level student standardized test data were available for their content area, so their results derived from math and reading tests their students took for other classes.

**Note:** Correlation analysis was performed using both linear Pearson coefficient and nonparametric Kendall coefficient. Both statistics produced the same results in terms of significance of correlations. Only Pearson coefficients are displayed.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

## The strength of correlation between observation items and student academic progress differed for higher and lower scoring teachers

The strength of correlation between observation item scores and student academic progress varied between teachers with high and low student academic progress scores (table 8). For example, the results for observation item 3d, Using Assessment in Instruction, correlated with differences in student academic progress scores only among teachers who scored relatively low on the item. None of the observation items differentiated between student academic progress scores across the full range of scoring differences (1–2, 2–3, and 3–4). The fact that only a few rows in table 8 contain more than one or two significant differences illustrates this point. The strength of this correlation also varied between the

**Table 8. Mean differences between student academic progress scores in the Arizona pilot teacher evaluation model, by observation item score across group and score ranges, 2012/13**

| Domain and observation item | All teachers | | | Group A[a] | | | Group B[b] | | |
|---|---|---|---|---|---|---|---|---|---|
| | Between 1 and 2 | Between 2 and 3 | Between 3 and 4 | Between 1 and 2 | Between 2 and 3 | Between 3 and 4 | Between 1 and 2 | Between 2 and 3 | Between 3 and 4 |
| 1. Planning and Preparation | | | | | | | | | |
| 1a | na | .02 | –.01 | na | –.18 | .01 | na | .12** | –.02 |
| 1b | –.13 | –.01 | .07** | na | –.14 | .09 | na | .04 | .08** |
| 1c | na | .00 | .07** | na | .01 | .09 | na | .00 | .05 |
| 1d | na | .00 | .02 | na | –.12 | .00 | na | .07 | .05 |
| 1e | na | –.04 | .08** | na | –.09 | .02 | na | –.02 | .12** |
| 1f | na | .03 | .11** | na | –.09 | .17** | na | .10 | .06 |
| 2. Classroom Environment | | | | | | | | | |
| 2a | na | –.01 | .07** | na | .16** | .02 | na | –.05 | .09** |
| 2b | na | –.03 | .05 | na | –.08 | .07 | na | .01 | .02 |
| 2c | na | .05 | .09** | na | .11 | .15** | na | .05 | .04 |
| 2d | na | –.02 | .12** | na | –.02 | .13 | na | –.03 | .11** |
| 2e | na | .06 | .05 | na | –.01 | .00 | na | .10** | .06 |
| 3. Instruction | | | | | | | | | |
| 3a | na | .06 | .04 | na | .02 | .04 | na | .08** | .04 |
| 3b | –.08 | .02 | .16** | na | –.05 | .19 | na | .07** | .12** |
| 3c | –.04 | .04 | .06 | na | –.07 | .09 | .03 | .10** | .04 |
| 3d | –.13 | .10** | .08 | na | .04 | –.10 | –.12 | .13** | .10 |
| 3e | na | .03 | .06 | na | –.03 | .05 | na | .07** | .06 |
| 4. Professional Responsibilities | | | | | | | | | |
| 4a | .06 | –.05 | .06 | –.09 | –.09 | .07 | .12 | –.02 | .05 |
| 4b | –.04 | –.01 | .09** | –.28 | –.03 | .11 | .19 | –.08 | .07** |
| 4c | .14 | –.06 | .09** | na | –.05 | .14** | .21** | –.09 | .06 |
| 4d | –.09 | –.04 | .09** | –.43** | .04 | .06 | .13 | –.02 | .11** |
| 4e | na | –.06 | .12** | na | –.13 | .07 | na | –.01 | .13** |
| 4f | –.15 | .02 | .09** | na | –.05 | .11** | –.08 | .04 | .09** |

** Significant at the .05 level.

na is not applicable because of an insufficient number of observations.

**Note:** Differences are on the 40-point student academic progress score scale.

**a.** Teachers for whom classroom-level student standardized test data were available.

**b.** Teachers for whom no classroom-level student standardized test data were available for their content area, so their results derived from math and reading tests their students took for other classes.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

two groups of teachers. Almost all items differentiated between at least one set of student academic progress scoring levels in group B, while only a few comparisons for group A teachers indicated significant differences in student academic progress.[10]

These findings suggest that summing observation item scores to produce a single metric (the aggregated observation component score) may not be appropriate. It is likely that some observation items can accurately evaluate teacher effectiveness only at the high end of the student academic progress component distribution, while other observation items can do so only at the low end. Consequently, some observation item score ranges should not contribute to a domain score.

### Student academic progress correlated with other metrics only among group B teachers

Despite efforts to align the scales, there are differences in correlation patterns between teachers' component scores in groups A and B (table 9). The most salient difference concerns the correlations between student academic progress and other component scores. This is expected, since group A teachers were evaluated primarily on student test results in their content area.

There are no significant correlations between student academic progress and other components or subcomponents for group A teachers. This lack of consistency suggests a need to revise the implementation of survey and observation instruments or the design of the student academic progress component, or both. For group B teachers, in contrast, student academic progress correlated significantly with domains 1 and 4 (Planning and Preparation and Professional Responsibilities). Both domains are based on observations outside the classroom and may reflect characteristics of the school environment rather than individual teacher classroom performance.

**Table 9. Correlations between component, subcomponent, and observation domain scores in the Arizona teacher evaluation model, by teacher group, 2012/13**

| | | Student academic progress | Domain 1 | Domain 2 | Domain 3 | Domain 4 | Observation total[a] | Student survey | Parent survey | Peer survey | Survey total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GROUP A | Student academic progress | × | | | | | | | | | |
| | Domain 1 | 0.02 | × | | | | | | | | |
| | Domain 2 | 0.10 | 0.67** | × | | | | | | | |
| | Domain 3 | 0.01 | 0.69** | 0.79** | × | | | | | | |
| | Domain 4 | 0.08 | 0.73** | 0.47** | 0.45** | × | | | | | |
| | Observation total[a] | 0.05 | 0.92** | 0.86** | 0.87** | 0.76** | × | | | | |
| | Student survey | 0.00 | 0.29** | 0.42** | 0.37** | 0.20** | 0.38** | × | | | |
| | Parent survey | –0.11 | 0.30** | 0.36** | 0.40** | 0.11 | 0.35** | 0.28** | × | | |
| | Peer survey | 0.06 | 0.12 | 0.15 | 0.12 | 0.09 | 0.14 | –0.05 | 0.02 | × | |
| | Survey total | 0.06 | 0.25** | 0.41** | 0.37** | 0.14 | 0.34** | 1.00** | 0.41** | 0.01 | × |
| GROUP B | Student academic progress | × | | | | | | | | | |
| | Domain 1 | 0.15** | × | | | | | | | | |
| | Domain 2 | 0.07 | 0.77** | × | | | | | | | |
| | Domain 3 | 0.04 | 0.72** | 0.81** | × | | | | | | |
| | Domain 4 | 0.18** | 0.78** | 0.59** | 0.55** | × | | | | | |
| | Observation total[a] | 0.12 | 0.93** | 0.90** | 0.87** | 0.82** | × | | | | |
| | Student survey | 0.10 | 0.24** | 0.27** | 0.28** | 0.29** | 0.31** | × | | | |
| | Parent survey | 0.11 | 0.21** | 0.20** | 0.24** | 0.24** | 0.26** | 0.01 | × | | |
| | Peer survey | 0.07 | 0.07 | 0.12 | 0.09 | 0.06 | 0.09 | 0.06 | 0.01 | × | |
| | Survey total | 0.18** | 0.22** | 0.28** | 0.29** | 0.27** | 0.30** | 0.99** | 0.27** | 0.16** | × |

**Significant at the .05 level.

**Note:** Group A teachers had classroom-level student standardized test data available; group B teachers had no classroom-level student standardized test data available for their content area, so their results derived from math and reading tests their students took for other classes.

**a.** Calculated as the sum of the four domain scores.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

Another notable difference concerns the stakeholder surveys. For group A teachers, the parent and student surveys were more strongly correlated with domains 2 and 3 (Classroom Environment and Instruction), which were uncorrelated with student academic progress.[11] The aggregated survey component score correlated with student academic progress in group B.

One consistency across the two teacher groups was the isolation of peer survey results, which had no significant correlations with any other metric, including the total survey component score. It appears, therefore, that peer survey results failed to convey relevant information. Removing this peer survey subcomponent would likely increase the efficiency of the teacher evaluation model.

***Teachers scoring in the top and bottom quartiles of the aggregated observation and survey component score distributions did not have significantly different mean student academic progress scores***

The study team also investigated whether there was a correspondence between component scores for teachers performing at the extremes of the distributions of each component score (who tend to be the focus of high-stakes decisions). The observation and survey component scores were tested for significant differences between the mean scores of teachers in the upper and lower quartiles of the student academic progress distribution (using *t*-tests and nonparametric Wilcoxon rank sum tests; table 10). None of the differences were statistically significant, which implies that large differences in teacher performance as measured by student academic progress were not reflected in teacher observation or survey component scores. In other words, high- and low-performing teachers did not look much different on average when evaluated using only the aggregated metrics.

Complementary analysis showed similar results: teachers scoring in the top and bottom quartiles of the aggregated observation and survey component score distributions did not have significantly different mean student academic progress scores (table 11).

The results in tables 10 and 11 suggest that there was not a strong enough correspondence between the three component scores in the Arizona model (observation, student academic progress, and survey) to allow inferring one score from another. The Arizona Department of Education might consider different ways to create teacher evaluation composite scores that would account for the statistical properties of component metrics (for example, their variance and correlation with other metrics).

**Table 10. Mean differences in component scores between teachers with high and low student academic progress scores in the Arizona pilot teacher evaluation model, 2012/13**

| Teacher scoring on student academic progress | Statistic | Mean aggregated observation component score | Mean aggregated survey component score |
|---|---|---|---|
| High (4th quartile) | Mean | .76 | .56 |
| | Standard deviation | .14 | .36 |
| Low (1st quartile) | Mean | .74 | .51 |
| | Standard deviation | .10 | .33 |
| Significance of mean difference, *p* value | *t*-test | .23 | .50 |
| | Rank sum test | .28 | .87 |

**Note:** Component scores were given as percentages of maximum possible and ranged from 0 to 1.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

**Table 11. Mean differences in student academic progress between teachers with high and low aggregated observation and survey component scores in the Arizona pilot teacher evaluation model, 2012/13**

| Teacher scoring on aggregated observation and survey components | Statistic | Observation (high- vs. low-scoring teachers) | Survey (high- vs. low-scoring teachers) |
|---|---|---|---|
| High (4th quartile) | Mean | .58 | .53 |
| | Standard deviation | .24 | .27 |
| Low (1st quartile) | Mean | .53 | .54 |
| | Standard deviation | .19 | .18 |
| Significance of mean difference, *p* value | *t*-test | .13 | .83 |
| | Rank sum test | .20 | .53 |

**Note:** Student academic progress scores were given as percentages of maximum possible and ranged from 0 to 1.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

## Implications of the study and suggestions for further research

The study findings have several implications for research and practice.

*The survey measures may deserve more attention if the state wishes to give greater weight to metrics that correlate more with student academic progress*

The analysis indicated that the ratings using the Framework for Teaching observation instrument tended to be underdispersed—often with more than 60 percent of the ratings on just one of the four possible levels. A lack of dispersion in scoring reduces the instrument's ability to differentiate between high- and low-performing teachers. Although it is possible that the teachers in the pilot evaluation model were just very homogeneous in their teaching performance, the state could test whether administrators could use the observation instrument to better differentiate performance by focusing additional observer training on items in domain 3 (Instruction), especially item 3d (Using Assessment in Instruction).

Only two domains in the observation instrument (Planning and Preparation and Professional Responsibilities), which are both assessed outside of the classroom, correlated significantly with the student academic progress component—and only for the group B teachers, who were evaluated on that component in large part based on student test results outside the subject they taught. The student academic progress component also correlated with the overall stakeholder survey component among the group B teachers. The survey measures represented the cumulative experience of the principal, peers, students, and parents with the teacher during the year. These measures may deserve more attention if the state wishes to give greater weight to metrics that correlate more with student academic progress.

All of the results for teachers evaluated on classroom-level standardized student test scores (group A) tended to differ from results for teachers evaluated based on math and reading tests their students took for other classes (group B). Given the different practical realities faced by the two groups of teachers, the state should ensure that every teacher's evaluation appropriately accounts for teaching environment, including grade level, classroom type, and student characteristics.

The structure of the Arizona model was very complex. For no single evaluation component could the results be predicted on the basis of another component's results. This finding can be viewed in several ways. It could indicate that each component contributed unique

information (although components uncorrelated with other metrics, such as the peer survey, may need to be substantially revised or eliminated). Alternatively, it could indicate that information was lost in aggregating results to the component level. The Arizona Department of Education should explore more efficient ways to create summative scores for teachers by further examining the structure underlying the item scores or employing strategies to account for the statistical properties of the components (for example, reliability, variance, correlation with other components) and potential differences across groups of teachers.

## Study limitations

This study has certain limitations. For example, the small sample size and the self-selection of small-town and rural districts limited the generalizability of the findings. However, the study's data collection focus produced results from an under-researched rural segment of K–12 schools and thus makes a useful addition to the field.

Another limitation is that the study relied on summative evaluation data, aggregated at the teacher level, rather than raw, disaggregated datasets. Access to raw data from the three components of the teacher evaluation model—observation, student academic progress, and surveys—would have allowed more sophisticated analyses of the instruments as well as a more in-depth look at the relationships between components.

Finally, only one year of data was available, limiting the models that could be applied and the conclusions that could be drawn. A teacher evaluation model should have an optimal weighting schema that can predict student outcomes based on results from current and prior years (Mihaly et al., 2013). This type of analysis requires access to at least two successive years of student academic progress results, which, in turn, require three successive years of student test data.

***The Arizona Department of Education should explore more efficient ways to create summative scores for teachers by further examining the structure underlying the item scores or employing strategies to account for the statistical properties of the components and potential differences across groups of teachers***

## Appendix A. Arizona teacher evaluation model

In May 2010 the Arizona legislature passed, and the governor signed, Senate Bill 1040, which requires all district and charter schools to evaluate teachers and principals every year. The enabling legislation empowered the Arizona State Board of Education to adopt a teacher and principal evaluation model that includes quantitative data on student performance. The State Board of Education adopted the Arizona Framework for Measuring Educator Effectiveness in April 2011. In April 2012 House Bill 2823 further clarified the State Board of Education's and districts' responsibilities related to the design and implementation of teacher and principal evaluations. Specifically, state law (A.R.S. 15–203[A] [38]; 15–537[D] [1]; and 15–537[G]) now requires that Arizona local education agencies use an instrument that meets the requirements established by the Framework for Measuring Educator Effectiveness to evaluate all teachers and principals every year, beginning in 2012/13. Beginning in 2013/14, Arizona local education agencies are required to describe in their policies how teacher and principal performance classifications will be used in making employment-related decisions.

### Data and scoring

In the summer of 2012 the Arizona Department of Education developed a teacher evaluation model (referred to here as the Arizona model) based on the Framework for Measuring Educator Effectiveness. The model includes three components: teaching observations; measures of student academic progress; and surveys of students, parents, and peer teachers. A teacher's composite evaluation score is a weighted average of the three component scores. The model's base formula assigns weights of 50, 33, and 17 to these components, respectively. All district evaluation models are supposed to include at least observations and student academic progress, but the evaluation details are at the discretion of local education agencies. The 2012/13 model documentation recommended various weighting schemas for various environments (Arizona Department of Education, 2012).

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 and associated training. Participating principals administered two classroom observations per teacher using Danielson's Framework for Teaching rubric (Danielson Group, 2011).

Due to the variations in the nature and extent of available test data, the Arizona model relies on a number of formulas to calculate student academic progress depending on the grade level and subject area taught. Possible elements contributing to a teacher's student academic progress score include:

- Arizona Instrument to Measure Standards—classroom, aggregate school, or grade-level results.
- Stanford 10—classroom, aggregate school, or grade-level results.
- Dynamic Indicators of Basic Early Literacy Skills.
- Other local education agency or school-level assessments.

The Arizona Department of Education 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 items from Cambridge Education's Tripod Student Perception Survey and asked students to rate their teacher on the extent to which the teacher Captivates

Students, Cares about Students, Challenges Students, Clarifies Lessons, Confers with Students, Consolidates Knowledge, Controls Behavior, and Engages Students (eight constructs).[12] The 15-question peer survey asked teachers to rate their peers' performance on a four-point scale. 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.

### Pilot participation

In the summer of 2012 the Arizona Department of Education reached out to all local education agencies in the state to participate in a teacher evaluation pilot, with the goal of gathering a sample of schools that were most likely to eventually adopt the Arizona model. All districts that expressed interest were accepted into the pilot, which included 297 teachers. These teachers came from 11 public schools in four school districts, as well as one charter agency (see appendix B for demographic information on the pilot districts and schools). This small sample is not representative of the public K–12 education system in Arizona, which includes more than 1,500 operating schools.

As anticipated, the pilot local education agencies included primarily smaller rural or town schools,[13] which lack the capacity to develop and implement their own teacher evaluation models and thus are most likely to eventually adopt the Arizona model. Larger districts in Arizona, given their assessment and research capacities, will likely develop their own local multiple-measure teacher evaluation models.

### Sample scoring schemas for 40-point student academic progress component from the Arizona 2012/13 teacher evaluation model

The Arizona Department of Education's Research and Evaluation Division sought to maintain consistency in how local education agencies calculated the school-, grade-, or classroom-level student academic progress component of teacher evaluations in the 2012/13 pilot. The department also trained the staff in charge of student data management on data entry and tabulation and monitored data management closely throughout the pilot year. The teacher data tables for 2013/14 are available at http://www.azed.gov/teacherprincipal-evaluation/teacher-rating-tables/. Table A1 was developed by the department and provided as references. It is included here to display the types of scoring schema adopted by the department for use during the pilot year (2012/13). These weightings are just examples, however; the exact weighting for any given teacher depends on the availability of that teacher's classroom-level student test data.

## Table A1. Sample rating table in the Arizona pilot teacher evaluation model, 2012/13

| Points/percent of school-level data | Category | Point value | Classroom-level data | Point value | Point determination | | |
|---|---|---|---|---|---|---|---|
| 40 points; 33 percent of total score | Achievement | 8 | Grades 3–8 and 10: Percent passing AIMS Reading | 4 | 4 points:<br>3 points:<br>2 points:<br>0 points: | 80–100%<br>60–79<br>20–59<br><20 | |
| | | | Grades 2 and 9: Percent at or above the fourth stanine on the Stanford 10 Language | | | | |
| | | | Grades 3–8 and 10: Percent passing AIMS Mathematics | 4 | 4 points:<br>3 points:<br>2 points:<br>0 points: | 80–100%<br>60–79<br>20–59<br><20 | |
| | | | Grades 2 and 9: Percent at or above the fourth stanine on the Stanford 10 Mathematics | | | | |
| | Classroom student learning objectives (to be implemented in 2013/14)[a] | | | | | | |
| | Growth | 24 | Grades 4–7: Catch Up mean ratio of student growth target—Reading | 8 | 8 points:<br>0 points: | ≥ 1.16<br>< 1.16 | Sum of points from both levels divided by 2 to total up to 8 points |
| | | | Grades 4–7: Keep Up mean ratio of student growth target—Reading | | 8 points:<br>0 points: | ≥ 5.97<br>< 5.97 | |
| | | | Grades 4–7: Catch Up mean ratio of student growth target—Mathematics | 8 | 8 points:<br>0 points: | ≥ .85<br>< .85 | Sum of points from both levels divided by 2 to total up to 8 points |
| | | | Grades 4–7: Keep Up mean ratio of student growth target—Mathematics | | 8 points:<br>0 points: | ≥ 2.18<br>< 2.18 | |
| | | | Grades 3–10: Mean student growth percentile (Reading and Mathematics) | 8 | 8 points:<br>6 points:<br>5 points:<br>4 points:<br>3 points:<br>2 points:<br>1 point:<br>0 points: | 58–100%<br>53–57<br>50–52<br>47–49<br>43–46<br>36–42<br>5–35<br><5 | |
| | Targeted student learning objectives (to be implemented in 2013/14) | | | | | | |
| | Career and college ready | 8 | Grades 3–7 and 10: AIMS college and career ready equivalent score—Reading | 4 | 4 points:<br>3 points:<br>2 points:<br>0 points: | 50–100%<br>30–49<br>5–29<br><5 | |
| | | | Grades 3–7 and 10: AIMS college and career ready equivalent score—Mathematics | 4 | 4 points:<br>3 points:<br>2 points:<br>0 points: | 50–100%<br>30–49<br>5–29<br><5 | |

AIMS is Arizona Instrument to Measure Standards.

**Note:** This table is a truncated version of the full model and was used only in the pilot year (2012/13). The information is part of a pilot teacher and principal evaluation program and has not been validated. The Arizona Department of Education recommended that local education agencies not wholly rely on this information for final 2012/13 teacher and principal evaluations. Data were aggregated for each teacher. If a teacher had multiple classrooms or grades, data from those classrooms or grades were combined before aggregation.

**a.** *Catch Up* refers to those who are currently below "Meets standards" performance level but are expected to reach "Meets standards" within the next three years or by grade 10, whichever comes sooner. *Keep Up* refers to those who are currently at or above "Meets standards" and are expected to remain at or above "Meets standards" within the next three years or by grade 10, whichever comes sooner. *Mean ratio* refers to the average ratio of current student growth percentile to targeted student growth percentile of students in the classroom. The formula is the sum of current student growth percentile and targeted student growth percentile divided by the number of students in the classroom.

**Source:** Arizona Department of Education.

## Appendix B. Pilot local education agency information

This appendix provides more detail on the pilot districts and schools (table B1) and on teachers in groups A and B (table B2).

**Table B1. Pilot school demographics compared with state averages in the Arizona pilot teacher evaluation model, 2010/11 (percent unless otherwise indicated)**

| District ID | Pilot school ID | American Indian/ Alaskan Native | Asian/ Pacific Islander | Black | Hispanic | White | Low-income[a] | English language leaner | Locale type | A–F rating |
|---|---|---|---|---|---|---|---|---|---|---|
| na | Arizona town/ rural average | 10 | 2 | 4 | 37 | 45 | 48 | — | na | na |
| na | Arizona state average | 5 | 3 | 6 | 42 | 43 | 45 | 7 | na | na |
| A | A1 | 2 | 4 | 11 | 33 | 49 | 51 | 4 | Rural: Distant | C |
| A | A2 | 22 | 5 | 8 | 27 | 37 | 58 | 6 | Rural: Distant | C |
| A | A3 | 3 | 4 | 17 | 33 | 42 | 49 | 2 | Rural: Distant | D |
| A | A4 | 13 | 4 | 13 | 39 | 30 | 60 | 2 | Rural: Distant | D |
| A | A5 | 6 | 4 | 18 | 33 | 39 | 47 | 1 | Rural: Distant | C |
| B | B1 | 6 | 1 | 2 | 38 | 53 | 71 | 9 | Rural: Remote | C |
| B | B2 | 3 | 2 | 3 | 40 | 50 | 52 | — | Rural: Remote | B |
| C | C1 | 1 | 1 | 1 | 60 | 35 | 43 | 3 | Town: Remote | C |
| C | C2 | 1 | 1 | 1 | 59 | 37 | 3 | 0 | Town: Remote | B |
| C | C3 | 1 | 1 | 3 | 65 | 30 | — | — | Rural: Fringe | C |
| D | D1 | 14 | 1 | 1 | 70 | 13 | 100 | 21 | Rural: Remote | C |
| E | E1 | 1 | 4 | 4 | 46 | 43 | 38 | 9 | Suburb: Large | A |
| na | Sample average | 7 | 3 | 10 | 42 | 37 | 57 | 5 | na | na |

— is unavailable; na is not applicable.

**a.** Eligible for free or reduced-price lunch.

**Source:** U.S. Department of Education, 2011; Arizona Department of Education, 2011.

**Table B2. Teacher demographics for group A and group B in the Arizona pilot teacher evaluation model, 2012/13 (percent of group total unless otherwise indicated)**

| Characteristic | Group A[a] | Group B[b] |
|---|---|---|
| Total number of teachers | 104 | 193 |
| Grade level | | |
| Elementary | 44 | 33 |
| Elementary and middle | 3 | 1 |
| Middle | 33 | 23 |
| Middle and high | 0 | 1 |
| High | 20 | 39 |
| Various | 0 | 3 |
| Highest degree attained | | |
| Bachelor's | 64 | 66 |
| Master's | 34 | 30 |
| Doctorate | 2 | <1 |
| Racial/ethnic minority | | |
| Yes | 25 | 14 |
| No | 75 | 86 |
| Gender | | |
| Male | 43 | 42 |
| Female | 57 | 58 |
| Years of experience | | |
| Mean | 9.3 | 7.4 |
| Standard deviation | 7.1 | 7.7 |

**a.** Teachers for whom classroom-level student standardized test data were available.

**b.** Teachers for whom no classroom-level student standardized test data were available for their content area.

**Source:** Arizona Department of Education.

## Appendix C. Study methods

This study examined the 2012/13 pilot of the Arizona Department of Education teacher evaluation model, focusing on the statistical properties of the model's components and the relationships between them. The study relied largely on descriptive statistics and correlation analysis.

### Statistical properties of the teacher observation instrument

In examining the statistical properties of the observation instrument, the primary concern was the distribution of scores by observation item and domain. Although the standard descriptive statistics (score means and standard deviations) were reported, the analysis focused on determining which scores were used most frequently by the observers and whether some score values or subranges were disproportionately more frequently assigned, leading to a skewed score distribution that weakened the instrument's effectiveness. For example, if all teachers are given scores of 3 or 4 on a four-point scale, the item effectively has only a two-point scale, and its utility in distinguishing between certain aspects of teaching practice is limited. This is particularly important given recent interest in the observed clustering of observation scores in the middle or high end of an observation metric's scale (Weisberg, Sexton, Mulhern, & Keeling, 2009). To help draw conclusions about the observation instrument's implementation quality, the study results were compared with benchmark Framework for Teaching data from the Measures of Effective Teaching database.

Analysis of correlation revealed the extent of interdependency among the observation items. The results can suggest whether some items that were strongly correlated with other items could be removed from the observation rubric without loss of information, thus making the instrument more efficient. The results can also show whether the instrument is internally consistent (that is, all items are positively correlated).

A principal component analysis then provided a summary characterization of the structure of correlations between observation items.[14] The question of particular interest is whether the first principal component—a linear combination of item scores estimated from the data—explains a sufficiently large share of the total variation in scores (exceeding the proportion explained by the second principal component by one order of magnitude). If the answer is positive, the information inherent in all the observation items could be reduced to a single number (first principal component) without a substantial loss of information—and therefore could be used as a single composite teaching effectiveness score. If two or more principal components have comparable contributions to the total variance but are associated with different items, a single composite score may not be an adequate metric, and the recommendation might be to calculate two scores based on the observation (using different formulas) instead of a single aggregated observation score. Principal component analysis can be performed for the observation instrument (as well as for the survey instrument, had the disaggregated data been made available) to derive an optimal compositing formula.

### Relationships between observation scores and the student academic progress component

Correlation analysis, significance testing, and estimation of nonparametric models were used to investigate the relationships between each of the 22 observation items and the student academic progress component. First, correlations were computed between each

observation item score and the student academic progress component to establish which correlations were statistically significant. Both conventional linear (Pearson) correlation coefficients and rank correlation coefficients (Kendall) were calculated. Then, for each observation item, the distribution of the student academic progress metric was analyzed. Significant differences between the mean academic progress metric at each level of the observation score were tested using both parametric (*t*-test) and nonparametric (Wilcoxon rank sum) tests. These analyses can reveal significant differences in mean student academic progress among teachers receiving different observation scores. If such differences are found only for some levels of an item, it can be concluded that the item is not effective in differentiating between all teachers but can be used to single out the most or least effective teachers. Observed irregularities of this kind may imply that the relationship between items is nonlinear. Although the small sample size did not allow for a full-scale analysis of nonlinear relationships in the data, an exploratory analysis was performed (detailed in appendix E).

### Statistical relationships among the components

This research area was addressed using correlation analysis, applied at the component score level. Correlations were computed separately for each group of teachers: group A teachers, who had classroom-level student standardized test data, and group B teachers, whose student academic progress results were derived from math and reading tests their students took for other classes. This analysis indicates the extent to which the teacher performance metrics align with one another. In addition, analyses were performed to establish correspondence between the score ranges (tiers), defined on the basis of the three aggregated component metrics. First, following Kane and Staiger (2012), the analysis examined whether teachers in the top and bottom of the student academic progress quartiles had significantly different average ratings on the aggregated teacher observation and survey scores. Next, a complementary analysis examined whether teachers in each pair of adjacent quartiles had significantly different average student outcomes. This complementary analysis allows conclusions to be drawn about the practical relevance of each component differentiating between teachers at various performance levels.

## Appendix D. Detailed Arizona teacher evaluation model pilot results, 2012/13

The following tables provide additional information about the various components of Arizona's teacher evaluation model.

**Table D1. Comparative correlations of observation item scores of teachers in the Arizona pilot teacher evaluation model and Measures of Effective Teaching project districts, 2012/13**

| | Observation item | 2a | 2b | 2c | 2d | 3a | 3b | 3c |
|---|---|---|---|---|---|---|---|---|
| ARIZONA MODEL | 2a. Creating an Environment of Respect and Rapport | 1.00 | | | | | | |
| | 2b. Establishing a Culture for Learning | .60 | 1.00 | | | | | |
| | 2c. Managing Classroom Procedures | .55 | .60** | 1.00 | | | | |
| | 2d. Managing Student Behavior | .60 | .54 | .57 | 1.00 | | | |
| | 3a. Communicating with Students | .51 | .60** | .52 | .50** | 1.00 | | |
| | 3b. Using Questioning and Discussion Techniques | .49 | .60** | .60** | .50** | .60** | 1.00 | |
| | 3c. Engaging Students in Learning | .49 | .62 | .50** | .50** | .59 | .70** | 1.00 |
| | 3d. Using Assessment in Instruction | .39 | .40** | .34 | .39 | .41 | .40** | .40** |
| MEASURES OF EFFECTIVE TEACHING DATA | 2a. Creating an Environment of Respect and Rapport | 1.00 | | | | | | |
| | 2b. Establishing a Culture for Learning | .54 | 1.00 | | | | | |
| | 2c. Managing Classroom Procedures | .55 | .47 | 1.00 | | | | |
| | 2d. Managing Student Behavior | .64 | .46 | .61 | 1.00 | | | |
| | 3a. Communicating with Students | .48 | .53 | .46 | .41 | 1.00 | | |
| | 3b. Using Questioning and Discussion Techniques | .42 | .52 | .35 | .32 | .50 | 1.00 | |
| | 3c. Engaging Students in Learning | .47 | .64 | .42 | .40 | .52 | .56 | 1.00 |
| | 3d. Using Assessment in Instruction | .42 | .52 | .39 | .35 | .47 | .56 | .57 |

** Significantly different from the Measures of Effective Teaching data benchmark at the 0.05 level.

**Note:** Scores were available only on these eight domains in the Measures of Effective Teaching data.

**Source:** Data from the Arizona Department of Education for 2012/13 and Measures of Effective Teaching Project (2010).

**Table D2. Descriptive statistics of student academic progress, survey composite, and individual survey (student, parent, peer) scores in the Arizona pilot teacher evaluation model, 2012/13**

| Measure | Group[a] | Minimum | Maximum | Median | Mean | Standard deviation |
|---|---|---|---|---|---|---|
| Student academic progress composite | A | 1 | 40 | 16 | 17.4 | 9.2 |
| | B | 10 | 40 | 19 | 22.5 | 8.4 |
| | No student survey | 9 | 44 | 26 | 27.7 | 8.5 |
| Survey composite | A | 2 | 20 | 6 | 9.9 | 6.9 |
| | B | 1 | 20 | 5 | 9.2 | 6.7 |
| | No student survey | 4 | 10 | 8 | 8.1 | 1.5 |
| Student survey | A | 0 | 15 | 1 | 5.9 | 6.7 |
| | B | 0 | 15 | 0 | 5.5 | 6.7 |
| | No student survey | na | na | na | na | na |
| Parent survey | A | 0 | 2 | 1 | 1.0 | .7 |
| | B | 0 | 2 | 1 | .8 | .7 |
| | No student survey | 0 | 5 | 3 | 3.3 | 1.2 |
| Peer survey | A | 1 | 2 | 2 | 2.0 | .2 |
| | B | 0 | 2 | 2 | 1.9 | .4 |
| | No student survey | 0 | 4 | 4 | 3.8 | .8 |

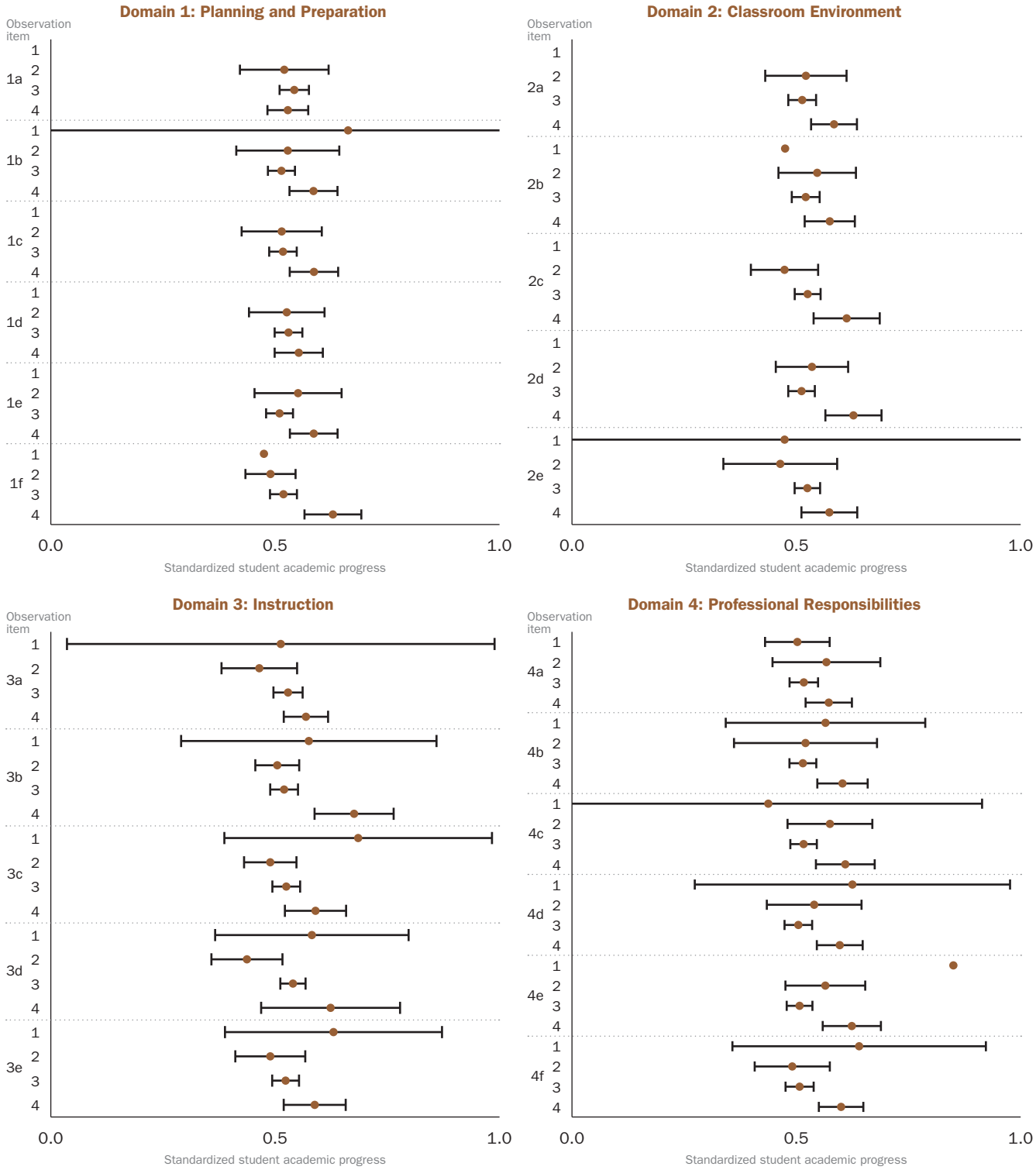
na is not applicable.

**Note:** In the Arizona Department of Education pilot teacher evaluation model, scales for each of the surveys were adjusted depending on the availability of student survey data. Due to this variability, statistics for group A and group B include only teachers with a student survey score. Those without student survey scores are shown separately.

**a.** Group A teachers had classroom-level student standardized test data; group B teachers had no classroom-level student standardized test data for their content area, so their results derived from math and reading tests their students took for other classes.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

**Figure D1. Mean student academic progress scores and confidence intervals by performance levels of observation items in the Arizona pilot teacher evaluation model, 2012/13**




Performance levels: 1: Unsatisfactory, 2: Basic, 3: Proficient, and 4: Distinguished.

**Note:** Group A teachers had classroom-level student standardized test data; group B teachers had no classroom-level student standardized test data for their content area. This figure includes all pilot teachers. Because student academic progress was measured differently for group A and group B, student academic progress data were standardized (using a z-transformation) within each group and then combined. A confidence interval indicates the degree of error present in the measurement of a data point, or the reliability of an estimate. This figure displays 95 percent confidence intervals, which means there is a 95 percent probability that the true value lies somewhere on the bracketed line. A larger confidence interval (or longer bracketed line) corresponds to a lower reliability or a higher degree of error.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

## Appendix E. Detecting nonlinear relationships between observation item scores and student academic progress metrics

Analysis of correlation can be taken one step further by asking whether the relationship between two metrics is really linear, as is conventionally presumed. This analysis focused on discovering the shape of the relationship between observation scores and the student academic progress component. The expected finding is that, as in earlier studies (Lazarev & Newman, 2013), some of these relationships will be nearly linear, some will have an inflection point (Γ shape), and others will be nonmonotonic. For evaluation system development, it is particularly important to isolate items that have a nonmonotonic relationship with student academic progress—for example, a U-shape (or an inverted U-shape) would imply that teachers in the middle of the range received the lowest (or highest) observation scores. Including such an ambiguous item (which, essentially, registers deviations from the mean) could be counterproductive. Identifying an item with these properties would warrant additional investigation of whether the item should be excluded from the evaluation system or whether the problem might be mitigated by additional observer training or revision of the scoring guidelines.

The analysis presented here is based on nonparametric regression, in which teacher effectiveness, as measured by student academic progress, was an arbitrary smooth (nonparametric) function of an observation indicator: $T_j = f_j(x_j) + \varepsilon_j$, where $f_j(x_j)$ needed to be estimated from data. The approach used allowed the true shape of the relationship (optimal degree of smoothing) to be established.[15] Due to the small number of observations, this analysis was likely to be underpowered, and the analysis was limited to the exploration of binary relationships, although the same approach could be used to derive a nonparametric model with multiple inputs.

One important diagnostic produced for each smooth component $f_j(x_j)$ was the estimated degree of freedom. This diagnostic helps determine which components can be reasonably well approximated by linear terms, and which require additional transformation. This information can help specify a simplified parametric model that best approximates the generalized additive model and that predicts the student academic progress metric by a number of parametric functions of observation items. An estimated degree of freedom close to unity implies that the relationship is linear, while larger numbers mean that the relationship between an item and the student academic progress metric can be approximated by a higher degree polynomial. Another characteristic is the explained proportion of variance in the student academic progress metric $R^2$. Of particular interest was the square root of this measure, which is on the same scale as the correlation coefficient. These numbers can be compared to evaluate the gain in explanatory power achieved by replacing a linear model with a nonlinear one. Because the shapes of relationships between Framework for Teaching items and student academic progress metrics varied substantially by grade level, the nonparametric analysis was done separately for teachers in group A (teachers who had classroom-level student standardized test data) and in group B (teachers who had no classroom-level student standardized test data) (Lazarev & Newman, 2013).

The analysis revealed a much larger number of significant associations between observation items and student academic progress (table E1) than did the conventional correlation analysis. Several relationships with relatively high linear correlations (greater than –.1) turned out to have stronger associations with student academic progress (greater than –.14)

**Table E1. Linear and nonlinear (nonparametric) estimates of the relationships between observation items and student academic progress metrics in the Arizona pilot teacher evaluation model, 2012/13**

| Domain and item | Correlation (Pearson) | | Nonparametric model (estimated degrees of freedom)[a] | | Nonparametric model ($R^2$) | |
|---|---|---|---|---|---|---|
| | Group A | Group B | Group A | Group B | Group A | Group B |
| 1. Planning and Preparation | | | | | | |
| 1a | –.05 | .02 | na | na | ns | ns |
| 1b | .04 | .15** | 2 | 1 | .05 (.22) | .02 (.14) |
| 1c | .14 | .07 | na | na | ns | ns |
| 1d | –.08 | .13 | na | na | ns | ns |
| 1e | –.04 | .19** | na | 2 | ns | .05 (.22) |
| 1f | .11 | .17** | 2 | 1 | .05 (.23) | .02 (.16) |
| 2. Classroom Environment | | | | | | |
| 2a | .06 | .06 | na | na | ns | ns |
| 2b | –.03 | .01 | na | na | ns | ns |
| 2c | .22** | .06 | 1 | na | .04 (.21) | ns |
| 2d | .13 | .10 | na | 2 | ns | .06 (.24) |
| 2e | –.01 | .10 | na | na | ns | ns |
| 3. Instruction | | | | | | |
| 3a | .03 | .06 | na | na | ns | ns |
| 3b | –.06 | .19** | na | 1 | ns | .03 (.19) |
| 3c | –.01 | .13 | na | na | ns | ns |
| 3d | .05 | .10 | na | na | ns | ns |
| 3e | .07 | .08 | na | na | ns | ns |
| 4. Classroom Responsibilities | | | | | | |
| 4a | .04 | .06 | na | na | ns | ns |
| 4b | .03 | .13 | na | na | ns | ns |
| 4c | .13 | .06 | 2 | na | .05 (.22) | ns |
| 4d | .00 | .21** | na | 1 | ns | .04 (.21) |
| 4e | .03 | .17** | na | 2 | ns | .06 (.24) |
| 4f | .11 | .16** | 2 | 2 | .06 (.24) | .03 (.17) |

** Significant at the .05 level.

ns is not significant; na is not applicable.

**Note:** Group A teachers had classroom-level student standardized test data; group B teachers had no classroom-level student standardized test data for their content area, so their results derived from math and reading tests their students took for other classes. Numbers in parentheses are square roots of the model's $R^2$. They can be compared with corresponding correlation coefficients in the two columns on the left.

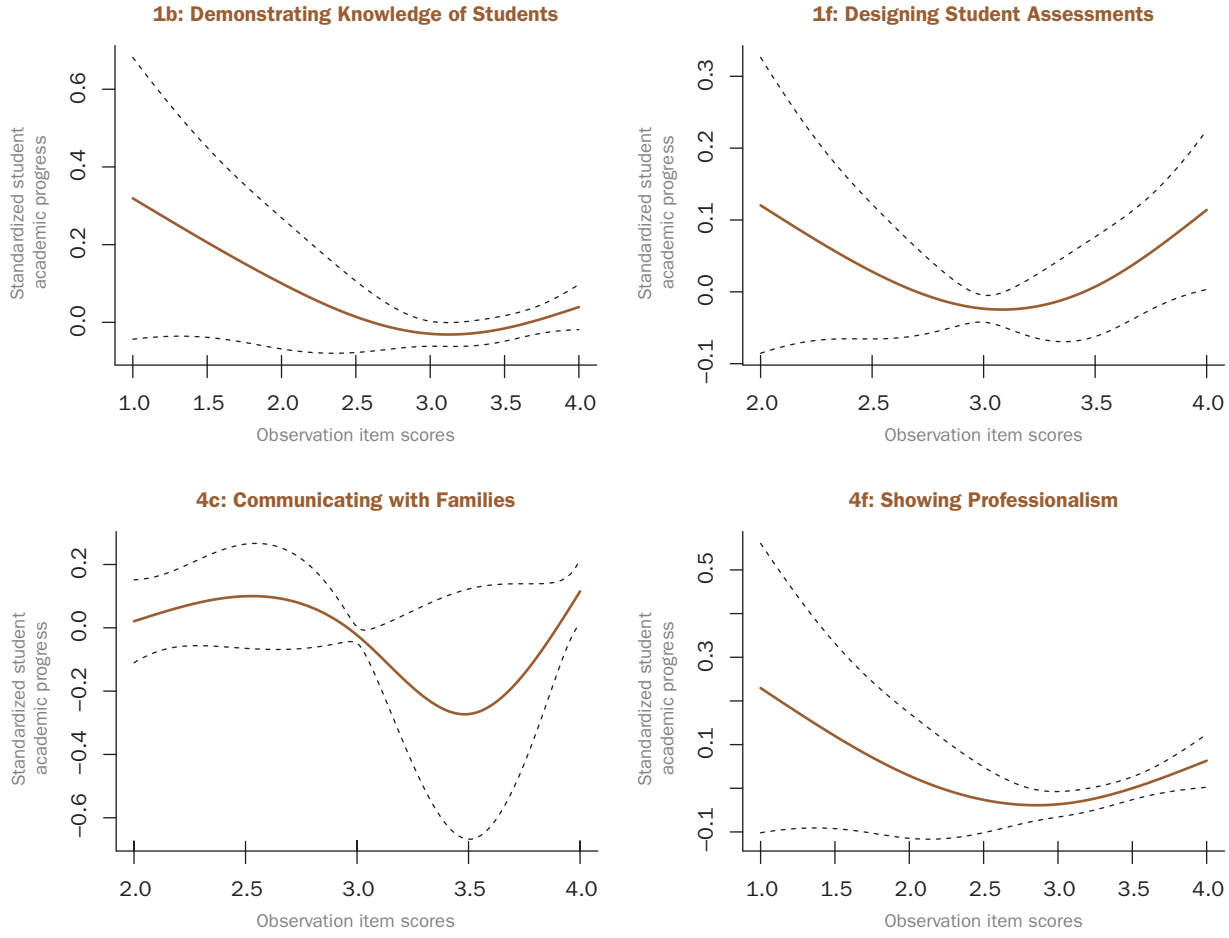
**a.** Estimated degrees of freedom rounded to the nearest integer. Values are not displayed for nonparametric analyses that were not significant.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

and nonlinear shape (approximate estimated degree of freedom equals 2). Examples include observation items 1f (Designing Student Assessments) in group A and 2d (Managing Student Behavior) in group B. Most relationships with significant linear correlations were confirmed to have an actual linear relationship with student academic progress (estimated degrees of freedom equal 1). Several relationships with significant linear correlations—observation items 1e (Designing Coherent Instruction), 4e (Growing and Developing Professionally), and 4f (Showing Professionalism) in group B—were estimated to have a nonlinear relationship and slightly higher strength of association than linear analysis yields.

Estimated relationship graphs for nonlinear items suggest that all or most relationships were U-shaped (figures E1 and E2). The lowest student academic progress is associated with teachers scoring in the middle of the observation score distributions (level 3), while higher student academic progress is associated with deviations in both directions. However, the scarcity of teachers scoring 1 or 2 on any of these items did not allow for reliable inference at the left end of the range, which was evidenced by wide and diverging confidence boundaries (dotted lines in the graphs). All that can be said with confidence is that the curves slope upward between levels 3 and 4. This is consistent with the findings presented in table 6 of the report; there are significant differences in student academic progress between teachers scoring a 3 and 4 on each of the items shown in figures E1 and E2. A larger sample size is needed to produce more accurate results. If further analyses confirm that these relationships are U-shaped, a substantial revision of the evaluation model may be warranted because the ambiguity inherent in such a shape negatively affects consistency and effectiveness. The results presented here and elsewhere in the report suggest only that the items presented in figures E1 and E2 can have limited use in evaluating teachers in the upper half of the score range.

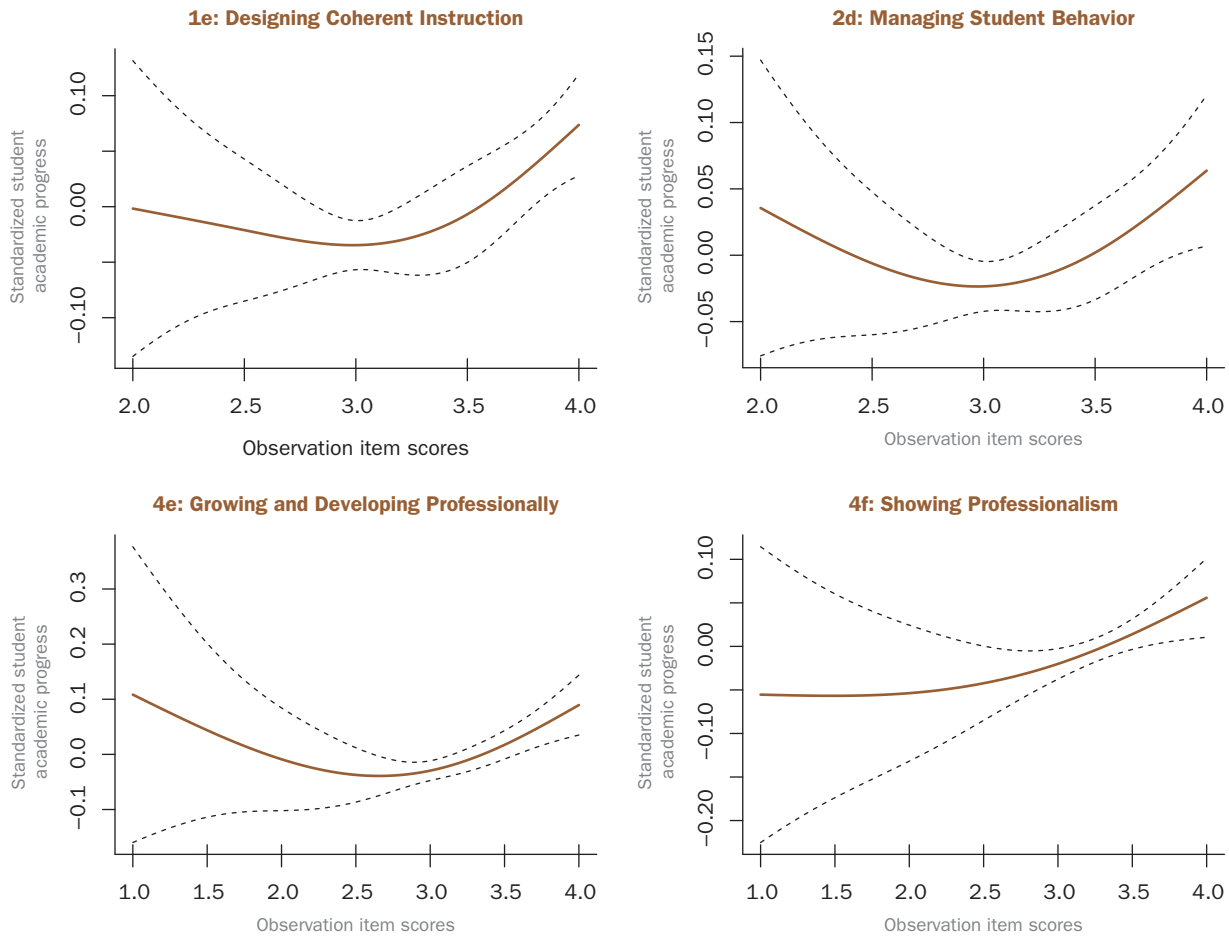
**Figure E1. Group A: Estimated relationships between observation items and student academic progress in the Arizona pilot teacher evaluation model, 2012/13**




**Note:** Group A teachers had classroom-level student standardized test data. Dotted lines show the .05 confidence bands.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

**Figure E2. Group B: Estimated relationships between observation items and student academic progress in the Arizona pilot teacher evaluation model, 2012/13**




**Note:** Group B teachers had no classroom-level student standardized test data available for their content area, so their results derived from math and reading tests their students took for other classes. Dotted lines show the .05 confidence bands.

**Source:** Authors' analysis based on 2012/13 data from the Arizona Department of Education.

## Notes

1. This report refers to the components of the Framework for Teaching as "items" to avoid confusion with the three components of the state teacher evaluation model.
2. In April 2011 the Arizona State Board of Education adopted the Arizona Framework for Measuring Educator Effectiveness, which requires that all the state's local education agencies base at least 50 percent of a teacher's evaluation on teaching performance (measured through periodic classroom observations).
3. Although some correlation between observation items is desirable, items that are too strongly correlated do not effectively evaluate distinct aspects of teaching performance.
4. A principal component is a combination of items or variables calculated using a particular method of statistical analysis (see appendix C).
5. The Measures of Effective Teaching project collected evaluation data from six school districts in six states during the 2010/11 and 2011/12 school years. The final sample contained 1,555 teachers. Teachers in the Measures of Effective Teaching studies were scored only on the eight items observable in the classroom (items 2a–d and 3a–d). The comparative analyses here are therefore limited to these items.
6. This threshold is often set at 95 percent (Joliffe, 1986).
7. This is a so-called scree test, first proposed by Cattell (1966).
8. The low item-level correlations for group A teachers with student academic progress in domain 4 suggest that the aggregation of item scores into domain scores may average out a portion of the measurement error inherent in these instruments, leading to a significant correlation at the domain level.
9. Correlations with Managing Student Behavior and Setting Instructional Outcomes were not significant at the .05 level, which may be due to the study's small sample size.
10. Some items, such as 4c (Communicating with Families) as well as most other items in domain 4, exhibited anomalies in the middle of the range—a lower student academic progress score at level 3 than at level 2 (that is, the corresponding mean differences are negative in the table). This suggests a possible nonmonotonic relationship between these items and the student growth metric. Additional exploratory analysis of this aspect of the relationship between observation and student academic progress scores is presented in appendix E.
11. Pairwise differences between corresponding correlation coefficients are significant at the .05 level.
12. The Colorado Department of Education's preliminary analyses of the surveys' item properties yielded promising results, with Cronbach's alpha values of .84 for the grade 3–5 survey and .92 for the grade 6–12 survey. Both surveys and a summary of their purpose, development, and intended use in Colorado are available at http://colegacy.org/educator-effectiveness-2/evaluation-systems/studentsurvey/.
13. There are 802 public (noncharter) schools in Arizona that are categorized by the National Center for Education Statistics as rural or town schools. Together, these schools educate 36 percent of Arizona's K–12 student population.
14. In principal component analysis, the term "component" has a special meaning, which is different from its use throughout this report. A principal component is a combination of items or variables calculated using a particular method of statistical analysis. It is similar to a factor in the context of factor analysis.
15. This study's approach is based on the use of penalized spline smoothing, an advanced method of nonparametric regression (see Wood, 2006 for an extensive overview). Most other methods of nonparametric regression require setting arbitrary parameters that

affect the degree of smoothing, so that the resulting shape of the nonlinear relationship between two variables reflects the decisions made by the researcher. By contrast, the method used here allows estimating the optimal shape of the relationship from data without using the researchers' expert knowledge. This feature makes it a preferred method in situations where little is known about the pattern of relationships between the variables of interest, as is the case in this study.

## References


Arizona Department of Education. (2011). *2010–2011 October 1 enrollment*. Retrieved April 22, 2013, from http://www.azed.gov/research-evaluation/files/2012/09/2011octenrollmentfinal1.xls

Arizona Department of Education. (2012). *Teacher evaluation process: An Arizona model for measuring educator effectiveness, based on the Arizona Framework for Measuring Educator Effectiveness*. Retrieved May 2013, from http://www.azed.gov/teacherprincipal–evaluation/files/2012/10/teacher-evaluation-web.pdf

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245–276.

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

Jerald, C. D. (2012). Movin' it and improvin' it! Using both education strategies to increase teaching effectiveness. Washington, DC: Center for American Progress. http://eric.ed.gov/?id=ED535645

Joliffe, I. T. (1986). *Principal components analysis*. New York: Springer Verlag.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (research report). Seattle, WA: Bill & Melinda Gates Foundation. http://eric.ed.gov/?id=ED540960

Lazarev, V., & Newman, D. (2013, September). *How non-linearity and grade-level differences complicate the validation of observation protocols*. Paper presented at the Fall 2013 *Society for Research on Educational Effectiveness* conference, Washington, DC.

Measures of Effective Teaching Project. (2010). *Validation engine for observation protocols*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved April 22, 2013, from http://metproject.org/downloads/Validation_Engine_concept_paper_09241.pdf

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved April 22, 2013, from http://metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf

Rothstein, J., & Mathis, W. J. (2013). *Review of "Have We Identified Effective Teachers?" and "A Composite Estimator of Effective Teaching: Culminating findings from the Measures of Effective Teaching Project."* Boulder, CO: National Education Policy Center. http://eric.ed.gov/?id=ED539299

U.S. Department of Education. (2010, April 9). Overview Information; Race to the Top Fund; Notice Inviting Applications for New Awards for Fiscal Year (FY) 2010, *Federal Register* 75(68), 18171.

U.S. Department of Education, National Center for Education Statistics. (2011). *Common Core of Data. Public Elementary/Secondary School Universe Survey, 2010–11, v.2a*. Retrieved April 22, 2013, from http://nces.ed.gov/ccd/pubschuniv.asp

U.S. Department of Education. (2012). *ESEA flexibility* [Web page]. Washington, DC. Retrieved from https://www.ed.gov/esea/flexibility

U.S. Department of Education. (2013a). *Race to the Top—Maryland Year 2: Year 2011–2012*. Washington, DC. Retrieved http://eric.ed.gov/?id=ED539238

U.S. Department of Education. (2013b). *Race to the Top—Tennessee Year 2: Year 2011–2012*. Washington, DC. Retrieved http://eric.ed.gov/?id=ED539244

U.S. Government Accountability Office. (2013). *Race to the Top: States implementing teacher and principal evaluation systems despite challenges*. GAO-13–777 Race to the Top Evaluation Systems. Retrieved December 2013, from http://www.gao.gov/assets/660/657936.pdf

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project. http://eric.ed.gov/?id=EJ857720

Wood, S. (2006). *Generalized additive models: An introduction with R*. Oxford, England: Taylor and Francis.

## 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